Electronic Colloquium on Computational Complexity, Report No. 64 (2025)

Linear Hashing Is Optimal

Michael Jaber* Vinayak M. Kumar* David Zuckerman†

Abstract

We prove that hashing n balls into n bins via a random matrix over $\mathbb{F}_2$ yields expected maximum load $O(\log n/\log\log n)$. This matches the expected maximum load of a fully random function and resolves an open question posed by Alon, Dietzfelbinger, Miltersen, Petrank, and Tardos (STOC '97, JACM '99). More generally, we show that the maximum load exceeds $r \cdot \log n/\log\log n$ with probability at most $O(1/r^2)$.

1 Introduction

In 1997, Alon, Dietzfelbinger, Miltersen, Petrank, and Tardos [ADM+97] asked "Is Linear Hashing Good?" We answer this question in the affirmative over $\mathbb{F}_2$.

Theorem 1. *Let $u \geq \ell \geq 1$ be integers, $n := 2^\ell$, and $\mathcal{H}$ the set of linear maps $\mathbb{F}_2^u \to \mathbb{F}_2^\ell$. For any $S \subseteq \mathbb{F}_2^u$ with cardinality n, we have*

$$\mathbb{E}_{h\sim\mathcal{H}}\left[\max_{y\in\mathbb{F}_2^\ell} |h^{-1}(y)\cap S|\right] \leq \frac{16\log n}{\log\log n}.$$

Due to the classical balls-and-bins result, this shows that the expected maximum load of a random linear map is within a constant factor of the maximum load of a fully random function. Our proof uses potential functions, which more generally allows us to show that the maximum load exceeds $\frac{r\log n}{\log\log n}$ with probability $O(1/r^2)$. We now provide context on our work, and cover prior results on linear hashing.

1.1 Motivation

Consider tossing n balls uniformly and independently into n bins. The maximum number of balls in any bin–the max load–is a well studied quantity critical to randomized algorithm design [MU05, Chapter 5]. A common application arises in hashing with chaining. In this scenario, keys are mapped to addresses via a random hash, and each address holds a linked list of all keys mapping to it. Consequently, retrieving a key might require sweeping through the largest linked list, and so the worst-case retrieval time would be the max-load.

A classical result says that tossing n balls randomly into n bins will have expected max-load $O\left(\frac{\log n}{\log\log n}\right)$, implying that a fully random hash will have only $O\left(\frac{\log n}{\log\log n}\right)$ expected retrieval time. However, placing balls independently and uniformly requires sampling a truly random function, which has near-maximal time and space complexity. This raises a central question:

*Department of Computer Science, University of Texas at Austin. {mjjaber, vmkumar}@cs.utexas.edu. Supported by NSF Grant CCF-2312573 and a Simons Investigator Award (#409864, David Zuckerman).

†Department of Computer Science, University of Texas at Austin. diz@cs.utexas.edu. Supported in part by NSF Grant CCF-2312573 and a Simons Investigator Award (#409864).

ISSN 1433-8092

What is the simplest hash family with optimal expected max-load?

An engineering approach to this question is to use any tools necessary to design hash functions that minimize the time and space complexity, say in the word-RAM model. Such an approach will naturally give low complexity hash functions, but at the cost of a more contrived construction that is hard to implement practically. Nevertheless, this approach has been extremely successful, and arguably started with Wegman and Carter's definition of k-wise independent hash functions [WC81]. One can easily verify that $O(\log n/\log\log n)$-wise independent hash functions have optimal max-load, and only require $O(\log n/\log\log n)$ evaluation time and $O(\log^2 n/\log\log n)$ description size. Recent works greatly optimized the use of k-wise independence, culminating in a construction of Meka, Reingold, Rothblum, and Rothblum [CRSW13, MRRR14] only needing $O(\log n \log\log n)$ bits to describe and $O((\log\log n)^2)$ time to evaluate. This gets us within a $\text{poly}(\log\log n)$ factor of optimal in both parameters. However, these function rely on concatenating hashes of gradually increasing independence, which will require implementing either prime fields or polynomial rings, and performing $\omega(1)$ multiplications over them. Although possible, it is quite complicated and slow to do in practice.

The approach we adopt in this paper is to only consider simple and practical hash functions and analyze their properties as well as possible. To this end, Chung, Mitzenmacher, and Vadhan [CMV13] showed that basic universal hash functions (e.g., linear congruence or multiplication schemes) can achieve optimal max-load if the balls are assumed to have high enough entropy, but say nothing about a worst-case choice of balls. In terms of worst-case results, a surprising result of Pătraşcu and Thorup [PT12] show that tabulation hashing has optimal max-load. This scheme is only 3-wise independent, is simple and practical to use, and has $O(1)$ evaluation time. However, it requires $O(n^\varepsilon)$ bits to describe.

Linear Hash Functions

In this paper, we consider an extremely simple hash family proposed in the first paper on universal hashing [CW79]: random matrices over $\mathbb{F}_2$. In particular, let $\ell \coloneqq \log n$, and say we arrange our n bins into a vector space $\mathbb{F}_2^\ell$. Further, arrange the universe set into a vector space $\mathbb{F}_2^u$. Our hash family is simply the set of linear maps $h : \mathbb{F}_2^u \to \mathbb{F}_2^\ell$. This family is arguably the easiest to implement over all constructions mentioned thus far. For $u = O(\ell)$, this family only needs to bitwise XOR $O(\log n)$ words together, and takes $O(\log^2 n)$ bits to describe. This family requires no multiplication, is not even 3-wise independent, and can be seen as the simplest version of tabulation hashing, where the lookup tables only stores two values.

Furthermore, linear hash functions can batch compute keys that are clustered together, which happens often in practice. This notion of efficiency, dubbed *incremental cryptography*, was first introduced in a pair of papers by Bellare, Goldreich, and Goldwasser [BGG94, BGG95]. The motivation behind incremental cryptography is that data to be hashed often consists of slight modifications of each other, whether it be consecutive frames of video footage or edited versions of files. Ideally, after computing the first hash, future hashes should have computation time proportional to the amount of modification made to this initial key. This is indeed the case for linear maps. Let $h(x) \coloneqq \sum_{i:x_i=1} c_i$ be a linear map whose matrix has columns $c_1, \ldots, c_u$. If x has Hamming weight w, $h(x)$ is simply a sum of w column vectors. Hence, if $x, x^1, \ldots x^{100}$ are keys such that each x^i differs from x in less than $w \ll u$ bits, we can compute all the hashes by computing $h(x)$, $h(x + x^i)$ for all i, and then computing $h(x^i) = h(x) + h(x + x^i)$. This gives a total of $u + 100w$ additions, which is better than the naive way of computing $h(x), h(x^1), \ldots, h(x^{100})$ directly, which can take up to $100u$ additions.

Computational considerations aside, it is a fundamental question to ask how much a random matrix behaves like a random function. From a technical standpoint, dealing with the correlations between linearly dependent balls has proven to be challenging. For example, the mere

existence of good linear seeded extractors is a longstanding open problem in pseudorandomness [Woo14, Question 7.6], and highlights our lack of understanding of random linear maps.

The Expected Max-Load of Linear Maps

Characterizing the max-load of $\mathcal{H}$, the set of linear maps, has remained elusive. As a preliminary bound, the expected max-load of $\mathcal{H}$ is $O(\sqrt{n})$ since it is universal [CW79]. Indeed, some universal hash functions saturate this bound [ADM+97]. However, Markowsky, Carter, and Wegman [MCW78] showed $\mathcal{H}$ achieves expected max-load $O(n^{1/4})$, significantly outperforming the generic universal bound [ADM+97]. This initiated further study of $\mathcal{H}$'s max-load, with Mehlhorn and Vishkin [MV84] (implicitly) showing a subpolynomial bound of $2^{O(\sqrt{\log n})}$, and later with Alon, Dietzfelbinger, Miltersen, Petrank, and Tardos [ADM+97] giving a bound of $O(\log n \log\log n)$. Since then, progress has largely stalled, except for a note by Babka [Bab18], who observed that an improved choice of parameters in the argument of [ADM+97] yields a bound of $O(\log n)$. Unfortunately, the argument in [ADM+97] has an inherent barrier at $O(\log n)$,[1] leaving open whether $\mathcal{H}$ can match the performance of a fully random map. Indeed, this natural question was explicitly posed in [ADM+97].

1.2 Our Results

We fully resolve this question by showing that a random linear map hashing n balls to n bins will have expected max-load $O\left(\frac{\log n}{\log\log n}\right)$. Our proof easily generalizes to m balls and n bins for $m \neq n$. In particular, define

$$\textsc{opt}(m,n) = \begin{cases} \frac{\log n}{\log\left(\frac{n\log n}{m}\right)} & m \leq \frac{1}{2} n \log n \\ \frac{m}{n} & m > \frac{1}{2} n \log n. \end{cases}$$

It is well known that a random function mapping m balls to n bins will have expected max-load $\Theta(\textsc{opt}(m,n))$ [RS98, Theorem 1]. We show that a random linear map performs just as well.

Theorem 2 (Theorem 1 generalized)**.** *Let $u \geq \ell, m$ be integers, $n := 2^\ell$, and $\mathcal{H}$ the set of linear maps $\mathbb{F}_2^u \to \mathbb{F}_2^\ell$. For any $S \subseteq \mathbb{F}_2^u$ with cardinality m,*

$$\mathbb{E}_{h\sim\mathcal{H}}\left[\max_{y\in\mathbb{F}_2^\ell} |h^{-1}(y) \cap S|\right] \leq 16 \cdot \textsc{opt}(m,n)$$

Our Theorem 2 is a simple corollary of the following theorem, which gives quadratically decaying tail bounds on the max-load.

Theorem 3. *Let $u \geq \ell \geq 1, m \geq 1$ be integers, $n := 2^\ell$, and $\mathcal{H}$ the set of linear maps $\mathbb{F}_2^u \to \mathbb{F}_2^\ell$. For any $S \subseteq \mathbb{F}_2^u$ with cardinality m and $r \geq 6$,*

$$\Pr_{h\sim\mathcal{H}}\left[\max_{y\in\mathbb{F}_2^\ell} |h^{-1}(y) \cap S| \geq r \cdot \textsc{opt}(m,n)\right] \leq \frac{49}{(r-2)^2}.$$

Our techniques arguably yield a more streamlined proof than [ADM+97], which first proves a coupon collector property, and then cleverly converts it to a max-load bound. In contrast, we directly analyze the max-load by defining a potential function and tracking its growth. We elaborate further in Section 1.3.

