# Linear Hash Functions

Noga Alon* Martin Dietzfelbinger† Peter Bro Miltersen‡ Erez Petrank§
Gábor Tardos¶

**Abstract**

Consider the set $\mathcal{H}$ of all linear (or affine) transformations between two vector spaces over a finite field $F$. We study how good $\mathcal{H}$ is as a class of hash functions, namely we consider hashing a set $S$ of size $n$ into a range having the same cardinality $n$ by a randomly chosen function from $\mathcal{H}$ and look at the expected size of the largest hash bucket. $\mathcal{H}$ is a universal class of hash functions for any finite field, but with respect to our measure different fields behave differently.

If the finite field $F$ has $n$ elements then there is a bad set $S \subset F^2$ of size $n$ with expected maximal bucket size $\Omega(n^{1/3})$. If $n$ is a perfect square then there is even a bad set with largest bucket size *always* at least $\sqrt{n}$. (This is worst possible, since with respect to a universal class of hash functions every set of size $n$ has expected largest bucket size below $\sqrt{n} + 1/2$.)

If, however, we consider the field of two elements then we get much better bounds. The best previously known upper bound on the expected size of the largest bucket for this class was $O(2^{\sqrt{\log n}})$. We reduce this upper bound to $O(\log n \log\log n)$. Note that this is not far from the guarantee for a random function. There, the average largest bucket would be $\Theta(\log n / \log\log n)$.

In the course of our proof we develop a tool which may be of independent interest. Suppose we have a subset $S$ of a vector space $D$ over $\mathbf{Z}_2$, and consider a random linear mapping of $D$ to a smaller vector space $R$. If the cardinality of $S$ is larger than $c_\epsilon |R| \log |R|$ then with probability $1 - \epsilon$, the image of $S$ will cover all elements in the range.

## 1 Introduction

Consider distributing $n$ balls in $s$ buckets, randomly and independently. The resulting distribution of the balls in the buckets is the object of occupancy theory.

*Dep. of Math., Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel and Institute for Advanced Study, Princeton, NJ 08540. Research supported in part by a USA-Israeli BSF grant, by the Sloan Foundation grant No. 96-6-2, by an NEC Research Institute grant and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University. E-mail: noga@math.tau.ac.il.

†Fakultät für Informatik und Automatisierung, Technische Universität Ilmenau, Postfach 100565, 98684 Ilmenau, Germany. This work was done while the author was affiliated with the University of Dortmund. Partially supported by DFG grant Di 412/5-1. E-mail: Martin.Dietzfelbinger@theoinf.tu-ilmenau.de.

‡BRICS, Centre of the Danish National Research Foundation, University of Aarhus, Ny Munkegade, Aarhus, Denmark. Supported by the ESPRIT Long Term Research Programme of the EU under project number 20244 (ALCOM-IT). E-mail: bromille@brics.dk. Part of this work was done while the author was at the University of Toronto.

§IBM Haifa Research Lab, MATAM, Haifa 31905, Israel. E-mail: erezp@haifa.vnet.ibm.com. Most of this work was done while the author was visiting the University of Toronto.

¶Rényi Institute of the Hungarian Academy of Sciences, Pf. 127, Budapest, H-1364 Hungary. Partially supported by DIMACS Center and the grants OTKA T-020914, T-030059 and FKFP 0607/1999. E-mail: tardos@cs.elte.hu. Part of this work was done while the author was visiting the University of Toronto and the Institute for Advanced Study, Princeton.

In the theory of algorithms and in complexity theory, it is often necessary and useful to consider putting balls in buckets without complete independence. More precisely, the following setting is studied: A class $\mathcal{H}$ of hash functions, each mapping a universe $U$ to $\{1, 2, \ldots, s\}$, is fixed. A set $S \subseteq U$ to be hashed is given by an adversary, a member $h \in \mathcal{H}$ is chosen uniformly at random, $S$ is hashed using $h$, and the distribution of the multi-set $\{h(x) | x \in S\}$ is studied. If the class $\mathcal{H}$ is the class of all functions between $U$ and $\{1, 2, \ldots, s\}$, we get the classical occupancy problems. Carter and Wegman defined a class $\mathcal{H}$ to be universal if

$$\forall x \neq y \in U : \text{Prob}(h(x) = h(y)) \leq 1/s.$$

We remark that a stricter definition is often used in the complexity theory literature.

For universal families, the following properties are well known; variations of them have been used extensively in various settings:

1. If $S$ of size $n$ is hashed to $n^2$ buckets, with probability more than $1/2$, no collision occurs.
2. If $S$ of size $2n^2$ is hashed to $n$ buckets, with probability more than $1/2$, every bucket receives an element.
3. If $S$ of size $n$ is hashed to $n$ buckets, the expected size of the largest bucket is less than $\sqrt{n}+\frac{1}{2}$.

The intuition behind universal hashing is that we often lose relatively little compared to using a completely random map. Note that for the property 1, this is true in a very strong sense; even with complete randomness, we do not expect $o(n^2)$ buckets to suffice (the birthday paradox), so nothing is lost by using a universal family instead. The bounds in the second and third properties, however, are rather coarse compared to what one would get with complete randomness. For property 2, with complete randomness, $\Theta(n \log n)$ balls would suffice to cover the buckets with good probability (the coupon collector's theorem), i.e. a polynomial improvement over $n^2$, and for property 3, with complete randomness, we expect the largest bucket to have size $\Theta(\log n / \log \log n)$, i.e. an exponential improvement over $\sqrt{n}$. In these last cases we do seem to lose quite a lot compared to using a completely random map and better bounds would seem desirable. However, it is rather easy to construct (unnatural) examples of universal families and sets to be hashed showing that size $\Theta(n^2)$ is necessary to cover $n$ buckets with non-zero probability, and that buckets of size $\sqrt{n}$ are in general unavoidable, when a set of size $n$ is hashed to $n$ buckets. This shows that the *abstract* property of universality does not allow for stronger statements. Now fix a *concrete* universal family of hash functions. We ask the following question: *To which extent are the finer occupancy properties of completely random maps preserved?*

We provide answers to these questions for the case of *linear maps* between two vector spaces over a finite field, a natural and well known class of universal (in the sense of Carter and Wegmen) hash functions. The general flavor of our results is that "large fields are bad", in the sense that the bounds become the worst possible for universal families, while "small fields are good" in the sense that the bounds become as good or almost as good as the ones for independently distributed balls.

More precisely, for the covering problem, we show the following (easy) theorem.

**Theorem 1** *Let $F$ be a field of size $n$ and let $\mathcal{H}$ be the class of linear maps between $F^2$ and $F$. There is a subset $S$ of $F^2$ of size $\Theta(|F|^2)$, so that for no $h \in \mathcal{H}$, $h(S) = F$.*

On the other hand, we prove the following harder theorem.

**Theorem 2** *Let $S$ be a subset of a vector space over $\mathbf{Z}_2$ and choose a random linear map to a smaller vector space $R$. If $|S| \geq c_\epsilon |R| \log |R|$ then with probability at least $1 - \epsilon$ the image of $S$ covers the entire range $R$.*

For the "largest bucket problem", let us first introduce some notation: Let $U$ be the universe from which the keys are chosen. We fix a class $\mathcal{H}$ of functions mapping $U$ to $\{1,\ldots,s\}$. Then, a set $S \subseteq U$ of size $n$ is chosen *by an adversary*, and we uniformly at random pick a hash function $h \in \mathcal{H}$, hash $S$ using $h$ and look at the size of the largest resulting hash bucket. We denote the expectation of this size by $L_n^s$. Formally,

$$L_n^s(\mathcal{H}) = \max_{S\subseteq U, |S|=n} E_{h\in\mathcal{H}}[\max_{y\in\{1,\ldots,s\}} |\{x \in S \mid h(x) = y\}|].$$

Usually we think of $s$ being of size close to $n$. Note that if $s = \Omega(n^2)$, any universal class yields $L_n^s = O(1)$.

