

Lemma 1 (Constant optimizations and duplication effects for Lemma 4.1 of [hmp01]). Suppose (f, g) are each functions from $[n]$ to $\{0, 1\}^r$. Let $F_x = \{i \in [n] : f(i) = x\}$ and $G_x = \{i \in [n] : g(i) = x\}$. Let $\mu : [n] \rightarrow [n]$ be a “deduplication” map, so that for all $x, y \in \{0, 1\}^r$, μ maps all elements of $U_{xy} := \{i \in [n] : f(i) = x \wedge g(i) = y\}$ to a single arbitrary element of U_{xy} . Then in $O(n \log n)$ deterministic time and $O(n \log n)$ bits of space, one can construct $d : \{0, 1\}^r \rightarrow \{0, 1\}^r$ for which, with function $h(x) = g(x) \oplus d(f(x))$, and $H_x = \{i \in [n] : h(i) = x\}$, we have:

1. $\sum_{x \in \{0, 1\}^r} \binom{|H_x|}{2} - \sum_{i \in [n]} \frac{1}{|\mu^{-1}(i)|} \binom{|\mu^{-1}(i)|}{2} \leq \frac{1}{2^r} \binom{n}{2}$
2. $\sum_{x \in \{0, 1\}^r} \binom{|H_x|}{2} - \sum_{i \in [n]} \binom{|\mu^{-1}(i)|}{2} \leq n \left\lfloor \frac{1}{2^r} \left(\max \left(n, \sum_{x: |F_x| \geq 2} |F_x|^2 \right) - 1 \right) \right\rfloor$.

Proof. This is derived from the proofs in Section 4 of [hmp01]. To construct d , select a permutation $\pi : \{0, 1\}^r \rightarrow \{0, 1\}^r$ for which $|F_{\pi(1)}| \geq |F_{\pi(2)}| \geq \dots \geq |F_{\pi(2^r)}|$. (The last sets in the sequence will all be empty if $n < 2^r$.) Then in order, for each $i \in [2^r]$, choose $d(\pi(i))$ to have value $a \in \{0, 1\}^r$ so that the multiset $S_{a,i} := (a \oplus g(j) : h \in F_{\pi(i)})$ has no more than the average number of collisions with preceding multisets $\{S_{d(\pi(j)),j}\}_{j < i}$. The number of collisions $c(A, B)$ between two multisets $A = (a_1, \dots, a_{|A|})$ and $B = (b_1, \dots, b_{|B|})$ is defined as $|\{x \in [|A|], y \in [|B|] : a_x = b_y\}|$. Specifically, we want:

$$\begin{aligned} \sum_{j < i} c(S_{a,i}, S_{d(\pi(j)),j}) &\leq \left\lfloor \frac{1}{2^r} \sum_{b \in \{0, 1\}^r} \sum_{j < i} c(S_{b,i}, S_{d(\pi(j)),j}) \right\rfloor \\ &= \left\lfloor \frac{1}{2^r} \sum_{j < i} \sum_{b \in \{0, 1\}^r} c(S_{b,i}, S_{d(\pi(j)),j}) \right\rfloor \\ &= \left\lfloor \frac{1}{2^r} \sum_{j < i} |F_{\pi(i)}| |F_{\pi(j)}| \right\rfloor \end{aligned}$$

where the last step follows because for each $x \in F_{\pi(i)}$, $y \in F_{\pi(j)}$, there is exactly one value of $b \in \{0, 1\}^r$ for which $b \oplus g(x) = d(\pi(j)) \oplus g(y)$. This can be done (even with multi-sets!) using a dynamic search table structure as described in Section 4.3 of [hmp01].

The quantity $\sum_{j < i} c(S_{a,i}, S_{d(\pi(j)),j})$ counts the total number of colliding pairs $a, b \in [n]$ where $f(a) \neq f(b)$ and $h(a) = h(b)$. Since $g(i) = h(i) \oplus d(f(i))$, the number of colliding pairs where $a, b \in [n]$ satisfy $f(a) = f(b)$ and $h(a) = h(b)$ is equal to $\sum_{i \in [n]} \binom{|\mu^{-1}(i)|}{2}$ (the number of collisions that (f, g) have.)

Consequently,

$$\begin{aligned} \sum_{x \in \{0,1\}^r} \binom{|H_x|}{2} - \sum_{i \in [n]} \frac{1}{|\mu^{-1}(i)|} \binom{|\mu^{-1}(i)|}{2} &= \sum_{i \in \{0,1\}^r} \sum_{j < i} c(S_{a,i}, S_{d(f(j)),j}) \\ &\leq \sum_{i \in \{0,1\}^r} \left[\frac{1}{2^r} \sum_{j < i} |F_{\pi(i)}| |F_{\pi(j)}| \right] \end{aligned}$$

There are two ways to bound this. First,

$$\begin{aligned} \sum_{i \in \{0,1\}^r} \left[\frac{1}{2^r} \sum_{j < i} |F_{\pi(i)}| |F_{\pi(j)}| \right] &\leq \frac{1}{2^r} \sum_{\{i,j\} \in \binom{\{0,1\}^r}{2}} |F_{\pi(i)}| |F_{\pi(j)}| \\ &\leq \frac{1}{2^r} \cdot \frac{1}{2} \left(\sum_{i,j \in \{0,1\}^r} |F_{\pi(i)}| |F_{\pi(j)}| - \sum_{i \in [n]} |F_{\pi(i)}|^2 \right) \\ &\leq \frac{1}{2^r} \cdot \frac{n^2 - n}{2} = \frac{1}{2^r} \binom{n}{2} \end{aligned}$$