In terms of matching lower bounds, we note Theorem 2 is tight up to constant factors. In particular, Celis, Reingold, Segev, and Wieder show that *any* family of hash functions will have max-load $\Omega(\textsc{opt}(m,n))$ with high probability [CRSW13, generalization of Theorem 5.1].

[1] Applying their argument to a purely random function only yields an $O(\log n)$ expectation bound on the max-load.

Interestingly, in the regime $m = \Omega(n \log n)$, we show that with high probability, *every* bin has load within a constant factor of $\textsc{opt}(m, n)$.

Theorem 4. *Let $u \geq \ell \geq 1, m \geq 1$ be integers, $S \subseteq \mathbb{F}_2^u$ a set of cardinality m, and $\mathcal{H}$ the set of linear maps $\mathbb{F}_2^u \to \mathbb{F}_2^\ell$. For every $0 < \varepsilon < 1/2$, there exists constants $C_1 < C_2$ depending on ε such that for $m \geq C_1^{-1} n \log n$,*

$$\Pr_{h \sim \mathcal{H}} \left[\forall y \in \mathbb{F}_2^\ell,\ C_1 \frac{m}{n} \leq |h^{-1}(y) \cap S| \leq C_2 \frac{m}{n} \right] \geq 1 - \varepsilon.$$

In particular, $C_1 = \Omega(\varepsilon^{74})$ and $C_2 = O(\varepsilon^{-1/2})$.

This can be seen as a generalization of the result of [ADM+97, Theorem 7], which states that $\mathcal{H}$ satisfies the *covering property*. The covering property, named after the fact that the cover time of a random walk on the complete graph on n vertices is $O(n \log n)$, says that if $\Omega(n \log n)$ balls are mapped to n bins by $h \sim \mathcal{H}$, every bin will be occupied with high probability. We show these hash functions actually satisfy a *blanketing property*: if $\Omega(n \log n)$ balls are mapped to n bins by $h \sim \mathcal{H}$, every bin will contain $\Omega(\log n)$ balls with high probability. A fully random h has this property by the fact that the blanket time [WZ96] of the complete graph is $O(n \log n)$. Interestingly, we use potential functions to prove this claim as well.

Comparison with Dhar and Dvir

A recent work of Dhar and Dvir [DD22] also established two-sided bounds on all bins for linear hash functions. They proved the following:

Theorem 5 ([DD22], Theorem 2.4)**.** *Let $u \geq \ell \geq 1$ be integers, $n := 2^\ell$, $\varepsilon, \tau \in (0, 1)$, and $S \subseteq \mathbb{F}_2^u$ a set of cardinality m. Let $\mathcal{H}$ be the set of linear maps $h : \mathbb{F}_2^u \to \mathbb{F}_2^\ell$. For $m = \Omega(n \log^4(n/\tau\varepsilon))$ we have*

$$\Pr_{h \sim \mathcal{H}} \left[\forall y \in \mathbb{F}_2^\ell,\ (1-\tau) \frac{m}{n} \leq |h^{-1}(y) \cap S| \leq (1+\tau) \frac{m}{n} \right] \geq 1 - \varepsilon.$$

Take $\varepsilon = O(1)$. We note that Theorem 5 gives strong deviation bounds from the optimal load, especially if one sets $\tau = o(1)$, whereas our Theorem 4 only establishes that all bins are within some constant multiplicative factor of optimal. However, Theorem 5 only holds in the regime $m = \Omega(n \log^4 n)$, whereas Theorem 4 can handle m as small as $\Theta(n \log n)$. This gap between m and n is optimal, as it was shown in [ADM+97] that there exists sets of size $0.69 n \log n$ such that no linear map will even occupy all bins. Our techniques are also quite different. Dhar and Dvir [DD22] first reduce the problem of bounding the max-load to bounding the size of Furstenberg sets and then apply the polynomial method, while our proof constructs a potential function and bounds its growth to yield the result.

In summary:

- when $m = O(n \log n)$, Theorem 2 gives us optimal upper bounds (up to a constant multiplicative factor) on all bins,
- when $m \in [\Omega(n \log n), O(n \log^4 n)]$, Theorem 4 gives us two-sided multiplicative bounds from the mean load on all bins,
- and when $m = \Omega(n \log^4 n)$, Theorem 5 gives very strong bounds on the additive deviation of all bins from the mean.

1.3 Proof Overview

For simplicity, assume $\mathcal{H}$ is a random *surjective* linear map $\mathbb{F}_2^u \to \mathbb{F}_2^\ell$. We would like to directly analyze the load distribution of $\mathcal{H}$, but this is quite complicated to do. Instead, for a set of balls S and hash h, we define a potential function Φ that measures how "imbalanced" the

allocation of S by h is. In particular, we want a potential function $\Phi := \Phi(S, h)$ such that if one preimage of h contains $\geq t$ elements in S, then $\Phi \geq f(t)$ for some function f. The hope is now that analyzing Φ will be much easier than analyzing the load distribution directly.

We use the potential $\mathbb{E}_y[b^{|h^{-1}(y)\cap S|}]$, which takes the average of the exponentials of all the bin loads with some base $b > 1$. To analyze this potential, we think of hashing our universe, $\mathbb{F}_2^u$, "one kernel vector at a time." In particular, any $h : \mathbb{F}_2^u \to \mathbb{F}_2^\ell$ can be decomposed into $h = h_1 \circ h_2 \circ \cdots \circ h_{u-\ell}$, where $h_i : \mathbb{F}_2^{u-i+1} \to \mathbb{F}_2^{u-i}$ is a random surjective map. Each h_i is much simpler to analyze, and allows us to show that if $h_{\leq i} := h_1 \circ \cdots \circ h_i$ and $\Phi_i := \Phi(S, h_{\leq i})$, then $\mathbb{E}[\Phi_{i+1}|\Phi_1, \ldots \Phi_i] \leq \Phi_i^2$.

It remains to show that these conditional expectation bounds can be leveraged to give a tail bound on $\Phi_{u-\ell} = \Phi$. A tempting approach is to use the conditional expectations to bound $\mathbb{E}[\Phi] \leq f(t)/2$, and then apply Markov's inequality to get a tail bound. Unfortunately, there exists random variables satisfying the conditional expectation bounds, but with $\mathbb{E}[\Phi]$ much larger than $f(t)$. Nevertheless, we prove a technical lemma showing that although $\mathbb{E}[\Phi]$ may be larger than $f(t)$, the conditional expectation bounds enforce that Φ *will still be smaller than* $f(t)$ *with good probability.* A tail bound of similar flavor was established in [ADM+97], but only works on random variables less than 1, and is inapplicable to our setting.

This lemma allows us to prove that the max-load exceeds $\frac{r \log n}{\log \log n}$ with probability $O(1/r)$. Unfortunately, this is not enough to deduce an $O(\log n / \log \log n)$ bound in expectation. In fact, only using the property $\mathbb{E}[\Phi_{i+1}|\Phi_1, \ldots \Phi_i] \leq \Phi_i^2$ provably cannot give a stronger tail bound. Thankfully, our potential functions have a strong monotonicity property: $\forall i, \Phi_{i+1} - 1 \geq 2(\Phi_i - 1)$. With some technical work, we can leverage the monotonicity and squared conditional expectation properties to obtain quadratically stronger tail bounds, from which optimal expected max-load follows straightforwardly.

Surprisingly, our potential functions also allow us to establish *lower bounds* on the loads of all bins. By setting the base $b < 1$, the potential now detects whether some bin is light. Combining this with the max-load analysis allows us to deduce our two-sided bounds on the bin loads.

To the best of our knowledge, our analysis provides the first proof of optimal max-load for a function that is universal, but not 3-wise independent. In fact, our proof technique applies to a broader class of hash functions mapping $[2^u] \to [2^\ell]$. Imagine starting with 2^u bins, where each of the 2^u universe elements are in their own bin. For $u - \ell$ iterations, pseudorandomly pair bins up in a pairwise independent fashion (i.e., for an arbitrary bin, the marginal distribution of the bin's partner is uniform among all bins) and consolidate each pair into one bin. Consequently, each iteration halves the number of bins, and at the end of this process, we will have hashed into n bins. Our techniques show that any such hashing scheme will have optimal expected max-load. The case of surjective linear maps is when in each round, a random vector v is picked, and each bin x is paired with bin $x + v$.

2 Preliminaries

2.1 Notation

We let log denote the base-2 logarithm, and ln denote the base-e logarithm. For a list of vectors $v_1, \ldots, v_n$, we will denote the tuple $v_{\leq i} := (v_1, v_2, \ldots, v_i)$ and $V_i = \text{Span}(v_{\leq i})$ (by convention, $V_0 = \{0\}$). We will use the convention $0^0 = 1$. Throughout this paper, we always have $\ell = \ell(n) := \log n$ and n a power of 2. $\mathbb{F}_2$ is the finite field over $\{0, 1\}$. For a set S, we write $s \sim S$ to denote that s is sampled uniformly at random from S.

Definition 1. *Let $h : A \to B$ and $S \subseteq A$ be a subset. We define the maximum load function*

$$M(S, h) := \max_{y \in B} |h^{-1}(y) \cap S|.$$

2.2 Inequalities

We will use the classic inequality $1 + x \le e^x$ for all x, and the following not-so-classic variant.

Fact 1. *For $x < 1$, we have $1 - x \ge e^{-\frac{x}{1-x}}$.*

Proof. Note $\frac{1}{1-x} = 1 + \frac{x}{1-x} \le e^{\frac{x}{1-x}}$. When $x < 1$, both sides of the inequality are positive, and taking reciprocals gives the fact. □

We will also use Bernoulli's Inequality.

Fact 2. *For an integer $n \ge 1$ and $1 + x \ge 0$, $(1 + x)^n \ge 1 + nx$.*

Finally, for the two-sided bounds, we will use standard facts about martingales, such as the Doob martingale and an elementary version of Azuma's inequality [MU05, Chapter 12].

