A Note On the Size of Largest Bins Using Placement With Linear Transformations

Martin Babka

October 11, 2018

Abstract

We study the placement of \( n \) balls into \( n \) bins where balls and bins are represented as two vector spaces over \( \mathbb{Z}_2 \). The placement is done according to a linear transformation between the two vector spaces. We analyze the expected size of a largest bin. The only currently known upper bound is \( O(\log n \log \log n) \) by Alon et al. and holds for placing \( n \log n \) balls into \( n \) bins. We show that this bound can be improved to \( O(\log n) \) in the case when \( n \) balls are placed into \( n \) bins. We use the same basic technique as Alon et al. but give a tighter analysis for this case.

1 Introduction

Research of hash function families is nowadays naturally focused on finding fast systems suitable for universal hashing, cuckoo hashing, linear probing, load balancing, etc. Each application has slightly different requirements on the system. For example universal hashing \[3\] requires families having small largest bins, for linear probing we have to provide at least a 5-independent family \[6\]. Additionally the time to compute the hash function should be small.

In this article we are dealing with the size of a largest bin in a balls-and-bins setting. It is known that if we place \( n \) balls into \( n \) bins randomly and independently, then with high probability the size of a largest bin is \( \Theta(\log n / \log \log n) \). There are non-trivial hash function families that achieve the sublogarithmic bound such as systems constructed by Siegel \[8\], the systems given in \[3\], tabulation hashing \[7\], and any \( \Theta(\log n / \log \log n) \)-independent hash function family. The hash function families with high degrees of independence provide asymptotically perfect results for other applications e.g. concentration bounds, Bloom filters, “two choices”, etc.

Unfortunately the systems with high degrees of independence are inefficient in practice either because of their size and/or speed according to Siegel’s lower bound \[8\]. So the research then focused on finding hash function families best fitting the needs of an application. There are systems designed to achieve the optimal size of a largest bin for balls-and-bins model that emerged in \[4\]. For cuckoo hashing there are known function families and modifications of the scheme which preserve the expected \( O(1) \) operation time such as cuckoo hashing with stash from \[5\] and \[2\] without using \( \Omega(\log n) \)-independent hash function family. For linear probing it is known that 5-independence is enough to achieve the expected constant probe sequence length \[6\].

The system of linear transformations between the binary vector spaces forms a natural two-wise independent system of functions. We show that using this system the size of a largest bin is nearly optimal despite its limited independence. Precisely if \( n = 2^b \) and \( n \) balls, chosen arbitrarily from
Proposition 1. Let \( \ell \in \mathbb{N} \) and \( \alpha < 1 \). Then \( \Pr_{\ell \in \mathbb{N}} [T(S) \neq \mathbb{Z}_2^k] \leq \alpha^{n - t - \log t + \log \log \frac{1}{\epsilon}} \).

Proposition 2 (\cite{Aho85} Theorem 7a, p. 7]). For each \( \epsilon > 0 \) there exists \( c_\epsilon > 0 \) depending solely on \( \epsilon \), such that for each \( t \in \mathbb{N}, S \subseteq \mathbb{Z}_2^k \) satisfying \( |S| \geq c_\epsilon t 2^t \) it holds \( \Pr_{T \in \mathcal{L}_u^b} [T(S) = \mathbb{Z}_2^k] \geq 1 - \epsilon \).

Let us note that from the proof in \cite{Aho85} it follows that \( c_\epsilon \) may be chosen as \( 4 \left( \frac{2}{t} \right)^{\frac{3}{4}} \).

Following \cite{Aho85} we define two events needed to estimate the probability of having a bin of size \( t \). The first event, \( E_1 \), occurs if there is a bin of size at least \( t \). The second one, \( E_2 \) is used to upper bound the probability of occurrence of \( E_1 \).