This bound does *not* use the permutation sort order; the following one does (and needs it, when $(F_{\pi(i)})_{i \in [n]}$ looks like $\sqrt{n}, \sqrt{n}, 1, 1, 1, \dots, 1$). Specifically:

$$\begin{aligned} \sum_{i \in \{0,1\}^r} \left[\frac{1}{2^r} \sum_{j < i} |F_{\pi(i)}| |F_{\pi(j)}| \right] &\leq n \max_{i \in \{0,1\}^r} \left[\frac{1}{2^r} \sum_{j < i} |F_{\pi(i)}| |F_{\pi(j)}| \right] \\ &\leq n \max_{i \in \{0,1\}^r} \begin{cases} \left\lfloor \frac{1}{2^r} \sum_{j < i} |F_{\pi(j)}|^2 \right\rfloor & \text{if } |F_{\pi(i)}| \geq 2 \\ \left\lfloor \frac{1}{2^r} \sum_{j < i} |F_{\pi(j)}| \right\rfloor & \text{if } |F_{\pi(i)}| \leq 1 \end{cases} \\ &\leq n \max_{i \in \{0,1\}^r} \begin{cases} \left\lfloor \frac{1}{2^r} \left(\sum_{x: |F_x| \geq 2} |F_x|^2 - \min_{|F_x| \geq 2} |F_x|^2 \right) \right\rfloor & \text{if } |F_{\pi(i)}| \geq 2 \\ \left\lfloor \frac{1}{2^r} (n-1) \right\rfloor & \text{if } |F_{\pi(i)}| \leq 1 \end{cases} \\ &\leq n \left\lfloor \frac{1}{2^r} \left(\max \left(n, \sum_{x: |F_x| \geq 2} |F_x|^2 \right) - 1 \right) \right\rfloor \end{aligned}$$

□

Lemma 2 (Deterministic double displacement.). *Applying Lemma 1 twice, with $r = \lceil \log_2(\alpha n) \rceil$, gives a perfect hash function mapping n unique pairs $(f_i, g_i)_{i=1}^n$ to values $\lambda_i \in \{0,1\}^r$, when $\alpha \geq \sqrt{2}$.*

Proof. First, apply Lemma 1 to $(f_i, g_i)_{i=1}^n$, producing $(h_i)_{i=1}^n$ with each $h_i \in \{0,1\}^r$ satisfying $h_i = g_i \oplus d_{1,i}(f_i)$ for some displacement function d_1 from $\{0,1\}^r \rightarrow \{0,1\}^r$. Then with $H_x := \{i \in [n] : h_i = x\}$ as defined in Lemma 1, we have $\sum_{x \in \{0,1\}^r} \binom{|H_x|}{2} \leq \frac{1}{2^r} \binom{n}{2}$. Next, apply to Lemma 1 to $(h_i, f_i)_{i=1}^n$,

producing $(\lambda_i)_{i=1}^n$ with each $\lambda_i \in \{0, 1\}^r$ satisfying $\lambda_i = f_i \oplus d_{2,i}(h_i)$ for some displacement function $d_2 : \{0, 1\}^r \rightarrow \{0, 1\}^r$. Define $\Lambda_x := \{i \in [n] : \lambda_i = x\}$. Then:

$$\frac{1}{2^r} \binom{n}{2} \geq \sum_{x \in \{0,1\}^r} \binom{|H_x|}{2} \geq \frac{1}{4} \sum_{x \in \{0,1\}^r : |H_x| \geq 2} |H_x|^2 \quad (1)$$

so by the second bound in Lemma 1:

$$\begin{aligned} \sum_{x \in \{0,1\}^r} \binom{|\Lambda_x|}{2} &\leq n \left\lfloor \frac{1}{2^r} \max \left(n-1, 4 \frac{1}{2^r} \binom{n}{2} \right) \right\rfloor \\ &= n \left\lfloor \frac{2}{2^{2r}} n(n-1) \right\rfloor && \text{if } 2^r \geq n \\ &= 0 && \text{if } 2^r \geq n\sqrt{2} \end{aligned}$$

□

Note 3. The first and second bounds of Lemma 1 do not fit together when $i \mapsto (f(i), g(i))$ is not one-to-one. It is *possible* that, when $\sum_{x \in \{0,1\}^r} \binom{|F_x|}{2} - \sum_{i \in [n]} \frac{1}{|\mu^{-1}(i)|} \binom{|\mu^{-1}(i)|}{2} \leq n$, the bound $\sum_{x \in \{0,1\}^r} \binom{|H_x|}{2} - \sum_{i \in [n]} \frac{1}{|\mu^{-1}(i)|} \binom{|\mu^{-1}(i)|}{2} = 0$ holds, but proving or disproving this may require going into the details of the search table procedure. (If there is a hard instance, it might have each nonempty multiset F_x contain two distinct g values (possibly with duplicates) structured to trick the search procedure into using a small branch of the table.)

Say that for some x , F_x has $k = |\mu(F_x)|$ equivalence classes by μ , of sizes a_1, \dots, a_k , with all $a_j \geq 1$. Because $\sum_{j \in [k]} (a_j - 1)^2 \leq \left(\sum_{j \in [k]} (a_j - 1) \right)^2$:

$$\begin{aligned} &\binom{|F_x|}{2} - \sum_{i \in F_x} \frac{1}{|\mu^{-1}(i)|} \binom{|\mu^{-1}(i)|}{2} \\ &= \binom{|F_x|}{2} - \sum_{j \in [k]} \binom{a_j}{2} \\ &= \binom{|F_x|}{2} - \frac{1}{2} \sum_{j \in [k]} [(a_j - 1)^2 + (a_j - 1)] \\ &= \binom{|F_x|}{2} - \frac{1}{2} (|F_x| - k) - \frac{1}{2} \sum_{j \in [k]} (a_j - 1)^2 \\ &\geq \binom{|F_x|}{2} - \frac{1}{2} (|F_x| - k) - \frac{1}{2} (|F_x| - k)^2 \\ &= \binom{|F_x|}{2} - \binom{|F_x| - k}{2} \\ &= \frac{1}{2} (|F_x|^2 - |F_x| - (|F_x| - k)^2 + (|F_x| - k)) \\ &= \frac{k}{2} (2|F_x| - k - 1) \end{aligned}$$

Therefore, the nontrivial collision count κ satisfies:

$$\begin{aligned}
\kappa &:= \sum_{x \in \{0,1\}^r} \binom{|F_x|}{2} - \sum_{i \in [n]} \frac{1}{|\mu^{-1}(i)|} \binom{|\mu^{-1}(i)|}{2} \\
&= \sum_{x \in \{0,1\}^r} \left(\binom{|F_x|}{2} - \sum_{i \in F_x} \frac{1}{|\mu^{-1}(i)|} \binom{|\mu^{-1}(i)|}{2} \right) \\
&\geq \sum_{x \in \{0,1\}^r} \frac{|\mu(F_x)|}{2} (2|F_x| - |\mu(F_x)| - 1)
\end{aligned}$$