Fact 3. *Let $(X_i)_{i\ge 0}$ be a martingale such that for all i, $|X_{i+1} - X_i| \le 1$. Then for all positive integers k and real number $\varepsilon > 0$, we have*

$$\Pr[X_k - X_0 \le -\varepsilon] \le e^{-\varepsilon^2/2k}.$$

2.3 Random Linear Maps

If $h : \mathbb{F}_2^u \to \mathbb{F}_2^\ell$ is a linear map with kernel $V \le \mathbb{F}_2^u$ and y is in the image of h, then $h^{-1}(y) = x + V$ for some $x \in \mathbb{F}_2^u$. A random *surjective* linear map $h : \mathbb{F}_2^u \to \mathbb{F}_2^\ell$ is equivalent to sampling a uniform $(u - \ell)$-dimensional subspace $V \le \mathbb{F}_2^u$, and then sampling a uniform linear h with kernel V. For any $u \ge t \ge \ell$ and surjective map $h_2 : \mathbb{F}_2^t \to \mathbb{F}_2^\ell$, if $h_1 : \mathbb{F}_2^u \to \mathbb{F}_2^t$ is a uniform random linear map, then $h_1 \circ h_2 : \mathbb{F}_2^u \to \mathbb{F}_2^\ell$ is a uniform random linear map.

3 Introducing the Potential Functions

Consider k vectors $v_{\le k} \in (\mathbb{F}_2^u)^k$ and a set of balls $S \subseteq \mathbb{F}_2^u$. Recall $V_i = \mathrm{Span}(v_{\le i})$. To $(v_{\le k}, S)$ we will associate functions $\{S_i : \mathbb{F}_2^u \to \mathbb{N}\}_{0\le i\le k}$, defined by

$$S_i(x) \coloneqq |(x + V_i) \cap S|. \tag{1}$$

In particular, S_i depends on S and $v_{\le i}$. Intuitively, S_0 is (the indicator of) S, and $S_i(x)$ is the number of balls in the same bin as x after hashing according to the kernel vectors $v_{\le i}$.

Now for any real number $b \ge 0$, we can define a sequence of potential functions

$$\Phi_i = \Phi_i(S; b; v_{\le i}) \coloneqq \mathbb{E}_{x\sim\mathbb{F}_2^u}[b^{S_i(x)}].$$

The following claim shows that if $b \ge 1$, a heavy bin implies large potential.

Lemma 1. *Let $v_{\le i}$ be a fixed tuple of linearly independent vectors in $\mathbb{F}_2^u$, and let $b \ge 1$. If there exists $x \in \mathbb{F}_2^u$ such that $|(x + V_i) \cap S| \ge m$, then $\Phi_i \ge \frac{b^m}{2^{u-i}}$.*

Proof. If $v \in V_i$, then $v + V_i = V_i$. Hence, for all $v \in V_i$, $S_i(x + v) = S_i(x) \ge m$. Therefore,

$$\Phi_i = \frac{1}{2^u}\sum_{x'\in\mathbb{F}_2^u} b^{S_i(x')} \ge \frac{1}{2^u}\sum_{x'\in x+V_i} b^{S_i(x')} \ge \frac{1}{2^u}|V_i|b^m = \frac{b^m}{2^{u-i}},$$

where we used the fact b^x is increasing for $b \ge 1$. □

Interestingly, if we use a base $b \le 1$, we can detect by the same test if some bin is light as well. This will help us establish two-sided bounds on the bins later on.

Lemma 2. *Let $v_{\le i}$ be a fixed tuple of linearly independent vectors in $\mathbb{F}_2^u$, and let $b \le 1$. If there exists $x \in \mathbb{F}_2^u$ with $|(x+V_i) \cap S| \le m$, then $\Phi_i \ge \frac{b^m}{2^{u-i}}$.*

Proof. The proof is essentially the same as Lemma 1. We observe that for all $v \in V_i$, $S_i(x+v) \le m$. Hence,

$$\Phi_i = \frac{1}{2^u} \sum_{x' \in \mathbb{F}_2^u} b^{S_i(x')} \ge \frac{1}{2^u} \sum_{x' \in x+V_i} b^{S_i(x')} \ge \frac{1}{2^u} |V_i| b^m = \frac{b^m}{2^{u-i}},$$

where we used the fact b^x is decreasing for $b \le 1$. □

The following claim relates S_i to S_{i+1}.

Claim 1. *For any vectors $v_{\le (i+1)}$, we have*

$$S_i(x) + S_i(x+v_{i+1}) = \begin{cases} 2S_i(x) & v_{i+1} \in V_i \\ S_{i+1}(x) & v_{i+1} \notin V_i \end{cases}.$$

Proof. If $v_{i+1} \in V_i$, then $V_i = v_{i+1} + V_i$, and so

$$S_i(x) + S_i(x+v_{i+1}) = S_i(x) + S_i(x) = 2S_i(x).$$

If $v_{i+1} \notin V_i$, then $V_{i+1} = V_i \sqcup (v_{i+1} + V_i)$. Thus, $(x+V_i) \cap S$ and $(x+v_{i+1}+V_i) \cap S$ partition $(x+V_{i+1}) \cap S$, implying that $S_i(x) + S_i(x+v_{i+1}) = S_{i+1}(x)$. □

Using the above claim, we can prove the following crucial lemma which upper bounds the conditional expectations of our potentials.

Lemma 3. *Let $v_{\le i} \in (\mathbb{F}_2^u)^i$, and let $v_{i+1} \sim \mathbb{F}_2^u \setminus V_i$. We have*

$$\mathbb{E}_{v_{i+1}}[\Phi_{i+1}] \le \Phi_i^2.$$

Proof. For $x, v_{i+1} \in \mathbb{F}_2^u$ picked uniformly and independently, x and $x+v_{i+1}$ are uniform and independent as well. Hence

$$\mathbb{E}_{x,v_{i+1}}[b^{S_i(x)+S_i(x+v_{i+1})}] = \mathbb{E}_{x,v_{i+1}}[b^{S_i(x)+S_i(v_{i+1})}] = \mathbb{E}_x[b^{S_i(x)}] \cdot \mathbb{E}_{v_{i+1}}[b^{S_i(v_{i+1})}] = \Phi_i^2. \tag{2}$$

Now for any fixed $v_{i+1} \in V_i$, we have

$$\mathbb{E}_x[b^{S_i(x)+S_i(x+v_{i+1})}] = \mathbb{E}_x[b^{2S_i(x)}] \ge \mathbb{E}_x[b^{S_i(x)}]^2 = \Phi_i^2 \tag{3}$$

by Claim 1 and convexity. By an averaging argument, (2) and (3) imply

$$\begin{aligned} \Phi_i^2 &\ge \mathbb{E}_{x,v_{i+1}}[b^{S_i(x)+S_i(x+v_{i+1})} | v_{i+1} \notin V_i] \\ &= \mathbb{E}_{x,v_{i+1}}[b^{S_{i+1}(x)} | v_{i+1} \notin V_i] \\ &= \mathbb{E}_{v_{i+1}}[\Phi_{i+1} | v_{i+1} \notin V_i], \end{aligned}$$

where the first equality follows from Claim 1. □

At a high level, we will use Lemma 3 to upper bound the potential, from which Lemma 1 implies a small max-load.

While the above suffices to get decent tail bounds on the max-load, for technical reasons we will need the following lemma, which is essential to establishing optimal expected max-load and quantitatively stronger two-sided bounds.

Lemma 4. *Let $v_{\le i} \in (\mathbb{F}_2^u)^i$, and $v_{i+1} \in \mathbb{F}_2^u \setminus V_i$. For any $b \ge 0$ we have $\Phi_{i+1} - 1 \ge 2(\Phi_i - 1)$. When $b \le 1$, we also have $\Phi_{i+1} \le \Phi_i$.*

Proof. By Claim 1, elementary manipulations, and linearity of expectation,

$$\begin{aligned}\Phi_{i+1} - 1 &= \mathbb{E}_x[b^{S_i(x)+S_i(x+v_{i+1})} - 1] \\ &= \mathbb{E}_x[b^{S_i(x)} - 1] + \mathbb{E}_x[b^{S_i(x+v_{i+1})} - 1] + \mathbb{E}_x[b^{S_i(x)+S_i(x+v_i)} - b^{S_i(x)} - b^{S_i(x+v_{i+1})} + 1] \\ &= 2\mathbb{E}_x[b^{S_i(x)} - 1] + \mathbb{E}_x[(b^{S_i(x)} - 1)(b^{S_i(x+v_{i+1})} - 1)] \\ &\ge 2(\Phi_i - 1),\end{aligned}$$

where the inequality follows from the fact for any $b, r, s \ge 0$, $b^r - 1$ and $b^s - 1$ have the same sign. When $b \le 1$, Claim 1 and the fact $f(r) = b^r$ is decreasing tells us

$$\Phi_{i+1} = \mathbb{E}_x[b^{S_i(x)+S_i(x+v_{i+1})}] \le \mathbb{E}_x[b^{S_i(x)}] = \Phi_i.$$

□

4 Warmup: Optimal Max-Load With .99 Probability

In this section, we will use the potentials we constructed to show that a random linear map has maximum load $O\left(\frac{\log n}{\log\log n}\right)$ with probability 0.99.

4.1 The Existential Case

To give intuition on how the potential functions will be used, we first prove a preliminary result. We will show that for any choice of n balls, there *exists* a linear hash map that has a maximum load of $\frac{2\ln n}{\ln\ln n}$. To the best of our knowledge, even this existential result was not known prior to our work. Recall Definition 1.

Theorem 6. *Let $u \ge \ell \ge 1$ be integers, $n := 2^\ell$, and $S \subseteq \mathbb{F}_2^u$ be of cardinality n. There exists a linear map $h : \mathbb{F}_2^u \to \mathbb{F}_2^\ell$ such that*

$$M(S, h) < \frac{2\ln n}{\ln\ln n}.$$

Proof. Define $k := u - \ell$. We will carefully pick linearly independent kernel vectors $v_{\le k}$ from $\mathbb{F}_2^u$, and then argue that any linear $h : \mathbb{F}_2^u \to \mathbb{F}_2^\ell$ with kernel $V := \mathrm{Span}(v_{\le k})$ will have small maximum load.

