Randomized Algorithms, Lecture 7

Jacob Holm (jaho@di.ku.dk)

May 14th 2019

Good Afternoon.

Today's Lecture

Data structures

- Hashing fundamentals
- Hash Table with Linear Probing
- k-independence
- Linear probing with 5-independence

Hash function

Given a (large) universe U of keys, and a positive integer t.

Definition

A hash function $h : U \to [t]$ is a random variable, whose values are functions from $U \to [t]$.
Equivalently, for each $x \in U$, $h(x) \in [t]$ is a random variable.

When discussing hash functions, we care about

1. Space (*seed size*) needed to represent h.
2. Time needed to calculate $h(x)$ given $x \in U$.
3. Properties of the random variable.

Hash function types

Definition

A hash function $h : U \to [t]$ is *truly random* if the variables $h(x)$ for $x \in U$ are independent and uniform.

Definition

A hash function $h : U \to [t]$ is *c-universal* if, for all $x \neq y \in U$: $\Pr[h(x) = h(y)] \leq \frac{c}{t}$.

For many purposes c-universal hash functions for some small constant c are enough. We will see examples of such functions a little later today.

Linear Probing

Want to maintain a subset $S \subseteq U$. (Let $n = |S|$).

Let $t \geq \frac{3}{2}n$, and let T be a table of size t where initially $T[i] = \textbf{nil}$ for $i \in [t]$. Pick hash function $h : U \to [t]$.

```
function INSERT(x)
    i ← h(x)
    while T[i] ≠ nil do
        i ← (i + 1) mod t
    T[i] ← x
```

```
function MEMBER(x)
    i ← h(x)
    while T[i] ∉ {nil, x} do
        i ← (i + 1) mod t
    return T[i] = x
```

Linear Probing

Want to maintain a subset $S \subseteq U$. (Let $n = |S|$).

Let $t \geq \frac{3}{2}n$, and let T be a table of size t where initially $T[i] = \textbf{nil}$ for $i \in [t]$. Pick hash function $h : U \to [t]$.

```
function INSERT(x)
    i ← h(x)
    while T[i] ≠ nil do
        i ← (i + 1) mod t
    T[i] ← x
```

```
function MEMBER(x)
    i ← h(x)
    while T[i] ∉ {nil, x} do
        i ← (i + 1) mod t
    return T[i] = x
```

Linear Probing

Want to maintain a subset $S \subseteq U$. (Let $n = |S|$).

Let $t \geq \frac{3}{2}n$, and let T be a table of size t where initially $T[i] = \textbf{nil}$ for $i \in [t]$. Pick hash function $h : U \to [t]$.

function $\textsc{Insert}(x)$
  $i \leftarrow h(x)$
  while $T[i] \neq \textbf{nil}$ **do**
    $i \leftarrow (i+1) \bmod t$
  $T[i] \leftarrow x$

function $\textsc{Member}(x)$
  $i \leftarrow h(x)$
  while $T[i] \notin \{\textbf{nil}, x\}$ **do**
    $i \leftarrow (i+1) \bmod t$
  return $T[i] = x$

Linear Probing

Want to maintain a subset $S \subseteq U$. (Let $n = |S|$).

Let $t \geq \frac{3}{2}n$, and let T be a table of size t where initially $T[i] = \textbf{nil}$ for $i \in [t]$. Pick hash function $h : U \to [t]$.

```
function INSERT(x)
    i ← h(x)
    while T[i] ≠ nil do
        i ← (i + 1) mod t
    T[i] ← x
```

```
function MEMBER(x)
    i ← h(x)
    while T[i] ∉ {nil, x} do
        i ← (i + 1) mod t
    return T[i] = x
```

Linear Probing

Delete is tricky. Why?

Linear Probing

Delete is tricky. Why? If not careful, MEMBER will stop working.

For $x \in S$ let $\ell(x) \in [t]$ such that $T[\ell(x)] = x$.

Invariant

For all $x \in S$, $T[h(x)], \ldots, T[\ell(x)]$ are full.

Linear Probing

Delete is tricky. Why? If not careful, MEMBER will stop working.

				$h(q)$				$\ell(q)$				
$\cdots$		x_3	x_7	x_4	x_5	x_{12}	x_9	q	x_8	x_{10}		$\cdots$

DELETE(x_{12})

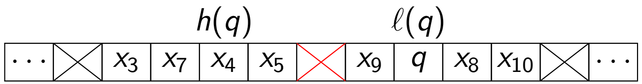
For $x \in S$ let $\ell(x) \in [t]$ such that $T[\ell(x)] = x$.

Invariant

For all $x \in S$, $T[h(x)], \ldots, T[\ell(x)]$ are full.

Linear Probing

Delete is tricky. Why? If not careful, MEMBER will stop working.



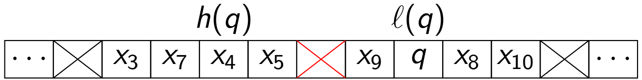
For $x \in S$ let $\ell(x) \in [t]$ such that $T[\ell(x)] = x$.

Invariant

For all $x \in S$, $T[h(x)], \ldots, T[\ell(x)]$ are full.

Linear Probing

Delete is tricky. Why? If not careful, MEMBER will stop working.



MEMBER(q) — fails

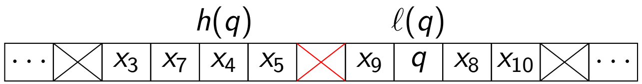
For $x \in S$ let $\ell(x) \in [t]$ such that $T[\ell(x)] = x$.

Invariant

For all $x \in S$, $T[h(x)], \ldots, T[\ell(x)]$ are full.

Linear Probing

Delete is tricky. Why? If not careful, MEMBER will stop working.



MEMBER(q) — fails

For $x \in S$ let $\ell(x) \in [t]$ such that $T[\ell(x)] = x$.

Invariant

For all $x \in S$, $T[h(x)], \ldots, T[\ell(x)]$ are full.

Linear Probing

The details of DELETE are not so important, only that it takes time proportional to the distance from $h(x)$ to the first **nil**.

```
function DELETE(x)
    i ← h(x)
    while T[i] ∉ {nil, x} do
        i ← (i + 1) mod t
    if T[i] = x then
        j ← i, i ← (i + 1) mod t
        while T[i] ≠ nil do
            k ← h(T[i])
            if (i − k) mod t ≥ (j − k) mod t then
                T[j] ← T[i], j ← i
            i ← (i + 1) mod t
        T[j] ← nil
```

Linear Probing

Define the *cost* of an element $x \in U$ to be the distance from $h(x)$ to the nearest **nil**.

Each of $\textsc{insert}(x)$, $\textsc{member}(x)$, and $\textsc{delete}(x)$ take time proportional to $\text{cost}(x)$.

Theorem (Knuth 1963)

If h is fully random, then $\mathbb{E}_h[\text{cost}(x)] \in \mathcal{O}(1)$.

Theorem (Pagh et al. 2007)

If h is 5-independent, then $\mathbb{E}_h[\text{cost}(x)] \in \mathcal{O}(1)$.

Linear Probing

Define the *cost* of an element $x \in U$ to be the distance from $h(x)$ to the nearest **nil**.

Each of INSERT(x), MEMBER(x), and DELETE(x) take time proportional to cost(x).

Theorem (Knuth 1963)

If h is fully random, then $\mathbb{E}_h[\text{cost}(x)] \in \mathcal{O}(1)$.

Theorem (Pagh et al. 2007)

If h is 5*-independent, then* $\mathbb{E}_h[\text{cost}(x)] \in \mathcal{O}(1)$.

A small remark on notation. I use a subscript h on the expectation symbol to emphasize that the expectation is wrt. the random choice of h, and *not* wrt some random choice of x.

Linear Probing

This is what we focus on for the rest of the talk. Starting with what it means for a hash function to be k-independent.

Define the *cost* of an element $x \in U$ to be the distance from $h(x)$ to the nearest **nil**.

Each of $\textsc{Insert}(x)$, $\textsc{Member}(x)$, and $\textsc{Delete}(x)$ take time proportional to $\mathrm{cost}(x)$.

Theorem (Knuth 1963)

If h is fully random, then $\mathbb{E}_h[\mathrm{cost}(x)] \in \mathcal{O}(1)$.