This can be used to bound the cost of *delayed* deduplication for the second displacement round. For example, for each F_x , one can by a variant of insertion sort construct a sorted list of unique elements in $O(|F_x| |\mu(F_x)|)$ time, which summed over all $x \in \{0,1\}^r$ is $O(n)$. Or the search table design can be modified so that, when it is time to update table frequencies after choosing $d(\pi(i))$, the leaves are updated to indicate just *whether* an element has been used, not *how many times* it has been used, and the remaining values derived from the leaves. Then the key quantity to bound would be:

$$\begin{aligned}
\max_{i \in \{0,1\}^r} |F_{\pi(i)}| \sum_{j < i} |\mu(F_{\pi(j)})| &\leq \max \left(n-1, \sum_{x \in \{0,1\}^r: |F_x| \geq 2} |F_x| |\mu(F_x)| \right) \\
&\leq \max(n-1, 4\kappa)
\end{aligned}$$

The last inequality follows because if $|F_x| > |\mu(F_x)|$, then $|\mu(F_x)| (2|F_x| - |\mu(F_x)| - 1) \geq |F_x| |\mu(F_x)|$, while if $|F_x| = |\mu(F_x)|$ and $|F_x| \geq 2$ then $|\mu(F_x)| (2|F_x| - |\mu(F_x)| - 1) \geq \frac{1}{2} |\mu(F_x)| |F_x|$.

Corollary 4. *For deterministic double displacement, f and g both map to $\{0,1\}^r$. The displacement procedure can also be used, just once, to convert a hash function with few collisions to one with none. Then we may have $f : [n] \rightarrow \{0,1\}^t$ and $g : [n] \rightarrow \{0,1\}^r$. We require $x \mapsto (f(x), g(x))$ to be a bijection. Define F_x and G_x as in [1](#); the total number of collisions for f is $\sum_x \binom{|F_x|}{2}$. Say this is $\leq c$. Then one can construct a displacement function $d : \{0,1\}^t \rightarrow \{0,1\}^r$ for which, with function $h(x) = g(x) \oplus d(f(x))$, and $H_x = \{i \in [n] : h(i) = x\}$, we have that all $|H_x| \leq 1$ if:*

$$\max \left(n, \sum_{x: |F_x| \geq 2} |F_x|^2 \right) \leq 2^r \quad \text{where} \quad \sum_{x: |F_x| \geq 2} |F_x|^2 \leq 4c.$$

Proof. Modify the proof of Lemma [1](#); the bound on $\sum_{x: |F_x| \geq 2} |F_x|^2$ follows from Eq. [1](#). \square

Definition 5. Odd-multiply-shift (OMS) hashing. Used by [\[r95\]](#), proved 2-approximately universal by [\[dhkp97\]](#), with the hash function type known at

latest since TAOCP Vol 3 in 1973. The hash family $\mathcal{H}_{b,s}$ maps $\{0, \dots, 2^u - 1\} \equiv \{0, 1\}^u$ to $\{0, \dots, 2^s - 1\} \equiv \{0, 1\}^s$ and is parameterized by odd integers $a \in \{0, \dots, 2^u - 1\}$; individual functions are given by

$$h_\alpha(x) = (ax \bmod 2^u) \operatorname{div} 2^{u-s}.$$

Lemma 6. *For any distinct $x, y \in \{0, 1\}^u$, with α drawn uniformly at random from the odd integers in $\{0, \dots, 2^u - 1\}$, we have the following two bounds:*

$$\begin{aligned} \Pr_a[h_a(x) = h_a(y)] &\leq \begin{cases} 0 & \text{if } (x - y \bmod 2^{u-s}) = 0 \\ 1 & \text{otherwise} \end{cases} \\ \Pr_a[h_a(x) = h_a(y)] &\leq \Pr_a[a(x - y) \bmod 2^u \operatorname{div} 2^{u-s} = 0] \\ &\quad + \Pr_a[a(y - x) \bmod 2^u \operatorname{div} 2^{u-s} = 0] = \frac{2}{2^s} \end{aligned}$$

Proof. We have

$$\Pr_a[h_a(x) = h_a(y)] = \Pr_a[(ax \bmod 2^u) \operatorname{div} 2^{u-s} = (ay \bmod 2^u) \operatorname{div} 2^{u-s}]$$

Let i be the largest integer for which $x \bmod 2^i = y \bmod 2^i$. Decompose $x = x_h 2^i + c$ and $y = y_h 2^i + c$ where $c = x \bmod 2^i = y \bmod 2^i$. If $i \geq u - s$, then

$$\begin{aligned} (ax \bmod 2^u) \operatorname{div} 2^{u-s} &= (ax_h 2^i + ac) \bmod 2^u \operatorname{div} 2^{u-s} \\ &= (((ax_h \bmod 2^{u-i}) 2^i + (ac \bmod 2^u)) \bmod 2^u) \operatorname{div} 2^{u-s} \\ &= \left((ax_h \bmod 2^{u-i}) 2^{i-(u-s)} + (ac \bmod 2^u \operatorname{div} 2^{u-s}) \right) \bmod 2^{u-s} \\ &= \left((ax_h \bmod 2^{u-i}) 2^{i-(u-s)} + d \right) \bmod 2^{u-s} \end{aligned}$$

where $d = (ac \bmod 2^u \operatorname{div} 2^{u-s})$. Since a is odd, for $z \in \{0, \dots, 2^{u-i}\}$, the map $z \rightarrow (az \bmod 2^{u-i})$ is a bijection; and since $x_h \neq y_h$, it follows $h_a(x) \neq h_a(y)$.