Definition 1 (\cite{Aho85} Event \( E_1 \), p. 11]). Let \( u, b, \ell, t \in \mathbb{N} \), \( T \in \mathcal{L}_u^b \). We put

\[
E_1(S, T, \ell) \equiv \exists y \in \mathbb{Z}_2^b : |T^{-1}(\{y\}) \cap S| \geq \ell.
\]
Figure 1: The decomposition of $T$, general case. Let us note that $F_A = T_1^{-1}(\tilde{y})$, $U_A = T^{-1}(\tilde{y})$ and $S_A = S \cap U_A$.

To define the second event, $E_2$, we decompose the chosen random linear map $T \in \mathcal{L}_u^b$ into $T_0 \in \mathcal{L}_u^f$ and a surjective $T_1 \in \mathcal{L}_S^b$ satisfying $T = T_1 \circ T_0$.

**Definition 2 (1 Event $E_2$, p. 11).** Let $u, f, b, \ell \in \mathbb{N}$, $f \geq b$, $S \subseteq \mathbb{Z}_2^u$, $T_0 \in \mathcal{L}_u^f$ and $T_1 \in \mathcal{L}_S^b$. The event $E_2(S, T_0, T_1)$ occurs when $\exists \tilde{y} \in \mathbb{Z}_2^b: T_1^{-1}(\tilde{y}) \subseteq T_0(S)$.

Refer to Fig. 1 for the general case of the decomposition and to Fig. 2 for the case when $E_2$ occurs. Now we show a relation between $E_1$ and $E_2$.

**Proposition 3 (1 Proposition 3.2, p. 11).** For each $\epsilon > 0$ there is $c_\epsilon > 0$ such that for each $u, f, b, \ell \in \mathbb{N}$ satisfying $u \geq f \geq b$, $\ell \geq c_\epsilon(f - b)2^f - b$ and for arbitrary $S \subseteq \mathbb{Z}_2^u$, it holds that $\Pr_{T \in \mathcal{L}_u^f} [E_1(S, T, \ell)] \leq \frac{1}{|T_0(S)|} \Pr_{T_0 \in \mathcal{L}_u^f, T_1 \in \mathcal{L}_S^b} [E_2(S, T_0, T_1)]$. In addition the value $c_\epsilon$ can be chosen according to Proposition 2 and depends only on $\epsilon$.

For completeness we provide a proof of Proposition 3 in the appendix.

Now we estimate the probability of $E_2$. Our Proposition 4 is a slight restatement of Proposition 3.1 from 1. It is similar to Proposition 3.1 in 1 but gives a slightly better bound. The proof is similar.

**Proposition 4.** Let $u, f, b \in \mathbb{N}$ such that $u \geq f \geq b$. If $S \subseteq \mathbb{Z}_2^u$, $|S| = 2^b$, $f > b$ and $\mu = \frac{2^b}{2^{2f}}$, then $\Pr_{T_0 \in \mathcal{L}_u^f, T_1 \in \mathcal{L}_S^b} [E_2(S, T_0, T_1)] \leq \mu^{\log b - \log \mu + \log \log \mu^{-1}}$.

**Proof.** Observe that $\exists \tilde{y} \in \mathbb{Z}_2^b: T_1^{-1}(\tilde{y}) \subseteq T_0(S)$ is equivalent to $\exists \tilde{y} \in \mathbb{Z}_2^b: \tilde{y} \notin T_1(\mathbb{Z}_2^f \setminus T_0(S))$. Hence $E_2(S, T_0, T_1)$ is equivalent to $T_1(\mathbb{Z}_2^f \setminus T_0(S)) \neq \mathbb{Z}_2^b$. Refer to Fig. 2 for more details of the situation when $E_2(S, T_0, T_1)$ occurs.