Theorem (Pagh et al. 2007)

If h is 5-independent, then $\mathbb{E}_h[\mathrm{cost}(x)] \in \mathcal{O}(1)$.

k-independence

Definition

A hash function $h : U \to [t]$ is *k-independent* if

1. any k distinct keys hash independently; and
2. each hash value is uniform in $[t]$.

A 2-independent hash function is sometimes called *strongly universal*.

k-independence

Definition

A hash function $h : U \to [t]$ is *k-independent* if

1. any k distinct keys hash independently; and
2. each hash value is uniform in $[t]$.

A 2-independent hash function is sometimes called *strongly universal*.

k-independence

Let p be a prime, let $a_0, \ldots, a_{k-1} \in [p]$ be chosen uniformly and independently at random. Define

$$h(x) := \left(\left(\sum_{i=0}^{k-1} a_i x^i\right) \bmod p\right) \bmod t$$

Then $h : [p] \to [t]$ is a k-independent hash function.

k-independence

Let p be a prime, let $a_0, \ldots, a_{k-1} \in [p]$ be chosen uniformly and independently at random. Define

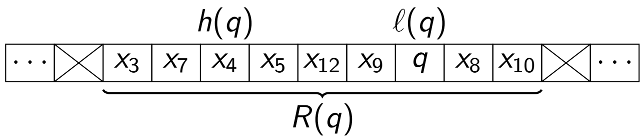
$$h(x) := \left(\left(\sum_{i=0}^{k-1} a_i x^i\right) \bmod p\right) \bmod t$$

Then $h : [p] \to [t]$ is a k-independent hash function.

Well, close enough. Any k keys hash independently, but each key is not entirely uniform. In Assignment 4 you'll prove this doesn't matter if $t \ll p$.

Linear probing with 5-independence

For $q \in U$ let $R(q) \subseteq [t]$ be the longest filled run containing $h(q)$.



Note $\text{cost}(q) \leq |R(q)|$ and that for any maximal filled run R, exactly $|R|$ elements in S hash to R.

Want to bound $\mathbb{E}_h[|R(q)|]$.

Linear probing with 5-independence

For $q \in U$ let $R(q) \subseteq [t]$ be the longest filled run containing $h(q)$.

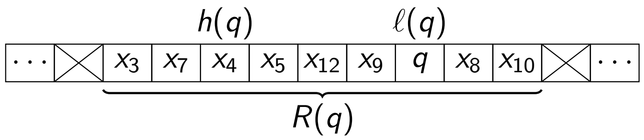
$$\boxed{\cdots} \boxed{\times} \underbrace{\boxed{x_3}\boxed{x_7}\overset{h(q)}{\boxed{x_4}}\boxed{x_5}\boxed{x_{12}}\boxed{x_9}\overset{\ell(q)}{\boxed{q}}\boxed{x_8}\boxed{x_{10}}}_{R(q)} \boxed{\times} \boxed{\cdots}$$

Note $\text{cost}(q) \leq |R(q)|$ and that for any maximal filled run R, exactly $|R|$ elements in S hash to R.

Want to bound $\mathbb{E}_h[|R(q)|]$.

Linear probing with 5-independence

For $q \in U$ let $R(q) \subseteq [t]$ be the longest filled run containing $h(q)$.

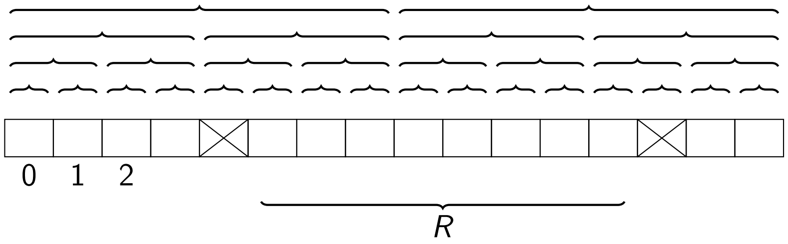


Note $\text{cost}(q) \leq |R(q)|$ and that for any maximal filled run R, exactly $|R|$ elements in S hash to R.

Want to bound $\mathbb{E}_h[|R(q)|]$.

Linear probing with 5-independence

Trick — Focus on *dyadic ℓ-intervals*: intervals of length 2^ℓ starting at positions that are 0 mod 2^ℓ.

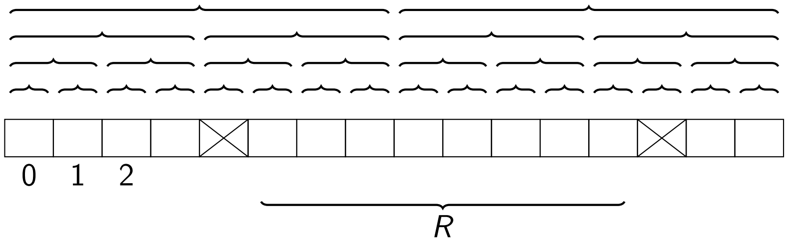


For simplicity, we will assume t is a power of 2.

We'll show that if $R(q)$ is large, a similarly sized dyadic interval "close to" $h(q)$ has many hits.

Linear probing with 5-independence

Trick — Focus on *dyadic ℓ-intervals*: intervals of length 2^ℓ starting at positions that are 0 mod 2^ℓ.

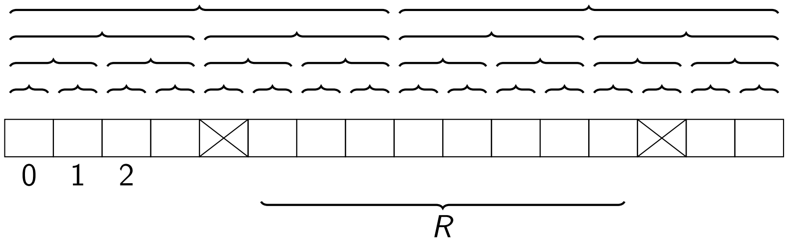


For simplicity, we will assume t is a power of 2.

We'll show that if $R(q)$ is large, a similarly sized dyadic interval "close to" $h(q)$ has many hits.

Linear probing with 5-independence

Trick — Focus on *dyadic ℓ-intervals*: intervals of length 2^ℓ starting at positions that are 0 mod 2^ℓ.



For simplicity, we will assume t is a power of 2.

We'll show that if $R(q)$ is large, a similarly sized dyadic interval "close to" $h(q)$ has many hits.

Linear probing with 5-independence

For any maximal filled run R, whether it contains $h(q)$ or not.

Let R be any maximal, filled run.

Lemma

If $|R| \geq 2^{\ell+2}$, one of first 4 ℓ-intervals intersecting R has at least $\frac{3}{4}2^{\ell}$ keys in $S \setminus \{q\}$ hashing into it.

Proof.

Let $L = \left(\cup_{i=0}^{3} I_i\right) \cap R$. $|L| \geq 3 \cdot 2^{\ell} + 1$. At least $|L|$ keys hash to L, so some I_i is hit by $\frac{3}{4}2^{\ell}$ keys from $S \setminus \{q\}$. $\square$

Linear probing with 5-independence

Let R be any maximal, filled run.

Lemma

If $|R| \geq 2^{\ell+2}$, one of first 4 ℓ-intervals intersecting R has at least $\frac{3}{4}2^{\ell}$ keys in $S \setminus \{q\}$ hashing into it.

Proof.

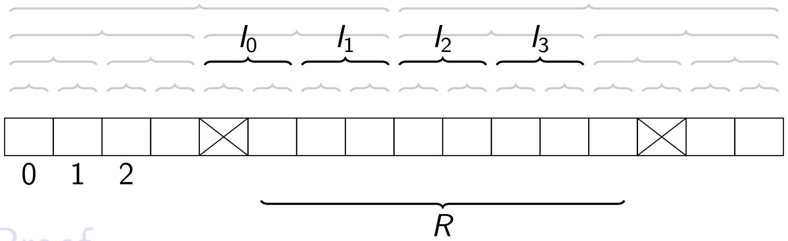
Let $L = \left(\cup_{i=0}^{3} I_i\right) \cap R$. $|L| \geq 3 \cdot 2^{\ell} + 1$. At least $|L|$ keys hash to L, so some I_i is hit by $\frac{3}{4}2^{\ell}$ keys from $S \setminus \{q\}$. $\square$

Linear probing with 5-independence

Let R be any maximal, filled run.