The class $\mathcal{H}$ we will consider is the set of linear maps between $F^m \to F^k$ for $m > k$. Here $F$ is a finite field and $s = |F|^k$. This class is universal for all values of the parameters.

When $k = 1$ and thus $|F| = s$ the expected largest bucket can be large.

**Theorem 3** *Let $F$ be a finite field with $|F| = s$. For the class $\mathcal{H}$ of all linear transformations $F^2 \to F$ we have*

$$L_s^s(\mathcal{H}) = \Omega(s^{1/3}).$$

*Furthermore if $|F|$ is a perfect square we have*

$$L_s^s(\mathcal{H}) > \sqrt{s}.$$

Note how close our lower bound for quadratic fields is to the upper bound of $\sqrt{s} + 1/2$ that holds for every universal class. We also mention that for the bad set we construct in Theorem 8 for a quadratic field there is no good linear hash function, since there *always* exists a bucket of size at least $\sqrt{s}$.

When the field is the field of two elements, the situation is completely different. Markowsky, Carter and Wegman [MCW78] showed that for this case $L_s^s(\mathcal{H}) = O(s^{1/4})$. Mehlhorn and Vishkin [MV84] improved on this result (although this is implicit in their paper) and showed that $L_s^s(\mathcal{H}) = O(2^{\sqrt{\log s}})$. We further improve the bound and show that:

**Theorem 4** *For the class $\mathcal{H}$ of all linear transformations between two vector spaces over $\mathbf{Z}_2$,*

$$L_s^s(\mathcal{H}) = O(\log s \log\log s).$$

Furthermore, we also show that even if the range is smaller than $|S|$ by a logarithmic factor, the same still holds:

**Theorem 5** *For the class $\mathcal{H}$ of all linear transformations between two vector spaces over $\mathbf{Z}_2$,*

$$L_{s\log s}^s(\mathcal{H}) = O(\log s \log\log s).$$

Note that even if one uses the class $\mathcal{R}$ of *all* functions one obtains only a slightly better result: the expected size of the largest bucket in this case is $L_s^s(\mathcal{R}) = \Theta(\log s/\log\log s)$ and $L_{s\log s}^s(\mathcal{R}) = \Theta(\log s)$, which is the best possible bound for any class. Interestingly, our upper bound is based on our upper bound for the covering property.

We do not have any non-trivial lower bounds on $L_s^s$ for the class of linear maps over $\mathbf{Z}_2$, i.e., it might be as good as $O(\log s/\log\log s)$. We leave this as an open question.

## 1.1 Motivation

There is no doubt that the method of implementing a dictionary by hashing with chaining, recommended in textbooks [CLR90, GBY90] especially for situations with many update operations, is a practically important scheme.

In situations in which a good bound on the cost of single operations is important, e. g., in real-time applications, the expected maximal bucket size as formed by all keys ever present in the dictionary during a time interval plays a crucial role. Our results show that, at least as long as the size of the hash table can be determined right at the start, using a hash family of linear functions over $\mathbf{Z}_2$ will perform very well in this respect. For other simple hash classes such bounds on the worst case bucket size are not available, or even fail to hold (see example in Section 4); other, more sophisticated hash families [S89, DM90, DGMP92] that do guarantee small maximal bucket sizes consist of functions with higher evaluation time. Of course, if worst case constant time for certain operations is absolutely necessary, the known two-level hashing schemes can be used, e. g., the FKS scheme [FKS84] for static dictionaries; dynamic perfect hashing [DKMHRT94] for the dynamic case with constant time lookups and expected time $O(n)$ for $n$ update operations; and the "real-time dictionaries" from [DM90] that perform each operation in constant time, with high probability. It should be noted, however, that a price is to be paid for the guaranteed constant lookup time in the dynamic schemes: the (average) cost of insertions is significantly higher than in simple schemes like chained hashing; the overall storage requirements are higher as well.

## 1.2 Related work

Another direction in trying to show that a specific class has a good bound on the expected size of the largest bucket is to build a class specifically designed to have such good property.

One immediate such result is obtained by looking at the class $\mathcal{H}$ of $d$-degree polynomials over finite fields, where $d = c\log n/\log\log n$ (see, e.g., [ABI86].) It is easy to see that this class maps each $d$ elements of the domain independently to the range, and thus, the bound that applies to the class of all functions also applies to this class. We can combine this with the following well known construction, found in, e.g., [FKS84], and sometimes called "collapsing the universe": There is a class $\mathcal{C}$ of size $2^{\Theta(\log n+\log\log|U|)}$ containing functions mapping $U$ to $\{1,\ldots,n^{k+2}\}$, so that, for any set $S$ of size $n$, a randomly chosen map from $\mathcal{C}$ will be one-to-one with probability $1-O(1/n^k)$.

The class consisting of functions obtained by first applying a member of $\mathcal{C}$, then a member of $\mathcal{H}$ is then a class with $L_n^n = \Theta(\log n/\log\log n)$ and size $2^{O(\log\log|U|+\log^2 n/\log\log n)}$ and with evaluation time $O(\log n/\log\log n)$ in a reasonable model of computation, say, a RAM with unit cost operations on members of the universe to be hashed.

More efficient (but much larger) families were given by Siegel [S89] and by Dietzfelbinger and Meyer auf der Heide [DM90]. Both provide families of size $|U|^{n^\epsilon}$ such that the functions can be evaluated in O(1) time on a RAM and with $L_n^n = \Theta(\log n/\log\log n)$. The families from [S89] and [DM90] are somewhat complex to implement while the class of linear maps requires only very basic bit operations (as discussed already in [CW79]). It is therefore desirable to study this class, and this is the main purpose of the present paper.

## 1.3 Notation

If $S$ is a subset of the domain $D$ of a function $h$ we use $h(S)$ to denote $\{h(s) \mid s \in S\}$. If $x$ is an element of the range we use $h^{-1}(x)$ to denote $\{s \in D \mid h(s) = x\}$. If $A$ and $B$ are subsets of a vector space $V$ and $x \in V$ we use the notations $A + B = \{a + b \mid a \in A \wedge b \in B\}$ and

$x + A = \{x + a \mid a \in A\}$. We use $\mathbf{Z}_2$ to denote the field with 2 elements. All logarithms in this paper are base two.

# 2 The covering property

## 2.1 Lower bounds for covering with a large field

We prove Theorem 1. Take any set $A \subset F$ of size $|A| = \lfloor |F|/2 \rfloor$ and consider $S = \{(x,y) \mid y \neq 0 \wedge x/y \in A \wedge (x-1)/y \notin A\}$. $S$ has density around one quarter. To see this, note that if $x$ and $y$ are picked randomly and independently in $F^*$, $(x/y, (x-1)/y)$ has the same distribution as $(x, x-y)$. Also, no linear map $g : F^2 \to F$ satisfies $g(S) = F$. To see this take a nonzero linear map $g : (x,y) \mapsto ax + by$ and note that if $0 \in g(S)$ then $a \neq 0$ and $-b/a \in A$ but in this case $a \notin g(S)$.