On the other hand, if $i < u - s$, then

$$\begin{aligned} \Pr_a[(ax \bmod 2^u) \operatorname{div} 2^{u-s} = (ay \bmod 2^u) \operatorname{div} 2^{u-s}] &\leq \Pr_a[((ax \bmod 2^u) - (ay \bmod 2^u)) \operatorname{div} 2^{u-s} = 0] \\ &\quad + \Pr_a[((ay \bmod 2^u) - (ax \bmod 2^u)) \operatorname{div} 2^{u-s} = 0] \\ &= \Pr_a[(a(x - y) \bmod 2^u) \operatorname{div} 2^{u-s} = 0] \\ &\quad + \Pr_a[(a(y - x) \bmod 2^u) \operatorname{div} 2^{u-s} = 0] \end{aligned}$$

The first inequality follows since $w \operatorname{div} 2^{u-s} = z \operatorname{div} 2^{u-s}$ implies both w and z lie in the same set $\{2^{u-s}k, \dots, 2^{u-s}(k+1) - 1\}$; if $w \leq z$, then $(z - w) \bmod 2^u < 2^{u-s}$ so $(z - w) \bmod 2^u \operatorname{div} 2^{u-s} = 0$, while if $w \geq z$, then $(w - z) \bmod 2^u \operatorname{div} 2^{u-s} = 0$. Note that $(x - y) \bmod 2^u = ((x_h - y_h) 2^i) \bmod 2^u$ and

and $x_h - y_h \bmod 2 = 1$. Multiplying by $x_h - y_h$ under $\bmod 2^{u-i}$ permutes odd values in $\{0, \dots, 2^{u-i} - 1\}$, so $(a(x_h - y_h)2^i) \bmod 2^u$ is uniformly distributed over *odd* multiples of 2^i in 2^u . Since $i < u - s$, and $\{0, \dots, 2^{u-s}\}$ contains $2^{u-s-i-1}$ odd multiples of 2^i (out of 2^{u-i-1} possible values).

$$\Pr[(a(x - y) \bmod 2^u) \in \{0, \dots, 2^{u-s}\}] = \frac{1}{2^s}.$$

The $\Pr_a[h_a(x) = h_a(y)] \leq \frac{1}{2^{s-1}}$ upper bound is tight. For example, with $u = 4, s = 3$, $\Pr[h_a(1) = h_a(y)] = \frac{1}{4}$ holds $y = 3$ and $y = 11$. \square

Note 7. For comparison, the upper bound from [r95] on $\Pr_a[h_a(x) = h_a(y)]$ is

$$\Pr_a[a(x - y) \bmod 2^u \in \{0, \dots, 2^{u-s} - 1\} \cup \{2^u - 2^{u-s} + 1, \dots, 2^{u-1}\}]$$

for which a $\leq \frac{1}{2^{s-1}}$ upper bound also holds. The bound of Lemma 6 has the advantage of being slightly easier to compute in the incremental a -finding hash construction approach of [r95]. Also, either bound can be used to find that the expected number of collisions when uniformly randomly hashing an n -element set n from $\{0, 1\}^s$ to $\{0, 1\}^b$ is $\leq \frac{1}{2^{s-1}} \binom{n}{2}$. Consequently, for a perfect hash function from $\mathcal{H}_{b,s}$ to exist for all sets of size n , one needs $\frac{1}{2^{s-1}} \binom{n}{2} < 1$ which holds iff $s \geq \lceil \log(n(n-1) + 1) \rceil$.

However, this is *not* a necessary condition, and one can do slightly better in some cases. When $n = 1$, $2^{\lceil \log(n(n-1)+1) \rceil} = 1$, but when $n = 2$, $2^{\lceil \log(n(n-1)+1) \rceil} = 4$. However, given any $x, y \in \{0, 1\}^u$, let $j \in \{0, \dots, u-1\}$ be the highest bit for which $x_j \neq y_j$. If $j = u-1$, then $h_1(x) = x_j \neq y_j = h_1(y)$ and choosing $a = 1$ separates the two. Otherwise, let $a = 2^{u-1-j}$ (which is admittedly *even*); then $ax = 2^{u-1-j}x + x$; the multiplication does not overflow 2^u so $ax \div 2^{u-1} = x_j$, and similarly $ay \div 2^{u-1} = y_j$, so $h_a(x) \neq h_a(y)$. Thus a perfect hash function for $n = 2$ with $s = 1$ always exists. It *may* be possible that similar minimal perfect hashing can be done with odd multiply-shift hashes for $n = 2, 3, 4, \dots$ up to some small set size threshold, although space lower bounds imply the threshold is at most $O(\log u)$.

Definition 8. A hash function h from A to B is c -colliding on a set $S \subseteq A$ of size n if $\sum_{\{x,y\} \in \binom{S}{2}} 1_{h(x)=h(y)} \leq c$.

Lemma 9. Say that $h : A \rightarrow B$ is c -colliding on the set S . Let b_1, \dots, b_k be the sizes of the equivalence classes of S under h , so that $\sum_i b_i = n = |S|$ and $\sum_i \binom{b_i}{2} \leq c$. Let

$$s(b) = \begin{cases} b & \text{if } b \leq 2 \\ 2^{\lceil \log(b_i(b_i-1)+1) \rceil} & \text{if } b \geq 3 \end{cases}$$

Then

$$\sum_i s(b_i) < n + 4c$$

$$\max \left(n, \sum_{b_i \geq 2} s(b_i) \right) \leq \max(n, 4c)$$

(These give bounds on the total output sizes for the FKS two-level hashing scheme ([fks84]) if placing leaf hash tables disjointly using offset values, and if packing size-1 tables into the gaps of the other tables. Size 2-tables are also contiguous, so one may be able to pack those into the $\max(\frac{1}{2}s(b) - b, 0)$ gaps of length 2 left when hashing b -sized equivalence classes; to what extent this works depends on the exact equivalence class size distribution.)