Using the vectors $v_{\le k}$ and set S, define $\{S_i(x) := |(x + V_i) \cap S|\}_{0 \le i \le k}$ and define potentials $\{\Phi_i := \mathbb{E}_{x \sim \mathbb{F}_2^u}[b^{S_i(x)}]\}_{0 \le i \le k}$ for $b = \ln n$. As $|S| = n$, we can compute

$$\Phi_0 = \mathbb{E}_x[b^{S_0(x)}] = \frac{n}{2^u} \cdot b + \left(1 - \frac{n}{2^u}\right) \cdot 1 \le 1 + \frac{\ln n}{2^k}.$$

By Lemma 3, for $i = 1, 2, \ldots, k$ we can iteratively pick $v_{i+1} \notin \mathrm{Span}(v_{\le i})$ such that $\Phi_{i+1} \le \Phi_i^2$. Upon picking these k linearly independent vectors $v_{\le k}$ in this manner, we claim any h with $\mathrm{Span}(v_{\le k}) =: V$ as the kernel will yield the desired result. To see this, first notice

$$\Phi_k \le \Phi_{k-1}^2 \le \cdots \le \Phi_0^{2^k} \le \left(1 + \frac{\ln n}{2^k}\right)^{2^k} < e^{\ln n} = n.$$

Assume there was y such that $|h^{-1}(y) \cap S| \ge \frac{2\ln n}{\ln\ln n}$. Then there must exist x such that $|(x + V) \cap S| \ge \frac{2\ln n}{\ln\ln n}$. But by Lemma 1, this implies

$$\Phi_k \ge \frac{(\ln n)^{\frac{2\ln n}{\ln\ln n}}}{2^{u-k}} = \frac{n^2}{2^\ell} = n,$$

absurd! □

4.2 A Tail Bound to Boost Existence to Abundance

The above theorem showed the existence of a choice of kernel vectors $v_1, \dots, v_k$ that minimized the potential function, thereby implying the existence of a load-balancing linear map. We would like to show most choices of $v_{\le k}$ will minimize the potential. To do so, we will prove a technical lemma that converts the conditional expectation upper bound guarantees from Lemma 3 into tail bounds on the final potential. In particular, while the last section showed the existence of $v_{\le k}$ such that $\Phi_k \le \Phi_0^{2^k}$, the following lemma shows that most choices of $v_{\le k}$ will have $\Phi_k \le (\Phi_0^{2^k})^{O(1)}$.

Lemma 5. *Let $X_0 \ge 1$ be a constant, and let $X_1, \dots, X_k \ge 1$ be random variables satisfying $\mathbb{E}[X_{i+1} | X_{\le i}] \le X_i^2$. For any $t > 1$,*

$$\Pr[X_k \ge t^{2^{k-1}}] \le \frac{X_0^2 - 1}{t - 1}.$$

Remark 1. *Lemma 5 is tight for the sequence*

- $X_1 = \begin{cases} t & \text{with probability } \frac{X_0^2-1}{t-1} \\ 1 & \text{otherwise} \end{cases}$,
- *For $i > 1$, $X_{i+1} = X_i^2$.*

Proof of Lemma 5. We proceed by induction on k. For $k = 1$, it follows by Markov's inequality and the fact $X_1 - 1$ is a nonnegative random variable that

$$\Pr[X_1 \ge t] = \Pr[X_1 - 1 \ge t - 1] \le \frac{X_0^2 - 1}{t - 1}.$$

Now assume we have the lemma for k and wish to prove it for $k + 1$. We can bound

$$\Pr[X_{k+1} \ge t^{2^k}] = \mathbb{E}_{X_1}\left[\Pr\left[X_{k+1} \ge (t^2)^{2^{k-1}}\right] | X_1\right] \le \mathbb{E}_{X_1}\left[\min\left(1, \frac{X_1^2 - 1}{t^2 - 1}\right)\right],$$

where the inequality follows from the inductive hypothesis and the fact that all probabilities are at most 1. To bound the above expression, we will first upper bound the argument of $\mathbb{E}_{X_1}[\cdot]$ in the domain $X_1 \ge 1$ by a linear function.

Claim 2. *Let $x \ge 1$ and $t > 1$. We have*

$$\min\left(1, \frac{x^2 - 1}{t^2 - 1}\right) \le \frac{x - 1}{t - 1}.$$

Proof. If $1 \le x \le t$, then $\frac{x^2-1}{t^2-1} \le 1$, $\frac{x+1}{t+1} \le 1$, and $\frac{x-1}{t-1} \ge 0$. Hence

$$\min\left(1, \frac{x^2 - 1}{t^2 - 1}\right) = \frac{x^2 - 1}{t^2 - 1} = \left(\frac{x - 1}{t - 1}\right)\left(\frac{x + 1}{t + 1}\right) \le \frac{x - 1}{t - 1}.$$

If $x > t$, we have $\frac{x-1}{t-1} > 1 \ge \min\left(1, \frac{x^2-1}{t^2-1}\right)$. □

With this linear upper bound, we can use the bound on $\mathbb{E}[X_1]$ and linearity of expectation to finish the inductive step as follows.

$$\Pr[X_{k+1} \ge t^{2^k}] \le \mathbb{E}_{X_1}\left[\min\left(1, \frac{X_1^2 - 1}{t^2 - 1}\right)\right] \le \mathbb{E}_{X_1}\left[\frac{X_1 - 1}{t - 1}\right] \le \frac{X_0^2 - 1}{t - 1}.$$

The desired result follows. □

4.3 Using the Tail Bound to Analyze Potentials

With Lemma 5 in hand, we can now revisit the proof strategy of Theorem 6, and use the tail bound to argue the potentials are minimized with high probability.

Theorem 7. *Let $\ell \geq 1, u \geq \ell + \log \ell$ be integers, and $n \coloneqq 2^\ell$. Let $S \subseteq \mathbb{F}_2^u$ be of cardinality n, and let $\mathcal{H}$ be the set of all surjective linear maps $h : \mathbb{F}_2^u \to \mathbb{F}_2^\ell$. For any $r \geq 1$,*

$$\Pr_{h \sim \mathcal{H}} \left[M(S,h) \geq r \cdot \frac{\log n}{\log \log n} \right] \leq \frac{3 \log e}{2(r-1)}.$$

In the next section, we will establish a stronger tail bound that decays quadratically (Theorem 9). Theorem 7 suffices for a high constant probability guarantee.

Proof. Let $k \coloneqq u - \ell$. To pick $h \sim \mathcal{H}$, we will pick linearly independent vectors $v_1, \ldots v_k \in \mathbb{F}_2^u$ uniformly at random, and then pick a uniformly random h with kernel $V \coloneqq \mathrm{Span}(v_{\leq k})$. In particular, for $i = 1, \ldots, k$ we will iteratively pick $v_i \sim \mathbb{F}_2^u \setminus V_{i-1}$. Using S and $v_{\leq k}$, define the functions $\{S_i(x) \coloneqq |(x + V_i) \cap S|\}_{0 \leq i \leq k}$, and the potentials $\{\Phi_i(x) \coloneqq \mathbb{E}_{x \sim \mathbb{F}_2^u}[\ell^{\overline{S}_i(x)}]\}_{i \in [k]}$. If some preimage of h had load $r\ell/\log \ell$, then $|(x+V) \cap S| \geq r\ell/\log \ell$ for some shift x. Lemma 1 then implies

$$\Phi_k \geq \frac{\ell^{r\ell/\log \ell}}{2^{u-k}} = \frac{2^{r\ell}}{2^\ell} = n^{r-1}.$$

We can easily compute

$$\Phi_0 = \mathbb{E}_x[\ell^{S(x)}] = 1 - \frac{n}{2^u} + \frac{n\ell}{2^u} < 1 + \frac{\ell}{2^k}.$$

Consequently, $\Phi_0^2 \leq 1 + 2\ell/2^k + (\ell/2^k)^2 \leq 1 + 3\ell/2^k$, where we used the assumption $u \geq \ell + \log \ell \iff \ell/2^k \leq 1$. We already know by Lemma 3 that $\mathbb{E}[\Phi_{i+1} | \Phi_{\leq i}] \leq \Phi_i^2$. Hence, by Lemma 5 (where all $\Phi_i \geq 1$ since $\ell \geq 1$), we have

$$\begin{aligned}
\Pr\left[M(S,h) \geq \frac{r\ell}{\log \ell} \right] \leq \Pr[\Phi_k \geq n^{r-1}] &\leq \frac{\Phi_0^2 - 1}{n^{(r-1)/2^{k-1}} - 1} \\
&\leq \frac{3\ell/2^k}{(r-1)(\ln n)/2^{k-1}} && (e^x - 1 \geq x) \\
&= \frac{3 \log e}{2(r-1)}.
\end{aligned}$$

□

5 Optimal Average Max-Load

In this section, we refine the techniques from Section 4 to prove tail bounds strong enough to establish our main result: optimal expected max-load. The tail bound in Theorem 7 decays too slowly to imply this expectation bound. At first glance, strengthening the tail bound appears difficult, as the technical lemma underpinning it, Lemma 5, is tight by Remark 1. Crucially, our potential functions satisfy a *strong* monotonicity property that Lemma 5 does not exploit: $X_{i+1} - 1 \geq 2(X_i - 1)$ for all $i \geq 1$ (Lemma 4). Under this assumption, we can prove a quadratically stronger version of Lemma 5 (see Theorem 8).

Interestingly, the conditions required for quadratically decaying appear to be quite delicate for two reasons.

- A stronger version of Lemma 5 is not possible under the weaker and more standard monotonicity property $X_{i+1} \geq X_i$. There exists a random sequence (X_i) with $\mathbb{E}[X_{i+1}|X_{\leq i}] \leq X_i^2$ and $X_{i+1} \geq X_i$ that asymptotically saturates the bound of Lemma 5.
- The tight example from Remark 1 satisfies $X_{i+1} - 1 \geq 2(X_i - 1)$ for all $i \geq 1$, but not for $i = 0$—showing that even a single exceptional timestep can eliminate any asymptotic improvement on the tails.

5.1 A Stronger Tail Bound

The main result we will show is the following.

Theorem 8. *Let $X_0 > 1$ be a constant, and let $X_1, \ldots, X_k$ be random variables satisfying $X_{i+1} - 1 \geq 2(X_i - 1)$ and $\mathbb{E}[X_{i+1}|X_{\leq i}] \leq X_i^2$ for all i. For $1 + 4(X_0 - 1) \leq t \leq 2$, we have*

$$\Pr[X_k \geq t^{2^{k-1}}] \leq 48 \left(\frac{X_0 - 1}{t - 1}\right)^2.$$

We would like to prove the above statement using induction, similar to the proof of Lemma 5. However, to make the induction go through, we need to strengthen our inductive hypothesis. For each $i \geq 0$, define the functions

$$\beta_i(t, \delta) = \frac{t^{2^i} - (1 + 2^i\delta)^2}{t^{2^i} - (1 + 2^{i+1}\delta)}.$$

The following stronger claim will be easier to prove via induction.