## 2.2 Upper bounds for covering with a small field - the existential case

We start by showing that if we have a subset $A$ of a vector space over $\mathbf{Z}_2$ and $|A|$ is sufficiently larger than another space $W$ then *there exists* a linear transformation $T$ mapping $A$ to the entire range $T(A) = W$. The constant $e$ below is the base of the natural logarithm.

**Theorem 6** *Let $A$ be a finite set of vectors in a vector space $V$ of an arbitrary dimension over $\mathbf{Z}_2$ and let $t > 0$ be an integer. If $|A| > t2^t/\log e$ then there exists a linear map $T : V \to \mathbf{Z}_2^t$, so that $T$ maps $A$ onto $\mathbf{Z}_2^t$.*

For the proof of this theorem we need the following simple lemma. Note that although we state the lemma for vector spaces, it holds for any finite group.

**Lemma 2.1** *Let $V$ be a finite vector space, $A \subseteq V$, $\alpha = 1 - |A|/|V|$. Then for a random $v \in V$,*

$$E_v(1 - |A \cup (v + A)|/|V|) = \alpha^2.$$

**Proof.** If $v$ and $u$ are both chosen uniformly independently at random from $V$ then both events $u \notin A$ and $u \notin v + A$ have probability $\alpha$ and they are independent. $\Box$

**Proof of Theorem 6.** Let $m$ be the dimension of $V$, $N = |A|$ and $\alpha = 1 - |A|/|V| = 1 - N/2^m$. Starting with $A_0 = A$, we choose a vector $v_1 \in V$ so that for $A_1 = A_0 \cup (v_1 + A_0)$

$$1 - \frac{|A_1|}{|V|} \leq \alpha^2.$$

Such a choice for $v_1$ exists by Lemma 2.1. Then, by the same procedure, we choose a $v_2$ so that for

$$A_2 = A_1 \cup (v_2 + A_1) = A + \mathrm{Span}\{v_1, v_2\},$$

$$1 - \frac{|A_2|}{|V|} \leq \alpha^4,$$

and so on up to $A_s = A + \mathrm{Span}\{v_1, \ldots, v_s\}$ with $s = m - t$ for which

$$1 - \frac{|A_s|}{|V|} \leq \alpha^{2^s}.$$

Note that one can assume that the vectors $v_1, \ldots, v_s$ are linearly independent since choosing a vector $v_i$ which linearly depends on the vectors formerly chosen makes $A_i = A_{i-1}$.

Let $W = \mathrm{Span}\{v_1, \ldots, v_s\}$. We have $A + W = V$ since for $x \in V \setminus (A+W)$ the sets $x + W$ and $A + W = A_s$ were disjoint, a contradiction as $|x + W| = |W|$ and $|A_s| \geq 2^m - 2^m \alpha^{2^s} \geq 2^m - 2^m e^{-N2^{-t}} > |V| - |W|$.

We choose an onto linear map $T : V \to \mathbf{Z}_2^t$ such that its kernel $T^{-1}(0)$ equals $W$. As $T(W) = \{0\}$ we have $T(A) = T(A+W) = T(V) = \mathbf{Z}_2^t$ as claimed. □

The bound in Theorem 6 is asymptotically tight as shown by the following proposition.

**Proposition 2.2** *For every large enough integer $t$ there is a set $A$ of at least $(t - 3\log t)2^t/\log e$ vectors in a vector space $V$ over $\mathbf{Z}_2$ so that for any linear map $T : V \to \mathbf{Z}_2^t$, $T$* does not *map $A$ onto $\mathbf{Z}_2^t$.*

**Proof:** Let $V = Z_2^{t+s}$ with $s = \lfloor t/10 \rfloor$ and let $A$ be chosen at random by picking each element of $V$ independently and with probability $p = 1 - 2^{-x}$ into the set with $x = (t - 2\log t)2^{-s}$. From Chebyshev's inequality we know that with probability at least $3/4$, $A$ has cardinality at least $pN - 2\sqrt{pN}$ for $N = 2^{s+t}$. Using $p > x/\log e - x^2/(2\log^2 e)$ one can show that this is as many as claimed in the proposition. Let us compute the probability that there exists a linear map $T : V \to \mathbf{Z}_2^t$ such that $T$ maps $A$ onto $\mathbf{Z}_2^t$. There are $2^{t(t+s)}$ possible maps $T$ and each of them satisfies $T(A) = \mathbf{Z}_2^t$ with probability at most $\left(1 - (1-p)^{2^s}\right)^{2^t} = \left(1 - 2^{-2^s x}\right)^{2^t} = (1 - t^2/2^t)^{2^t} < e^{-t^2}$. So with probability almost $3/4$, $A$ is not small and still no $T$ maps $A$ onto $\mathbf{Z}_2^t$. □

### 2.3 Choosing the linear map at random

In this subsection we strengthen Theorem 6 and prove that if $A$ is bigger than what is required there by only a constant factor, then almost all choices of the linear transformation $T$ work. This may seem immediate at first glance since Lemma 2.1 tells us that a random choice for the next vector is good on average. In particular, it might seem that for a random choice of $v_1$ and $v_2$ in the proof of Theorem 6, $E_{v_1,v_2}(1 - |A + \mathrm{Span}\{v_1, v_2\}|/|V|) \leq \alpha^4$. Unfortunately this is not the case: For example, think of $A$ being a linear subspace containing half of $V$. In this case, the ratio $\alpha$ of points that are not covered is $1/2$. As random vectors $v_i$ are chosen to be added to $A$, vectors in $A$ are chosen with probability $1/2$. Thus, after $i$ steps, $\alpha$ remains $1/2$ with probability $1/2^i$ and becomes $0$ otherwise. Thus, the expected value of $\alpha_i$ is $2^{-i-1}$ which is much bigger than $2^{-2^i}$.

Our first lemma is technical in nature.

**Lemma 2.3** *Let $\alpha_i$ for $1 \leq i \leq k$ be random variables and let $0 < \alpha_0 < 1$ be a constant. Suppose that for $0 \leq i < k$ we have $0 \leq \alpha_{i+1} \leq \alpha_i$ and conditioned on any set of values for $\alpha_1, \ldots, \alpha_i$ we have $E[\alpha_{i+1}|\alpha_1, \ldots, \alpha_i] = \alpha_i^2$. Then for any threshold $0 < t < 1$ we have*

$$\mathrm{Prob}[\alpha_k \geq t] \leq \alpha_0^{k - \log\log(1/t) + \log\log(1/\alpha_0)}.$$

**Proof:** The proof is by induction on $k$. The $k = 0$ base case is trivial.

We assume the statement of the lemma for $k$ and prove it for $k+1$. Let $c = k - \log\log(1/t)$. We may suppose $c + 1 + \log\log(1/\alpha_0) \geq 0$ since otherwise the bound in the lemma is greater than 1.

After the choice of $\alpha_1$, the rest of the random variables form a random process of length $k$ satisfying the conditions of the lemma (unless $\alpha_1 = 0$); thus we can apply the inductive hypothesis to get

$$\mathrm{Prob}[\alpha_{k+1} \geq t] = E_{\alpha_1}[\mathrm{Prob}[\alpha_{k+1} \geq t \mid \alpha_1]] \leq E[f(\alpha_1)],$$

where we define $f_0(x) = x^{c+\log\log(1/x)}$ for $0 < x < 1$ and take $f(x) = \min(1, f_0(x))$ in the same interval and $f(0) = 0$. The value $f(\alpha_1)$ is clearly an upper bound on $\mathrm{Prob}[\alpha_{k+1} \ge t \mid \alpha_1]$.