We prove the estimate for arbitrary fixed $T_0$ and uniform choice of $T_1$. From Proposition 3 it follows that $\Pr_{T_1 \in \mathcal{L}_S^b} [T_1(\mathbb{Z}_2^f \setminus T_0(S)) \neq \mathbb{Z}_2^b] \leq \alpha f - b \log b + \log \log \alpha^{-1}$ where $\alpha = 1 - \frac{|\mathbb{Z}_2^f \setminus T_0(S)|}{|\mathbb{Z}_2^b|} = \frac{|T_0(S)|}{2^f} \leq \frac{|S|}{2^{2f}} = \mu = 2^{2b - f} < 1$. Since the function $\alpha f - b \log b + \log \log \alpha^{-1}$ is increasing w.r.t. $\alpha$ in $(0, 1)$ we get that $\Pr_{T_1 \in \mathcal{L}_S^b} [E_2(S, T_0, T_1)] \leq \mu^{-\log \mu - \log b + \log \log \mu^{-1}}$. 


The following theorem gives an upper bound for the tail distribution of the random variable $\text{lbin}(T, S)$. The theorem is similar to Corollary 3.3 from [1], however the stated estimate is slightly different because it is adapted to our setting. The substantial difference between them is that we obtain non-trivial estimates for the logarithmic size of a largest bin whereas in [1] they get them for super-logarithmic sizes. The theorem in turn implies Theorem 1.

**Theorem 2.** For each $\epsilon > 0$ there exists $c_\epsilon > 0$ such that for each $u, b \in \mathbb{N}, u \geq b, r \geq 4$ it holds that

$$
\Pr_{T \in \mathbb{L}^b_u} [\text{lbin}(T, S) \geq 2c_\epsilon \ell] \leq \frac{1}{1 - \epsilon} \left( \frac{\log r}{r} \right)^{-\log b - \log \frac{\log r}{r} + \log \log \frac{\log r}{r}}.
$$

Moreover $c_\epsilon$ depends solely on $\epsilon$ and may be chosen according to Proposition 2.

**Proof.** Let $\epsilon, u, b, r$ be given so that they meet the requirements of the theorem. We put $f = b + \log r - \log \log r + 1$ and $\ell = 2c_\epsilon r$ where $c_\epsilon$ comes from Proposition 2. Recall that $\text{lbin}(T, S) \geq \ell$ is equivalent to the occurrence of event $E_1(S, T, \ell)$.

Proposition 3 implies that $\Pr_{T \in \mathbb{L}^b_u} [E_1(S, T, \ell)] \leq \frac{1}{1 - \epsilon} \Pr_{T \in \mathbb{L}^b_u} \{ T \in \mathbb{L}^b_u, T_1 \in \mathbb{L}^b_u, |E_2(S, T_0, T_1)| \}. We have to verify that $\ell \geq c_\epsilon (f - b)2f - b$. From the requirement $r \geq 4$ it follows that $c_\epsilon (f - b)2f - b \leq c_\epsilon (\log r - \log \log r + 1)2\log r - \log \log r + 1 \leq 2c_\epsilon r (\log r - \log \log r + 1) \leq 2c_\epsilon r \leq \ell$.

To bound the probability of $E_2(S, T_0, T_1)$ we use Proposition 3. Observe that the choice of $f$ from the beginning of the proof satisfies $f > b$. This also means that $\mathbb{L}^b_u$ is nonempty. We put $\mu = b^f$. Since $\mu \leq 2^{-\log r - \log \log r} = \frac{\log r}{r}$ and the function $g(x) := x^{-\log b + \log x - \log \log x}$ is increasing in $(0, 1)$, from Proposition 3 it follows that $\Pr_{T \in \mathbb{L}^b_u} \{ T \in \mathbb{L}^b_u, T_1 \in \mathbb{L}^b_u, |E_2(S, T_0, T_1)| \} \leq g(\mu) \leq g \left( \frac{\log r}{r} \right).$ 

Now we show the proof of the main theorem.

**Proof of Theorem 1.** We show the theorem for $|S| = 2^b$. If $|S| < 2^b$, the theorem follows from the proved case. Put $n = 2^b = |S|$. We split $\sum_{\ell=1}^n \Pr_{T \in \mathbb{L}^b_u} [\text{lbin}(T, S) \geq \ell]$ into two sums according to $\ell$ being lower or greater than $8c_\epsilon \log n$. We show that in the second case the probability of $\text{lbin}(T, S) \geq \ell$ is $O(\ell^{-3/2})$. 

