Eksamensdisposition - Kapitel 3

Søren Mulvad, rbn601

17. juni 2019

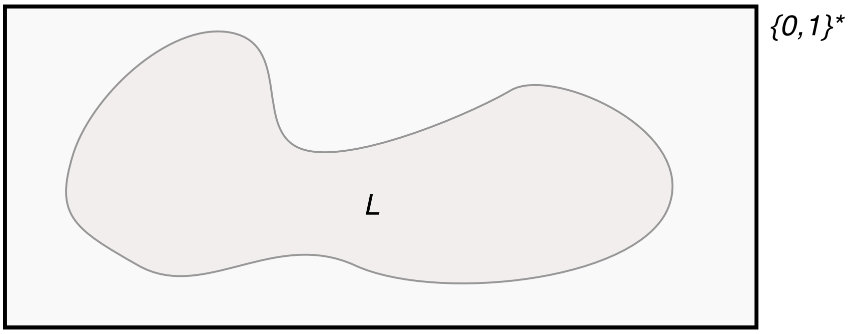
- **Two-point sampling**
  - Algoritme 1
  - Algoritme 2
  - Sandsynlighed for algoritme 2 fejler
- **The Coupon Collector's Problem**
  - Forventet antal runder
  - Sandsynlighed for flere end r runder
- **Markovs ulighed og Chebyshevs ulighed (hvis tid)**

Eksamensdisposition - Kapitel 3

Two-point Sampling

Et decision-problem kan repræsenteres om en delmængde (et sprog) $L \subseteq \{0,1\}^* = \{\emptyset, 0, 1, 00, 01, \dots\}$. En korrekt algoritme A^* for L skal opfylde

$$x \in L \to A^*(x) = 1$$
$$x \notin L \to A^*(x) = 0$$



Figur 1: En delmængde (sprog) $L \subseteq \{0,1\}^*$

Lad os nu betragte en Monte Carlo algoritme A (one-sided error), så:

$$x \in L \to A(x) = 1 \text{ med sandsynlighed } p \geq 1/2$$
$$x \notin L \to A(x) = 0 \text{ med sandsynlighed } 1$$

Antag at algoritme A bruger $\lg n$ tilfældige bits repræsenteret som et tal $r \in \{0, ..., n-1\}$ hvor n er et primtal. I følgende bruger vi notationen $A(x, r)$ for at beskrive outputtet af A på input x, hvor A vælger den tilfældige bitstreng r. Og lad os i fejlsandsynlighederne antage, at vores konkrete $x \in L$ så det korrekte svar er 1.

Algoritme 1 - $t \lg n$ random bits
Vælg t tal $r_0, \dots, r_{t-1} \in [n]$ uafhængigt og uniformt tilfældigt.
Beregn $A(x, r_0), \dots, A(x, r_{t-1})$. Hvis vi en enkelt gang ser tallet 1 er det bevis på $x \in L$, ellers hvis vi *alle* gange får 0 vælger vi det som output.

Så vil fejlsandsynligheden være $< \left(\frac{1}{2}\right)^t = 1/2^t$.

Problemet ved denne tilgang er, at vi skal vælge $t \lg n$ random bits. Hvis vi f.eks. vælger $t = 2$ skal vi bruge $2 \ln n$ random bits for en fejlsandsynlighed $< 1/4$.

Algoritme 2 - $2 \lg n$ random bits
Vælg $a, b \in [n]$ uafhængigt og uniformt tilfældigt.

Da vi antager n er et primtal, så ved vi at såfremt vi lades $r_i = (a \cdot i + b) \bmod n$, så vil r_i og r_j hvor $i \neq j$ være uniformt distribueret i $[n]$ og parvist uafhængige (kan blot antages, skal ikke bevises).

Igen beregner vi $A(x, r_0), \dots, A(x, r_{t-1})$ og vælger 1 såfremt den optræder bare én gang, ellers 0.

Nu bruger vi kun $2 \lg n$ random bits.

Sandsynlighed for at algoritme 2 fejler

For $i = 0, \ldots, t-1$ lader vi $Y_i = A(x, r_i)$. Lad nu $Y = \sum_{i \in [t]} Y_i$.
Da kan vi beregne den forventede værdi:

$$\mathbb{E}[Y] = \sum_{i \in [t]} \mathbb{E}[Y_i] = tp \geq \frac{t}{2} \tag{1}$$

Idet vi lader symbolet $p = \mathbb{P}[Y_i = 1] \geq \frac{1}{2}$.

