Karlsruher Institut für Technologie

Stefan Walzer
ITI Prof. Sanders
Wintersemester 2023/2024

Übungsblatt 7 – Classic Hash Tables

Randomisierte Algorithmik

Aufgabe 1 – 2–Unabhängigkeit vs. 1-Universalität

Sei $\mathcal{H} \subseteq [m]^D$ eine Familie von Hashfunktionen von D nach $[m]$. Zeige oder widerlege, dass folgende Implikationen gelten:

(a) $\mathcal{H}$ ist 2-unabhängig $\Rightarrow$ $\mathcal{H}$ ist 1-universell.

(b) $\mathcal{H}$ ist 1-universell $\Rightarrow$ $\mathcal{H}$ ist 2-unabhängig.

Hinweis: In einem Fall lässt sich die Implikationen leicht zeigen. Im anderen Fall kann man dämliche Gegenbeispiele finden.

Lösung 1

(a) Die Implikation ist richtig. Für jedes $x \neq y \in D$ gilt nach Definition von 2-Unabhängigkeit:

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

Damit können wir für $\mathcal{H}$ die Kollisionswahrscheinlichkeit von x und y folgendermaßen abschätzen:

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

Damit ist die Bedingung für 1-Universalität gezeigt.

(b) Die Implikation gilt nicht. Allerdings vor allem aus „dummen“ Gründen.

Beispiel 1. Man nehme $D = [m]$ und $\mathcal{H} = \{\mathrm{id}\}$. Natürlich erzeugt die Identitätsfunktion niemals Kollisionen, also ist $\mathcal{H}$ sogar 0-universell (und damit auch 1-universell). Natürlich ist $\mathcal{H}$ aber nicht 2-unabhängig, weil die Hashwerte nicht gleichverteilt in $[m]$ liegen (die sind ja nicht mal zufällig).

Beispiel 2. Man nehme die Klasse $\mathcal{H} = \mathcal{H}^{\text{lin}}_{p,m}$ aus der Vorlesung mit Parametern p und m, sodass m kein Teiler von $p \cdot (p-1)$ ist. Nach Vorlesung ist $\mathcal{H}$ eine 1-universelle Klasse. Weil $|\mathcal{H}| = p \cdot (p-1)$ ist, sind alle relevanten Wahrscheinlichkeiten (alle Zahlen der Form $\Pr_{h \sim \mathcal{H}}[\ldots]$) aber Vielfache von $\frac{1}{p \cdot (p-1)}$. Nun ist aber $\frac{1}{m}$ kein solches Vielfaches. Also kann $\Pr_{h \sim \mathcal{H}}[h(x) = 0] = \frac{1}{m}$ für kein x gelten. Der Hashwert von x ist also nicht gleichverteilt in $[m]$.

Aufgabe 2 – d-Unabhängigkeit ohne Unabhängigkeit

Alice und Bob drehen beide an einem Glücksrad, das 10 gleich große Segmente mit den Zahlen von 0 bis 9 hat. Seien A und B die Ergebnisse von Alice und Bob. Sei $C = (A + B) \bmod 10$.

(a) Zeige: A, B, C sind paarweise unabhängig.

(b) Zeige: A, B, C sind nicht unabhängig.

(c) Finde für beliebiges $d \in \mathbb{N}$ eine Familie von Zufallsvariablen, die d-unabhängig ist aber nicht unabhängig ist.

Lösung 2

Wir lösen die Aufgabe direkt für beliebiges $d, m \in \mathbb{N}$ (statt für $d = 2$ und $m = 10$). Das heißt wir haben d unabhängige Zufallsvariablen $A_1, A_2, \ldots, A_d \sim \mathcal{U}([m])$ und eine weitere Zufallsvariable $C := (A_1 + \cdots + A_d) \bmod m$. Es ist leicht zu sehen, dass auch $C \sim \mathcal{U}([m])$ gilt indem man sich vorstellt, dass $A_1, \ldots, A_d$ nacheinander gewählt werden: Ganz egal was für $A_1, \ldots, A_{d-1}$ herausgekommen ist, die m Möglichkeiten die noch für A_d bleiben machen noch jeden Wert in $[m]$ für C mit gleicher Wahrscheinlichkeit möglich. Wir zeigen zwei Eigenschaften:

Die Familie $\{A_1, \ldots, A_d, C\}$ ist nicht unabhängig. Wir haben $\Pr[\forall i \in [d] : A_i = 0] = m^{-d}$ sowie $\Pr[C = 0] = m^{-1}$. Gleichzeitig impliziert aber das erste Ereignis das zweite (wenn alle $A_i = 0$ sind, dann auch $C = 0$). Also gilt $\Pr[C = 0 \wedge \forall i \in [d] : A_i = 0] = m^{-d}$. Wäre $\{A_1, \ldots, A_d, C\}$ eine unabhängige Familie von Zufallsvariablen hätte $m^{-(d+1)}$ herauskommen müssen.

Die Familie $\{A_1, \ldots, A_d, C\}$ ist d-fach unabhängig. Wir betrachten eine beliebige Auswahl von d der Variablen und das Ereignis, dass diese eine bestimmte Kombination von Werten annehmen. Es ist zu zeigen, dass die Wahrscheinlichkeit für dieses Ereignis sich wie erwartet als Produkt von Wahrscheinlichkeiten ergibt. Vernachlässigt man symmetrische Fälle (die Variablen $A_1, \ldots, A_d$ haben alle die gleiche Rolle) hat man nur zwei Fälle zu unterscheiden. Entweder C ist nicht ausgewählt oder C ist ausgewählt:

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

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

Die „$\overset{!}{=}$“ sind zu zeigen. Im ersten Fall ist das klar, weil $A_1, \ldots, A_d$ schon als unabhängig definiert sind. Es bleibt der zweite Fall. Wir betrachten also ein Ereignis der Form:

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

wobei $a_1, \ldots, a_{d-1}, c$ feste Zahlen sind. Dieses Ereignis lässt sich aber nach Definition von C auch äquivalent schreiben als:

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

Weil wir aber die Werte von A_1 bis A_{d-1} bereits vorschreiben ist das aquivalent zu:

$$E = \{A_1 = a_1 \wedge \ldots \wedge A_{d-1} = a_{d-1} \wedge A_d = c - a_1 - a_2 - \ldots - a_{d-1}\}.$$

In dieser Form ist klar, dass E Wahrscheinlichkeit m^{-d} hat, weil $A_1, \ldots, A_d$ nach Definition unabhängig und gleichverteilt sind.

Ein Nachtrag. Schaut man sehr genau auf unsere Definitionen, könnte man noch eine Sorge haben. Wir haben zwar oben gezeigt, dass egal welche d Variablen wir wählen, diese d Variablen zulassen, dass wir Wahrscheinlichkeiten von zusammengesetzten Ereignisse als Produkte auseinanderziehen. Die Definition von d-Unabhängigkeit spricht aber von „bis zu“ d Variablen. Was ist also, wenn wir k Variablen wählen, für ein $k < d$? Folgt dann automatisch dass diese k Variablen auch unabhängig sind? Die Antwort ist „ja“.

Betrachten wir beispielsweise das Ereignis:

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

Wir müssen zeigen, dass für dieses E gilt $\Pr[E] \stackrel{!}{=} \Pr[C = c] \cdot \prod_{i=1}^{k-1} \Pr[A_i = a_i] = m^{-k}$. Das geht hier indem man einfach zusätzliche Fallunterscheidungen über andere Zufallsvariablen macht und das Ergebnis von vorher verwendet:

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

Aufgabe 3 – Konzentrationsschranken für Summen d-unabhängiger Zufallsvariablen

Sei $d \in \mathbb{N}$ gerade und $\{X_1, \ldots, X_n\}$ eine d-unabhängige Familie von Zufallsvariablen, die alle Verteilung $\mathrm{Ber}(p)$ mit $p = \Omega(1/n)$ haben. Wir betrachten die Summe $X = \sum_{i=1}^{n} X_i$. Beachte: Weil $X_1, \ldots, X_n$ nicht unabhängig sind ist X nicht unbedingt binomialverteilt!