Proof. Let $r(b) = \frac{s(b)}{b(b-1)/2}$ (i.e, the ratio of table size to collision count for an equivalence class). We have $r(1) = \infty$, $r(2) = 2$, $r(3) = r(4) = \frac{8}{3}$, $r(5) = \frac{16}{5}$, and $\sup_{b \geq 6} r(b) = 4$ (although *exactly* 4 is never reached; as $b \rightarrow \infty$, $r(b)$ oscillates between 2 and 4 depending on how much the $\lceil \cdot \rceil$ operation in $s(b)$ adds.)

$$\sum_{b_i \geq 2} s(b_i) < 4 \sum_{b_i \geq 2} \frac{b_i(b_i - 1)}{2} \leq 4c$$

Since $s(1) = 1$, $\sum_{b_i=1} s(b_i) = |\{i : b_i = 1\}| \leq n$, so

$$\sum_i s(b_i) = \sum_{b_i=1} s(b_i) + \sum_{b_i=2} s(b_i) \leq n + 4c$$

(This is roughly tight; if all collisions are concentrated in a single equivalence class of size $d = O(\sqrt{c})$, then the large class will contribute $s(d) = \Theta(c)$ while the small ones add $n - d$ for the singleton equivalence classes. \square)

Remark 10. If the FKS hashing scheme ([fks84]) is implemented using OMS hashing (Definition 5), with hash functions picked to have a sub-average number of collisions, using size-2 leaf tables for size-2 leaves, tightly packing size-1 subtables in the free space of leaf tables, using α and β bits per primary and secondary table entries, and with main hash output bit count s , the total space usage is

$$\alpha 2^s + \beta \max \left(n, \frac{4}{2^{s-1}} \binom{n}{2} \right)$$

which is minimized when

$$s = \min \left(\lceil \log(n-1) \rceil, \left\lceil \frac{1}{2} \log \left(\frac{4\beta \binom{n}{2}}{\alpha} \right) \right\rceil \right)$$

(Choosing $s = \lceil \frac{1}{2} \log(\frac{x}{2}) \rceil$ minimizes $2^s + \frac{x}{2^s}$ with value $\leq \frac{3}{\sqrt{2}}\sqrt{x}$, so the total space usage for the table is $\leq \max \left(\frac{6\sqrt{\beta\alpha}}{\sqrt{2}}n, \beta n + 2\alpha n \right)$.)

Lemma 11 (Parameter selection for an odd-multiply-shift + displacement table perfect hash construction.). *Let n and s be integers, and α, β nonnegative reals. Under the constraints $r + t \geq s$ and*

$$2^r \geq \max(n, 4c) \quad \text{where} \quad c = \frac{n(n-1)}{2^t}.$$

we have that the minimum value of $\alpha 2^r + \beta 2^t$ occurs when:

$$r = \min \left(\max \left(\left\lceil \frac{1}{2} \left(\log \left(\frac{\beta}{2\alpha} \right) + w \right) \right\rceil, \lceil \log n \rceil \right), w \right) \\ \text{where} \quad w = \max(s, \lceil \log(4(n(n-1))) \rceil)$$

Proof. First, note that $2^r \geq 4 \frac{n(n-1)}{2^t}$ is true iff $r+t \geq \lceil \log(4(n(n-1))) \rceil$. Since reducing either of r or t non-increases $\alpha 2^r + \beta 2^t$, the optimal value of

$$r + t = \max(s, \lceil \log(4(n(n-1))) \rceil) =: w.$$

The quantity $2^r + \frac{\beta}{\alpha} 2^w$ is minimized when $r = \left\lceil \frac{1}{2} \log \left(\frac{\beta}{\alpha} 2^w \right) \right\rceil = \left\lceil \frac{1}{2} \log \left(\frac{\beta}{2\alpha} 2^w \right) \right\rceil$, but r is also subject to the constraints $r \in [\lceil \log n \rceil, w]$. The space usage then is

$$\begin{aligned} & \min \left(\max \left(\alpha 2^{\lceil \log n \rceil} + \beta 2^{w - \lceil \log n \rceil}, \frac{3}{\sqrt{2}} \sqrt{\beta \alpha 2^w} \right), \alpha 2^w + \beta \right) \\ & \leq \min \left(\max \left(2\alpha n + \beta \max \left(\frac{2^s}{n}, 8(n-1) \right), \frac{3}{\sqrt{2}} \sqrt{\beta \alpha \max(2^s, 8n(n-1))} \right), \alpha 2^w + \beta \right) \\ & \leq \min \left(\max \left(2\alpha n + \beta \frac{2^s}{n}, 2\alpha n + 8\beta n, \frac{3}{\sqrt{2}} \sqrt{\beta \alpha 2^{s/2}}, 6\sqrt{\beta \alpha n} \right), \alpha \max(8n^2, 2^s) + \beta \right) \end{aligned}$$

□

Lemma 12 (Evaluating conditional expectation for multiply-shift collision probability.). *(This is similar to [r95, Lemma 4], but instead optimizing the simpler bounds of Lemma 6.) Let $x, y \in \{0, \dots, 2^u - 1\}$ be distinct with $(x - y) \bmod 2^u = z 2^i$ for z odd and i a nonnegative integer. Fix $\alpha \in \{0, 1\}^*$ of length α , and let $E(\alpha)$ be the set of integers $a \in \{0, \dots, 2^u - 1\}$ whose $|\alpha|$ least significant bits match α . Then for integer $s \in [0, \dots, u]$ with $i < u - s$:*

$$\Pr_{a \sim E(\alpha)} [a(x - y) \bmod 2^u \text{ div } 2^{u-s} = 0] = \begin{cases} 2^{-s} & i + |\alpha| \leq u - s \\ 1_{\{(\alpha(x-y)+2^{u-s}-1) \bmod \min(2^u, 2^{|\alpha|-i}) < 2^{u-s}\}} \frac{1}{2^{u-|\alpha|-i}} & i + |\alpha| > u - s \end{cases}$$

Proof. As in the proof of [r95, Lemma 4], we observe that a can be written as $a' 2^{|\alpha|} + \alpha$ where α is interpreted as an integer in $\{0, \dots, 2^{|\alpha|} - 1\}$ and a' is an integer in $\{0, \dots, 2^{u-|\alpha|} - 1\}$. Then:

$$\begin{aligned} p &:= \Pr_{a \sim E(\alpha)} [a(x - y) \bmod 2^u \text{ div } 2^{u-s} = 0] \\ &= \Pr_{a' \sim \{0, \dots, 2^{u-|\alpha|} - 1\}} [a' z 2^{|\alpha|+i} + \alpha z 2^i \bmod 2^u \text{ div } 2^{u-s} = 0] \end{aligned}$$

