Karlsruhe Institute of Technology

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

Exercise Sheet 7 – Classic Hash Tables

Probability and Computing

Exercise 1 – 2-Independence vs. 1-Universality

Let $\mathcal{H} \subseteq [m]^D$ be a family of hash functions mapping D to $[m]$. Prove or disprove the following implications:

(a) $\mathcal{H}$ is 2-independent $\Rightarrow$ $\mathcal{H}$ is 1-universal.

(b) $\mathcal{H}$ is 1-universal $\Rightarrow$ $\mathcal{H}$ is 2-independent.

Hint: In one case, the implication is straightforward. In the other, trivial counterexamples exist.

Solution 1

(a) The implication holds. For any $x \neq y \in D$, by the definition of 2-independence:

$$\forall i, j \in [m] : \Pr_{h \sim \mathcal{U}(\mathcal{H})}[h(x) = i \wedge h(y) = j] = \frac{1}{m^2}.$$

Consequently, the collision probability for x and y under $\mathcal{H}$ is bounded as follows:

$$\begin{aligned}\Pr_{h \sim \mathcal{U}(\mathcal{H})}[h(x) = h(y)] &= \sum_{i=1}^{m} \Pr_{h \sim \mathcal{U}(\mathcal{H})}[h(x) = i \wedge h(y) = i] \\ &= \sum_{i=1}^{m} \frac{1}{m^2} = \frac{1}{m}.\end{aligned}$$

This verifies the condition for 1-universality.

(b) The implication does not hold. This is primarily due to “trivial” counterexamples.

Example 1. Take $D = [m]$ and $\mathcal{H} = \{\mathrm{id}\}$. The identity function never causes collisions; hence, $\mathcal{H}$ is even 0-universal (and thus 1-universal). However, $\mathcal{H}$ is not 2-independent: the hash values are not uniformly distributed in $[m]$—in fact, they are deterministic.

Example 2. Consider the class $\mathcal{H} = \mathcal{H}^{\text{lin}}_{p,m}$ from lecture, parameterized by p and m, such that m does not divide $p(p-1)$. As shown in lecture, $\mathcal{H}$ is 1-universal. Since $|\mathcal{H}| = p(p-1)$, all relevant probabilities (of the form $\Pr_{h\sim\mathcal{H}}[\ldots]$) are multiples of $\frac{1}{p(p-1)}$. However, $\frac{1}{m}$ is not such a multiple. Consequently, $\Pr_{h\sim\mathcal{H}}[h(x) = 0] = \frac{1}{m}$ cannot hold for any x. Hence, the hash value of x is not uniformly distributed in $[m]$.

Exercise 2 – d-Independence without Mutual Independence

Alice and Bob each spin a roulette wheel with 10 equally sized segments labeled 0 to 9. Let A and B denote Alice's and Bob's outcomes, respectively. Define $C = (A + B) \bmod 10$.

(a) Show that A, B, and C are pairwise independent.

(b) Show that A, B, and C are not mutually independent.

(c) For any $d \in \mathbb{N}$, construct a family of random variables that is d-independent but not fully independent.

Solution 2

We solve the problem for arbitrary $d, m \in \mathbb{N}$ (instead of $d = 2$ and $m = 10$). Specifically, let $A_1, A_2, \ldots, A_d \sim \mathcal{U}([m])$ be d mutually independent uniform random variables over $[m]$, and define $C := (A_1 + \cdots + A_d) \bmod m$. It is straightforward to verify $C \sim \mathcal{U}([m])$: regardless of the values $A_1, \ldots, A_{d-1}$, the m possible values of A_d yield each residue modulo m for C with equal probability. We prove two properties:

The family $\{A_1, \ldots, A_d, C\}$ is not mutually independent. We have $\Pr[\forall i \in [d] : A_i = 0] = m^{-d}$ and $\Pr[C = 0] = m^{-1}$. However, the event $\forall i \in [d] : A_i = 0$ implies $C = 0$, so $\Pr[C = 0 \wedge \forall i \in [d] : A_i = 0] = m^{-d}$. Had mutual independence of $\{A_1, \ldots, A_d, C\}$ held, we would have obtained $m^{-(d+1)}$.