Ziel dieser Aufgabe ist es, eine Konzentrationsschranke für X zu zeigen, nämlich, dass für beliebige $\delta > 0$ gilt:

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

Dazu schauen wir die „zentrierten“ Zufallsvariablen $Y_i := X_i - p$ an, deren Summe $Y = \sum_{i=1}^{n} Y_i$ und den Erwartungswert $\mathbb{E}[Y^d]$.

(i) Zum Aufwärmen: Sei $d \geq 3$ und $n \geq 3$. Überzeuge dich, dass Folgendes gilt und erkläre kurz warum.

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

Im weiteren Verlauf darfst du die zugrundeliegenden Einsichten ohne weitere Begründung verallgemeinert verwenden.

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

(iii) Seien $i_1, \ldots, i_d \in [n]$ (nicht notwendig verschieden) sowie $S = \{i_1, \ldots, i_d\}$. Zeige:

- Falls $|S| > d/2$ dann gilt $\mathbb{E}[Y_{i_1} \cdot \ldots \cdot Y_{i_d}] = 0$.
- Andernfalls gilt $\mathbb{E}[Y_{i_1} \cdot \ldots \cdot Y_{i_d}] \leq p^{|S|}$.

(iv) Zeige $\mathbb{E}[Y^d] = O((np)^{d/2})$. Du darfst annehmen, dass $d = O(1)$ gilt.
Hinweis: Multipliziere $(\sum_{i=1}^n Y_i)^d$ aus. Ja, das ergibt n^d Terme.

(v) Beweise das ursprüngliche Ziel dieser Aufgabe indem du die Markov Ungleichung auf Y^d anwendest.

Lösung 3

(i) Wegen $d \geq 3$ gilt für beliebige verschiedene $i_1, i_2, i_3 \in [n]$, dass $X_{i_1}, X_{i_2}, X_{i_3}$ unabhängig sind. Insbesondere sind X_1, X_2, X_3 unabhängig. Weil sich Y_1^5, Y_2^{42} und Y_3 jeweils als Funktion aus X_1,X_2 bzw. X_3 ergeben sind auch Y_1^5, Y_2^{42}, Y_3 unabhängig.

(a) Der Erwartungswert des Produktes unabhängiger Zufallsvariablen ist das Produkt der Erwartungswerte dieser Zufallsvariablen nach Definition von Unabhängigkeit.

(b) Auch hier kann die Erwartungswerte auseinanderziehen und dann $\mathbb{E}[Y_3] = \mathbb{E}[X_3 - p] = \mathbb{E}[X_3] - p = p - p = 0$ verwenden.

(c) Weil $|Y_1| \leq 1$ ist und x^i für $x \in [0, 1]$ monoton fallend in i ist folgt:

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

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

(iii) Die zentrale Frage für diesen Aufgabenteil ist, ob es einen Index gibt, der in der Liste $i_1, \ldots, i_d$ nur einmal vorkommt.

- $|S| > d/2$ bedeutet, dass die Liste der d Indizes mehr als $d/2$ verschiedene Werte abdeckt, also muss es mindestens einen Index j geben, der nicht doppelt vorkommt. Dann hat der Erwartungswert $\mathbb{E}[Y_{i_1} \cdot \ldots \cdot Y_{i_d}]$ die Form wie in Aufgabenteil (i) (b). Nach dem Auseinanderziehen kommt also $\mathbb{E}[Y_j] = 0$ als Faktor vor.

- $|S| \leq d/2$ heißt, dass höchstens $d/2$ verschiedene Variablen im Produkt vorkommen. Die kann man wie in Aufgabenteil (i) (a) auseinanderziehen. Wenn eine Variable mit Potenz 1 vorkommt, bekommt man wieder 0 als Gesamtergebnis. Andernfalls ist jede Potenz mindestens 2. Man kann dann Aufgabenteil (i) (c) sowie (ii) verwenden, um jeden der $|S|$ Terme mit p abzuschätzen.