We claim that in the interval $0 \le x \le \alpha_0$ we have $f(x) \le f_0(\alpha_0)x/\alpha_0$. To prove this simply observe that $f_0(x)/x$ is first increasing then decreasing on $(0, 1)$. To see this compute the derivative $(f_0(x)/x)' = (c+\log e - 1 + \log\log(1/x))f_0(x)/x^2$. If $\alpha_0$ is still in the increasing phase then we have $f(x)/x \le f_0(x)/x \le f_0(\alpha_0)/\alpha_0$ for $0 < x \le \alpha_0$. Suppose now that $\alpha_0$ is already in the decreasing phase and define $x' = 2^{-2^{-c-1}}$. Notice that we assumed $\alpha_0 \le x'$ in the beginning of the proof, so we have $f_0(\alpha_0)/\alpha_0 \ge f_0(x')/x'$. Let us define $x'' = x'^2 = 2^{-2^{-c}}$ and notice that we have $f(x) = 1$ if and only if $x \ge x''$. It is easy to check that $x''$ must still be in the increasing phase of $f_0(x)/x$ thus we have $f(x)/x = f_0(x)/x \le f_0(x'')/x'' = 1/x''$ for $0 < x \le x''$. For $x'' \le x < 1$ we simply have $f(x)/x = 1/x \le 1/x''$. Thus we must have $f(x)/x \le 1/x'' = f_0(x')/x' \le f_0(\alpha_0)/\alpha_0$ for $0 < x < 1$. We have thus proved the claim in all cases for $0 < x \le \alpha_0$. The claim is trivial for $x = 0$.

Using the claim we can finish the proof writing:

$$\mathrm{Prob}[\alpha_{k+1} \ge t] \le E[f(\alpha_1)] \le E[f_0(\alpha_0)\alpha_1/\alpha_0] = f_0(\alpha_0)E[\alpha_1]/\alpha_0 =$$

$$f_0(\alpha_0)\alpha_0 = \alpha_0^{c+1+\log\log(1/\alpha_0)}.$$

□

We remark that the bound in the lemma is achievable for $t = \alpha_0^{2^j}$ with an integer $0 \le j \le k$. The optimal process has $\alpha_i = \alpha_{i-1}$ or $\alpha_i = 0$ for $1 \le i \le k-j$, while $\alpha_i = \alpha_{i-1}^2$ for $k-j < i \le k$.

**Theorem 7** *a) For every $\epsilon > 0$ there is a constant $c_\epsilon > 0$ such that the following holds. Let $A$ be a finite set of vectors in a vector space $V$ of an arbitrary dimension over $\mathbf{Z}_2$, let $t > 0$ be an integer. If $|A| \ge c_\epsilon t 2^t$ then for a uniform random linear transformation $T : V \to \mathbf{Z}_2^t$*

$$\mathrm{Prob}(T(A) = \mathbf{Z}_2^t) \ge 1 - \epsilon.$$

*b) If $A$ is a subset of the vector space $\mathbf{Z}_2^u$ of density $|A|/2^u = 1 - \alpha < 1$ and $0 \le t < u$ is an integer then for a uniform random onto linear transformation $T : \mathbf{Z}_2^u \to \mathbf{Z}_2^t$*

$$\mathrm{Prob}(T(A) \ne \mathbf{Z}_2^t) \le \alpha^{u-t-\log t+\log\log(1/\alpha)}.$$

**Proof:** We start with proving part b) of the theorem. In order to pick the onto map $T$ we use the following process (similar to the one in the proof of Theorem 6). Pick $s = u - t$ vectors $v_1, \ldots, v_s$ uniformly at random from the vectors in $\mathbf{Z}_2^u$ and choose $T$ to be a random onto linear transformation $T : \mathbf{Z}_2^u \to \mathbf{Z}_2^t$ with the constraints $T(v_i) = 0$ $(i = 1, \ldots, s)$, i.e. the vectors $v_1, \ldots, v_s$ are in the kernel of $T$. Note that the $v_i$'s are not necessarily linearly independent and that they do not necessarily span the kernel. Still, the transformation $T$ is indeed distributed uniformly at random amongst all onto linear maps of $\mathbf{Z}_2^u$ onto $\mathbf{Z}_2^t$.

Using notations similar to the ones used in the proof of Theorem 6, let $A_0 = A$, $A_i = A_0 + \mathrm{Span}\{v_1, \ldots, v_i\}$ and $\alpha_i = 1 - |A_i|/2^u$ for $i = 0, \ldots, s$. Clearly $\alpha_i$ is nonnegative and monotone decreasing in $i$ with $\alpha_0 = \alpha$. The equation $E[\alpha_{i+1} \mid \alpha_1, \ldots, \alpha_i] = \alpha_i^2$ is guaranteed by Lemma 2.1 since $A_{i+1} = A_i \cup (A_i + v_{i+1})$ and $v_{i+1}$ is independent of $\alpha_j$ for $j \le i$. Thus all the conditions of Lemma 2.3 are satisfied and we have

$$\mathrm{Prob}[\alpha_s \ge 2^{-t}] \le \alpha^{s-\log t+\log\log(1/\alpha)}.$$

By the definition of $s$ the right hand side here is equal to the estimate in the theorem. Finally note that (as in the proof of Theorem 6) when $\alpha_s < 2^{-t}$ then $T(A) = \mathbf{Z}_2^t$ since for $x \in \mathbf{Z}_2^t \setminus T(A)$ the sets

$T^{-1}(x)$ and $A_s$ were disjoint with sizes $2^{u-t}$ and $(1-\alpha_s)2^u > 2^u - 2u - t$, a contradiction. Thus we have the claimed upper bound for the probability that $T(A) \neq \mathbf{Z}_2^t$.

Now we turn to part a) of the theorem and prove it using part b). Part a) is about a random linear transformation, not necessarily onto, but this difference from the claim just proved poses less of a problem, the difficulty is that we do not have an a priori bound on $1 - |A|/|V|$. In fact, this ratio can be arbitrarily small. To solve this, we choose the transformation $T$ in two steps, the first step ensuring that the density of the covered set is substantial, then applying part b) for the second step.

Let $W = \mathbf{Z}_2^u$, with $u = \lceil \log(2|A|/\epsilon) \rceil$. We factor $T$ through $W$. First, we pick uniformly at random a linear transformation $T_0 : V \to W$. Then, we pick a random onto linear map $T_1 : W \to \mathbf{Z}_2^t$, and set $T = T_0 \circ T_1$. This results in a uniformly chosen linear map $T : V \to \mathbf{Z}_2^t$. This is true even for a fixed onto $T_1$ and a random $T_0$, since the values $T_0(e_i)$ for a basis $e_1, e_2, \ldots$ of $V$ are independent and uniformly distributed in $W$, thus the values $T(e_i)$ are also independent and uniformly distributed in $\mathbf{Z}_2^t$.