Lemma 6. *Let $\delta > 0$ and $X_0 > 1$ be constants, and let $X_1, \ldots, X_k$ be random variables such that $\forall i \geq 0$, $X_i \geq 1 + 2^i\delta$, and $\mathbb{E}[X_{i+1}|X_{\leq i}] \leq X_i^2$. For any $t > 1 + 2\delta$,*

$$\Pr[X_k \geq t^{2^{k-1}}] \leq 1 - \frac{t - X_0^2}{t - 1 - 2\delta} \prod_{i=1}^{k-1} \beta_i(t, \delta).$$

Remark 2. *Lemma 6 is tight for the sequence*

- *$X_0 = 1 + \delta$, and $X_1 = \begin{cases} 1 + 2\delta & \text{with probability } \frac{t - X_0^2}{t - 1 - 2\delta} \\ t & \text{otherwise} \end{cases}$,*
- *for $i \geq 1$,*
  - *if $X_i = t^{2^{i-1}}$, set $X_{i+1} = t^{2^i}$*
  - *if $X_i = 1 + 2^i\delta$, set $X_{i+1} = \begin{cases} 1 + 2^{i+1}\delta & \text{with probability } \beta_i(t, \delta) \\ t^{2^i} & \text{otherwise} \end{cases}$.*

Remark 3. *Setting $\delta = 0$ recovers Lemma 5.*

Before we start the proof, we will establish some preliminary properties of $\beta_i(t, \delta)$.

Claim 3. $\beta_i(t^2, 2\delta) = \beta_{i+1}(t, \delta)$

Proof.

$$\beta_i(t^2, 2\delta) = \frac{(t^2)^{2^i} - (1 + 2^i(2\delta))^2}{(t^2)^{2^i} - (1 + 2^{i+1}(2\delta))} = \frac{t^{2^{i+1}} - (1 + 2^{i+1}\delta)^2}{t^{2^{i+1}} - (1 + 2^{i+2}\delta)} = \beta_{i+1}(t, \delta).$$

□

Claim 4. *For all $i \geq 1$ and $t > 1 + 2\delta$, $0 \leq \beta_i(t, \delta) \leq 1$.*

Proof. For the lower bound, note $t^{2^i} \geq 1 + 2^{i+1}\delta$ for all i (by Fact 2). Consequently, the numerator and denominator of $\beta_i(t, \delta)$ are both positive. For the upper bound, we see $1 - \beta_i(t, \delta) = \frac{4^i\delta^2}{t^{2^i} - (1+2^{i+1}\delta)}$. The numerator is trivially nonnegative, and the denominator is positive by Fact 2. □

With these tools, we are ready to prove Lemma 6.

Proof of Lemma 6. We apply induction on k. For the base case $k = 1$, we apply Markov's inequality on the nonnegative random variable $X_1 - 1 - 2\delta$ to yield

$$\Pr[X_1 \geq t] = \Pr[X_1 - 1 - 2\delta \geq t - 1 - 2\delta] \leq \frac{X_0^2 - 1 - 2\delta}{t - 1 - 2\delta} = 1 - \frac{t - X_0^2}{t - 1 - 2\delta}.$$

Note the implicit use of $t > 1 + 2\delta$ in the application of Markov's inequality.

Now assume the lemma for k. We will now prove it for $k + 1$. Write the tail probability as

$$\Pr[X_{k+1} \geq t^{2^k}] = \mathbb{E}_{X_1}[\Pr[X_{k+1} \geq (t^2)^{2^{k-1}} | X_1]]. \tag{4}$$

For fixed X_1, it follows that $X_2, \ldots, X_{k+1}$ is a sequence of random variables of length k satisfying $X_{i+1} \geq 1 + 2^i(2\delta)$ and $\mathbb{E}[X_{i+2} | X_1, X_2, \ldots, X_{i+1}] \leq X_{i+1}^2$ for all $i \in [k]$. Furthermore, $t^2 \geq 1 + 2(2\delta)$ by Fact 2. Hence, all assumptions of the inductive hypothesis are satisfied with the instantiation $\delta \leftarrow 2\delta$, $t \leftarrow t^2$. Utilizing this as well as the fact all probabilities are bounded by 1, we obtain

$$\begin{aligned}
\mathbb{E}_{X_1}[\Pr[X_{k+1} \geq (t^2)^{2^{k-1}} | X_1]] &\leq \mathbb{E}_{X_1}\left[\min\left(1, 1 - \frac{t^2 - X_1^2}{t^2 - 1 - 4\delta}\prod_{i=1}^{k-1}\beta_i(t^2, 2\delta)\right)\right] \\
&= 1 - \mathbb{E}_{X_1}\left[\max\left(0, \frac{t^2 - X_1^2}{t^2 - 1 - 4\delta}\prod_{i=2}^{k}\beta_i(t, \delta)\right)\right] \\
&= 1 - \mathbb{E}_{X_1}\left[\max\left(0, \frac{t^2 - X_1^2}{t^2 - 1 - 4\delta}\right)\right]\prod_{i=2}^{k}\beta_i(t, \delta) \qquad (5)
\end{aligned}$$

where the last equality used the fact $\prod_{i=2}^{k}\beta_i(t, \delta) \geq 0$. To bound this expression, we would like to mimic the intuition of Claim 2 and bound the argument of $\mathbb{E}_{X_1}[\cdot]$ by a linear function in the domain $X_1 \geq 1 + 2\delta$. The following claim does so.

Claim 5. *Let $\delta \geq 0$, $x \geq 1 + 2\delta$, and $t > 1 + 2\delta$. Then*

$$\max\left(0, \frac{t^2 - x^2}{t^2 - 1 - 4\delta}\right) \geq \left(\frac{t - x}{t - 1 - 2\delta}\right)\beta_1(t, \delta).$$

Assuming this claim is true for now (see Remark 4 for geometric intuition), we can use the fact $X_1 \geq 1 + 2\delta$, $\beta_i(t, \delta) \geq 0$, Claim 5, and linearity of expectation to bound

$$\begin{aligned}
1 - \mathbb{E}_{X_1}\left[\max\left(0, \frac{t^2 - X_1^2}{t^2 - 1 - 4\delta}\right)\right]\prod_{i=2}^{k}\beta_i(t, \delta) &\leq 1 - \mathbb{E}_{X_1}\left[\frac{t - X_1}{t - 1 - 2\delta} \cdot \beta_1(t, \delta)\right]\prod_{i=2}^{k}\beta_i(t, \delta) \\
&\leq 1 - \frac{t - X_0^2}{t - 1 - 2\delta}\prod_{i=1}^{k}\beta_i(t, \delta). \qquad (6)
\end{aligned}$$

Combining (4),(5), and (6) gives us

$$\Pr[X_{k+1} \geq t^{2^k}] \leq 1 - \frac{t - X_0^2}{t - 1 - 2\delta} \prod_{i=1}^{k} \beta_i(t, \delta).$$

The desired result follows by induction. □

We now prove Claim 5. Notice when $\delta = 0$ this is exactly Claim 2.

Proof of Claim 5. If $x > t$, then $\frac{t-x}{t-1-2\delta} < 0$ as $t > 1 + 2\delta$. Since we know $\beta_i(t, \delta) \geq 0$ by Claim 4, it follows

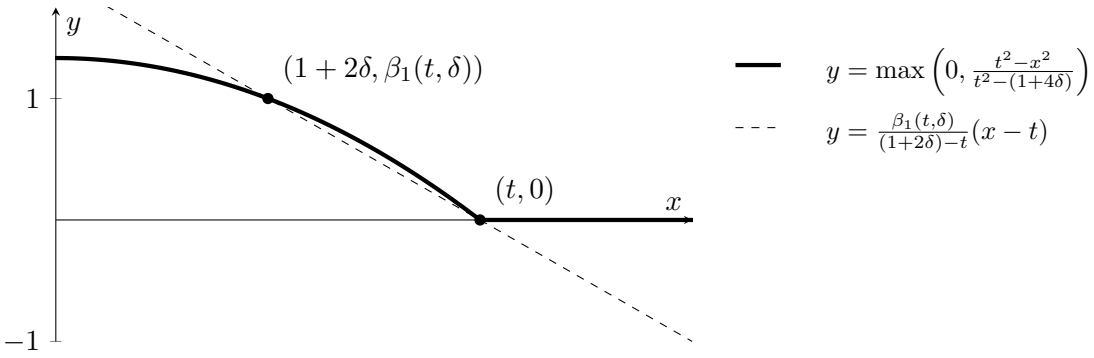
$$\left(\frac{t - x}{t - 1 - 2\delta}\right) \beta_1(t, \delta) < 0 \leq \max\left(0, \frac{t^2 - x^2}{t^2 - 1 - 4\delta}\right).$$

If $x \leq t$, then $\frac{t-x}{t-1-2\delta} \geq 0$ and $\frac{t^2-x^2}{t^2-1-4\delta} \geq 0$, since $t > 1 + 2\delta$ and $t^2 > 1 + 4\delta$ (by Fact 2). Hence,

$$\begin{aligned}
\max\left(0, \frac{t^2 - x^2}{t^2 - 1 - 4\delta}\right) = \frac{t^2 - x^2}{t^2 - 1 - 4\delta} &= \left(\frac{t - x}{t - 1 - 2\delta}\right)\left(\frac{(t + x)(t - 1 - 2\delta)}{t^2 - 1 - 4\delta}\right) \\
&\geq \left(\frac{t - x}{t - 1 - 2\delta}\right) \frac{(t + 1 + 2\delta)(t - 1 - 2\delta)}{t^2 - 1 - 4\delta} \\
&= \left(\frac{t - x}{t - 1 - 2\delta}\right)\left(\frac{t^2 - (1 + 2\delta)^2}{t^2 - 1 - 4\delta}\right) \\
&= \left(\frac{t - x}{t - 1 - 2\delta}\right) \beta_1(t, \delta).
\end{aligned}$$

□

Remark 4. *As per the diagram below, the LHS of Claim 5 is a downward-facing parabola which flattens out to 0 for $x \geq t$. By convexity, a lower bound for $x \geq 1 + 2\delta$ will be the line that intersects the parabola at $x = 1 + 2\delta$ and $x = t$. The equation of this line is exactly the RHS.*



At this point, we have unconditionally proven Lemma 6. Using Lemma 6, we can now show Theorem 8 is true.