(iv) Wir rechnen drauf los und kommentieren unten die einzelnen Schritte:

$$\begin{aligned}
\mathbb{E}[Y^d] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} Y_i\right)^d\right] = \mathbb{E}\left[\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_d=1}^{n} Y_{i_1}\cdot Y_{i_2}\cdot\ldots\cdot Y_{i_d}\right] \\
&\overset{(1)}{=} \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_d=1}^{n}\mathbb{E}[Y_{i_1}\cdot Y_{i_2}\cdot\ldots\cdot Y_{i_d}] \\
&\overset{(2)}{=} \sum_{i_1,\ldots,i_d}\mathbb{E}[Y_{i_1}\cdot Y_{i_2}\cdot\ldots\cdot Y_{i_d}] \overset{(3)}{=} \sum_{r=1}^{d}\sum_{\substack{S\subseteq[n]\\|S|=r}}\sum_{i_1,\ldots,i_d}\mathbb{1}_{\{i_1,\ldots,i_d\}=S}\cdot\mathbb{E}[Y_{i_1}\cdot Y_{i_2}\cdot\ldots\cdot Y_{i_d}] \\
&\overset{(4)}{\leq} \sum_{r=1}^{d/2}\sum_{\substack{S\subseteq[n]\\|S|=r}}\sum_{i_1,\ldots,i_d}\mathbb{1}_{\{i_1,\ldots,i_d\}=S}\cdot p^{|S|} \overset{(5)}{=} \sum_{r=1}^{d/2}\sum_{\substack{S\subseteq[n]\\|S|=r}} p^{|S|}\sum_{i_1,\ldots,i_d}\mathbb{1}_{\{i_1,\ldots,i_d\}=S} \\
&\overset{(6)}{\leq} \sum_{r=1}^{d/2}\sum_{\substack{S\subseteq[n]\\|S|=r}} p^{|S|}|S|^d \overset{(7)}{=} \sum_{r=1}^{d/2}\binom{n}{r}p^r r^d \overset{(8)}{\leq} (d/2)^d\sum_{r=1}^{d/2} n^r p^r \overset{(9)}{\leq} O(n^{d/2}p^{d/2}).
\end{aligned}$$

(1) Linearität des Erwartungswertes.

(2) Kompaktere Schreibweise.

(3) Gruppierung der Terme nach der Menge $S = \{i_1, \ldots, i_d\}$.

(4) Nach (iv) sind die Terme mit $|S| > d/2$ gleich 0, können also weggelassen werden. Die übrigen sind höchstens $p^{|S|}$.

(5) Ausklammern.

(6) Damit die $\mathbb{1}$-Variable 1 liefern kann müssen alle $i_1, \ldots, i_d$ in S gewählt werden (das ist notwendig, aber nicht hinreichend). Dafür gibt es $|S|^d$ Möglichkeiten.

(7) Die innere Summe hängt nicht mehr von S ab sondern nur noch von $r = |S|$. Es gibt $\binom{n}{r}$ Summenglieder.

(8) Wir verwenden $\binom{n}{r} \leq n^r$ und $r \leq d/2$.

(9) Wenn $d = O(1)$ dann ist auch $(d/2)^d = O(1)$ und fällt weg. Wegen $p = \Omega(1/n)$ ist $pn = \Omega(1)$. Also dominiert der Summand mit $r = d/2$ die anderen (konstant vielen) Summanden.

(v) Wieder erst eine Rechnung, dann die Begründungen:

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

(1) Verwendet Definition von Y und Erwartungswert von X.

(2) Weil d gerade ist, können die Betragsstriche weggelassen werden.

(3) Markov Ungleichung angewendet auf Y^d. Beachte: Weil d gerade ist gilt $Y^d \geq 0$.

(4) Das Resultat der letzten Teilaufgabe.