Any pair of vectors $v \neq w \in A$ collide (due to $T_0$) with probability $\mathrm{Prob}[T_0(v) = T_0(w)] = 1/|W|$. Thus the expected number of collisions is $\binom{|A|}{2}/|W|$. Since $|T_0(A)| \leq |A|/2$ implies at least $|A|/2$ such collisions, Markov's inequality gives $\mathrm{Prob}[|T_0(A)| \leq |A|/2] \leq 2\binom{|A|}{2}/(|A||W|) < |A|/|W| \leq \epsilon/2$. For any fixed $T_0$, part b) of the theorem gives

$$\mathrm{Prob}[T(A) \neq \mathbf{Z}_2^t] \leq \alpha^{u-t-\log t+\log\log(1/\alpha)},$$

where $\alpha = 1 - |T_0(A)|/|W|$. In case $|T_0(A)| > |A|/2$ we have $\alpha < 1 - |A|/(2|W|) < e^{-\epsilon/8}$, thus using the monotonicity of the bound above we get

$$\mathrm{Prob}[T(A) \neq \mathbf{Z}_2^t] \leq e^{-\epsilon(u-t-\log t+\log(\log e^{\epsilon/8}))/8}. \tag{1}$$

Choosing $c_\epsilon = 4(2/\epsilon)^{8/\epsilon}$ we have that $|A| \geq c_\epsilon t 2^t$ implies $u = \lceil \log(2|A|/\epsilon) \rceil > t + \log t + \log(4/\epsilon) + (4/\epsilon)\log(2/\epsilon)$. This implies that the bound in Equation 1 is less than $\epsilon/2$, thus we get $\mathrm{Prob}[T(A) \neq \mathbf{Z}_2^t] \leq \mathrm{Prob}[|T_0(A)| \leq |A|/2] + \epsilon/2 < \epsilon$ as claimed. $\square$

We remark that a more careful analysis gives $c_\epsilon$ that is a small polynomial of $1/\epsilon$.

# 3 The largest bucket

## 3.1 Lower bound for the largest bucket with a large field

We start by showing why linear hashing over a large finite field is bad with respect to the expected largest bucket size measure. This natural example shows that universality of the class is not enough to assure small buckets. For a finite field $F$ we prove the existence of a bad set $S \subset F^2$ of size $|S| = |F|$ such that the expected largest bucket in $S$ with respect to a random linear map $F^2 \to F$ is big. We prove the results in Theorem 3 separately for quadratic and non-quadratic fields.

We start with an intuitive description of the constructions. Linear hashing of the plane collapses all straight lines of a random direction. Thus, a bad set in the plane must contain many points on at least one line in many different directions. It is not hard to come up with bad sets that contain many points of many different lines, however the obvious constructions (subplane or grid) yield sets where many of the "popular lines" tend to be parallel and thus they only cover a few directions. This problem can be solved by a projective transformation: the transformed set has many popular lines, but they are no longer parallel.

For the non-quadratic case, it is convenient to explicitly use the concept of the projective plane over a field $F$. Recall that the *projective plane* $P$ over $F$ is defined as $(F^3 - \{(0,0,0)\})/\sim$ , where $\sim$ is the equivalence relation $(x, y, z) \sim (cx, cy, cz)$ for all $c \neq 0$. The affine plane $F^2$ is embedded in $P$ by the one-to-one map $(x, y) \mapsto (x, y, 1)$. A *line* in $P$ is given by an equation $\{(x, y, z) | ax + by + cz = 0\}$, i.e., a projective line corresponds to a plane in $F^3$ containing the origin. All projective lines are extensions (by one new point) of lines in the affine plane $F^2$, except for the *ideal* line, given by $\{(x, y, z) | z = 0\}$. A *projective transformation* mapping the ideal line to another projective line $L$ is a map $\tilde{f} : P \to P$ obtained as the $\sim$-quotient of a nonsingular linear map $f : F^3 \to F^3$ mapping the plane corresponding to the ideal line into the plane corresponding to $L$.

Projective geometry is useful for understanding the behavior of linear hash functions due to the following fact which is easily verified: Picking a random non-trivial linear map $F^2 \to F$ as a hash function and partitioning a subset $S \subset F^2$ into hash buckets accordingly, corresponds *exactly* to picking a random point $p$ on the ideal line and partitioning the points of $S$ according to which line through $p$ they are on. This observation will be used explicitly in the proof of Theorem 9.

**Theorem 8** *Let $F$ be a finite field with $|F|$ being a perfect square. There exists a set $S \subset F^2$ of size $|S| = |F|$ such that for every linear map $h : F^2 \to F$, $S$ has a large bucket, i.e. there exists a value $y \in F$ with $|h^{-1}(y)| \geq \sqrt{|F|}$.*

**Proof.** We have a finite field $F_0$ of which $F$ is a quadratic extension. Let $|F_0| = m$ and $|F| = m^2 = n$. Let $a$ be an arbitrary element in $F \setminus F_0$ and define $S = \{(\frac{1}{x+a}, \frac{y}{x+a}) \mid x, y \in F_0\}$. Note that $|S| = m^2 = |F|$. Notice also, that $S$ is the image of the subplane $F_0^2$ under the projective transformation $(x, y) \mapsto (\frac{1}{x+a}, \frac{y}{x+a})$.

Fix $A, B \in F$ and consider the function $h : F^2 \to F$ defined by $h(x, y) = Ax + By$. We must show that there is some $C \in F$ such that $|h^{-1}(C) \cap S| \geq m$. If $B = 0$ then $h$ maps all the $m$ elements of $S' = \{(1/a, y/a) \mid y \in F_0\}$ to $C = A/a$, as needed. Otherwise, we claim that there is a $C \in F$ such that both $\frac{C}{B}$ and $\frac{aC-A}{B}$ are in $F_0$. To see this observe that if $g_1$ and $g_2$ are two distinct members of $F_0$, then $ag_1$ and $ag_2$ lie in distinct additive cosets of $F_0$ in $F$, since otherwise their difference, $a(g_1 - g_2)$ would have to be in $F_0$, contradicting the fact that $a \notin F_0$. Thus, as $g$ ranges over all members of $F_0$, $ag$ intersects distinct additive cosets of $F_0$ in $F$, and hence $aF_0$ intersects all those cosets. In particular, there is some $g \in F_0$ so that $ag \in F_0 + \frac{A}{B}$, implying that $C = gB$ satisfies the assertion of the claim. For the above $C$, define $y(x) = (C/B)x + (aC - A)/B$; it follows that $y(x) \in F_0$ for every $x \in F_0$. We have now $A\frac{1}{a+x} + B\frac{y(x)}{a+x} = C$, showing that $h$ maps all the $m$ elements of $S' = \{(\frac{1}{a+x}, \frac{y(x)}{a+x}) \mid x \in F_0\} \subset S$ to $C$. $\square$

**Theorem 9** *Let $F$ be a finite field. There exists a set $S \subset F^2$ of size $|S| = |F|$ such that for more than half of the linear maps $h : F^2 \to F$, $S$ has a large bucket, i.e. there exists a value $y \in F$ with $|h^{-1}(y)| \geq |F|^{1/3}/3 - 1$.*

**Proof.** First we construct a set $S' \subset F^2$ such that $|S'| \leq |F| = n$ and there are $n$ distinct lines in the plane $F^2$ each containing at least $m \geq n^{1/3}/3$ points of $S'$.

Let us first consider the case when $n$ is a prime, so $F$ consists of the integers modulo $n$. We let $A = \{i \mid 1 \leq i < \sqrt{n}\}$ and consider the square grid $S' = A \times A$. Clearly $|S'| < n$. It is well known that each of the $n$ most popular lines contains at least $m \geq n^{1/3}/3$ points of $S'$. This is usually proved for the same grid in the Euclidean plane (see e.g. [PA95], pp. 178–179) but that result implies the same for our grid in $F^2$.