Derudover har vi jf. at de forskellige Y_i er parvist uafhængige, at:

$$\sigma_Y{}^2 = \sum_{i \in [t]} \sigma_{Y_i}{}^2 = \sum_{i \in [t]} p(1-p) \leq \frac{t}{4} \tag{2}$$

$$\Updownarrow$$

$$\sigma_Y = \frac{\sqrt{t}}{2} \tag{3}$$

I (2) bruger vi at de stokastiske variable Y_i er Bernoulli trials som har variansen $\sigma_{Y_i}^2 = p \cdot (1-p)$ hvor $p = \mathbb{P}[Y_i = 1]$ og at udtrykket $p(1-p)$ er størst når $p = 1/2$ og da bliver $1/4$.
I (3) tager vi kvadratroden og får herved standardafvigelsen.
Da kan vi beregne sandsynligheden for at algoritme 2 fejler til:

$$\begin{aligned}
\mathbb{P}[\text{Algoritme 2 fejler}] &= \mathbb{P}[Y = 0] && (4)\\
&\leq \mathbb{P}\left[|Y - \mu_Y| \geq \frac{t}{2}\right] && (5)\\
&= \mathbb{P}\left[|Y - \mu_Y| \geq \sqrt{t}\frac{\sqrt{t}}{2}\right] \\
&\leq \frac{1}{(\sqrt{t})^2} && (6)\\
&\leq \frac{1}{t}
\end{aligned}$$

I (4) bruger vi, at algoritmen fejler (givet vores antagelse at svaret *er* 1) når vi får 0 i alle vores trials.
I (5) indsætter vi ulighed (1) for μ_Y. Grunden til vi får '$\leq$'-tegnet i denne ligning er pga. vi tager den absolutte værdi, hvorved der kommer "to muligheder" for at det udtryk kan være større end brøken, hvilket der ikke gjorde i (1).
I (6) benytter vi Chebyshev's ulighed.

Hermed har vi altså bestemt, at vi kan få en relativt lav sandsynlighed for fejl selvom vi kun bruger $2 \lg n$ random bits.

The Coupon Collector's Problem

Betragt følgende eksperiment. Vi har n unikke unikke kupontyper:

$$\underbrace{\square\ \square\ \square\ \square\ \square\ \cdots\ \square}_{n}$$

I hver runde vælges en kupon-type uafhængigt og uniformt tilfældigt. Vi stopper når alle kupon-typer er valgt. Hvor mange runder vil der være i dette eksperiment?

For at besvare dette skal vi først definere hvad en epoke er. For $i = 0, \dots, n-1$ består den i'te epoke af de runder, der starter lige efter den i'te succes og slutter i runden med $(i+1)$'te succes, hvor en succes er defineret som at vælge en kupontype vi ikke har set før. Eksempelvis kunne vi have:

$$\underbrace{C_2,}_{\text{Epoke 0}} \underbrace{C_2, C_1,}_{\text{Epoke 1}} \underbrace{C_2, C_2, C_3}_{\text{Epoke 2}}, \dots$$

Forventet antal runder

For $i = 0, \dots, n-1$ lader vi Y_i være længden af epoke i. Lad nu $Y = \sum_{i=0}^{n-1} Y_i$. Vi har, at sandsynligheden i den i'te epoke for at finde en ny kupon er antallet af ufundne kuponer $n - i$ over alle de forskellige kupontyper n:

$$p_i = \frac{n-i}{n}$$

Bruger vi, at dette er geometrisk distribueret får vi:

$$\mathbb{E}[Y_i] = \frac{1}{p_i} = \frac{n}{n-i}$$

Da kan vi beregne:

$$\mu_Y = \sum_{i=0}^{n-1} \mathbb{E}[Y_i] = \sum_{i=0}^{n-1} \frac{n}{n-i} = n \sum_{i=1}^{n} \frac{1}{i} = nH_n = n \ln n + \Theta(n) = O(n \ln n)$$

Sandsynlighed for flere end r runder

For $i = 1, \dots, n$ og $r \in \mathbb{N}_0$ defineres følgende begivenhed:
$\mathcal{E}_i^r$: Kupontype i vælges *ikke* i de første r runder.