Since z is odd, $a'z \bmod 2^{u-|\alpha|-i}$ is uniformly distributed over $\{0, \dots, 2^{u-|\alpha|-i}\}$. Let $b = a'z$. Consequently,

$$p = \Pr_{b \sim \{0, \dots, 2^{u-|\alpha|-i}-1\}} \left[b2^{|\alpha|+i} \bmod 2^u \in \{\alpha z 2^i, \dots, \alpha z 2^i + 2^{u-s} - 1\} \bmod 2^u \right]$$

We now have two cases: if $|\alpha|+i \leq u-s$, then the set $\{\alpha z 2^i, \dots, \alpha z 2^i + 2^{u-s} - 1\}$ contains exactly $\frac{2^{u-s}}{2^{|\alpha|+i}}$ distinct values of $b2^{|\alpha|+i}$, in which case $p = \frac{2^{u-s}}{2^{|\alpha|+i}} \cdot \frac{1}{2^{u-|\alpha|-i}} = \frac{1}{2^s}$. On the other hand, if $|\alpha|+i > u-s$, then $\{\alpha z 2^i, \dots, \alpha z 2^i + 2^{u-s} - 1\}$ contains *at most one* multiple of $b2^{|\alpha|+i}$. If this occurs, then for some $c \in \{0, \dots, 2^{u-s} - 1\}$, we have need $(\alpha z 2^i + c) \bmod \min(2^u, 2^{|\alpha|-i}) = 0$. This occurs iff

$$(\alpha z 2^i + 2^{u-s} - 1) \bmod \min(2^u, 2^{|\alpha|-i}) < 2^{u-s}$$

In that case, $p = \frac{1}{2^{u-|\alpha|-i}}$. \square

Definition 13. Odd-Multiply-add-shift (OMA) hashing. A *variant* of this was proven by proven by [thorup15] to be 2-independent/strongly universal, but that is stronger than needed for universe reduction or the finding of low-collision hash functions: for both tasks approximate universality suffices. Define the hash family $\mathcal{G}_{b,s}$ mapping $\{0, \dots, 2^u - 1\} \equiv \{0, 1\}^u$ to $\{0, \dots, 2^s - 1\} \equiv \{0, 1\}^s$ to contain, over all odd integers $a \in \{1, 3, \dots, 2^u - 1\}$, and integers $b \in \{0, 1, 2, \dots, 2^{u-s} - 1\}$, the functions

$$h_{a,b} = (ax + b) \bmod 2^u \text{ div } 2^{u-s}.$$

We will show in the following lemma that this family is (1-approximately) universal, a factor two better than the Odd-multiply-shift family.

Lemma 14 (Universality of OMA hash functions.). *Fix $u \geq s > 0$. For any distinct $x, y \in \{0, 1\}^u$, and $h_{a,b}$ drawn uniformly at random from $\mathcal{G}_{b,s}$ (see Definition 13),*

$$\Pr[h_{a,b}(x) = h_{a,b}(y)] = \begin{cases} 0 & \text{if } (x - y) \bmod 2^{u-s} = 0 \\ \frac{1}{2^s} & \text{otherwise} \end{cases}$$

Proof. Define for brevity $p := \Pr[h_{a,b}(x) = h_{a,b}(y)]$. Let i be the integer for which $(x - y) \bmod 2^u = z2^i$ for z odd. Decompose $x = x_h 2^i + x_l$, $y = y_h 2^i + y_l$, where $x_h, y_h \in \{0, \dots, 2^{u-i} - 1\}$ and $x_l, y_l \in \{0, \dots, 2^i - 1\}$. There are two cases:

- If $i \geq u - s$, then

$$(ax + b) - (ay + b) \bmod 2^u = (a(x_h - y_h) \bmod 2^{u-i}) 2^i$$

which is always nonzero because both a and $x_h - y_h = z$ are odd. Consequently, $h_{a,b}(x)$ and $h_{a,b}(y)$ differ by at least $2^{i-(u-s)}$, so $p = 0$.

- If $i < u - s$, let $c = x_i = y_i$, and observe:

$$p = \Pr_{a,b} [(ax_h 2^i + ac + b \bmod 2^u) \operatorname{div} 2^{u-s} = (ay_h 2^i + ac + b \bmod 2^u) \operatorname{div} 2^{u-s}]$$

For any $u, v \in \{0, \dots, 2^u - 1\}$, consider

$$\Pr_b [(u + b) \bmod 2^u \operatorname{div} 2^{u-s} = (v + b) \bmod 2^u \operatorname{div} 2^{u-s}]$$

Let d be the signed difference $u - v \bmod 2^u$ so that $d \in \{-2^{u-1} + 1, \dots, -1, 0, 1, \dots, 2^{u-1} - 1\}$, with the special case $u = (v + 2^{u-1}) \bmod 2^u$, mapped arbitrarily to either $+2^{u-1}$ or -2^{u-1} (either case will lead to the same calculated value because $s \geq 1$). If $|d| > 2^{u-s}$, then $(u + b) \bmod 2^u \operatorname{div} 2^{u-s}$ and $(v + b) \bmod 2^u \operatorname{div} 2^{u-s}$ always differ by at least one – for them to fall in the same set $\{k2^{u-s}, \dots, k2^{u-s} + 2^{u-s} - 1\}$ for some k requires $|d| < 2^{u-s}$. If $|d| \leq 2^{u-s}$, then exactly $2^{u-s} - |d|$ of the possible values for $b \in \{0, \dots, 2^{u-s} - 1\}$ will make $u + b$ and $v + b$ fall in the same set $\{k2^{u-s}, \dots, k2^{u-s} + 2^{u-s} - 1\}$ for some k . (If $d \geq 0$, then all d values of b which for which $(v + b) \bmod 2^{u-s} < 2^{u-s} - (u - v)$ work, and a symmetric condition holds if $d \leq 0$.) Thus

$$\Pr_b [(u + b) \bmod 2^u \operatorname{div} 2^{u-s} = (v + b) \bmod 2^u \operatorname{div} 2^{u-s}] = \max \left(0, \frac{2^{u-s} - |d|}{2^{u-s}} \right)$$

