Karlsruhe Institute of Technology

Stefan Walzer, Stefan Hermann
ITI Prof. Sanders
winter term 2025/2026

Exercise Sheet 17 – Perfect Hashing

Probability and Computing

Exercise 1 – Minimal Perfect Hash Function with Brute Force

Consider the following pair of algorithms for the construction and evaluation of minimal perfect hash functions[1]. Let $S \subseteq D$ and $n = |S|$. According to the SUHA, we assume that $h_1, h_2, h_3, \ldots \sim \mathcal{U}([n]^D)$ are independent, fully random hash functions that themselves require no storage space.[2]

Algorithm construct(S)**:**
  for seed $= 1$ **to** ∞ **do**
    if $|\{h_{\text{seed}}(x) \mid x \in S\}| = n$ **then**
      return seed

Algorithm eval(seed, x)**:**
  return $h_{\text{seed}}(x)$

(a) Argue: Each loop iteration in construct succeeds with probability $\frac{n!}{n^n}$.

(b) Argue: The return value seed of construct satisfies $\mathbb{E}[\text{seed}] = \frac{n^n}{n!}$.

(c) The space requirement is space $= \lceil \log_2(\text{seed}) \rceil$ bits. Show: $\mathbb{E}[\text{space}] \leq n \log_2(e) + 1$.
Hint: Use Jensen's inequality (Exercise Sheet 10) as well as Stirling's approximation of the factorial function. (see en.wikipedia.org/wiki/Stirling's_approximation).

(d) Comment on the expected space requirement and the expected construction time of the presented method.

Exercise 2 – Lower Space Bounds for MPHF

We will show that, in general, a minimal perfect hash function cannot be represented with fewer than $\log_2(e) \approx 1.44$ bits per element.

Recall the definition of a perfect hash function from the lecture. Let $\varepsilon = 0$, $n = m \in \mathbb{N}$, $d = |D|$. Let $\mathcal{I} := \{S \subseteq D \mid |S| = n\}$ denote the set of all possible inputs of size n. We consider the number of inputs for which a given data structure P can simultaneously serve as a perfect hash function. Formally:

$$\text{cov}(P) := \left\{ S \in \mathcal{I} \;\middle|\; \{\text{eval}_P(x) \mid x \in S\} = [n] \right\}.$$

[1] Note: The "data structure" here consists solely of the number "seed".

[2] Alternatively, one may imagine a single hash function $h : \mathbb{N} \times D \to [n]$ that is fully random and, in addition to the actual key, accepts a natural number as a seed.

(a) As a warm-up: Suppose $n = 3$, $D = \{\mathsf{a}, \mathsf{b}, \mathsf{c}, \mathsf{d}, \mathsf{e}, \mathsf{f}, \mathsf{g}, \mathsf{h}\}$ and P is a data structure for which eval_P behaves as shown in the following table. Argue: P can serve as an MPHF for $|\text{cov}(P)| = 12$ different $S \in \mathcal{I}$.

$x \in D$	a	b	c	d	e	f	g	h
$\text{eval}_P(x)$	3	3	2	1	1	3	1	3

(b) Now let n and d as well as P be arbitrary. Show $\text{cov}(P) \leq (\frac{d}{n})^n$.
Hint: Recall that among all rectangles with a given sum of edge lengths the square has maximum area. Use without proof an analogous claim in n dimensions, namely that $\max_{\substack{0 \leq c_1, \ldots, c_n \leq d \\ c_1 + \cdots + c_n = d}} c_1 \cdot \ldots \cdot c_n = (d/n)^n$.

Let now $\mathcal{P} = \{\text{construct}(S) \mid S \in \mathcal{I}\}$ be the set of all distinct data structures produced by construct.

(c) Argue $\bigcup_{P \in \mathcal{P}} \text{cov}(P) = \mathcal{I}$ and conclude $|\mathcal{P}| \geq \binom{d}{n} / (\frac{d}{n})^n$.

(d) Show $\binom{d}{n} / (\frac{d}{n})^n \geq \frac{n^n}{n!} \cdot (1 - o(1))$ if $d = \omega(n^2)$.
Hint: In an intermediate step, show that $(1 - \frac{n}{d})^n \geq 1 - \frac{n^2}{d} = 1 - o(1)$.

(e) Argue: If the data structures produced by construct are each encoded as a bit string of length ℓ, then $\ell \geq \lceil \log_2(|\mathcal{P}|) \rceil$.

(f) Show that $\ell \geq \log_2(e) \cdot n - o(n)$ if $d = \omega(n^2)$.
Hint: Combine (c)-(e) and use Stirling's approximation of the factorial function.