Lemma

If $|R| \geq 2^{\ell+2}$, one of first 4 ℓ-intervals intersecting R has at least $\frac{3}{4}2^{\ell}$ keys in $S \setminus \{q\}$ hashing into it.



Proof.

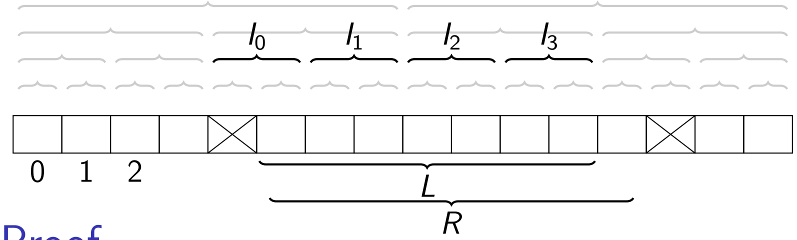
Let $L = \left(\cup_{i=0}^{3} I_i\right) \cap R$. $|L| \geq 3 \cdot 2^{\ell} + 1$. At least $|L|$ keys hash to L, so some I_i is hit by $\frac{3}{4}2^{\ell}$ keys from $S \setminus \{q\}$. $\square$

Linear probing with 5-independence

Let R be any maximal, filled run.

Lemma

If $|R| \geq 2^{\ell+2}$, one of first 4 ℓ-intervals intersecting R has at least $\frac{3}{4}2^{\ell}$ keys in $S \setminus \{q\}$ hashing into it.



Proof.

Let $L = \left(\cup_{i=0}^{3} I_i\right) \cap R$. $|L| \geq 3 \cdot 2^{\ell} + 1$. At least $|L|$ keys hash to L, so some I_i is hit by $\frac{3}{4}2^{\ell}$ keys from $S \setminus \{q\}$. $\square$

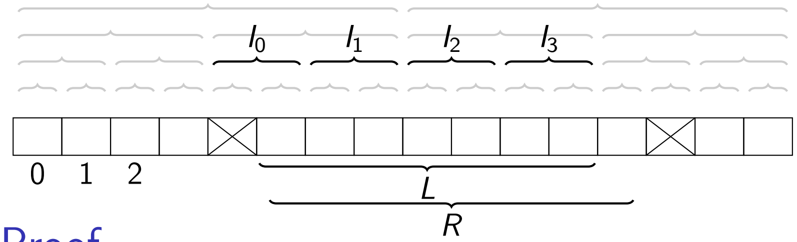
Linear probing with 5-independence

By definition, $|I_0 \cap R| \geq 1$ and $|I_i \cap R| = 2^\ell$ for $i \in \{1, 2, 3\}$.

Let R be any maximal, filled run.

Lemma

If $|R| \geq 2^{\ell+2}$, one of first 4 ℓ-intervals intersecting R has at least $\frac{3}{4}2^\ell$ keys in $S \setminus \{q\}$ hashing into it.



Proof.

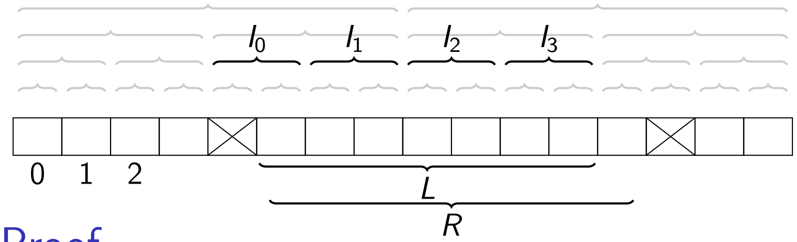
Let $L = \left(\cup_{i=0}^{3} I_i\right) \cap R$. $|L| \geq 3 \cdot 2^\ell + 1$. At least $|L|$ keys hash to L, so some I_i is hit by $\frac{3}{4}2^\ell$ keys from $S \setminus \{q\}$. □

Linear probing with 5-independence

Let R be any maximal, filled run.

Lemma

If $|R| \geq 2^{\ell+2}$, one of first 4 ℓ-intervals intersecting R has at least $\frac{3}{4}2^{\ell}$ keys in $S \setminus \{q\}$ hashing into it.



Proof.

Let $L = \left(\cup_{i=0}^{3} I_i\right) \cap R$. $|L| \geq 3 \cdot 2^{\ell} + 1$. At least $|L|$ keys hash to L, so some I_i is hit by $\frac{3}{4}2^{\ell}$ keys from $S \setminus \{q\}$. □

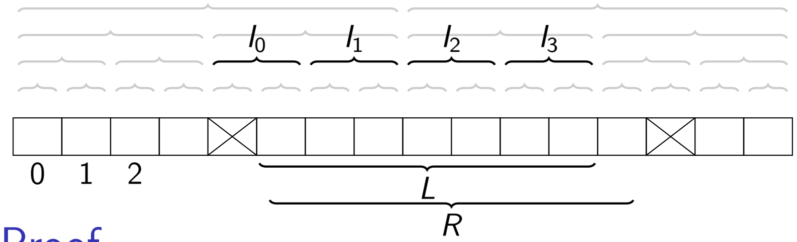
Because the preceding cell is **nil**.

Linear probing with 5-independence

Let R be any maximal, filled run.

Lemma

If $|R| \geq 2^{\ell+2}$, one of first 4 ℓ-intervals intersecting R has at least $\frac{3}{4}2^{\ell}$ keys in $S \setminus \{q\}$ hashing into it.



Proof.

Let $L = \left(\cup_{i=0}^{3} I_i\right) \cap R$. $|L| \geq 3 \cdot 2^{\ell} + 1$. At least $|L|$ keys hash to L, so some I_i is hit by $\frac{3}{4}2^{\ell}$ keys from $S \setminus \{q\}$. □

There are at least $3 \cdot 2^{\ell}$ keys in L that are not q. On average, each of $I_0, \ldots, I_3$ is hit by $\frac{1}{4}$ of these. Thus at least one must be hit by this many.

Linear probing with 5-independence

Lemma

If $2^{\ell+2} \leq |R(q)| < 2^{\ell+3}$, one of the 12 following ℓ-intervals has at least $\frac{3}{4}2^{\ell}$ keys in $S \setminus \{q\}$ hashing into it:

- *The ℓ-interval I_q containing $h(q)$; or one of*
- *the 8 ℓ-intervals to the left of I_q; or one of*
- *the 3 ℓ-intervals to the right of I_q.*

Proof.

Since $|R(q)| < 8 \cdot 2^{\ell}$, first ℓ-interval intersecting $R(q)$ is at most 8 before I_q. The first 4 ℓ-intervals intersecting $R(q)$ are therefore among the 12 intervals mentioned. Use previous lemma. □

Linear probing with 5-independence

Lemma

If $2^{\ell+2} \leq |R(q)| < 2^{\ell+3}$, one of the 12 following ℓ-intervals has at least $\frac{3}{4}2^{\ell}$ keys in $S \setminus \{q\}$ hashing into it:

- *The ℓ-interval I_q containing $h(q)$; or one of*
- *the 8 ℓ-intervals to the left of I_q; or one of*
- *the 3 ℓ-intervals to the right of I_q.*

Proof.

Since $|R(q)| < 8 \cdot 2^{\ell}$, first ℓ-interval intersecting $R(q)$ is at most 8 before I_q. The first 4 ℓ-intervals intersecting $R(q)$ are therefore among the 12 intervals mentioned. Use previous lemma. □

Linear probing with 5-independence

Lemma

If $2^{\ell+2} \leq |R(q)| < 2^{\ell+3}$, one of the 12 following ℓ-intervals has at least $\frac{3}{4}2^{\ell}$ keys in $S \setminus \{q\}$ hashing into it:

- *The ℓ-interval I_q containing $h(q)$; or one of*
- *the 8 ℓ-intervals to the left of I_q; or one of*
- *the 3 ℓ-intervals to the right of I_q.*

Proof.