Proof of Theorem 8. Let $\delta \coloneqq X_0 - 1$. Notice that by composing the assumed inequality, the random variables satisfy $X_i - 1 \geq 2^i(X_0 - 1) = 2^i\delta$ for each i. Hence, we can apply Lemma 6 and use the fact $\beta_0(t, \delta) = \frac{t-(1+\delta)^2}{t-1-2\delta} = \frac{t-X_0^2}{t-1-2\delta}$ to yield

$$\Pr[X_k \geq t^{2^{k-1}}] \leq 1 - \left(\frac{t - X_0^2}{t - 1 - 2\delta}\right) \prod_{i=1}^{k} \beta_i(t, \delta) = 1 - \prod_{i=0}^{k} \beta_i(t, \delta).$$

Since Claim 4 tells us $0 \le \beta_i(t,\delta) \le 1$ for all i, we can treat these quantities as probabilities. Consider $k+1$ independent and biased coins, where coin i has probability $\beta_i(t,\delta)$ of showing heads. By the union bound on the event that at least one tail shows, we have

$$1 - \prod_{i=0}^{k} \beta_i(t,\delta) \le \sum_{i=0}^{k} (1 - \beta_i(t,\delta)) = \sum_{i=0}^{k} \frac{4^i \delta^2}{t^{2^i} - 1 - 2^{i+1}\delta} \le 2\delta^2 \sum_{i=0}^{k} \frac{4^i}{t^{2^i} - 1},$$

where we used the fact that for all i, $t^{2^i} - 1 \ge 2^i(t-1) \ge 2^{i+2}\delta$ by Fact 2 and the theorem assumption. For $i \le \log\left(\frac{5}{t-1}\right)$, we have

$$\begin{aligned}
\sum_{i \le \log\left(\frac{5}{t-1}\right)} \frac{4^i}{t^{2^i} - 1} = \sum_{i \le \log\left(\frac{5}{t-1}\right)} \frac{2^i}{\ln t} \cdot \frac{2^i \ln t}{e^{2^i \ln t} - 1} &\le \frac{1}{\ln t} \sum_{i \le \log\left(\frac{5}{t-1}\right)} 2^i && \left(\tfrac{x}{e^x - 1} \le 1\right) \\
&\le \frac{1}{\ln t} \cdot \frac{10}{t-1} \\
&\le \frac{t}{t-1} \cdot \frac{10}{t-1} && \left(1 + x \ge e^{\frac{x}{1+x}}\right) \\
&\le \frac{20}{(t-1)^2},
\end{aligned}$$

For $i > \log\left(\frac{5}{t-1}\right)$, we can bound

$$\begin{aligned}
\sum_{i > \log\left(\frac{5}{t-1}\right)} \frac{4^i}{t^{2^i} - 1} = \sum_{j \ge 1} \frac{4^{\log(5/(t-1))} 4^j}{(t^{5/(t-1)})^{2^j} - 1} &\le \left(\frac{5}{t-1}\right)^2 \sum_{j \ge 1} \frac{4^j}{(1 + \frac{5}{t-1}(t-1))^{2^j} - 1} && \text{(Fact 2)} \\
&\le \left(\frac{5}{t-1}\right)^2 \sum_{j \ge 1} (4/35)^j \\
&\le \frac{4}{(t-1)^2}.
\end{aligned}$$

Hence we have

$$\Pr[X_k \ge t^{2^{k-1}}] \le 2\delta^2 \sum_{i=0}^{k} \frac{4^i}{t^{2^i} - 1} \le 2\delta^2 \left(\frac{20}{(t-1)^2} + \frac{4}{(t-1)^2}\right) = 48 \left(\frac{X_0 - 1}{t-1}\right)^2.$$

□

5.2 Applying the Tail Bound To Our Potentials

We will strengthen the tail bounds of Theorem 7, and also generalize Theorem 7 to the setting of m balls and n bins, for $m \ne n$. For brevity, define

$$\textsc{opt}(m,n) = \begin{cases} \frac{\log n}{\log\left(\frac{n \ln n}{m}\right)} & m \le \frac{1}{2} n \log n \\ \frac{m}{n} & m > \frac{1}{2} n \log n \end{cases}$$

to be the function that outputs the maximum load obtained when a fully random hash maps m balls to n bins.

Theorem 9. *Let $u, \ell, m \ge 1$ be integers such that $u \ge \ell + 2\log(\ell m)$, and let $n := 2^\ell$. Let $\mathcal{H}$ be the set of surjective linear maps $\mathbb{F}_2^u \to \mathbb{F}_2^\ell$. Let $S \subseteq \mathbb{F}_2^u$ be a subset of size m. Then for any real $r \ge 6$,*

$$\Pr_{h \sim \mathcal{H}} [M(S,h) \ge r \cdot \textsc{opt}(m,n)] \le \frac{48}{(r-2)^2}$$

Proof. Define $k \coloneqq u - \ell$. We assume $r \le m$, since $M(S,h) \le |S| = m$ for any h. We will pick $h \sim \mathcal{H}$ by iteratively sampling $v_{i+1} \sim \mathbb{F}_2^u \setminus V_i$, and then picking random h with $\ker(h) = V$. Define the associated functions $\{S_i(x) \coloneqq |(x + V_i) \cap S|\}_{0 \le i \le k}$. We will split into two cases.

Case 1: $m \le \frac{1}{2} n \log n$.

Set $b = n\ell/m$ and $\Phi_i \coloneqq \mathbb{E}_{x \sim \mathbb{F}_2^u}[b^{S_i(x)}]$ for $0 \le i \le k$. Then

$$\Phi_0 = \left(1 - \frac{m}{2^u}\right) \cdot 1 + \frac{m}{2^u} \cdot b \le 1 + \frac{bm}{2^u} = 1 + \frac{\ell}{2^k}.$$

By Lemma 1 and the fact $\textsc{opt}(m,n) = \ell / \log b$ in this case, we have

$$\Pr[M(S,h) \ge r \cdot \textsc{opt}(m,n)] \le \Pr\left[\Phi_k \ge \frac{b^{r \cdot \textsc{opt}(m,n)}}{n}\right] = \Pr[\Phi_k \ge n^{r-1}].$$

Note that for $t = 1 + r\ell/2^k$, $t^{2^{k-1}} \le e^{r\ell/2} < n^{r-1}$. Furthermore, as $4 \le r \le m$ and $u \ge \ell + \log(\ell m)$, we have $t - 1 \ge 4\ell/2^k \ge 4(\Phi_0 - 1)$ and $t - 1 \le mn\ell/2^u \le 1$. Finally, we have $\Phi_{i+1} - 1 \ge 2(\Phi_i - 1)$ for all i by Lemma 4. Hence, we can use Theorem 8 to bound

$$\Pr[\Phi_k \ge n^{r-1}] \le \Pr[\Phi_k \ge t^{2^{k-1}}] \le 48 \left(\frac{\Phi_0 - 1}{t - 1}\right)^2 \le 48 \left(\frac{\ell/2^k}{r\ell/2^k}\right)^2 = \frac{48}{r^2}.$$

Case 2: $m > \frac{1}{2} n \log n$.

Define $\Phi_i \coloneqq \mathbb{E}_{x \sim \mathbb{F}_2^u}[2^{S_i(x)}]$ for $0 \le i \le k$. Then

$$\Phi_0 = 1 - \frac{m}{2^u} + \frac{2m}{2^u} = 1 + \frac{m}{2^u}.$$

By Lemma 1 and the fact $\textsc{opt}(m,n) = m/n$ in this regime,

$$\begin{aligned} \Pr[M(S,h) \ge r \cdot \textsc{opt}(m,n)] &= \Pr[M(S,h) \ge rm/n] \\ &\le \Pr[\Phi_k \ge 2^{rm/n - \ell}] \\ &\le \Pr[\Phi_k \ge 2^{(r-2)m/n}] && (m \ge \tfrac{1}{2} n\ell) \end{aligned}$$

For $t \coloneqq 1 + (r-2)m/2^u$, we have $t^{2^{k-1}} \le e^{(r-2)m/2n} \le 2^{(r-2)m/n}$. Since $6 \le r \le m$ and $u \ge \ell + 2\log m$, we can deduce $t - 1 \ge 4m/2^u \ge 4(\Phi_0 - 1)$ and $t - 1 \le m^2/2^u \le 1$. Furthermore, we have $\Phi_{i+1} - 1 \ge 2(\Phi_i - 1)$ for all i by Lemma 4. Hence, by Lemma 6, we have

$$\Pr[\Phi_k \ge t^{2^{k-1}}] \le 48 \left(\frac{\Phi_0 - 1}{t - 1}\right)^2 \le 48 \left(\frac{m/2^u}{(r-2)m/2^u}\right)^2 = \frac{48}{(r-2)^2}.$$

□

With a simple argument (whose proof is deferred to Appendix A.1), we can remove the artificial lower-bound condition on u and the surjectivity condition on h.

Theorem 10. *Let $u \ge \ell \ge 1$ and $m \le 2^u$ be integers. Let $n \coloneqq 2^\ell$. Let $h : \mathbb{F}_2^u \to \mathbb{F}_2^\ell$ be a random linear map. For any $S \subseteq \mathbb{F}_2^u$ of size m and $r \ge 6$, we have*

$$\Pr_{h \sim \mathcal{H}}[M(S,h) \ge r \cdot \textsc{opt}(m,n)] \le \frac{49}{(r-2)^2}.$$

With these strong tails, optimal expected max-load is a simple corollary.

Theorem 11. *Let $u \ge \ell \ge 1$ be integers, and $n \coloneqq 2^\ell$. For uniformly random linear map $h : \mathbb{F}_2^u \to \mathbb{F}_2^\ell$,*

$$\mathbb{E}_h[M(S,h)] \le 16 \cdot \textsc{opt}(m,n).$$

Proof.

$$\begin{aligned}
\mathbb{E}_h[M(S,h)] &= \int_0^\infty \Pr[M(S,h) \geq t]dt \\
&\leq \int_0^{9\cdot \textsc{opt}(m,n)} 1dt + \int_{9\cdot \textsc{opt}(m,n)}^\infty \Pr[M(S,h) \geq t]dt \\
&\leq 9\textsc{opt}(m,n) + \textsc{opt}(m,n)\int_9^\infty \Pr[M(S,h) \geq r\textsc{opt}(m,n)]dr \\
&\leq 9\textsc{opt}(m,n) + 49\textsc{opt}(m,n)\int_9^\infty \frac{1}{(r-2)^2}dr \qquad \text{(Theorem 10)} \\
&= 16 \cdot \textsc{opt}(m,n).
\end{aligned}$$

□

6 Two-Sided Bounds

In the regime of $m = \Omega(n \log n)$, we can give two-sided bounds on all bins. In particular, we can show for any set of m balls, a random linear map will hash $\Theta(m/n)$ balls to each bin with high probability.