The family $\{A_1, \ldots, A_d, C\}$ is d-wise independent. Consider any selection of d variables and the event that they attain specific values. We must show the probability of this event equals the product of individual probabilities. By symmetry (since $A_1, \ldots, A_d$ play identical roles), we only distinguish two cases: whether C is selected or not:

$$\Pr[A_1 = a_1 \wedge \cdots \wedge A_d = a_d] \overset{!}{=} \prod_{i=1}^{d} \Pr[A_i = a_i] = m^{-d},$$

$$\Pr[A_1 = a_1 \wedge \cdots \wedge A_{d-1} = a_{d-1} \wedge C = c] \overset{!}{=} \Pr[C = c] \cdot \prod_{i=1}^{d-1} \Pr[A_i = a_i] = m^{-d}.$$

The "$\overset{!}{=}$" must be proven. In the first case, this holds trivially by the independence of $A_1, \ldots, A_d$. For the second case, consider the event:

$$E = \{A_1 = a_1 \wedge \cdots \wedge A_{d-1} = a_{d-1} \wedge C = c\},$$

where $a_1, \ldots, a_{d-1}, c$ are fixed. By definition of C,

$$E = \{A_1 = a_1 \wedge \cdots \wedge A_{d-1} = a_{d-1} \wedge A_1 + \cdots + A_{d-1} + A_d = c\}.$$

Since $A_1, \ldots, A_{d-1}$ are fixed, this is equivalent to:

$$E = \{A_1 = a_1 \wedge \cdots \wedge A_{d-1} = a_{d-1} \wedge A_d = c - a_1 - \cdots - a_{d-1}\}.$$

In this form, it is clear that $\Pr[E] = m^{-d}$, since $A_1, \ldots, A_d$ are independent and uniformly distributed.

A Remark. If one examines our definitions very closely, one might still have a concern. Although we showed above that for any selection of d variables, we can factor the probability of a joint event into a product, the definition of d-independence refers to "up to" d variables. What if we select only k variables, where $k < d$? Does it automatically follow that these k variables are also independent? The answer is "yes".

Consider, for example, the event:

$$E = \{A_1 = a_1 \wedge \cdots \wedge A_{k-1} = a_{k-1} \wedge C = c\}.$$

We must show that for this event E,

$$\Pr[E] \stackrel{!}{=} \Pr[C = c] \cdot \prod_{i=1}^{k-1} \Pr[A_i = a_i] = m^{-k}.$$

This is achieved by introducing additional case distinctions over the remaining random variables and reusing the prior result:

$$\begin{aligned}
\Pr[E] &= \Pr[A_1 = a_1 \wedge \cdots \wedge A_{k-1} = a_{k-1} \wedge C = c] \\
&= \sum_{a_k=0}^{m-1} \sum_{a_{k+1}=0}^{m-1} \cdots \sum_{a_{d-1}=0}^{m-1} \Pr[A_1 = a_1 \wedge \cdots \wedge A_{d-1} = a_{d-1} \wedge C = c] \\
&= \sum_{a_k=0}^{m-1} \sum_{a_{k+1}=0}^{m-1} \cdots \sum_{a_{d-1}=0}^{m-1} m^{-d} = m^{d-k} \cdot m^{-d} = m^{-k}.
\end{aligned}$$

Exercise 3 – Find the Error

Let p be prime, $\mathbb{F}_p = \{0, \ldots, p-1\}$ and $m \in \mathbb{N}$. Consider the following class of hash functions from $\mathbb{F}_p$ to $[m]$, also mentioned in the lecture.

$$\mathcal{H} = \{x \mapsto ((a \cdot x) \bmod p) \bmod m \mid a \in \mathbb{F}_p^*\}.$$

Consider the following argument that $\mathcal{H}$ is 1-universal. Find the mistake in the proof.

The proof considers arbitrary $x, y \in \mathbb{F}_p$ with $x \neq y$. It has six steps.

$$\begin{aligned}
\Pr_{h\sim\mathcal{H}}[h(x) = h(y)] &\overset{1}{=} \Pr_{a\sim\mathcal{U}(\mathbb{F}_p^*)}[(ax \bmod p) \bmod m = (ay \bmod p) \bmod m] \\
&\overset{2}{=} \Pr_{a\sim\mathcal{U}(\mathbb{F}_p^*)}[((ax \bmod p) - (ay \bmod p)) \bmod m = 0] \\
&\overset{3}{=} \Pr_{a\sim\mathcal{U}(\mathbb{F}_p^*)}[((ax - ay) \bmod p) \bmod m = 0] \\
&\overset{4}{=} \Pr_{a\sim\mathcal{U}(\mathbb{F}_p^*)}[(a(x - y) \bmod p) \bmod m = 0] \\
&\overset{5}{=} \Pr_{u\sim\mathcal{U}(\mathbb{F}_p^*)}[u \bmod m = 0] \\
&\overset{6}{=} \frac{|\{m, 2m, 3m, \ldots, \} \cap \mathbb{F}_p^*|}{|\mathbb{F}_p^*|} \\
&\overset{7}{\leq} \frac{1}{m}.
\end{aligned}$$

In Step 5 we use that the function $a \mapsto az \bmod p$ is a bijection on $\mathbb{F}_p^*$ for any fixed $z \in \mathbb{F}_p^*$. Therefore, if $a \sim \mathcal{U}(\mathbb{F}_p^*)$ and $u := az$ then $u \sim \mathcal{U}(\mathbb{F}_p^*)$.