Since $|R(q)| < 8 \cdot 2^{\ell}$, first ℓ-interval intersecting $R(q)$ is at most 8 before I_q. The first 4 ℓ-intervals intersecting $R(q)$ are therefore among the 12 intervals mentioned. Use previous lemma. $\square$

Linear probing with 5-independence

Lemma

If $2^{\ell+2} \leq |R(q)| < 2^{\ell+3}$, one of the 12 following ℓ-intervals has at least $\frac{3}{4}2^{\ell}$ keys in $S \setminus \{q\}$ hashing into it:

- *The ℓ-interval I_q containing $h(q)$; or one of*
- *the 8 ℓ-intervals to the left of I_q; or one of*
- *the 3 ℓ-intervals to the right of I_q.*

Proof.

Since $|R(q)| < 8 \cdot 2^{\ell}$, first ℓ-interval intersecting $R(q)$ is at most 8 before I_q. The first 4 ℓ-intervals intersecting $R(q)$ are therefore among the 12 intervals mentioned. Use previous lemma. $\square$

Linear probing with 5-independence

Lemma

If $2^{\ell+2} \leq |R(q)| < 2^{\ell+3}$, one of the 12 following ℓ-intervals has at least $\frac{3}{4}2^{\ell}$ keys in $S \setminus \{q\}$ hashing into it:

- *The ℓ-interval I_q containing $h(q)$; or one of*
- *the 8 ℓ-intervals to the left of I_q; or one of*
- *the 3 ℓ-intervals to the right of I_q.*

Proof.

Since $|R(q)| < 8 \cdot 2^{\ell}$, first ℓ-interval intersecting $R(q)$ is at most 8 before I_q. The first 4 ℓ-intervals intersecting $R(q)$ are therefore among the 12 intervals mentioned. Use previous lemma. $\square$

Linear probing with 5-independence

Lemma

If $2^{\ell+2} \leq |R(q)| < 2^{\ell+3}$, one of the 12 following ℓ-intervals has at least $\frac{3}{4}2^{\ell}$ keys in $S \setminus \{q\}$ hashing into it:

- *The ℓ-interval I_q containing $h(q)$; or one of*
- *the 8 ℓ-intervals to the left of I_q; or one of*
- *the 3 ℓ-intervals to the right of I_q.*

Proof.

Since $|R(q)| < 8 \cdot 2^{\ell}$, first ℓ-interval intersecting $R(q)$ is at most 8 before I_q. The first 4 ℓ-intervals intersecting $R(q)$ are therefore among the 12 intervals mentioned. Use previous lemma. $\square$

Linear probing with 5-independence

Lemma

If $2^{\ell+2} \le |R(q)| < 2^{\ell+3}$, one of the 12 following ℓ-intervals has at least $\frac{3}{4}2^{\ell}$ keys in $S \setminus \{q\}$ hashing into it:

- *The ℓ-interval I_q containing $h(q)$; or one of*
- *the 8 ℓ-intervals to the left of I_q; or one of*
- *the 3 ℓ-intervals to the right of I_q.*

Proof.

Since $|R(q)| < 8 \cdot 2^{\ell}$, first ℓ-interval intersecting $R(q)$ is at most 8 before I_q. The first 4 ℓ-intervals intersecting $R(q)$ are therefore among the 12 intervals mentioned. Use previous lemma. $\square$

Linear probing with 5-independence

Corollary

Let P_ℓ be the probability that any given ℓ-interval has at least $\frac{3}{4}2^\ell$ keys in $S \setminus \{q\}$ hashing into it. Then $\Pr[2^{l+2} \leq |R(q)| < 2^{\ell+3}] \leq 12P_\ell$.

Thus

$$\mathbb{E}_h[|R(q)|] \leq 3 + \sum_{\ell=0}^{\log_2 t} 2^{\ell+3} \cdot 12P_\ell$$

$$\in \mathcal{O}\Big(1 + \sum_{\ell=0}^{\log_2 t} 2^\ell \cdot P_\ell\Big)$$

We now want to upper bound P_ℓ.

Because the probability of a union of events is at most the sum of the probabilities of each event.

Linear probing with 5-independence

Corollary

Let P_ℓ be the probability that any given ℓ-interval has at least $\frac{3}{4}2^\ell$ keys in $S \setminus \{q\}$ hashing into it. Then $\Pr[2^{l+2} \leq |R(q)| < 2^{\ell+3}] \leq 12P_\ell$.

Thus

$$\mathop{\mathbb{E}}_h[|R(q)|] \leq 3 + \sum_{\ell=0}^{\log_2 t} 2^{\ell+3} \cdot 12P_\ell$$

$$\in \mathcal{O}\left(1 + \sum_{\ell=0}^{\log_2 t} 2^\ell \cdot P_\ell\right)$$

We now want to upper bound P_ℓ.

The corollary only tells us about runs of length at least 4, however with probability at most 1 we have a run of length at most 3, so for an upper bound it is sufficient to just add 3.

Linear probing with 5-independence

Corollary

Let P_ℓ be the probability that any given ℓ-interval has at least $\frac{3}{4}2^\ell$ keys in $S \setminus \{q\}$ hashing into it. Then $\Pr[2^{l+2} \leq |R(q)| < 2^{\ell+3}] \leq 12P_\ell$.

Thus

$$\mathop{\mathbb{E}}_{h}[|R(q)|] \leq 3 + \sum_{\ell=0}^{\log_2 t} 2^{\ell+3} \cdot 12P_\ell$$

$$\in \mathcal{O}\left(1 + \sum_{\ell=0}^{\log_2 t} 2^\ell \cdot P_\ell\right)$$

We now want to upper bound P_ℓ.

Linear probing with 5-independence

Corollary

Let P_ℓ be the probability that any given ℓ-interval has at least $\frac{3}{4}2^\ell$ keys in $S \setminus \{q\}$ hashing into it. Then $\Pr[2^{l+2} \leq |R(q)| < 2^{\ell+3}] \leq 12P_\ell$.

Thus

$$\mathop{\mathbb{E}}_h[|R(q)|] \leq 3 + \sum_{\ell=0}^{\log_2 t} 2^{\ell+3} \cdot 12P_\ell$$
$$\in \mathcal{O}\left(1 + \sum_{\ell=0}^{\log_2 t} 2^\ell \cdot P_\ell\right)$$

We now want to upper bound P_ℓ.

Linear probing with 5-independence

To get $\mathcal{O}(1)$ expected cost, we assume h is 5-independent.

Given an ℓ-interval I, for $x \in S \setminus \{q\}$ let $X_x = [h(x) \in I]$. Then $X = \sum_{x \in S \setminus \{q\}} X_x$ is the number of keys in $S \setminus \{q\}$ that hash into I, and $\mu = \mathbb{E}[X] \leq n\frac{2^\ell}{t} \leq \frac{2}{3}2^\ell$.

Since h is 5-independent, the variables X_x for $x \in S \setminus \{q\}$ are 4-wise independent.

To analyze, we use a 4th moment bound.

Linear probing with 5-independence

To get $\mathcal{O}(1)$ expected cost, we assume h is 5-independent.

Given an ℓ-interval I, for $x \in S \setminus \{q\}$ let $X_x = [h(x) \in I]$. Then $X = \sum_{x \in S \setminus \{q\}} X_x$ is the number of keys in $S \setminus \{q\}$ that hash into I, and $\mu = \mathbb{E}[X] \leq n\frac{2^\ell}{t} \leq \frac{2}{3}2^\ell$.

Since h is 5-independent, the variables X_x for $x \in S \setminus \{q\}$ are 4-wise independent.

To analyze, we use a 4th moment bound.

Linear probing with 5-independence

To get $\mathcal{O}(1)$ expected cost, we assume h is 5-independent.

Given an ℓ-interval I, for $x \in S \setminus \{q\}$ let $X_x = [h(x) \in I]$. Then $X = \sum_{x \in S \setminus \{q\}} X_x$ is the number of keys in $S \setminus \{q\}$ that hash into I, and $\mu = \mathbb{E}[X] \leq n\frac{2^\ell}{t} \leq \frac{2}{3}2^\ell$.

Since h is 5-independent, the variables X_x for $x \in S \setminus \{q\}$ are 4-wise independent.

To analyze, we use a 4th moment bound.