Da er begivenheden at mindst én kupontype ikke vælges i de r første runder $\bigcup_{i=1}^{n} \mathcal{E}_i^r$. Denne sandsynlighed er naturligvis ækvivalent med sandsynligheden for, at der totalt set vil være flere end r runder før vi er færdig. Da kan vi bestemme sandsynligheden for flere end r runder til:

$$\begin{aligned}
\mathbb{P}\left[\bigcup_{i \in [n]} \mathcal{E}_i^r\right] &\le \sum_{i \in [n]} \mathbb{P}\left[\mathcal{E}_i^r\right] \tag{7} \\
&= \sum_{i \in [n]} \left(\frac{n-1}{n}\right)^r \tag{8} \\
&= n\left(1 - \frac{1}{n}\right)^r \\
&\le n\left(e^{-1/n}\right)^r \tag{9} \\
&= ne^{-r/n}
\end{aligned}$$

Hvor vi i (7) bruger Union Bound, i (8) har at $\mathbb{P}\left[\mathcal{E}_i^1\right] = \frac{n-1}{n}$ hvor det skal ske r gange og i (9) bruger regnereglen $1 + x \le e^x$ for alle $x \in \mathbb{R}$.

Eksempel på beregning

Lad os vælge $r = \beta n \ln n$. Da vil

$$\mathbb{P}\left[\text{Mere end } r \text{ runder}\right] \leq n \cdot e^{-\beta \ln n} = n \cdot n^{-\beta} = n^{1-\beta}$$

For $\beta = 2$ får vi:

$$\mathbb{P}\left[\text{Mere end } 2n \ln n \text{ runder}\right] \leq \frac{1}{n}$$

Altså er sandsynligheden for at laver flere end dobbelt så mange runder som forventet relativt lille.

Markovs ulighed

Givet en tilfældig variabel $X \geq 0$ og $t > 0$, så:

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t} \tag{10}$$

Såfremt $\mathbb{E}[X] \neq 0$ og $k > 0$ kan vi omskrive det til:

$$\mathbb{P}\Big[X \geq \underbrace{k \cdot \mathbb{E}[X]}_{t}\Big] \leq \frac{1}{k}$$

Bevis:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x} x\mathbb{P}[X = x] \\
&\geq \sum_{x \geq t} x\mathbb{P}[X = x] && (11) \\
&\geq \sum_{x \geq t} t\mathbb{P}[X = x] && (12) \\
&= t\mathbb{P}[X \geq t] && (13)
\end{aligned}$$

Hvor uligheden i (11) gælder da vi antager $X \geq 0$ og vi summerer over potentielt færre led, uligheden i (12) gælder da $t \leq x$ i vores summering, og (13) gælder da vi blot indsætter $x \geq t$ fra vores summering i selve sandsynligheden.

Chebyshevs ulighed

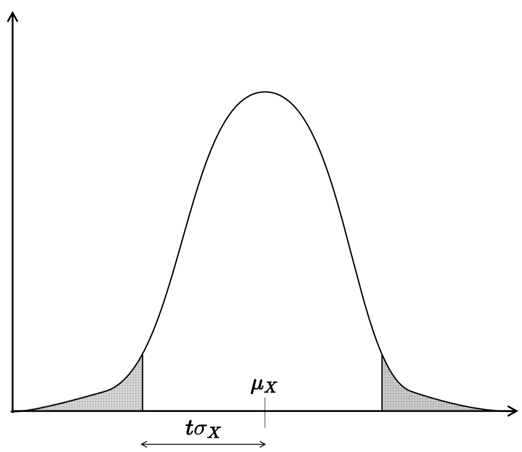
Givet en tilfældig variabel X med forventet værdi μ_X, hvor $\sigma_X > 0$ og $t > 0$, så:

$$\mathbb{P}[|X - \mu_X| \geq t\sigma_X] \leq \frac{1}{t^2} \tag{14}$$

Bevis:

$$\begin{aligned}
\mathbb{P}[|X - \mu_X| \geq t\sigma_X] &= \mathbb{P}\left[(X - \mu_X)^2 \geq t^2{\sigma_X}^2\right] \\
&\leq \frac{\mathbb{E}\left[(X - \mu_X)^2\right]}{t^2{\sigma_X}^2} && (15) \\
&= \frac{{\sigma_X}^2}{t^2{\sigma_X}^2} && (16) \\
&= \frac{1}{t^2}
\end{aligned}$$

Her benytter vi Markovs ulighed i (15) og selve definitionen på varians σ^2 i (16).



Figur 2: Illustration af Chebyshevs ulighed. Summen af de skraverede områder er $\leq 1/t^2$.