Solution 3

The error is in Step 3. In general it is not true that

$$(c \bmod p) - (d \bmod p) = (c - d) \bmod p.$$

The left hand side may even produce negative values! It is true however that

$$(c \bmod p) - (d \bmod p) \in \{(c - d) \bmod p, (c - d) \bmod p - p\}.$$

Remark: Adapting the argument to track through both cases we can get an upper bound of $\frac{2}{m} + \frac{1}{p-1}$. This almost proves 2-universality (assuming $p \gg m$). The details are somewhat annoying.

Exercise 4 – Bonus: Concentration Bounds for Sums of d-wise Independent Random Variables

Let $d \in \mathbb{N}$ be even, and $\{X_1, \ldots, X_n\}$ be a d-wise independent family of random variables, each distributed as $\mathrm{Ber}(p)$ with $p = \Omega(1/n)$.

Define $X = \sum_{i=1}^n X_i$. Note: X is not necessarily binomially distributed since the X_i are not mutually independent.

The goal is to prove the concentration bound: for any $\delta > 0$,

$$\Pr[X - \mathbb{E}[X] \geq \delta\mathbb{E}[X]] = O(\delta^{-d}(np)^{-d/2}).$$

To this end, consider the "centered" random variables $Y_i := X_i - p$, their sum $Y = \sum_{i=1}^n Y_i$, and the moment $\mathbb{E}[Y^d]$.

(i) Warm-up: Let $d \geq 3$ and $n \geq 3$. Verify and briefly explain why the following hold:

(a) $\mathbb{E}[Y_1^5 Y_2^{42}] = \mathbb{E}[Y_1^5]\mathbb{E}[Y_2^{42}]$

(b) $\mathbb{E}[Y_1^5 Y_2^{42} Y_3] = 0$

(c) $\mathbb{E}[Y_1^5] \leq \mathbb{E}[Y_1^2]$

In subsequent steps, you may apply these insights without further justification.

(ii) Show: $\mathbb{E}[Y_1^2] \leq p$.

(iii) Let $i_1, \ldots, i_d \in [n]$ (not necessarily distinct) and $S = \{i_1, \ldots, i_d\}$. Prove:

- If $|S| > d/2$, then $\mathbb{E}[Y_{i_1} \cdots Y_{i_d}] = 0$.
- Otherwise, $\mathbb{E}[Y_{i_1} \cdots Y_{i_d}] \leq p^{|S|}$.

(iv) Show: $\mathbb{E}[Y^d] = O((np)^{d/2})$. You may assume $d = O(1)$. **Hint:** Expand $(\sum_{i=1}^n Y_i)^d$. Yes, this yields n^d terms.

(v) Prove the original goal by applying Markov's inequality to Y^d.

Solution 4

(i) Since $d \geq 3$, for any distinct $i_1, i_2, i_3 \in [n]$, the random variables $X_{i_1}, X_{i_2}, X_{i_3}$ are mutually independent. Hence, Y_1^5, Y_2^{42}, and Y_3 (as functions of X_1, X_2, X_3) are also mutually independent.

(a) For independent random variables, the expectation of the product equals the product of expectations, by definition.

(b) Factor the expectation: $\mathbb{E}[Y_1^5 Y_2^{42} Y_3] = \mathbb{E}[Y_1^5]\mathbb{E}[Y_2^{42}]\mathbb{E}[Y_3]$. Since $\mathbb{E}[Y_3] = \mathbb{E}[X_3 - p] = p - p = 0$, the product is zero.

(c) Since $|Y_1| \leq 1$ and x^i is non-increasing for $x \in [0, 1]$ as i increases,

$$\mathbb{E}[Y_1^5] \leq \mathbb{E}[|Y_1^5|] = \mathbb{E}[|Y_1|^5] \leq \mathbb{E}[|Y_1|^2] = \mathbb{E}[Y_1^2].$$

(ii) $\mathbb{E}[Y_1^2] = \mathbb{E}[(X_1 - p)^2] = p(1-p)^2 + (1-p)(0-p)^2 = p(1-p)(1-p+p) \leq p$.