Linear probing with 5-independence

To get $\mathcal{O}(1)$ expected cost, we assume h is 5-independent.

Given an ℓ-interval I, for $x \in S \setminus \{q\}$ let $X_x = [h(x) \in I]$. Then $X = \sum_{x \in S \setminus \{q\}} X_x$ is the number of keys in $S \setminus \{q\}$ that hash into I, and $\mu = \mathbb{E}[X] \leq n\frac{2^\ell}{t} \leq \frac{2}{3}2^\ell$.

Since h is 5-independent, the variables X_x for $x \in S \setminus \{q\}$ are 4-wise independent.

To analyze, we use a 4th moment bound.

Linear probing with 5-independence

To get $\mathcal{O}(1)$ expected cost, we assume h is 5-independent.

Given an ℓ-interval I, for $x \in S \setminus \{q\}$ let $X_x = [h(x) \in I]$. Then $X = \sum_{x \in S \setminus \{q\}} X_x$ is the number of keys in $S \setminus \{q\}$ that hash into I, and $\mu = \mathbb{E}[X] \leq n\frac{2^\ell}{t} \leq \frac{2}{3}2^\ell$.

Since h is 5-independent, the variables X_x for $x \in S \setminus \{q\}$ are 4-wise independent.

To analyze, we use a 4th moment bound.

Linear probing with 5-independence

To get $\mathcal{O}(1)$ expected cost, we assume h is 5-independent.

Given an ℓ-interval I, for $x \in S \setminus \{q\}$ let $X_x = [h(x) \in I]$. Then $X = \sum_{x \in S \setminus \{q\}} X_x$ is the number of keys in $S \setminus \{q\}$ that hash into I, and $\mu = \mathbb{E}[X] \leq n\frac{2^\ell}{t} \leq \frac{2}{3}2^\ell$.

Since h is 5-independent, the variables X_x for $x \in S \setminus \{q\}$ are 4-wise independent.

To analyze, we use a 4th moment bound.

Linear probing with 5-independence

Now since $\frac{2}{3}2^\ell \geq \mu$, $\sqrt{\frac{2}{3}2^\ell} \geq \sqrt{2^\ell \mu}$, and

$$X \geq \frac{3}{4}2^\ell \implies X - \mu \geq \frac{3}{4}2^\ell - \mu \geq (\frac{3}{4} - \frac{2}{3})2^\ell$$
$$\implies X - \mu \geq \frac{1}{12}2^\ell > \frac{\sqrt{2^\ell}}{10}\sqrt{\mu}$$
$$\implies |X - \mu| \geq \frac{\sqrt{2^\ell}}{10}\sqrt{\mu}$$

So

$$P_\ell = \Pr[X \geq \frac{3}{4}2^\ell] \leq \Pr[|X - \mu| \geq \frac{\sqrt{2^\ell}}{10}\sqrt{\mu}]$$

Linear probing with 5-independence

Now since $\frac{2}{3}2^\ell \geq \mu$, $\sqrt{\frac{2}{3}2^\ell} \geq \sqrt{2^\ell \mu}$, and

$$
\begin{aligned}
X \geq \tfrac{3}{4}2^\ell &\implies X - \mu \geq \tfrac{3}{4}2^\ell - \mu \geq (\tfrac{3}{4} - \tfrac{2}{3})2^\ell \\
&\implies X - \mu \geq \tfrac{1}{12}2^\ell > \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu} \\
&\implies |X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}
\end{aligned}
$$

So

$$P_\ell \quad = \quad \Pr[X \geq \tfrac{3}{4}2^\ell] \quad \leq \quad \Pr[|X-\mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}]$$

Linear probing with 5-independence

Now since $\frac{2}{3}2^\ell \geq \mu$, $\sqrt{\frac{2}{3}2^\ell} \geq \sqrt{2^\ell \mu}$, and

$$
\begin{aligned}
X \geq \tfrac{3}{4}2^\ell \implies & X - \mu \geq \tfrac{3}{4}2^\ell - \mu \geq (\tfrac{3}{4} - \tfrac{2}{3})2^\ell \\
\implies & X - \mu \geq \tfrac{1}{12}2^\ell > \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu} \\
\implies & |X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}
\end{aligned}
$$

So

$$P_\ell \quad = \quad \Pr[X \geq \tfrac{3}{4}2^\ell] \quad \leq \quad \Pr[|X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}]$$

Linear probing with 5-independence

Now since $\frac{2}{3}2^\ell \geq \mu$, $\sqrt{\frac{2}{3}}2^\ell \geq \sqrt{2^\ell \mu}$, and

$$X \geq \tfrac{3}{4}2^\ell \implies X - \mu \geq \tfrac{3}{4}2^\ell - \mu \geq (\tfrac{3}{4} - \tfrac{2}{3})2^\ell$$

$$\implies X - \mu \geq \tfrac{1}{12}2^\ell > \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}$$

$$\implies |X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}$$

So

$$P_\ell \quad = \quad \Pr[X \geq \tfrac{3}{4}2^\ell] \quad \leq \quad \Pr[|X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}]$$

Linear probing with 5-independence

Now since $\frac{2}{3}2^\ell \geq \mu$, $\sqrt{\frac{2}{3}}2^\ell \geq \sqrt{2^\ell \mu}$, and

$$X \geq \tfrac{3}{4}2^\ell \implies X - \mu \geq \tfrac{3}{4}2^\ell - \mu \geq (\tfrac{3}{4} - \tfrac{2}{3})2^\ell$$
$$\implies X - \mu \geq \tfrac{1}{12}2^\ell > \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}$$
$$\implies |X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}$$

So

$$P_\ell \quad = \quad \Pr[X \geq \tfrac{3}{4}2^\ell] \quad \leq \quad \Pr[|X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}]$$

Linear probing with 5-independence

$$0.8\overline{3} = \frac{10}{12} = \sqrt{\frac{100}{144}} = \sqrt{\frac{2}{2.88}} > \sqrt{\frac{2}{3}} \approx 0.8164965809$$

Now since $\frac{2}{3}2^\ell \geq \mu$, $\sqrt{\frac{2}{3}}2^\ell \geq \sqrt{2^\ell \mu}$, and

$$\begin{aligned}
X \geq \tfrac{3}{4}2^\ell &\implies X - \mu \geq \tfrac{3}{4}2^\ell - \mu \geq (\tfrac{3}{4} - \tfrac{2}{3})2^\ell \\
&\implies X - \mu \geq \tfrac{1}{12}2^\ell > \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu} \\
&\implies |X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}
\end{aligned}$$

So

$$P_\ell \quad = \quad \Pr[X \geq \tfrac{3}{4}2^\ell] \quad \leq \quad \Pr[|X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}]$$

Linear probing with 5-independence

Now since $\frac{2}{3}2^\ell \geq \mu$, $\sqrt{\frac{2}{3}2^\ell} \geq \sqrt{2^\ell \mu}$, and

$$\begin{aligned} X \geq \tfrac{3}{4}2^\ell &\implies X - \mu \geq \tfrac{3}{4}2^\ell - \mu \geq (\tfrac{3}{4} - \tfrac{2}{3})2^\ell \\ &\implies X - \mu \geq \tfrac{1}{12}2^\ell > \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu} \\ &\implies |X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu} \end{aligned}$$

So

$$P_\ell \quad = \quad \Pr[X \geq \tfrac{3}{4}2^\ell] \quad \leq \quad \Pr[|X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}]$$

Linear probing with 5-independence

By definition of P_ℓ.