Figure 2: Decomposition of $T$ when event $E_2(S, T_0, T_1)$ occurs, i.e. $F_A \subseteq T_0(S)$. 

Theorem 1 implies that $F_A \subseteq T_0(S)$. For each super-logarithmic $T_0(S)$, the theorem in turn implies Theorem 1.
First we fix $\epsilon \in (0, 1)$ arbitrarily, assume that $\ell \geq 8\epsilon n$ and choose $r$ so that $\ell = 2\epsilon r$. Hence $r \geq 4 \log n$. We claim that if $n$ is large enough, then the estimate obtained by Theorem 2 is below $\frac{r^{-1.5}}{1 - \epsilon}$. To prove this we bound the exponent of the estimate from below as follows.

$$- \log b - \log \log r + \log r + \log(\log r - \log \log r) \geq - \log \log r + 2 + \log \left(\frac{3 \log r}{4}\right)$$

$$= \log(3) > \frac{3}{2}.$$

Hence when $n$ is large enough we get that $\left(\frac{\log r}{r}\right)^{\log 3} < r^{-3/2}$ and

$$\sum_{\ell=8\epsilon, \log n+1}^{n} \Pr_{T \in \mathcal{L}^b_n} \left[ \bin(T, S) \geq \ell \right] \leq \int_{8\epsilon, \log n}^{n} \Pr_{T \in \mathcal{L}^b_n} \left[ \bin(T, S) \geq \ell \right] d\ell$$

$$= 2\epsilon \int_{4 \log n}^{n/2\epsilon} \Pr_{T \in \mathcal{L}^b_n} \left[ \bin(T, S) \geq 2\epsilon r \right] dr$$

$$\leq 2\epsilon \int_{1}^{\infty} r^{-1.5} dr = O \left( \frac{\epsilon}{1 - \epsilon} \right).$$

The whole sum may be estimated as $\sum_{\ell=1}^{n} \Pr_{T \in \mathcal{L}^b_n} \left[ \bin(T, S) \geq \ell \right] \leq 2\epsilon \left( 4 \log n + \frac{O(1)}{1 - \epsilon} \right)$. \qed

4 The special case when balls form a vector subspace

Let us note that when $S$ is a subspace of the universe, then the expected size of the largest bin is constant.

Theorem 3. Let $b, u \in \mathbb{N}$ and $S$ be a subspace of $\mathbb{Z}_2^u$ of dimension $b$. Then

$$E_{T \in \mathcal{L}^b_u} \left[ \bin(T, S) \right] = O(1).$$

Proof. We first observe that the non-empty bins have a simple structure – all of them are formed by elements which are affine subspaces of the universe. This in turns means that all the non-empty bins have the same size. Since the bin containing $\vec{0}$ in $\mathbb{Z}_2^b$ is always non-empty and has a constant expected size, the theorem follows.

Assume that $T \in \mathcal{L}^b_u$ is fixed. Let $K = S \cap \ker(T)$. If $T(\vec{v}) = \vec{y}$ for some $\vec{v} \in S$, then $T^{-1}(\vec{y}) \cap S = \vec{v} + K$. Hence for each $\vec{y} \in \mathbb{Z}_2^b$ it holds that $|T^{-1}(\vec{y}) \cap S| = 0$ or $|T^{-1}(\vec{y}) \cap S| = |K|$. By $\bin(T, S, \vec{y})$ we denote $|T^{-1}(\vec{y}) \cap S|$ and it holds that $|K| = \bin(T, S, \vec{0})$. From this it follows that $E_{T \in \mathcal{L}^b_u} \left[ \bin(T, S) \right] = E_{T \in \mathcal{L}^b_u} \left[ \bin(T, S, \vec{0}) \right] = O(1)$. \qed

5 Acknowledgment

We would like to thank Václav Koubek and Michal Koucký for advices, consultations and time spent verifying this note.
References

1. Noga Alon, Martin Dietzfelbinger, Peter Bro Miltersen, Erez Petrank, and Gábor Tardos. Linear hash functions. *J. ACM*, 46(5):667–683, September 1999.