Now let $n = p^k$ and let $F_0$ be the subfield in $F$ of $p$ elements. Let $x \in F$ be a primitive element, then every element of $F$ can be uniquely expressed as a polynomial of $x$ of degree below $k$ with coefficients from $F_0$. Let $k_1 = \left\lfloor \frac{k+1}{3} \right\rfloor$, $k_2 = k - k_1 = \left\lfloor \frac{2k+1}{3} \right\rfloor$ and let $A_1 = \{f(x) \mid \deg(f) < k_1\}$, $A_2 = \{f(x) \mid \deg(f) < k_2\}$ where the polynomials $f$ have coefficients from $F_0$. Finally we take $S' = A_1 \times A_2$. Clearly $|S'| = n$. For $a \in A_1$ and $b \in A_2$ we consider the line $L_{a,b} = \{(y, ay + b) \mid y \in F\}$ in $F^2$. Notice that there are $n$ such lines and we have $ay + b \in A_2$ whenever $y \in A_1$. Thus, we have $n$ distinct lines each containing $m = |A_1| = p^{k_1}$ points of $S'$. We have $m \geq n^{1/3}$ as claimed unless $k \equiv 1 \pmod 3$. Notice that for $k \equiv 2 \pmod 3$ our $m$ is much higher than $n^{1/3}$. For the bad case $k \equiv 1 \pmod 3$ we apply the construction below instead.

Finally suppose $n = p^k$, $p$ is a prime and $k \equiv 1 \pmod 3$. To get our set $S'$ in this case we have to merge the two constructions above. Let $F_0$ be the $p$ element subfield of $F$, then $F_0$ consists of the integers modulo $p$. We set $A = \{i \mid 1 \leq i < \sqrt{p}\}$. Let $k_1 = (k+2)/3$ and $k_2 = (2k+1)/3$ and let $x \in F$ be a primitive element, so we can express any element of $F$ uniquely as a polynomial of $x$ of degree less then $k$ with coefficients from $F_0$. We set $A_1 = \{f(x) \mid \deg(f) < k_1 \wedge f(0) \in A\}$, $A_2 = \{f(x) \mid \deg(f) < k_2 \wedge f(0) \in A\}$ where the polynomials $f$ have coefficients from $F_0$. Finally we set $S' = A_1 \times A_2$. Clearly $|S'| < n$. For $j, j' \in F_0$ let $L_{j,j'} = \{(i, ji + j') \mid i \in F_0\}$. Let $a$ and $b$ be polynomials with coefficients from $F_0$ with $\deg(a) < k_1$ and $\deg(b) < k_2$. Consider the line $L_{a,b} = \{(y, a(x)y + b(x)) \mid y \in F\}$. We now compute the value of $|L_{a,b} \cap S'|$. Note that a point $(y, a(x)y + b(x))$ of $L_{a,b}$ is in $S'$ if and only if $y = f(x)$ for some polynomial $f$ so that $\deg(f) < k_1$, $f(0) \in A$ and $a(0)f(0) + b(0) \in A$. The number of such polynomials $f$ is exactly $p^{k_1-1}|L_{a(0),b(0)} \cap (A \times A)|$. Thus, $|L_{a,b} \cap S'|$ is exactly $p^{k_1-1}|L_{a(0),b(0)} \cap (A \times A)|$. Thus, from knowing that the $p$ most popular lines in $F_0^2$ contain at least $m_0 \geq p^{1/3}/3$ points from $A \times A$ we conclude that there exist $n$ distinct lines each containing at least $m = m_0 p^{k_1-1} \geq n^{1/3}/3$ points of $S'$; namely, the lines $L_{a,b}$ for those choices of $a$ and $b$ for which $L_{a(0),b(0)}$ is a popular line in $F_0^2$.

In all cases we have constructed our set $S' \subset F^2$ of size $|S'| \leq n$ with $n$ distinct popular lines each containing at least $m > n^{1/3}/3$ points of $S'$. Let $P$ be the projective plane containing $F^2$. Out of the $n^2 + n + 1$ points in $P$ every popular line covers $n + 1$. The $i$th popular line ($1 \leq i \leq n$) can only have $i - 1$ intersections with earlier lines, thus it covers at least $n + 2 - i$ points previously uncovered. Therefore a total of at least $\binom{n+2}{2} - 1$ points are covered by popular lines. Simple counting gives the existence of a line $L$ in $P$ not among the popular lines, such that more than half of the points on $L$ are covered by popular lines. Let $f$ be a projective transformation taking the ideal line $L' = P \setminus F^2$ to $L$. We define $S = \{x \in F^2 \mid f(x) \in S'\} = f^{-1}(S') \cap F^2$. Clearly $|S| \leq |S'| \leq n$.

One linear hash function $h : F^2 \to F$ is constant zero (and thus all of $S$ is a single bucket), for the rest there is a point $x_h \in L'$ such that $h$ collapses the points of $F^2$ of each single line going through $x_h$, as we observed at the beginning of the section. Furthermore, if the linear non-zero map is picked at random, all such points $x_h$ are equally likely. Thus, the statement of the theorem follows, if we show that for at least half the points $x_h$ on the ideal line, it holds that some line through $x_h$ intersects $S$ in at least $n^{1/3}/3 - 1$ points. But some line through $x_h$ intersects $S$ in at least $n^{1/3}/3 - 1$ points if and only if some line through $f(x_h)$ intersects $f(S)$ in at least $n^{1/3}/3 - 1$ (projective) points. For this, it is sufficient that some line through $f(x_h)$ intersects $S'$ in at least $(n^{1/3}/3 - 1) + 1 = n^{1/3}/3$ points (the $+1$ comes from the possibility of $f(x_h) \in S'$), i.e., that some line through $f(x_h)$ is popular, in the sense we used above. But by definition of $f$, this is true for at least half of the points $x_h$ on the ideal line, and we are done. □

### 3.2 Upper bound for the largest bucket with a small field

Let us now recall and prove our main result.

For convenience here we speak about hashing $n \log n$ keys to $n$ values. Also, we assume that $n$

is a power of 2.

**Theorem 5**: Let $\mathcal{H}$ be the class of linear transformations between two vector spaces over $\mathbf{Z}_2$, then

$$L^n_{n\log n}(\mathcal{H}) = O(\log n \log\log n).$$

This theorem implies Theorem 4.

We have to bound the probability of the event that many elements in the set $S$ are mapped to a single element in the range. Denote this bad event by $E_1$. The overall idea is to present another (less natural) event $E_2$ and show that the probability of $E_2$ is small, yet the probability of $E_2$ given $E_1$ is big. Thus, the probability of $E_1$ must be small. We remark here that a somewhat similar line of reasoning was used in the seminal paper of Vapnik and Chervonenkis [VC71].

For the proof we fix the domain to be $D = \mathbf{Z}_2^m$, the range (the buckets) to be $B = \mathbf{Z}_2^{\log n}$, and $S \subset D$ of size $|S| = n \log n$.

Let us choose arbitrary $\ell \geq \log n$ and consider the space $A = \mathbf{Z}_2^\ell$. We construct the linear transformation $h : D \to B$ through the intermediate range $A$ in the following way. We choose uniformly at random a linear transformation $h_1 : D \to A$ and uniformly at random an onto linear transformation $h_2 : A \to B$. Now we define $h \stackrel{\text{def}}{=} h_1 \circ h_2$. Note that as mentioned in the proof of part a) of Theorem 7 this yields an $h$ which is uniformly chosen from among all linear transformations from $D$ to $B$.