Now since $\frac{2}{3}2^\ell \geq \mu$, $\sqrt{\frac{2}{3}2^\ell} \geq \sqrt{2^\ell \mu}$, and

$$\begin{aligned} X \geq \tfrac{3}{4}2^\ell &\implies X - \mu \geq \tfrac{3}{4}2^\ell - \mu \geq (\tfrac{3}{4} - \tfrac{2}{3})2^\ell \\ &\implies X - \mu \geq \tfrac{1}{12}2^\ell > \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu} \\ &\implies |X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu} \end{aligned}$$

So

$$P_\ell \quad = \quad \Pr[X \geq \tfrac{3}{4}2^\ell] \quad \leq \quad \Pr[|X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}]$$

Linear probing with 5-independence

Now since $\frac{2}{3}2^\ell \geq \mu$, $\sqrt{\frac{2}{3}2^\ell} \geq \sqrt{2^\ell \mu}$, and

$$\begin{aligned} X \geq \tfrac{3}{4}2^\ell &\implies X - \mu \geq \tfrac{3}{4}2^\ell - \mu \geq (\tfrac{3}{4} - \tfrac{2}{3})2^\ell \\ &\implies X - \mu \geq \tfrac{1}{12}2^\ell > \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu} \\ &\implies |X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu} \end{aligned}$$

So

$$P_\ell \quad = \quad \Pr[X \geq \tfrac{3}{4}2^\ell] \quad \leq \quad \Pr[|X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}]$$

4th moment bound

We will prove this in a minute.

Theorem

If $X_0, \ldots, X_{n-1} \in \{0,1\}$ *are 4-wise independent,* $X = \sum_{i \in [n]} X_i$*, and* $\mu = \mathbb{E}[X] \geq 1$*, then for* $d > 0$

$$\Pr[|X - \mu| \geq d\sqrt{\mu}] \leq \frac{4}{d^4}$$

Thus

$$P_\ell \leq \Pr[|X - \mu| \geq \frac{\sqrt{2^\ell}}{10}\sqrt{\mu}] \leq \frac{4}{\left(\frac{\sqrt{2^\ell}}{10}\right)^4} = \frac{40000}{2^{2\ell}}$$

4th moment bound

Theorem

If $X_0, \ldots, X_{n-1} \in \{0,1\}$ are 4-wise independent, $X = \sum_{i \in [n]} X_i$, and $\mu = \mathbb{E}[X] \geq 1$, then for $d > 0$

$$\Pr[|X - \mu| \geq d\sqrt{\mu}] \leq \tfrac{4}{d^4}$$

Thus

$$P_\ell \leq \Pr[|X - \mu| \geq \tfrac{\sqrt{2^\ell}}{10}\sqrt{\mu}] \leq \frac{4}{\left(\frac{\sqrt{2^\ell}}{10}\right)^4} = \tfrac{40000}{2^{2\ell}}$$

Linear probing with 5-independence

Finally

$$\begin{aligned}
\mathop{\mathbb{E}}_{h}[\mathrm{cost}(q)] &\leq \mathop{\mathbb{E}}_{h}[|R(q)|] \\
&\in \mathcal{O}\left(1 + \sum_{\ell=0}^{\log_2 t} 2^{\ell} P_{\ell}\right) \\
&\subseteq \mathcal{O}\left(1 + \sum_{\ell=0}^{\log_2 t} 2^{\ell} \cdot \tfrac{40000}{2^{2\ell}}\right) \\
&= \mathcal{O}(1) \qquad \square
\end{aligned}$$

2nd moment bound

Let $X_0, \dots, X_{n-1} \in \{0,1\}$ be 2-independent.
Let $p_i = \Pr[X_i = 1] = \mathbb{E}_h[X_i]$, $X = \sum_{i \in [n]} X_i$,
$\mu = \mathbb{E}[X] = \sum_{i \in [n]} p_i$, and $d > 0$

$$\begin{aligned}
\sigma_i^2 &= \mathbb{E}[(X_i - p_i)^2] \\
&= p_i(1-p_i)^2 + (1-p_i)p_i^2 = p_i(1-p_i) \le p_i \\
\sigma^2 &= \sum_{i \in [n]} \sigma_i^2 \le \sum_{i \in [n]} p_i = \mu \qquad \text{(By 2-independence)}
\end{aligned}$$

$$\begin{aligned}
\Pr[|X-\mu| \ge d\sqrt{\mu}] &= \Pr[(X-\mu)^2 \ge d^2\mu] \\
&\le \frac{\mathbb{E}[(X-\mu)^2]}{d^2\mu} \qquad \text{(By Markov)} \\
&= \frac{\sigma^2}{d^2\mu} \le \frac{\mu}{d^2\mu} = \frac{1}{d^2}
\end{aligned}$$

This is just to recall what a 2nd moment bound looks like.
We *can* use this to get the bound $P_\ell \le \frac{100}{2^\ell}$, which then gives $\mathbb{E}[\mathrm{cost}(q)] \in \mathcal{O}(\log n)$, but we want to do better.

4th moment bound

Let $X_0, \ldots, X_{n-1} \in \{0, 1\}$ be 4-independent.
Let $p_i = \Pr[X_i = 1] = \mathbb{E}_h[X_i]$, $X = \sum_{i \in [n]} X_i$,
$\mu = \mathbb{E}[X] = \sum_{i \in [n]} p_i$, and $d > 0$

$$\begin{aligned} \sigma_i^2 &= \mathbb{E}[(X_i - p_i)^2] \\ &= p_i(1-p_i)^2 + (1-p_i)p_i^2 = p_i(1-p_i) \le p_i \\ \sigma^2 &= \sum_{i \in [n]} \sigma_i^2 \le \sum_{i \in [n]} p_i = \mu \qquad \text{(By } \ge \text{ 2-independence)} \end{aligned}$$

$$\begin{aligned} \Pr[|X - \mu| \ge d\sqrt{\mu}] &= \Pr[(X-\mu)^4 \ge d^4 \mu^2] \\ &\le \frac{\mathbb{E}[(X-\mu)^4]}{d^4 \mu^2} \qquad \text{(By Markov)} \end{aligned}$$

Our 4th moment bound starts exactly the same.

4th moment bound, proof

$$(X-\mu)=\sum_{i\in[n]}(X_i-p_i)$$

Note

$$\mathbb{E}[(X-\mu)^4]=\mathbb{E}\Big[\Big(\sum_{i\in[n]}(X_i-p_i)\Big)^4\Big]$$

$$=\sum_{i,j,k,l\in[n]}\mathbb{E}[(X_i-p_i)(X_j-p_j)(X_k-p_k)(X_l-p_l)]$$

If e.g. $i\notin\{j,k,l\}$ then (by 4-independence)

$$\begin{aligned}&\mathbb{E}[(X_i-p_i)(X_j-p_j)(X_k-p_k)(X_l-p_l)]\\&\quad=\mathbb{E}[(X_i-p_i)]\cdot\mathbb{E}[(X_j-p_j)(X_k-p_k)(X_l-p_l)]\\&\quad=0\cdot\mathbb{E}[(X_j-p_j)(X_k-p_k)(X_l-p_l)]\\&\quad=0\end{aligned}$$

So nonzero $\implies$ either all 4 terms equal, or two pairs of two terms (chosen in any one of $\binom{4}{2}$ ways) equal.

4th moment bound, proof

Note

$$\mathbb{E}[(X-\mu)^4] = \mathbb{E}\Big[\Big(\sum_{i\in[n]}(X_i - p_i)\Big)^4\Big]$$
$$= \sum_{i,j,k,l\in[n]} \mathbb{E}[(X_i - p_i)(X_j - p_j)(X_k - p_k)(X_l - p_l)]$$

If e.g. $i \notin \{j,k,l\}$ then (by 4-independence)

$$\begin{aligned}
&\mathbb{E}[(X_i - p_i)(X_j - p_j)(X_k - p_k)(X_l - p_l)] \\
&\quad = \mathbb{E}[(X_i - p_i)] \cdot \mathbb{E}[(X_j - p_j)(X_k - p_k)(X_l - p_l)] \\
&\quad = 0 \cdot \mathbb{E}[(X_j - p_j)(X_k - p_k)(X_l - p_l)] \\
&\quad = 0
\end{aligned}$$

So nonzero $\implies$ either all 4 terms equal, or two pairs of two terms (chosen in any one of $\binom{4}{2}$ ways) equal.

By linearity of expectation.