2. Martin Anmüller, Martin Dietzfelbinger, and Philipp Woelfel. Explicit and efficient hash families suffice for cuckoo hashing with a stash. *Algorithmica*, 70(3):428–456, 2014.

3. J.Lawrence Carter and Mark N. Wegman. Universal classes of hash functions. *Journal of Computer and System Sciences*, 18(2):143 – 154, 1979.

4. L. Elisa Celis, Omer Reingold, Gil Segev, and Udi Wieder. Balls and bins: Smaller hash families and faster evaluation. *SIAM Journal on Computing*, 42(3):1030–1050, 2013.

5. Adam Kirsch, Michael Mitzenmacher, and Udi Wieder. More robust hashing: Cuckoo hashing with a stash. *SIAM Journal on Computing*, 39(4):1543–1561, 2010.

6. Anna Pagh, Rasmus Pagh, and Milan Ružić. Linear probing with constant independence. *SIAM Journal on Computing*, 39(3):1107–1120, 2009.

7. Mihai Pătrașcu and Mikkel Thorup. The power of simple tabulation hashing. *J. ACM*, 59(3):14:1–14:50, June 2012.

8. Alan Siegel. On universal classes of extremely random constant-time hash functions. *SIAM Journal on Computing*, 33(3):505–543, 2004.

A  Proof of Proposition 3

We give the full proof of Proposition 3 along with the necessary claims.

**Proposition 5.** Let $T_1 \in L_{b,1}^{S_2}$ be fixed. Then the uniform choice of $T_0 \in L_u^f$ yields the uniform choice of $T \in L_{b,u}^f$ where $T = T_1 \circ T_0$.

**Proof.** The proof of the claim may be found in [1] in the proof of Theorem 7b. Let $\vec{e}_1, \ldots, \vec{e}_u$ be a basis of $\mathbb{Z}_2^u$. Recall that the uniform choice of $T \in L_{b,u}^f$ is equivalent to random and independent choice of $T(\vec{e}_i) \in \mathbb{Z}_2$ for $i \in \{1, \ldots, u\}$. Since $T_1$ is onto, for each $\vec{y} \in \mathbb{Z}_2^u$ we have that $|T_1^{-1}(\vec{y})| = 2^{f-b}$. Hence the uniform independent choice of values $T_0(\vec{e}_i) \in \mathbb{Z}_2^f$ yields uniform independent choice of values $T(\vec{e}_i) = T_1(T_0(\vec{e}_i)) \in \mathbb{Z}_2^f$ for $i \in \{1, \ldots, u\}$.

**Proposition 6** ([1 Proposition 3.4, p. 13]). Let $u, f, b \in \mathbb{N}$, such that $u \geq f \geq b$. For a fixed $T \in L_{b,u}^f$ and $T_1 \in L_{b,1}^{S_2}$ there is a bijection between $\{T_0 \in L_u^f \mid T = T_1 \circ T_0\}$ and linear maps from $\text{Ker}(T)$ to $\text{Ker}(T_1)$.

**Proof.** We show that when $T$ and $T_1$ are fixed, then each restriction of $T_0$ to $\text{Ker}(T)$ can be uniquely extended to $\mathbb{Z}_2^u$. Thus the bijection is defined as $T_0|_{\text{Ker}(T)} = T_K$ where $T_K$ is a linear map from $\text{Ker}(T)$ to $\text{Ker}(T_1)$.

Let $B$ be a orthogonal basis of $\text{Ker}(T)$ and $B^E$ be an orthogonal extension of $B$ to $\mathbb{Z}_2^u$. Similarly let $B_1$ be a orthogonal basis of $\text{Ker}(T_1)$ and $B_1^E$ be an orthogonal extension of $B_1$ to $\mathbb{Z}_2^f$.