Let us fix a threshold $t$. We define two events. $E_1$ is the existence of a bucket of size at least $t$:

**Event $E_1$:** There exists an element $\alpha \in B$ such that

$$\left| h^{-1}(\alpha) \cap S \right| > t.$$

We are going to limit the probability of $E_1$ through the seemingly unrelated event $E_2$:

**Event $E_2$:** There exists an element $\alpha \in B$ such that

$$h_2^{-1}(\alpha) \subseteq h_1(S).$$

Consider the distribution space in which $h_1$ and $h_2$ are uniformly chosen as above. We shall show that

**Proposition 3.1** *If $d = 2^\ell/(n \log n) > 1$ we have*

$$\mathrm{Prob}[E_2] \leq d^{-\log d - \log\log d}.$$

**Proposition 3.2** *If $t > c_{1/2}(2^\ell/n)\log(2^\ell/n)$ (with $c_{1/2}$ from Theorem 7a)) then*

$$\mathrm{Prob}[E_2|E_1] \geq \frac{1}{2}.$$

From Propositions 3.1 and 3.2 we deduce that the probability of $E_1$ must be small:

**Corollary 3.3** *There is a constant $C$, so that for all $r > 4$ and every power of two $n$, the following holds: If a subset $S$ of size $|S| = n \log n$ of a vector space over $\mathbf{Z}_2$ is hashed by a random linear transformation to $\mathbf{Z}_2^{\log n}$, we have*

$$\mathrm{Prob}[\mathit{maximum\ bucket\ size} > rC \log n \log\log n] \leq 2(r/\log r)^{-\log(r/\log r) - \log\log(r/\log r)}.$$

**Proof:** Given $r > 4$, let $l = \lfloor \log n + \log\log n + \log r - \log\log r + 1 \rfloor$ and let $t = 4c_{1/2} r \log n \log\log n$. Letting $d = 2^l/(n \log n)$, we have $d = 2^l/(n\log n) \geq 2^{\log n + \log\log n + \log r - \log\log r}/(n \log n) = r/\log r > 1$ and $2^l/n \leq 2^{\log n + \log\log n + \log r - \log\log r + 1}/n = 2\log n(r/\log r)$, so

$$\begin{aligned} c_{1/2}(2^l/n)\log(2^l/n) &< c_{1/2}(2\log n(r/\log r))(1 + \log\log n + \log r) \\ &< c_{1/2} 2\log n(r/\log r)(2\log\log n \log r) \\ &= 4c_{1/2} r \log n \log\log n \\ &= t, \end{aligned}$$

so the conditions of Proposition 3.1 and 3.2 are satisfied, and, combining their conclusions, we get

$$\Pr[E_1] \leq 2\Pr[E_2] \leq 2d^{-\log d - \log\log d}.$$

But the event $E_1$ is the event that the biggest bucket is bigger than $t = 4c_{1/2} r \log n \log\log n$ and since $d \geq r/\log r$, the statement of the corollary follows, by putting $C = 4c_{1/2}$. $\Box$

Let us now prove the propositions above.

**Proof of Proposition 3.1:** Note first that an alternative way to describe $E_2$ is

$$h_2(A \setminus h_1(S)) \neq B.$$

We will prove that Proposition 3.1 holds for any specific $h_1$, and thus it also holds for a randomly chosen $h_1$. So fix $h_1$ and consider the distribution in which $h_2$ is chosen uniformly amongst all full rank linear transformation from $A$ to $B$.

We use part b) of Theorem 7 for the set $A \setminus h_1(S) \subset A$. Its density is clearly $1 - \alpha$ for $\alpha = |h_1(S)|/|A| \leq |S|/|A| = 1/d$. Thus the theorem gives $\text{Prob}[E_2] \leq \alpha^{\ell - \log n - \log\log n + \log\log(1/\alpha)} \leq d^{-\log d - \log\log d}$ as claimed. $\Box$

**Proof of Proposition 3.2:** Fix $h$ for which $E_1$ holds, and fix any full rank $h_2$. We will show that the probability of event $E_2$ is at least $1/2$ even when these two are fixed and thus the conditional probability is also at least $1/2$.

Now since $E_1$ holds there is a subset $S' \subseteq S$ of cardinality at least $t$ mapped by $h$ to a single element $\alpha \in \mathbf{Z}_2^{\log n}$. Fix this $\alpha$ and define $D' \stackrel{\text{def}}{=} h^{-1}(\alpha)$ and $A' \stackrel{\text{def}}{=} h_2^{-1}(\alpha)$. Consider the distribution of $h_1$ satisfying $h = h_1 \circ h_2$. When we restrict $h_1$ to $D'$, we get that the distribution implied by such $h_1$ is a uniform choice of an affine or linear map from $D'$ into $A'$ (we show this in Proposition 3.4 below). For event $E_2$ to hold it is enough to have $A' \subseteq h_1(S)$. We will show that $h_1(S')$ covers all the points in $A'$ with probability at least $1/2$ and thus we get that event $E_2$ happens with probability $1/2$. Since $h_2$ is onto we have $|A'| = 2^\ell/n$. On the other hand, $D' \cap S$ has cardinality at least $t = \lceil c_{1/2}(2^\ell/n)\log(2^\ell/n) \rceil$. By part a) of Theorem 7, the probability that a set of cardinality $t$ mapped by a random linear transformation will cover a range of cardinality $2^\ell/n$ is at least $1/2$. (Note that Theorem 7, part a) clearly applies to a random affine transformation too.) $\Box$

At this point, we have proven Corollary 3.3. This limits the probability of large buckets with linear hashing. It is straightforward to deduce Theorem 5 from that corollary:

**Proof of Theorem 5:** $L^n_{n \log n}$ is the expectation of the distribution of the largest bucket size. Corollary 3.3 limits the probability of the tail of this distribution, thus yielding the desired bound on the expectation. The constant $C$ is from Corollary 3.3 and we set $K = C \log n \log\log n$.

$$E[\max S\text{-bucket size}] = \int_0^\infty \text{Prob}[\max S\text{-bucket size} > t]dt$$

$$
\begin{aligned}
&\leq \quad 4K + \int_{4K}^{\infty} \text{Prob}[\max S\text{-bucket size} > t]dt \\
&= \quad 4K + K \int_{4}^{\infty} \text{Prob}[\max S\text{-bucket size} > rK]dr \\
&\leq \quad 4K + K \int_{4}^{\infty} 2(r/\log r)^{-\log(r/\log r) - \log\log(r/\log r)} dr \\
&= \quad O(K) = O(\log n \log\log n).
\end{aligned}
$$

□

In order for the paper to be self-contained we include a proof of the simple statement about random linear transformations used above.

**Proposition 3.4** *Let $D$, $A$ and $B$ be vector spaces over $\mathbf{Z}_2$. Let $h : D \to B$ be an arbitrary linear map, and let $h_2 : A \to B$ be an arbitrary onto linear map. Let $\alpha$ be any point in $B$ and denote $D' \stackrel{\text{def}}{=} h^{-1}(\alpha)$ and $A' \stackrel{\text{def}}{=} h_2^{-1}(\alpha)$. Then, choosing a uniform linear map $h_1 : D \to A$ such that $h = h_1 \circ h_2$ and restricting the domain to $D'$ we get a uniformly chosen linear map from $D'$ to $A'$ if $\alpha = 0$ or uniformly chosen affine map from $D'$ to $A'$ otherwise.*

**Proof:** Consider $D_0 \stackrel{\text{def}}{=} h^{-1}(0)$ and $A_0 \stackrel{\text{def}}{=} h_2^{-1}(0)$. Let us choose a complement space $D_1$ to $D_0$ in $D$, i.e. $D_0 \cap D_1 = \{0\}$ and $D_0 + D_1 = D$. Let us call $x$ the unique vector in $D' \cap D_1$. We have $D' = D_0 + x$. A linear transformation $h_1 : D \to A$ is determined by its two restrictions $h' : D_0 \to A$ and $h'' : D_1 \to A$. Clearly the uniform random choice of $h_1$ corresponds to uniform and independent choices for $h'$ and $h''$. The restriction $h = h_1 \circ h_2$ means that $h'(D_0) \subseteq A_0$ and $h'' \circ h_2$ is the restriction of $h$ to $D_1$. Thus, after the restriction the random choices of $h'$ and $h''$ are still independent. Note now that if $\alpha = 0$ then the restriction of $h_1$ in question is exactly $h' : D' \to A'$. If $\alpha \neq 0$ then use $h_1(u + x) = h'(u) + h''(x)$ for $u \in D_0$ to note that the restriction in question is again $h'$, this time translated by the random value $h''(x) \in A'$. □

# 4 Remarks and open questions

We have discussed the case of a very small field (size 2) and a very large field (size $n$). What happens with intermediate sized fields? Some immediate rough generalizations of our bounds are the following: If we hash an adversely chosen subset of $F^m$ of size $n = |F|^k$ to $F^k$ by a randomly chosen linear map, the expected size of the largest bucket is at most $O((\log n \log\log n)^{\log|F|})$ and at least $\Omega(|F|^{1/3})$. Tighter bounds should be possible.