4th moment bound, proof

Note

$$\mathbb{E}[(X-\mu)^4] = \mathbb{E}\Big[\Big(\sum_{i\in[n]}(X_i - p_i)\Big)^4\Big]$$
$$= \sum_{i,j,k,l\in[n]} \mathbb{E}[(X_i-p_i)(X_j-p_j)(X_k-p_k)(X_l-p_l)]$$

If e.g. $i \notin \{j,k,l\}$ then (by 4-independence)

$$\mathbb{E}[(X_i-p_i)(X_j-p_j)(X_k-p_k)(X_l-p_l)]$$
$$= \mathbb{E}[(X_i-p_i)] \cdot \mathbb{E}[(X_j-p_j)(X_k-p_k)(X_l-p_l)]$$
$$= 0 \cdot \mathbb{E}[(X_j-p_j)(X_k-p_k)(X_l-p_l)]$$
$$= 0$$

So nonzero $\implies$ either all 4 terms equal, or two pairs of two terms (chosen in any one of $\binom{4}{2}$ ways) equal.

By 4-independence.

4th moment bound, proof

Note

$$\mathbb{E}[(X-\mu)^4] = \mathbb{E}\Big[\Big(\sum_{i\in[n]}(X_i - p_i)\Big)^4\Big]$$
$$= \sum_{i,j,k,l\in[n]} \mathbb{E}[(X_i - p_i)(X_j - p_j)(X_k - p_k)(X_l - p_l)]$$

If e.g. $i \notin \{j,k,l\}$ then (by 4-independence)

$$\mathbb{E}[(X_i - p_i)(X_j - p_j)(X_k - p_k)(X_l - p_l)]$$
$$= \mathbb{E}[(X_i - p_i)] \cdot \mathbb{E}[(X_j - p_j)(X_k - p_k)(X_l - p_l)]$$
$$= 0 \cdot \mathbb{E}[(X_j - p_j)(X_k - p_k)(X_l - p_l)]$$
$$= 0$$

So nonzero $\implies$ either all 4 terms equal, or two pairs of two terms (chosen in any one of $\binom{4}{2}$ ways) equal.

4th moment bound, proof

Note

$$\mathbb{E}[(X-\mu)^4] = \mathbb{E}\Big[\Big(\sum_{i\in[n]}(X_i - p_i)\Big)^4\Big]$$
$$= \sum_{i,j,k,l\in[n]} \mathbb{E}[(X_i - p_i)(X_j - p_j)(X_k - p_k)(X_l - p_l)]$$

If e.g. $i \notin \{j,k,l\}$ then (by 4-independence)

$$\mathbb{E}[(X_i - p_i)(X_j - p_j)(X_k - p_k)(X_l - p_l)]$$
$$= \mathbb{E}[(X_i - p_i)] \cdot \mathbb{E}[(X_j - p_j)(X_k - p_k)(X_l - p_l)]$$
$$= 0 \cdot \mathbb{E}[(X_j - p_j)(X_k - p_k)(X_l - p_l)]$$
$$= 0$$

So nonzero $\implies$ either all 4 terms equal, or two pairs of two terms (chosen in any one of $\binom{4}{2}$ ways) equal.

4th moment bound, proof

Note

$$\mathbb{E}[(X-\mu)^4] = \mathbb{E}\Big[\Big(\sum_{i\in[n]}(X_i - p_i)\Big)^4\Big]$$
$$= \sum_{i,j,k,l\in[n]} \mathbb{E}[(X_i-p_i)(X_j-p_j)(X_k-p_k)(X_l-p_l)]$$

If e.g. $i \notin \{j,k,l\}$ then (by 4-independence)

$$\begin{aligned}
&\mathbb{E}[(X_i-p_i)(X_j-p_j)(X_k-p_k)(X_l-p_l)] \\
&\quad = \mathbb{E}[(X_i-p_i)]\cdot\mathbb{E}[(X_j-p_j)(X_k-p_k)(X_l-p_l)] \\
&\quad = 0\cdot\mathbb{E}[(X_j-p_j)(X_k-p_k)(X_l-p_l)] \\
&\quad = 0
\end{aligned}$$

So nonzero $\implies$ either all 4 terms equal, or two pairs of two terms (chosen in any one of $\binom{4}{2}$ ways) equal.

4th moment bound, proof

Note

$$\mathbb{E}[(X-\mu)^4] = \mathbb{E}\Big[\Big(\sum_{i\in[n]}(X_i - p_i)\Big)^4\Big]$$
$$= \sum_{i,j,k,l\in[n]} \mathbb{E}[(X_i-p_i)(X_j-p_j)(X_k-p_k)(X_l-p_l)]$$

If e.g. $i \notin \{j,k,l\}$ then (by 4-independence)

$$\begin{aligned}
&\mathbb{E}[(X_i-p_i)(X_j-p_j)(X_k-p_k)(X_l-p_l)]\\
&\quad = \mathbb{E}[(X_i-p_i)]\cdot\mathbb{E}[(X_j-p_j)(X_k-p_k)(X_l-p_l)]\\
&\quad = 0\cdot\mathbb{E}[(X_j-p_j)(X_k-p_k)(X_l-p_l)]\\
&\quad = 0
\end{aligned}$$

So nonzero $\implies$ either all 4 terms equal, or two pairs of two terms (chosen in any one of $\binom{4}{2}$ ways) equal.

4th moment bound, proof

Note

$$\begin{aligned}
\mathbb{E}[(X-\mu)^4] &= \sum_{a\in[n]} \mathbb{E}[(X_a - p_a)^4] \\
&\quad + \binom{4}{2} \sum_{\substack{a,b\in[n] \\ a<b}} \mathbb{E}[(X_a - p_a)^2]\,\mathbb{E}[(X_b - p_b)^2] \\
&\le \sum_{a\in[n]} \mathbb{E}[(X_a - p_a)^2] \\
&\quad + \binom{4}{2} \sum_{\substack{a,b\in[n] \\ a<b}} \mathbb{E}[(X_a - p_a)^2]\,\mathbb{E}[(X_b - p_b)^2] \\
&= \sum_{a\in[n]} \sigma_a^2 + 6 \sum_{\substack{a,b\in[n] \\ a<b}} \sigma_a^2 \sigma_b^2
\end{aligned}$$

4th moment bound, proof

Note

$$\mathbb{E}[(X-\mu)^4] = \sum_{a\in[n]} \mathbb{E}[(X_a - p_a)^4] + \binom{4}{2} \sum_{\substack{a,b\in[n]\\ a<b}} \mathbb{E}[(X_a-p_a)^2]\,\mathbb{E}[(X_b-p_b)^2]$$

$$\leq \sum_{a\in[n]} \mathbb{E}[(X_a - p_a)^2] + \binom{4}{2} \sum_{\substack{a,b\in[n]\\ a<b}} \mathbb{E}[(X_a-p_a)^2]\,\mathbb{E}[(X_b-p_b)^2]$$

$$= \sum_{a\in[n]} \sigma_a^2 + 6 \sum_{\substack{a,b\in[n]\\ a<b}} \sigma_a^2 \sigma_b^2$$

$X_a, p_a \in [0,1]$, so $|X_a - p_a| \leq 1$, and thus $(X_a - p_a)^4 \leq (X_a - p_a)^2$

4th moment bound, proof

Note

$$
\begin{aligned}
\mathbb{E}[(X-\mu)^4] &= \sum_{a\in[n]} \mathbb{E}[(X_a-p_a)^4] \\
&\qquad + \binom{4}{2} \sum_{\substack{a,b\in[n]\\ a<b}} \mathbb{E}[(X_a-p_a)^2]\,\mathbb{E}[(X_b-p_b)^2] \\
&\leq \sum_{a\in[n]} \mathbb{E}[(X_a-p_a)^2] \\
&\qquad + \binom{4}{2} \sum_{\substack{a,b\in[n]\\ a<b}} \mathbb{E}[(X_a-p_a)^2]\,\mathbb{E}[(X_b-p_b)^2] \\
&= \sum_{a\in[n]} \sigma_a^2 + 6 \sum_{\substack{a,b\in[n]\\ a<b}} \sigma_a^2\sigma_b^2
\end{aligned}
$$

4th moment bound, proof

Note

$$
\begin{aligned}
\mathbb{E}[(X-\mu)^4] &= \sum_{a\in[n]} \sigma_a^2 + 6 \sum_{\substack{a,b\in[n]\\ a<b}} \sigma_a^2\sigma_b^2 \\
&\le \sum_{a\in[n]} p_a + 6 \sum_{\substack{a,b\in[n]\\ a<b}} p_a p_b && (\text{Since } \sigma_i^2 \le p_i) \\
&\le \sum_{a\in[n]} p_a + 3 \sum_{a,b\in[n]} p_a p_b \\
&= \sum_{a\in[n]} p_a + 3\Big(\sum_{i\in[n]} p_i\Big)^2 \\
&\le \mu + 3\mu^2 \\
&\le 4\mu^2 && (\text{Since } \mu \ge 1)
\end{aligned}
$$