Theorem 12. *Let $0 < \varepsilon < 1/2$ be a constant. There exists constants $C_1 < 1 < C_2$ depending on ε such that for $m \geq C_1^{-1} n \log n$ and any $S \subseteq \mathbb{F}_2^u$ of cardinality m, a uniformly random linear map $h : \mathbb{F}_2^u \to \mathbb{F}_2^\ell$ satisfies*

$$\Pr\left[\forall y \in \mathbb{F}_2^\ell, \frac{C_1 m}{n} \leq |h^{-1}(y) \cap S| \leq \frac{C_2 m}{n}\right] \geq 1 - \varepsilon.$$

Furthermore, $C_1 = \Omega(\varepsilon^{74})$ and $C_2 = O(\varepsilon^{-1/2})$.

We reiterate that the condition $m = \Omega(n \log n)$ is necessary: there exists S of size at least $0.69n \log n$ such that *every* linear map has at least one empty bin ([ADM+97], Proposition 2.2). Interestingly, this two-sided bound is also proven using potential functions. To prove this, we will require the following tail bound.

Lemma 7. *Let $0 < \varepsilon < 1$ and $0 \leq X_0 < 1$ be fixed, and let $1 > X_1 \geq X_2 \geq \cdots > 0$ be random variables satisfying $X_{i+1} \geq 2X_i - 1$ and $\mathbb{E}[X_{i+1}|X_{\leq i}] \leq X_i^2$. Then for $C_\varepsilon = (1 - X_0)^{25}(\varepsilon/2)^{50} \log(1/\varepsilon)$ and any s,*

$$\Pr\left[X_s \geq 2^{-C_\varepsilon 2^s}\right] \leq \varepsilon.$$

We will first prove Theorem 12 assuming this tail bound, and then prove the tail bound afterwards.

Proof of Theorem 12. We first focus on the lower bound. Set $t = \log(4m/\varepsilon)$ and factor $h = h_1 \circ h_2$, where $h_1 : \mathbb{F}_2^u \to \mathbb{F}_2^t$ is a uniformly random linear map, and $h_2 : \mathbb{F}_2^t \to \mathbb{F}_2^\ell$ is a uniformly random *surjective* linear map. h_1 will collide any fixed pair of elements in S with probability $1/2^t$. Hence the expected number of pairwise collisions among S is $\binom{m}{2}\frac{1}{2^t} \leq \frac{m^2}{2^{t+1}}$. Let $\mathcal{E}$ denote the event in which there are at least $m/2$ pairs that collide. By Markov's inequality,

$$\Pr[\mathcal{E}] \leq \frac{m^2}{2^{t+1}} \cdot \frac{2}{m} = \frac{m}{2^t} = \frac{\varepsilon}{4}.$$

Letting $k \coloneqq t - \ell$, we will sample random linearly independent vectors $v_1, \ldots, v_k \in \mathbb{F}_2^t$, and consider h_2 with kernel $V \coloneqq \text{Span}(v_{\leq k})$. Using the set $h_1(S)$ and vectors $v_{\leq k}$, we

construct the associated functions $\{S_i(x) \coloneqq |(x+V_i) \cap h_1(S)|\}_{0\le i\le k}$ and potentials $\{\Phi_i \coloneqq \mathbb{E}_{x\sim\mathbb{F}_2^u}[(1/2)^{S_i(x)}]\}_{0\le i\le k}$.

Conditioning on $\neg\mathcal{E}$, we have $|h_1(S)| \ge m - m/2 = m/2$, and so the density of $h_1(S) \subseteq \mathbb{F}_2^t$ is lower bounded by $\frac{m/2}{4m/\varepsilon} = \varepsilon/8$. Consequently, we have

$$\Phi_0 \le \left(1-\frac{\varepsilon}{8}\right)\cdot 1 + \frac{\varepsilon}{8}\cdot\frac{1}{2} = 1 - \frac{\varepsilon}{16}.$$

Lemma 3 and Lemma 4 show that (Φ_i) satisfy the premise of Lemma 7. Hence, we can use Lemma 7 and the fact $\ell \le C_1 m/n$ to deduce

$$\Pr[\Phi_k > 2^{-C_1 m/n-\ell}] \le \Pr[\Phi_k > 2^{-2C_1 m/n}] = \Pr[\Phi_k > 2^{-(\varepsilon/2)C_1 2^k}] \le \varepsilon/4$$

for $C_1 \coloneqq \frac{2}{\varepsilon}(1-\Phi_0)^{25}(\varepsilon/8)^{50}\log(4/\varepsilon) = \Omega(\varepsilon^{74})$. Now note for all $y \in \mathbb{F}_2^\ell$, we have

$$h^{-1}(y) = \bigsqcup_{z\in h_2^{-1}(y)} h_1^{-1}(z).$$

Hence,

$$|h^{-1}(y)\cap S| = \sum_{z\in h_2^{-1}(y)} |h_1^{-1}(z)\cap S| \ge \sum_{z\in h_2^{-1}(y)} \mathbf{1}(z\in h_1(S)) = |h_2^{-1}(y)\cap h_1(S)|.$$

Therefore, it follows

$$\begin{aligned}
\Pr\left[\forall y\in\mathbb{F}_2^\ell, |h^{-1}(y)\cap S| < C_1\frac{m}{n}\right] &\le \Pr\left[\forall y\in\mathbb{F}_2^\ell, |h_2^{-1}(y)\cap h_1(S)| < C_1\frac{m}{n}\right] \\
&\le \frac{\varepsilon}{4} + \Pr\left[\forall y\in\mathbb{F}_2^\ell, |h_2^{-1}(y)\cap h_1(S)| < C_1\frac{m}{n}|\neg\mathcal{E}\right] \\
&\le \frac{\varepsilon}{4} + \Pr[\Phi_k > 2^{-C_1 m/n-\ell}|\neg\mathcal{E}] \qquad \text{(Lemma 2)} \\
&\le \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}. \qquad (7)
\end{aligned}$$

We already have from Theorem 10 that there exists $C_2 = 7\varepsilon^{-1/2} + 2$ such that $\Pr[M(S,h) > C_2 m/n] \le \varepsilon/2$. Taking a union bound over this and Equation (7) gives the desired result.

□

Proof of Lemma 7. The proof will resemble that of Theorem 7 in [ADM+97]. Call i a *stride* if $X_{i-1} > 1/2$ and $(1-X_i) \ge \frac{5}{4}(1-X_{i-1})$, or if $X_{i-1}\le 1/2$ and $X_i \le \frac{3}{4}X_{i-1}$. We will show on any conditioning of $X_{\le i}$, $i+1$ is a stride with probability $\ge 1/3$. We split into two cases.

- If $X_i \le 1/2$, then $\mathbb{E}[X_{i+1}] \le X_i^2 \le \frac{1}{2}X_i$. Therefore, by Markov's inequality, $\Pr[X_{i+1} > \frac{3}{4}X_i] < \frac{X_i/2}{3X_i/4} = 2/3$, so $i+1$ is a stride with probability at least $1/3$.
- If $X_i > 1/2$, we will instead apply Markov's inequality on $1-2X_i+X_{i+1}$. Note

  $$\mathbb{E}[1-2X_i+X_{i+1}|X_{\le i}] = 1-2X_i+X_i^2 = (1-X_i)^2.$$

  Hence, conditioned on $X_{\le i}$, we have

  $$\Pr_{X_{i+1}}\left[1-X_{i+1} < \frac{5}{4}(1-X_i)\right] = \Pr_{X_{i+1}}\left[1-2X_i+X_{i+1} > \frac{3}{4}(1-X_i)\right] < \frac{(1-X_i)^2}{3(1-X_i)/4} < \frac{2}{3},$$

  so $i+1$ is a stride with probability $\ge 1/3$ in this case as well.

Let j be the first integer such that $X_{j+1} < 1/2$, and let k be the first integer such that $X_{k+1} < \varepsilon^2/8$. Let s_1 be the number of strides in $[j]$, and let s_2 be the number of strides in $\{j+1, \dots, k\}$. We observe that $(1 - X_0)(\frac{5}{4})^{s_1} \le \frac{1}{2}$, and so $s_1 \le \log_{5/4}(1/2(1 - X_0))$. Similarly, we must have $\frac{1}{2}\left(\frac{3}{4}\right)^{s_2} \ge \frac{\varepsilon^2}{8}$, implying that $s_2 \le \log_{4/3}(4/\varepsilon^2)$. Therefore, there must be at most $s_1 + s_2 \le \log_{5/4}(4/(1 - X_0)\varepsilon^2)$ strides in $[k]$. Define $k^* := 8\log_{5/4}(4/(1 - X_0)\varepsilon^2)$. Let $f(X_0, X_1, \dots, X_{k^*})$ evaluate the number of strides in the first k^* steps, and define $Y_i := \mathbb{E}[f(X_{\le k^*})|X_{<i}]$. Notice (Y_i) is a Doob martingale with $|Y_{i+1} - Y_i| \le 1$ for all i. Furthermore, since we showed each index is a stride with probability $\ge 1/3$,

$$Y_0 \ge k^*/3 = (8/3)\log_{5/4}(4/(1 - X_0)\varepsilon^2) \ge 16\log(2/\varepsilon)$$

Hence, by Azuma's inequality (Fact 3) and the above string of inequalities,

$$\begin{aligned}\Pr[k \ge k^*] &\le \Pr\left[f(X_{\le k^*}) \le \log_{5/4}\left(\frac{4}{(1 - X_0)\varepsilon^2}\right)\right] \\ &\le \Pr\left[Y_{k^*} \le (3/8)Y_0\right] \\ &\le e^{-\frac{(5Y_0/8)^2}{2k^*}} \le e^{-\frac{(5/8)^2}{6}(16\log(2/\varepsilon))} \le \varepsilon/2.\end{aligned}$$