(iii) The key question is whether any index appears exactly once in the multiset $\{i_1, \ldots, i_d\}$.

- If $|S| > d/2$, then at least one index j appears exactly once. Then $\mathbb{E}[Y_{i_1} \cdots Y_{i_d}]$ factors such that $\mathbb{E}[Y_j] = 0$ appears as a multiplicative factor, so the expectation is zero.
- If $|S| \leq d/2$, then the product involves at most $d/2$ distinct variables. Factor as in part (i)(a). If any variable appears with exponent 1, the entire product vanishes (since $\mathbb{E}[Y_j] = 0$). Otherwise, each exponent is at least 2. Then apply (i)(c) and (ii) to bound each distinct factor by p, giving $p^{|S|}$.

(iv) We compute:

$$\begin{aligned}
\mathbb{E}[Y^d] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} Y_i\right)^d\right] = \mathbb{E}\left[\sum_{i_1=1}^{n}\cdots\sum_{i_d=1}^{n} Y_{i_1}\cdots Y_{i_d}\right] \\
&\overset{(1)}{=} \sum_{i_1=1}^{n}\cdots\sum_{i_d=1}^{n}\mathbb{E}[Y_{i_1}\cdots Y_{i_d}] \\
&\overset{(2)}{=} \sum_{i_1,\ldots,i_d}\mathbb{E}[Y_{i_1}\cdots Y_{i_d}] \overset{(3)}{=} \sum_{r=1}^{d}\sum_{\substack{S\subseteq[n]\\|S|=r}}\sum_{i_1,\ldots,i_d} \mathbb{1}_{\{i_1,\ldots,i_d\}=S}\cdot\mathbb{E}[Y_{i_1}\cdots Y_{i_d}] \\
&\overset{(4)}{\le} \sum_{r=1}^{d/2}\sum_{\substack{S\subseteq[n]\\|S|=r}}\sum_{i_1,\ldots,i_d} \mathbb{1}_{\{i_1,\ldots,i_d\}=S}\cdot p^{|S|} \overset{(5)}{=} \sum_{r=1}^{d/2}\sum_{\substack{S\subseteq[n]\\|S|=r}} p^{|S|}\sum_{i_1,\ldots,i_d}\mathbb{1}_{\{i_1,\ldots,i_d\}=S} \\
&\overset{(6)}{\le} \sum_{r=1}^{d/2}\sum_{\substack{S\subseteq[n]\\|S|=r}} p^{|S|}\cdot|S|^d \overset{(7)}{=} \sum_{r=1}^{d/2}\binom{n}{r}p^r r^d \overset{(8)}{\le} (d/2)^d\sum_{r=1}^{d/2} n^r p^r \overset{(9)}{\le} O(n^{d/2}p^{d/2}).
\end{aligned}$$

(1) Linearity of expectation.

(2) Compact notation.

(3) Group terms by the set $S = \{i_1, \ldots, i_d\}$.

(4) By part (iii), terms with $|S| > d/2$ vanish; the others are bounded by $p^{|S|}$.

(5) Factor out $p^{|S|}$.

(6) For the indicator to be 1, all indices $i_1, \ldots, i_d$ must lie in S, which can occur in at most $|S|^d$ ways.

(7) The inner sum depends only on $r = |S|$, and there are $\binom{n}{r}$ such sets.

(8) Use $\binom{n}{r} \le n^r$ and $r \le d/2$.

(9) Since $d = O(1)$, $(d/2)^d = O(1)$. Since $p = \Omega(1/n)$, we have $np = \Omega(1)$, so the term at $r = d/2$ dominates the constant number of other terms.

(v) First, the calculation:

$$\begin{aligned}
\Pr[X - \mathbb{E}[X] \ge \delta\mathbb{E}[X]] &\overset{(1)}{=} \Pr[Y \ge \delta np] \le \Pr[|Y| \ge \delta np] = \Pr[|Y|^d \ge (\delta np)^d] \\
&\overset{(2)}{=} \Pr[Y^d \ge (\delta np)^d] \overset{(3)}{\le} \frac{\mathbb{E}[Y^d]}{(\delta np)^d} \overset{(4)}{\le} O\left(\frac{n^{d/2}p^{d/2}}{(\delta np)^d}\right) = O\left(\delta^{-d}(np)^{-d/2}\right).
\end{aligned}$$

(1) By definition of Y and linearity, $\mathbb{E}[X] = np$.

(2) Since d is even, $|Y|^d = Y^d$.

(3) Apply Markov's inequality to Y^d. Note that $Y^d \ge 0$ because d is even.

(4) Substitute the result from part (iv).