4. Průměrná časová složitost *(R. Cinkais, R. Barczi a R. Vansa)*

První pomoc z teorie pravděpodobnosti

Definice: *Diskrétní pravděpodobnostní prostor* je uspořádaná dvojice (Ω, P), kde Ω je nejvýše spočetná množina *elementárních jevů* a $P : \Omega \to [0, 1]$ je funkce, pro kterou platí:

$$\sum_{x \in \Omega} P(x) = 1.$$

Hodnotám $P(x)$ budeme říkat *pravděpodobnosti* jednotlivých elementárních jevů.

Příklad: Množina elementárních jevů při vrhu kostkou je $\Omega = \{1, 2, 3, 4, 5, 6\}$ a pokud je kostka spravedlivá, platí $\forall x \in \Omega : P(x) = 1/6$.

Definice: *Jevem* A rozumíme libovolnou množinu $A \subseteq \Omega$. *Pravděpodobnost* jevu A počítáme jako:

$$\Pr[A] = \sum_{a \in A} P(a).$$

Příklad: Jaká je pravděpodobnost složeného jevu „na kostce padne sudé číslo“? Označíme $A = \{2, 4, 6\} \subseteq \Omega$. Hledanou pravděpodobností je pak $P(A) = 1/6 + 1/6 + 1/6 = 1/2$.

Definice: *Náhodná proměnná* X je funkce $X : \Omega \to \mathbb{R}$. *Střední hodnotou* náhodné proměnné X rozumíme výraz

$$\mathbb{E}X = \sum_{x \in \Omega} P(x) \cdot X(x).$$

Příklad: Nechť H je náhodná proměnná přiřazující vrhu kostkou dosaženou hodnotu, tedy $\forall x \in \Omega : H(x) = x$. Střední (očekávanou) hodnotou této proměnné je $\mathbb{E}H = 1/6 \cdot 1 + 1/6 \cdot 2 + 1/6 \cdot 3 + 1/6 \cdot 4 + 1/6 \cdot 5 + 1/6 \cdot 6 = 7/2$.

Věta: (*Linearita střední hodnoty*) Nechť $A, B : \Omega \to \mathbb{R}$ jsou náhodné proměnné a $\alpha \in \mathbb{R}$ je libovolná konstanta. Potom:

1. $\mathbb{E}(A + B) = \mathbb{E}(A) + \mathbb{E}(B)$,
2. $\mathbb{E}(\alpha \cdot A) = \alpha \cdot \mathbb{E}(A)$.

Důkaz:

1. Podle definice střední hodnoty je:

$$\begin{aligned}\mathbb{E}(A + B) = \sum_{x \in \Omega} (A(x) + B(x)) \cdot P(x) &= \sum_{x \in \Omega} (A(x) \cdot P(x) + B(x) \cdot P(x)) = \\ &= \sum_{x \in \Omega} A(x) \cdot P(x) + \sum_{x \in \Omega} B(x) \cdot P(x) = \mathbb{E}A + \mathbb{E}B.\end{aligned}$$

2. Opět přímo z definice střední hodnoty:

$$\mathbb{E}(\alpha \cdot A) = \sum_{x \in \Omega} \alpha \cdot A(x) \cdot P(x) = \alpha \cdot \sum_{x \in \Omega} A(x) \cdot P(x) = \alpha \cdot \mathbb{E}A.$$

♡

Příklad: Házíme dvěmi odlišnými hracími kostkami. Uvažovaný prostor elementárních jevů je proto $\Omega = \{(1,1),(1,2),\ldots,(6,5),(6,6)\}$ a $\forall (a,b) \in \Omega : P((a,b)) = 1/36$. Mějme náhodnou proměnnou S přiřazující vrhu kostkami součet dosažených hodnot, tedy $\forall (a,b) \in \Omega : S((a,b)) = a+b$. Chceme určit $\mathbb{E}S$.

Označme A náhodnou proměnnou, která při vrhu kostkami vrací hodnotu na první kostce, tedy $\forall (a,b) \in \Omega : A((a,b)) = a$. Podobně označme B náhodnou proměnnou, která vrací hodnotu na druhé kostce, tedy $\forall (a,b) \in \Omega : B((a,b)) = b$. V předcházejícím příkladě jsme si ukázali, že $\mathbb{E}A = \mathbb{E}B = 7/2$. Protože $S((a,b)) = a+b = A((a,b)) + B((a,b))$, použitím věty o linearitě střední hodnoty dostáváme $\mathbb{E}S = \mathbb{E}(A+B) = \mathbb{E}A + \mathbb{E}B = 7/2 + 7/2 = 7$.

Poznámka: Pro každou náhodnou proměnnou Y platí:

$$\mathbb{E}Y = \sum_{x \in \mathbb{R}} x \cdot \Pr[Y = x].$$

Důkaz: Uvedený vztah plyne okamžitě z definice střední hodnoty, sloučíme-li dohromady všechny členy $x \in \Omega$ se stejnou hodnotou $Y(x)$. ♡

Průměrná složitost algoritmu

Průměrnou složitost algoritmu můžeme počítat přes všechny vstupy nebo přes náhodné volby vstupů. V prvním případě je algoritmus pravděpodobnostní („hází mincí“) a složitost počítáme jako průměr přes všechny vstupy. Ve druhém je zase algoritmus deterministický a zajímá nás pouze složitost v nejhorším případě. Pokud jde o konečný výsledek, jsou oba uvedené způsoby ekvivalentní.