Since a is chosen uniformly at random from odd values in $\{0, \dots, 2^u - 1\}$, and $x_h - y_h$ is odd, the product $a(x_h - y_h) \bmod 2^{u-i}$ is distributed uniformly at random over odd values in $\{0, \dots, 2^{u-i} - 1\}$, and the signed difference $ax_h 2^i - ay_h 2^i$ is distributed uniformly over the odd multiples of 2^i in $\{-2^{u-1} + 1, \dots, 0, \dots, 2^{u-1} - 1\}$. (Since $i < u - s$, these odd multiples never contain the value $\pm 2^{u-1}$.) By the law of total probability,

$$\begin{aligned} p &:= \frac{1}{2^{u-i-1}} \sum_{k \in \{+1, -1\}} \sum_{j=0}^{2^{u-i-1}} \max \left(0, \frac{2^{u-s} - |k(2j+1)2^i|}{2^{u-s}} \right) \\ &= \frac{1}{2^{u-i-1}} \cdot 2 \sum_{j=0}^{2^{u-s-i-1}} \frac{2^{u-s-i} - (2j+1)}{2^{u-s-i}} \\ &= \frac{1}{2^{u-i-1}} \cdot 2 \cdot 2^{u-s-i-1} \cdot \frac{1}{2} = \frac{1}{2^s}. \end{aligned}$$

□

Lemma 15 (Conditional probabilities for bit-by-bit selection of OMAS hash functions.). *Let $x, y \in \{0, \dots, 2^u - 1\}$ be distinct with $(x - y) \bmod 2^u = z2^i$ for z odd and i a nonnegative integer. Fix $\alpha \in \{0, 1\}^*$ of length α , and let $E(\alpha)$ be the set of integers $a \in \{0, \dots, 2^u - 1\}$ whose $|\alpha|$ least significant bits match α . Let $s \in [0, \dots, u]$. As argued in the proof of Lemma 6, if $i \geq u - s$, then any*

$h_{a,b} \in \mathcal{G}_{b,s}$ separates x and y . Assume $i < u - s$. Then if $|\alpha| + i < u - s$, we have:

$$\Pr_{h_{a,b} \in \mathcal{G}_{b,s}: a \in E(\alpha)} [h_{a,b}(x) = h_{a,b}(y)] = \frac{1}{2^s}$$

while if $|\alpha| + i \geq u - s$, we have:

$$\Pr_{h_{a,b} \in \mathcal{G}_{b,s}: a \in E(\alpha)} [h_{a,b}(x) = h_{a,b}(y)] = \frac{1}{2^{u-|\alpha|-i-1}} \cdot \max \left(0, \frac{2^{u-s} - |2^{|\alpha|+i} - (\alpha(x-y) \bmod 2^{|\alpha|+i+1})|}{2^{u-s}} \right)$$

Furthermore, for fixed $a, \beta \in \{0,1\}^*$, and $F(\beta)$ the set of integers $b \in \{0, \dots, 2^u - 1\}$ whose $|\beta|$ most significant bits match β , $\Pr_{b \in F(\beta)} [h_{a,b}(x) = h_{a,b}(y)]$ equals:

$$\begin{cases} \frac{\iota(u) - \iota(v)}{2^{u-s-|\beta|}} & v - u \bmod 2^u < 2^{u-s} \wedge u \bmod 2^u > v \bmod 2^u \\ 1 + \frac{\iota(u) - \iota(v)}{2^{u-s-|\beta|}} & v - u \bmod 2^u < 2^{u-s} \wedge u \bmod 2^u \leq v \bmod 2^u \\ \frac{\iota(v) - \iota(u)}{2^{u-s-|\beta|}} & u - v \bmod 2^u < 2^{u-s} \wedge v \bmod 2^u > u \bmod 2^u \\ 1 + \frac{\iota(v) - \iota(u)}{2^{u-s-|\beta|}} & u - v \bmod 2^u < 2^{u-s} \wedge v \bmod 2^u \leq u \bmod 2^u \\ 0 & \text{otherwise} \end{cases}$$

where $\iota(w) := \max(0, w \bmod 2^{u-s} - (2^{u-s} - 2^{u-s-|\beta|}))$.

Proof. Let $p := \Pr_{h_{a,b} \in \mathcal{G}_{b,s}: a \in E(\alpha)} [h_{a,b}(x) = h_{a,b}(y)]$. Partition random variable a into $a'2^{|\alpha|} + \alpha$ where α is interpreted as an integer in $\{0, \dots, 2^{|\alpha|} - 1\}$; the random variable b is uniform over $\{0, \dots, 2^{u-s} - 1\}$. Decompose $x = x_h 2^i + x_l$ and $y = y_h 2^i + y_l$, and $c = x_l = y_l$. Then

$$\begin{aligned} p &= \Pr [(ax_h 2^i + ac + b) \bmod 2^u \operatorname{div} 2^{u-s} = (ay_h 2^i + ac + b) \bmod 2^u \operatorname{div} 2^{u-s}] \\ &= \Pr \left[(a'x_h 2^{|\alpha|+i} + \alpha x_h 2^i + ac + b) \bmod 2^u \operatorname{div} 2^{u-s} \right. \\ &\quad \left. = (a'y_h 2^{|\alpha|+i} + \alpha y_h 2^i + ac + b) \bmod 2^u \operatorname{div} 2^{u-s} \right] \end{aligned}$$

First, we address the easy case, when $|\alpha| + i < u - s$. Then $a'(x_h - y_h) \bmod 2^{u-|\alpha|-i}$ is uniformly distributed over odd multiples of $2^{|\alpha|+i}$. Let

$$\begin{aligned} u &= a'x_h 2^{|\alpha|+i} + \alpha x_h 2^i + ac \bmod 2^u \\ v &= a'y_h 2^{|\alpha|+i} + \alpha y_h 2^i + ac \bmod 2^u \end{aligned}$$

Then the distribution of $u - v$ matches that of $\delta 2^{|\alpha|+i} + \alpha(x_h - y_h) 2^i \bmod 2^u$, where δ ranges over the odd integers $\{1, 3, \dots, 2^{u-|\alpha|-i} - 1\}$. As proven in Lemma 14, for any given signed difference $u - v \in \{-2^{u-1}, \dots, 2^{u-1}\}$,

$$\Pr [u + b \bmod 2^u \operatorname{div} 2^{u-s} = v + b \bmod 2^u \operatorname{div} 2^{u-s}] = \max \left(0, \frac{2^{u-s} - |u - v|}{2^{u-s}} \right)$$