Let $\vec{x} \in \mathbb{Z}_2^f$. There exists a unique decomposition of $\vec{x}$ into two vectors $\vec{x}_K \in \text{Span}(B)$ and $\vec{x}_C \in \text{Span}(B^E \setminus B)$ such that $\vec{x} = \vec{x}_K + \vec{x}_C$. Analogically there is a unique vector $\vec{q}_E \in \text{Span}(B_1^E \setminus B_1)$ satisfying that $T_1(\vec{q}_E) = T_1(\vec{x}_K) + T_1(\vec{x}_C)$. We put $T_0(\vec{x}) = \vec{q}_E + T_K(\vec{x}_K)$.
Proposition 3. (Proposition 3.2, p. 11). For each \( \epsilon > 0 \) there is \( c_\epsilon > 0 \) such that for each \( u, f, b, \ell \in \mathbb{N} \) satisfying \( u \geq f \geq b, \ell \geq c_\epsilon (f - b)2^{f-b} \) and for arbitrary \( S \subseteq \mathbb{Z}_2^n \), it holds that \( \Pr_{T \in U} [E_1(S, T, \ell)] \leq \frac{1}{1 - \epsilon} \Pr_{T_0 \in U} [E_2(S, T_0, T_1)] \). In addition the value \( c_\epsilon \) can be chosen according to Proposition 2 and depends only on \( \epsilon \).

Proof of Proposition 3. Fix \( \epsilon \in (0,1) \). First we show that \( 1 - \epsilon \leq \Pr_{T_0, T_1} [E_2 | E_1] \). Assume that \( E_1(S, T, \ell) \) occurs, i.e. there is \( \vec{y} \in \mathbb{Z}_2^u \) such that \( |T^{-1}(\vec{y}) \cap S| \geq \ell \). Put \( U_A = T^{-1}(\vec{y}), S_A = U_A \cap S \) and \( F_A = T_1^{-1}(\vec{y}) \). If \( T_0(S_A) = F_A \), then \( T_0(S) \supseteq T_0(S_A) = F_A = T_1^{-1}(\vec{y}) \) and by definition \( E_2(S, T_0, T_1) \) occurs. See Fig. 2 for a better picture of the situation when \( E_2(S, T_0, T_1) \) occurs. Thus it is sufficient to estimate \( \Pr_{T_0, T_1} [T_0(S_A) = F_A | E_1(S, T, \ell)] \). To do so we further assume that \( T_1 \) is fixed, \( T = T_1 \circ T_0 \) and \( E_1(S, T, \ell) \) occurs.

Since \( T_1 \) is onto, it holds that \( |F_A| = 2^{f-b} \). Also notice that \( U_A \) and \( F_A \) are affine subspaces of \( \mathbb{Z}_2^u \) and \( \mathbb{Z}_2^f \) and \( |S_A| \geq \ell \geq c_\epsilon (f - b)2^{f-b} \). Let \( T_A \) be an affine linear map from \( U_A \) to \( F_A \). From Proposition 2 used for \( T_A \), \( U_A \), \( S_A \), \( F_A \) we get that \( \Pr_{T_A} [T_A(S_A) = F_A | E_1] \geq 1 - \epsilon \). Notice that Proposition 2 may be used for affine linear transformations as well. Since the previous estimate holds for arbitrary fixed \( T \) and \( T_1 \), it holds for the uniform choice of the two transformations. Thus \( 1 - \epsilon \leq \Pr_{T_A} [T_A(S_A) = F_A | E_1] = \Pr_{T_A, T_0, T_1} [T_A(S_A) = F_A | E_1] \). From previous and Proposition 6 we get that \( 1 - \epsilon \leq \Pr_{T_0, T_1} [E_2 | E_1] \). From the previous inequality and Proposition 5 it follows that \( \Pr_{T} [E_1] = \Pr_{T_0, T_1} [E_1] \leq \frac{1}{1 - \epsilon} \Pr_{T_0, T_1} [E_2] \).