Našim cílem bude odvodit průměrnou složitost algoritmu *Select*(k,X) z minulé přednášky. K tomu si ukážeme algoritmus sloužící k výběru pivota pro *Select*(k,X) (*Select*(k,X) jej při svém běhu bude spouštět jako podprogram). Bez újmy na obecnosti předpokládejme, že všechny prvky množiny X jsou navzájem různé.

Algoritmus: (*Hledání lžimediánu v množině X*)

1. Zvol náhodně $x \in X$ (všechna x se stejnou pravděpodobností).
2. Spočítej $A = |\{y \in X : y < x\}|$ a $B = |\{y \in X : y > x\}|$.
3. Pokud $A \geq |X|/4$ a zároveň $B \geq |X|/4$, vrať x, jinak skoč na 1.

Pozorování: Uvažujme, že házíme spravedlivou mincí. Aby nedošlo k žádnému mezinárodnímu sporu, označme strany mince jako tučňák (T) a ryba (R). Tedy pravděpodobnost pádu na nějakou stranu je 1/2. Budeme zkoumat, jaká je střední hodnota čekání na pád tučňáka $\mathbb{E}T$. Množinu všech jevů (posloupností vrhů), které mohou nastat, označme jako $\Omega = \{T, RT, RRT, \ldots, RRR\ldots\}$. Platí:

$$\Pr[T] = \frac{1}{2}, \Pr[RT] = \frac{1}{4}, \Pr[R^kT] = \frac{1}{2^{k+1}}, \Pr[RRR\ldots] = \lim_{k\to\infty} \frac{1}{2^k} = 0.$$

Podle pozorování máme:

$$\mathbb{E}T = \sum_{k=1}^{\infty} k \cdot \underbrace{\Pr[T = k]}_{\frac{1}{2^k}} = \sum_{k=1}^{\infty} \frac{k}{2^k}.$$

Tuto sumu je však trochu obtížné spočítat (kdo chce si to může zkusit), proto provedeme následující trik:

$$\mathbb{E}T = 1 + \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \mathbb{E}T,$$

to znamená, že po jednom tahu jsme s poloviční pravděpodobností vyhráli a s poloviční pravděpodobností se situace nezměnila. Jednoduchou úpravou tedy zjistíme, že:

$$\mathbb{E}T = 2.$$

Lemma: Čekání na jev, který nastane s pravděpodobností p, trvá průměrně $1/p$.

Důkaz: $\mathbb{E}T = 1 + (1-p)\mathbb{E}T$, tedy $\mathbb{E}T = 1/p$. ♡

Při hledání lžimediánu máme poloviční šanci, že se trefíme při náhodném výběru do prostředních dvou čtvrtin – průměrně tedy vykonáme dva průchody cyklem, přičemž jeden průchod trvá $\Theta(n)$. Celkem bude tedy hledání lžimediánu trvat průměrně $\Theta(n)$.

Věta: *Select* s náhodnou volbou pivota má v průměru časovou složitost $\Theta(n)$.

Důkaz: Rozdělíme algoritmus na *fáze* $\equiv$ posloupnosti iterací končící dobrou volbou pivota (jako lžimedián). Všimneme si, že jedna fáze zmenší vstup na nejvýše $\frac{3}{4} \cdot n$ a trvá v průměru $\Theta(n)$ (průměrně 2 iterace po $\Theta(n)$). Tedy $\mathbb{E}T = \mathbb{E}T_1 + \mathbb{E}T_2 + \ldots = \Theta(n)$, kde $\mathbb{E}T_i$ označuje i-tou fázi. ♡

Věta: *Select* s pivotem rovným prostřednímu prvku má v průměru přes permutace na vstupu časovou složitost $\Theta(n)$.

Důkaz: Ukážeme, že algoritmus s pevnou volbou pivota se na náhodném vstupu chová stejně, jako algoritmus s náhodnou volbou pivota na pevném vstupu, jehož průměrnou složitost odhaduje předchozí věta.

Pokud je na vstupu náhodná permutace π, jsou všechny hodnoty pivota p stejně pravděpodobné. Zbývá ukázat, že další iterace algoritmu dostane na vstupu prvky větší nebo menší než p opět v náhodném pořadí (s rovnoměrným rozdělením pravděpodobností). K tomu sestrojíme bijekci mezi množinou všech permutací na $\{1,\ldots,n\}$ a množinou čtveřic (p, L, π_M, π_V), kde:

- $p \in \{1,\ldots,n\}$ je prostřední prvek permutace π (pivot), tedy $\pi[l]$ ležící na pozici $l = \lfloor (1+n)/2 \rfloor$,
- $L \in \binom{\{1,\ldots,l-1,l+1,\ldots,n\}}{p-1}$ je množina pozic, na nichž jsou umístěny prvky menší než p (ty, které náš algoritmus umístí do množiny M),
- π_M permutace na $\{1,\ldots,p-1\}$ popisující pořadí menších prvků a
- π_V permutace na $\{p+1,\ldots,n\}$, tedy pořadí větších prvků.

Jakmile zvolíme p, jsou volby L, π_M a π_V navzájem nezávislé. Každé volbě π_M tedy odpovídá stejný počet permutací π, a proto jsou všechny π_M stejně pravděpodobné. Analogicky pro π_V.

♡