If $|\alpha| = 0$, then $u - v$ will always (over δ) be an odd multiple of 2^i because δ is odd; and otherwise $u - v$ will always be an odd multiple of 2^i because $\alpha(x_h - y_h)$ is odd. The number of intersections between $u - v$ and $\{-2^{u-s}, \dots, 2^{u-s}\}$ will be $\frac{2^{u-s+1}}{2^{|\alpha|+i+1}} \geq 2$, and one can pair up intersections d_1 with $u - v < 0$ with intersections d_2 for $u - v > 0$ so that $|d_1| + |d_2| = 2^{u-s}$. Each intersection will on average have collision probability $\frac{1}{2}$, and since there are $\frac{2^{u-s+1}}{2^{|\alpha|+i+1}}$ intersecting values out of $2^{u-|\alpha|-i-1}$ values of δ ,

$$p = \frac{1}{2} \cdot \frac{2^{u-s+1}}{2^{|\alpha|+i+1}} \cdot \frac{1}{2^{u-|\alpha|-i-1}} = \frac{1}{2^s}$$

In the hard case, $|\alpha| + i \geq u - s$, and there will be at most one intersection between $u - v$ and $\{-2^{u-s}, \dots, 2^{u-s}\}$. As before, the distribution of $u - v$ matches that of $\delta 2^{|\alpha|+i} + \alpha(x_h - y_h) 2^i \pmod{2^u}$, where δ ranges over odd integers in $\{1, 3, \dots, 2^{u-|\alpha|-i-1} - 1\}$. Write $\delta = 2\gamma + 1$ for $\gamma \in \{0, \dots, 2^{u-|\alpha|-i-1} - 1\}$. We have an intersection if $d' = (\alpha(x_h - y_h) 2^i + 2^{|\alpha|+i}) \pmod{2^{|\alpha|+i+1}}$ is in $\{-2^{u-s}, \dots, 2^{u-s}\}$, and in that case it will have value $\frac{2^{u-s} - |d'|}{2^{u-s}}$. Since $|\alpha| + i \geq u - s$, equivalently for $d'' = \alpha(x_h - y_h) 2^i \pmod{2^{|\alpha|+i+1}}$ (a value in $\{0, \dots, 2^{|\alpha|+i+1} - 1\}$) the value of the intersection is $\max\left(0, \frac{2^{u-s} - |2^{|\alpha|+i} - d''|}{2^{u-s}}\right)$, so

$$p = \frac{1}{2^{u-|\alpha|-i-1}} \cdot \max\left(0, \frac{2^{u-s} - |2^{|\alpha|+i} - (\alpha(x - y) \pmod{2^{|\alpha|+i+1}})|}{2^{u-s}}\right).$$

(Compared to the odd-multiply-shift hash, we evaluate a triangular function instead of the sum of two step functions.)

Finally, we consider the case of fixed a , where the offset b is to be determined. Write $b = \beta 2^{u-s-|\beta|} + b'$ where $b' \in \{0, \dots, 2^{u-s-|\beta|} - 1\}$, $u = ax + \beta 2^{u-s-|\beta|}$ and $v = ay + \beta 2^{u-s-|\beta|}$, so that

$$\begin{aligned} p &:= \Pr[(ax + b) \pmod{2^u \text{ div } 2^{u-s}} = (ay + b) \pmod{2^u \text{ div } 2^{u-s}}] \\ &= \Pr[(u + b') \pmod{2^u \text{ div } 2^{u-s}} = (v + b') \pmod{2^u \text{ div } 2^{u-s}}] \end{aligned}$$

Consider the case where $d = v - u \pmod{2^u} < 2^{u-s}$. (If $v - u \pmod{2^u} \geq 2^{u-s}$ and $u - v \pmod{2^u} \geq 2^{u-s}$, then $p = 0$, so this is one half of the nontrivial case.) Define $\iota(w) = \max(0, w \pmod{2^{u-s}} - (2^{u-s} - 2^{u-s-|\beta|}))$ to count the $b' \in \{0, \dots, 2^{u-s-|\beta|} - 1\}$ for which $(w + b') \text{ div } 2^{u-s} > w \text{ div } 2^{u-s}$. Then we have a two cases: \square

- $p = \frac{\iota(u) - \iota(v)}{2^{u-s-|\beta|}}$: holds if $u \pmod{2^u} > v \pmod{2^u}$ - at $b' = 0$ we have no collision, and increasing $u \text{ div } 2^u$ but not $v \text{ div } 2^u$ will create a collision
- $p = 1 + \frac{\iota(u) - \iota(v)}{2^{u-s-|\beta|}}$: holds if $u \pmod{2^u} \leq v \pmod{2^u}$ - at $b' = 0$ we have a collision, and increasing $v \text{ div } 2^u$ but not $u \text{ div } 2^u$ will prevent a collision

Note: if $u - v \pmod{2^u} \geq 2^{u-s}$, then the cases have probabilities $\frac{\iota(v) - \iota(u)}{2^{u-s-|\beta|}}$ and $1 + \frac{\iota(v) - \iota(u)}{2^{u-s-|\beta|}}$.

References

- [fks84] Fredman, Komlós, Szemerédi, “Storing a Sparse Table with $O(1)$ Worst-case Access Time”, 1984. <https://doi.org/10.1145/828.1884>
- [hmp01] Hagerup, Miltersen, Pagh, “Deterministic Dictionaries”, 2001, <https://doi.org/10.1006/jagm.2001.1171>.
- [r95] Raman, “Improved data structures for predecessor queries in integer sets”, 1995, technical report.
- [dhkp97] Dietzfelbinger, Hagerup, Katajainen, Penttonen, A Reliable Randomized Algorithm for the Closest-Pair Problem”, 1997, <https://doi.org/10.1006/jagm.1997.0873>
- [thorup15] Thorup, “High Speed Hashing for Integers and Strings”, 2015., <https://doi.org/10.1006/jagm.1997.0873>