4th moment bound, proof

Note

$$
\begin{aligned}
\mathbb{E}[(X-\mu)^4] &= \sum_{a\in[n]} \sigma_a^2 + 6 \sum_{\substack{a,b\in[n] \\ a<b}} \sigma_a^2 \sigma_b^2 \\
&\le \sum_{a\in[n]} p_a + 6 \sum_{\substack{a,b\in[n] \\ a<b}} p_a p_b && \text{(Since } \sigma_i^2 \le p_i\text{)} \\
&\le \sum_{a\in[n]} p_a + 3 \sum_{a,b\in[n]} p_a p_b \\
&= \sum_{a\in[n]} p_a + 3\Big(\sum_{i\in[n]} p_i\Big)^2 \\
&\le \mu + 3\mu^2 \\
&\le 4\mu^2 && \text{(Since } \mu \ge 1\text{)}
\end{aligned}
$$

4th moment bound, proof

Note

$$\begin{aligned}
\mathbb{E}[(X-\mu)^4] &= \sum_{a\in[n]} \sigma_a^2 + 6 \sum_{\substack{a,b\in[n]\\ a<b}} \sigma_a^2\sigma_b^2 \\
&\leq \sum_{a\in[n]} p_a + 6 \sum_{\substack{a,b\in[n]\\ a<b}} p_a p_b && (\text{Since } \sigma_i^2 \leq p_i) \\
&\leq \sum_{a\in[n]} p_a + 3 \sum_{a,b\in[n]} p_a p_b \\
&= \sum_{a\in[n]} p_a + 3\Big(\sum_{i\in[n]} p_i\Big)^2 \\
&\leq \mu + 3\mu^2 \\
&\leq 4\mu^2 && (\text{Since } \mu \geq 1)
\end{aligned}$$

4th moment bound, proof

Note

$$\begin{aligned}
\mathbb{E}[(X-\mu)^4] &= \sum_{a\in[n]} \sigma_a^2 + 6 \sum_{\substack{a,b\in[n]\\a<b}} \sigma_a^2\sigma_b^2 \\
&\le \sum_{a\in[n]} p_a + 6 \sum_{\substack{a,b\in[n]\\a<b}} p_a p_b && \text{(Since } \sigma_i^2 \le p_i\text{)} \\
&\le \sum_{a\in[n]} p_a + 3 \sum_{a,b\in[n]} p_a p_b \\
&= \sum_{a\in[n]} p_a + 3\Big(\sum_{i\in[n]} p_i\Big)^2 \\
&\le \mu + 3\mu^2 \\
&\le 4\mu^2 && \text{(Since } \mu \ge 1\text{)}
\end{aligned}$$

4th moment bound, proof

Note

$$
\begin{aligned}
\mathbb{E}[(X-\mu)^4] &= \sum_{a\in[n]} \sigma_a^2 + 6 \sum_{\substack{a,b\in[n]\\ a<b}} \sigma_a^2\sigma_b^2 \\
&\leq \sum_{a\in[n]} p_a + 6 \sum_{\substack{a,b\in[n]\\ a<b}} p_a p_b && \text{(Since } \sigma_i^2 \leq p_i\text{)} \\
&\leq \sum_{a\in[n]} p_a + 3 \sum_{a,b\in[n]} p_a p_b \\
&= \sum_{a\in[n]} p_a + 3\Big(\sum_{i\in[n]} p_i\Big)^2 \\
&\leq \mu + 3\mu^2 \\
&\leq 4\mu^2 && \text{(Since } \mu \geq 1\text{)}
\end{aligned}
$$

4th moment bound, proof

Note

$$\begin{aligned}
\mathbb{E}[(X-\mu)^4] &= \sum_{a\in[n]} \sigma_a^2 + 6\sum_{\substack{a,b\in[n]\\ a<b}} \sigma_a^2\sigma_b^2 \\
&\le \sum_{a\in[n]} p_a + 6\sum_{\substack{a,b\in[n]\\ a<b}} p_a p_b && (\text{Since } \sigma_i^2 \le p_i) \\
&\le \sum_{a\in[n]} p_a + 3\sum_{a,b\in[n]} p_a p_b \\
&= \sum_{a\in[n]} p_a + 3\Big(\sum_{i\in[n]} p_i\Big)^2 \\
&\le \mu + 3\mu^2 \\
&\le 4\mu^2 && (\text{Since } \mu \ge 1)
\end{aligned}$$

4th moment bound, proof

Finally

$$\begin{aligned}
\Pr[|X-\mu| \geq d\sqrt{\mu}] &\leq \frac{\mathbb{E}[(X-\mu)^4]}{d^4\mu^2} \\
&\leq \frac{4\mu^2}{d^4\mu^2} \qquad \text{(By last slide)} \\
&= \frac{4}{d^4} \qquad \square
\end{aligned}$$

4th moment bound, proof

Finally

$$\begin{aligned}
\Pr[|X-\mu| \geq d\sqrt{\mu}] &\leq \frac{\mathbb{E}[(X-\mu)^4]}{d^4\mu^2} \\
&\leq \frac{4\mu^2}{d^4\mu^2} \qquad \text{(By last slide)} \\
&= \frac{4}{d^4} \qquad \square
\end{aligned}$$

4th moment bound, proof

Finally

$$\Pr[|X-\mu| \geq d\sqrt{\mu}] \leq \frac{\mathbb{E}[(X-\mu)^4]}{d^4\mu^2}$$
$$\leq \frac{4\mu^2}{d^4\mu^2} \qquad \text{(By last slide)}$$
$$= \frac{4}{d^4} \qquad \square$$

Summary

- We have learned what a k-independent hash function is, and seen one example of such a function.
- Hashing with linear probing is one of the most efficient ways in practice to implement hash tables (due to cache friendliness).
- With a 5-independent hash function, it is also theoretically good, but 4-independence by itself is not enough.
- Simple tabulation hashing is only 3-independent, but is also theoretically good with linear probing.
- Next time: The probabilistic method

Summary

- We have learned what a k-independent hash function is, and seen one example of such a function.
- Hashing with linear probing is one of the most efficient ways in practice to implement hash tables (due to cache friendliness).
- With a 5-independent hash function, it is also theoretically good, but 4-independence by itself is not enough.
- Simple tabulation hashing is only 3-independent, but is also theoretically good with linear probing.
- Next time: The probabilistic method

Summary

- We have learned what a k-independent hash function is, and seen one example of such a function.
- Hashing with linear probing is one of the most efficient ways in practice to implement hash tables (due to cache friendliness).
- With a 5-independent hash function, it is also theoretically good, but 4-independence by itself is not enough.
- Simple tabulation hashing is only 3-independent, but is also theoretically good with linear probing.
- Next time: The probabilistic method

Summary

- We have learned what a k-independent hash function is, and seen one example of such a function.
- Hashing with linear probing is one of the most efficient ways in practice to implement hash tables (due to cache friendliness).
- With a 5-independent hash function, it is also theoretically good, but 4-independence by itself is not enough.
- Simple tabulation hashing is only 3-independent, but is also theoretically good with linear probing.
- Next time: The probabilistic method

Summary

- We have learned what a k-independent hash function is, and seen one example of such a function.
- Hashing with linear probing is one of the most efficient ways in practice to implement hash tables (due to cache friendliness).
- With a 5-independent hash function, it is also theoretically good, but 4-independence by itself is not enough.
- Simple tabulation hashing is only 3-independent, but is also theoretically good with linear probing.
- Next time: The probabilistic method

Summary

- We have learned what a k-independent hash function is, and seen one example of such a function.
- Hashing with linear probing is one of the most efficient ways in practice to implement hash tables (due to cache friendliness).
- With a 5-independent hash function, it is also theoretically good, but 4-independence by itself is not enough.
- Simple tabulation hashing is only 3-independent, but is also theoretically good with linear probing.
- Next time: The probabilistic method