If $s \le k^*$, we have just shown with probability at most $\varepsilon/2$ that $X_s \ge X_k \ge \varepsilon^2/8 \ge 2^{-C_\varepsilon 2^s}$ for any $C_\varepsilon \le \log(8/\varepsilon^2)$. Henceforth, we will assume $s > k^*$. Conditioned on $k < k^*$, we have $X_{k^*} < \varepsilon^2/8$. Let $\mathcal{E}_i$ be the event $X_{k^*+i} < \varepsilon^{2^i+1}/2^{i+3}$ and denote $\mathcal{E}_{\le t} = \bigwedge_{i=1}^t \mathcal{E}_i$. Notice that $\mathbb{E}[X_{k^*+i+1}|\mathcal{E}_{\le i}] \le \varepsilon^{2^{i+1}+2}/2^{2i+6}$, so by Markov's inequality,

$$\Pr[\neg\mathcal{E}_{i+1}|\mathcal{E}_{\le i}] \le \frac{\varepsilon^{2^{i+1}+2}/2^{2i+6}}{\varepsilon^{2^{i+1}+1}/2^{i+4}} \le \varepsilon/2^{i+2}.$$

Hence for any t we have

$$\Pr\left[\neg\mathcal{E}_{\le t}\right] = \sum_{i=0}^{t}\Pr\left[\neg\mathcal{E}_{i+1} \wedge \mathcal{E}_{\le i}\right] \le \sum_{i=0}^{t}\Pr[\neg\mathcal{E}_{i+1}|\mathcal{E}_{\le i}] \le \sum_{i\ge 0}\varepsilon/2^{i+2} \le \varepsilon/2.$$

Consequently by a union bound, $k < k^*$ and $\mathcal{E}_{\le(s-k^*)}$ occurs with probability at least $1 - \varepsilon$. In the event this happens, and noting $k^* = 8\log_{5/4}(4/(1 - X_0)\varepsilon^2) \le 25\log(4/(1 - X_0)\varepsilon^2)$, it follows

$$X_s \le \frac{\varepsilon^{2^{s-k^*}+1}}{2^{s-k^*+3}} \le \varepsilon^{2^s((1-X_0)\varepsilon^2/4)^{25}+1} \le 2^{-C_\varepsilon 2^s}$$

for $C_\varepsilon = (1 - X_0)^{25}(\varepsilon/2)^{50}\log(1/\varepsilon)$. The desired result follows.

□

Acknowledgements

We thank Raghu Meka for discussions, one of which led to this question. We also thank Jesse Goodman for discussions, as well as Jeffrey Champion, Sabee Grewal, and Lin Lin Lee for comments that improved our presentation.

References

[ADM+97] Noga Alon, Martin Dietzfelbinger, Peter Bro Miltersen, Erez Petrank, and Gábor Tardos. Is linear hashing good? In *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing (STOC)*, page 465–474, 1997. doi:10.1145/258533.258639. [pp. 1, 3, 4, 5, 16, 17]

[Bab18] Martin Babka. A note on the size of largest bins using placement with linear transformations. 2018. arXiv:1810.04161. [p. 3]

[BGG94] Mihir Bellare, Oded Goldreich, and Shafi Goldwasser. Incremental cryptography: The case of hashing and signing. In Yvo G. Desmedt, editor, *Advances in Cryptology — CRYPTO 1994*, volume 839 of *Lecture Notes in Computer Science*, pages 216–233, 1994. doi:10.1007/3-540-48658-5_22. [p. 2]

[BGG95] Mihir Bellare, Oded Goldreich, and Shafi Goldwasser. Incremental cryptography and application to virus protection. In *Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing (STOC)*, page 45–56, 1995. doi:10.1145/225058.225080. [p. 2]

[CMV13] Kai-Min Chung, Michael Mitzenmacher, and Salil Vadhan. Why simple hash functions work: Exploiting the entropy in a data stream. *Theory of Computing*, 9(30):897–945, 2013. doi:10.4086/toc.2013.v009a030. [p. 2]

[CRSW13] 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. doi:10.1137/120871626. [pp. 2, 3]

[CW79] J.Lawrence Carter and Mark N. Wegman. Universal classes of hash functions. *Journal of Computer and System Sciences*, 18(2):143–154, 1979. doi:10.1016/0022-0000(79)90044-8. [pp. 2, 3]

[DD22] Manik Dhar and Zeev Dvir. Linear hashing with ℓ_∞ guarantees and two-sided kakeya bounds. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 419–428, 2022. doi:10.1109/FOCS54457.2022.00047. [p. 4]

[MCW78] George Markowsky, J. Lawrence Carter, and Mark N. Wegman. Analysis of a universal class of hash functions. In *Mathematical Foundations of Computer Science 1978*, volume 64 of *Lecture Notes in Computer Science*, pages 345–354. Springer, 1978. doi:10.1007/3-540-08921-7_82. [p. 3]

[MRRR14] Raghu Meka, Omer Reingold, Guy N. Rothblum, and Ron D. Rothblum. Fast pseudorandomness for independence and load balancing. In *Proceedings of the 41st International Colloquium on Automata, Languages, and Programming (ICALP)*, volume 8572 of *Lecture Notes in Computer Science*, pages 859–870. Springer, 2014. doi:10.1007/978-3-662-43948-7_71. [p. 2]

[MU05] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, Cambridge, 2005. doi:10.1017/CBO9780511813603. [pp. 1, 6]

[MV84] Kurt Mehlhorn and Uzi Vishkin. Randomized and deterministic simulations of prams by parallel machines with restricted granularity of parallel memories. *Acta Informatica*, 21(4):339–374, 1984. doi:10.1007/BF00264615. [p. 3]

[PT12] Mihai Pătraşcu and Mikkel Thorup. The power of simple tabulation hashing. *J. ACM*, 59(3), June 2012. doi:10.1145/2220357.2220361. [p. 2]

[RS98] Martin Raab and Angelika Steger. Balls into bins – a simple and tight analysis. In *Randomization and Approximation Techniques in Computer Science (RANDOM 1998)*, volume 1518 of *Lecture Notes in Computer Science*, pages 159–170. Springer, 1998. doi:10.1007/3-540-49543-6_13. [p. 3]

[WC81] Mark N. Wegman and J. Lawrence Carter. New hash functions and their use in authentication and set equality. *Journal of Computer and System Sciences*, 22(3):265–279, 1981. doi:10.1016/0022-0000(81)90033-7. [p. 2]

[Woo14] Mary Katherine Wootters. *Any Errors in This Dissertation Are Probably Fixable: Topics in Probability and Error Correcting Codes.* Ph.d. dissertation, University of Michigan, 2014. URL: https://deepblue.lib.umich.edu/handle/2027.42/108844. [p. 3]

[WZ96] Peter Winkler and David Zuckerman. Multiple cover time. *Random Structures & Algorithms*, 9(4):403–411, 1996. doi:10.1002/(SICI)1098-2418(199612)9:4<403::AID-RSA4>3.0.CO;2-0. [p. 4]

A Deferred Proofs

A.1 Proof of Theorem 10: Removing the Surjectivity Assumption

Proof of Theorem 10. We will first show the result for u large enough, i.e. $2^{u-\ell} \geq \max\{(m-2)^2, (m\ell)^2\}$. At the end, we will remove this assumption on u. Let $\mathcal{H}$ be the set of linear maps $\mathbb{F}_2^u \to \mathbb{F}_2^\ell$, and let $\mathcal{K}$ be the distribution of the nullity of $h \sim \mathcal{H}$. If k is the nullity of h, then the number of universe elements $v \in \mathbb{F}_2^u$ such that $h(v) = 0$ is 2^k. For $h \sim \mathcal{H}$ we have $\Pr[h(v) = 0] = 2^{-\ell}$ for fixed $v \neq 0$. Hence, we can compute by linearity that

$$\mathbb{E}_{k\sim K}[2^k] = \mathbb{E}_{h\sim\mathcal{H}}\left[\sum_{v\in\mathbb{F}_2^u} \mathbf{1}\{h(v)=0\}\right] = \sum_{v\in\mathbb{F}_2^u} \Pr_h[h(v)=0] = 1 + 2^{-\ell}(2^u - 1) \leq 2^{u-\ell} + 1.$$

Let $\mathcal{E}$ be the event that h is surjective, i.e. $k = u - \ell$. By Markov's inequality on the random variable $2^k - 2^{u-\ell}$ (as $k \geq u - \ell$), we have

$$\Pr[\neg\mathcal{E}] = \Pr_{k\sim\mathcal{K}}[k \geq u - \ell + 1] = \Pr_{k\sim\mathcal{K}}\left[2^k - 2^{u-\ell} \geq 2^{u-\ell}\right] \leq \frac{1}{2^{u-\ell}} \leq \frac{1}{(m-2)^2} \leq \frac{1}{(r-2)^2}.$$

For brevity, set $M \coloneqq r \cdot \textsc{opt}(m,n)$. We can use Theorem 9, the above observation, and the fact $u \geq \ell + 2\log(\ell m)$ to deduce

$$\Pr_{h\sim\mathcal{H}}[M(S,h) \geq M] \leq \Pr[\neg\mathcal{E}] + \Pr_{h\sim\mathcal{H}}[M(S,h) \geq M|\mathcal{E}] \leq \frac{1}{(r-2)^2} + \frac{48}{(r-2)^2} = \frac{49}{(r-2)^2}.$$

We will now remove the lower bound assumption on u. Intuitively, for any $\ell \leq u' < u$ we can simply embed $\mathbb{F}_2^{u'}$ into $\mathbb{F}_2^u$, and then port in the max-load result for $\mathbb{F}_2^u$.

More formally, let $\mathcal{H}'$ be the set of linear maps $\mathbb{F}_2^{u'} \to \mathbb{F}_2^\ell$. Take an arbitrary u'-dimensional subspace $V \leq \mathbb{F}_2^u$. There is an isomorphism $\iota : \mathbb{F}_2^{u'} \to V$. Denote $T \coloneqq \iota(S)$. Since $T \subseteq V$, we have for any $h : \mathbb{F}_2^u \to \mathbb{F}_2^\ell$ that

$$M(T, h|_V) = \max_{y\in\mathbb{F}_2^\ell} |(h|_V)^{-1}(y) \cap T| = \max_{y\in\mathbb{F}_2^\ell} |h^{-1}(y) \cap V \cap T| = \max_{y\in\mathbb{F}_2^\ell} |h^{-1}(y) \cap T| = M(T,h).$$

For $h \sim \mathcal{H}$, $h|_V : V \to \mathbb{F}_2^\ell$ will be a uniform random linear map. Consequently,

$$\Pr_{h'\sim\mathcal{H}'}[M(S,h') \geq M] = \Pr_{h\sim\mathcal{H}}[M(T,h|_V) \geq M] = \Pr_{h\sim\mathcal{H}}[M(T,h) \geq M] \leq \frac{49}{(r-2)^2}$$

since T has cardinality m. □

ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il