Another question is which properties other well known hash families have. Examples of the families we have in mind include: Arithmetic over $\mathbf{Z}_p$ [CW79, FKS84] (with $h_{a,b}(x) = (ax + b \bmod p) \bmod n$), integer multiplication [DHKP97, AHNR95] (with $h_a(x) = (ax \bmod 2^k) \operatorname{div} 2^{k-l}$), Boolean convolution [MNT93] (with $h_a(x) = a \circ x$ projected to some subspace).

An example of a natural non-linear scheme for which the assertion of Theorem 6 fails is the scheme that maps integers between 1 and $p$, for some large prime $p$, to integers between 0 and $n-1$ for $n = \lceil p/m \rceil$, by mapping $x \in \mathbf{Z}_p$ to $(ax + b \bmod p) \operatorname{div} m$, where $a, b$ are two randomly chosen elements of $\mathbf{Z}_p$. For this scheme, there are primes $p$ and choices of $n$ and a subset $S$ of cardinality $\Omega(n \log n \log\log\log n)$ of $\mathbf{Z}_p$, which is not mapped by the above mapping onto $[0, n-1]$ under any choice of $a$ and $b$.

To see this, let $p$ be a prime satisfying $p \equiv 3 \pmod 4$ and consider the set

$$S = \{j^2 \bmod p \mid j \in \mathbf{Z}_p \setminus \{0\}\},$$

of all quadratic residues modulo $p$. Note that for every nonzero element $a \in \mathbf{Z}_p$, the set $aS \pmod p$ is either the set of all quadratic residues or the set of all quadratic non-residues modulo $p$. The main result of Graham and Ringrose [GR90] asserts that for infinitely many primes $p$, the smallest quadratic nonresidue modulo $p$ is at least $\Omega(\log p \log\log\log p)$ (this result holds for primes $p \equiv 3 \pmod 4$ as well, as follows from the remark at the end of [GR90]). Since for such primes $p$, $-1$ is a quadratic nonresidue, it follows that for the above $S$ and for any choice of $a, b \in \mathbf{Z}_p$, the set $aS + b$ (computed in $\mathbf{Z}_p$) avoids intervals of length at least $\Omega(\log p \log\log\log p)$. Choosing $m = c \log p \log\log\log p$ for an appropriate (small) constant $c$, and defining $n = \lceil p/m \rceil$, it follows that $|S| = (p-1)/2 = \Omega(n \log n \log\log\log n)$ is not mapped onto $[0, n-1]$ under any choice of $a$ and $b$.

A final question is whether there exists a class $\mathcal{H}$ of size only $2^{O(\log\log|U| + \log n)}$ and with $L_n^n(\mathcal{H}) = O(\log n / \log\log n)$. Note that linear maps over $\mathbf{Z}_2$, even combined with collapsing the universe, use $O(\log\log|U| + (\log n)^2)$ random bits while the simple scheme using higher degree polynomials uses $O(\log\log|U| + (\log n)^2/\log\log n)$.

## Acknowledgment

We thank Sanjeev Arora for helpful remarks.

## References

[ABI86] N. Alon, L. Babai and A. Itai, A fast and simple randomized parallel algorithm for the maximal independent set problem. *J. Algorithms* **7** (1986) 567–583.

[AHNR95] A. Andersson, T. Hagerup, S. Nilsson, and R. Raman, Sorting in linear time?, in: *Proc. 27th ACM Symposium on Theory of Computing*, 1995, pp. 427–436.

[CW79] J. L. Carter and M. N. Wegman, Universal classes of hash functions, *J. Comput. Syst. Sci.* **18** (1979) 143–154.

[CLR90] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, Introduction to Algorithms, MIT Press, 1990.

[DHKP97] M. Dietzfelbinger, T. Hagerup, J. Katajainen, and M. Penttonen, A reliable randomized algorithm for the closest-pair problem, *J. Algorithms* 25(1997), 19–51.

[DM90] M. Dietzfelbinger and F. Meyer auf der Heide, Dynamic hashing in real time, in: J. Buchmann, H. Ganzinger, W. J. Paul (Eds.): *Informatik · Festschrift zum 60. Geburtstag von Günter Hotz*, Teubner-Texte zur Informatik, Band 1, B. G. Teubner, 1992, pp. 95–119. (A preliminary version appeared under the title "A New Universal Class of Hash Functions and Dynamic Hashing in Real Time" in *ICALP'90*.)

[DKMHRT94] M. Dietzfelbinger, A. Karlin, K. Mehlhorn, F. Meyer Auf Der Heide, H. Rohnert, R.E. Tarjan, Dynamic perfect hashing: upper and lower bounds, *SIAM J. Comput.* **23** (1994) 738–761.

[DGMP92] M. Dietzfelbinger, J. Gil, Y. Matias, and N. Pippenger, Polynomial hash functions are reliable, ICALP'92, Springer LNCS 623, pp. 235–246.

[FKS84] M. L. Fredman, J. Komlós, and E. Szemerédi, Storing a sparse table with $O(1)$ worst case access time, *J. Ass. Comput. Mach.* **31** (1984) 538–544.

[GBY90] G. Gonnet and R. Baeza-Yates, *Handbook of Algorithms and Data Structures*, Addison-Wesley, 1991.

[GR90] S. W. Graham and C. J. Ringrose, Lower bounds for least quadratic nonresidues, in: *Analytic Number Theory: Proceedings of a Conference in Honor of P.T. Bateman*, B. C. Berndt et al. (Eds.), Birkhäuser, Boston, 1990.

[MCW78] G. Markowsky, J.L. Carter, and M. N. Wegman, Analysis of a universal class of hash functions, in: *Proc. 7th Conference on Math. Found. of Computer Science (MFCS)*, 1978, Springer LNCS 64, pp. 345–354.

[MV84] K. Mehlhorn and U. Vishkin, Randomized and deterministic simulations of PRAMs by parallel machines with restricted granularity of parallel memories., *Acta Informatica* **21** (1984) 339–374.

[MNT93] Y. Mansour, N. Nisan, and P. Tiwari, The computational complexity of universal hashing. *Theoretical Computer Science* **107** (1993) 121–133.

[PA95] J. Pach and P. K. Agarwal, *Combinatorial Geometry*, Wiley 1995.

[S89] A. Siegel, On universal classes of fast high performance hash functions, their time-space tradeoff, and their application, in: *Proc. 30th IEEE Symposium on Foundations of Computer Science*, 1989, pp. 20–25.

[VC71] V. A. Vapnik and A. Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Theory of Prob. Applications* **16** (1971) 264–280.