Eksamensdisposition - Algebraiske Teknikker

Søren Mulvad, rbn601

17. juni 2019

- **Freivalds teknik: Matrixprodukt verificering**
- **Strenglighed 1: (mod p) fingerprint**
- **Strenglighed 2: Polynomial identitet med fast p**
- **Sætlighed med produkter**
- **Pattern Matching: Efficient Fingerprints**

Eksamensdisposition - Algebraic Techniques

Freivalds teknik: Matrixprodukt verificering

Givet tre $n \times n$ matricer $\mathbf{A}, \mathbf{B}$ og $\mathbf{C}$, verificer da om $\mathbf{AB} = \mathbf{C}$. Naivt kan vi gøre det ved selv at udregne matrixproduktet, hvilket tager omkring $O(n^{2.4})$ tid og er meget kompliceret.

Nu ønsker vi i stedet at lave en verificering om $\mathbf{C}$ er korrekt i $O(n^2)$ tid.

Algoritme

Generer en tilfældig vektor $\mathbf{r} \in \{0,1\}^n$.
Da siger vi "$\mathbf{C} = \mathbf{AB}$" hvis $\mathbf{Cr} = \mathbf{A}(\mathbf{Br})$.

$$\underset{\mathbf{C}}{\begin{bmatrix} & \\ & \\ & \end{bmatrix}} \times \underset{\mathbf{r}}{\begin{bmatrix} \\ \\ \end{bmatrix}} \stackrel{?}{=} \underset{\mathbf{A}}{\begin{bmatrix} & \\ & \\ & \end{bmatrix}} \times \left(\underset{\mathbf{B}}{\begin{bmatrix} & \\ & \\ & \end{bmatrix}} \times \underset{\mathbf{r}}{\begin{bmatrix} \\ \\ \end{bmatrix}} \right)$$

Det smarte er, at vi kan beregne produktet af en matrix og en vektor i $O(n^2)$ tid.

Sandsynlighed for false positive

Vi ønsker nu at kigge på sandsynligheden for en false positive (algoritmen verificerer en forkert $\mathbf{C}$). Det har vi når $\mathbf{D} = \mathbf{AB} - \mathbf{C} \neq 0^{n \times n}$, men vi samtidig i vores algoritme fik $\mathbf{Dr} = 0^n$.

Hvis $\mathbf{D}$ er forskellig fra 0-matricen må der findes minimum et koordinat som ikke er 0:

$$\exists i, j : D_{ij} \neq 0$$

Såfremt vektor $\mathbf{Dr} = 0^n$, altså at alle bits er 0, må der samtidig gælde at den i'te bit er 0:

$$(\mathbf{Dr})_i = \sum_{k \in [n]} D_{ik} r_k = 0$$

Antag $\mathbf{D}$ ikke er 0-matricen. Da beregner vi sandsynligheden for en false positive til:

$$\begin{aligned}
\mathbb{P}\left[\mathbf{Cr} = \mathbf{A}(\mathbf{Br})\right] &= \mathbb{P}\left[\mathbf{Dr} = 0^n\right] && (1)\\
&\leq \mathbb{P}\left[(\mathbf{Dr})_i = 0\right] && (2)\\
&= \mathbb{P}\left[\sum_{k \in [n]} \mathbf{D}_{ik}\mathbf{r}_k = 0\right] && (3)\\
&= \mathbb{P}\left[\mathbf{D}_{ij}\mathbf{r}_j + \sum_{\substack{k \in [n] \\ k \neq j}} \mathbf{D}_{ik}\mathbf{r}_k = 0\right] && (4)\\
&= \mathbb{P}\left[\mathbf{r}_j = -\frac{1}{\mathbf{D}_{ij}} \sum_{\substack{k \in [n] \\ k \neq j}} \mathbf{D}_{ik}\mathbf{r}_k\right] && (5)\\
&\leq \frac{1}{2} && (6)
\end{aligned}$$

I (1) har vi $\mathbf{Cr} = \mathbf{A}(\mathbf{Br}) \iff \mathbf{Cr} - (\mathbf{AB})\mathbf{r} = 0^n \iff (\mathbf{C} - \mathbf{AB})\mathbf{r} = 0^n$.
I (2) benytter vi, at hvis $\mathbf{Dr}$ er 0-vektoren, så må alle koordinater være 0.
I (3) benytter vi blot definition for hvordan man udregner matrix-vektor produktet.

I (4) splitter vi summen i det ene led hvor $k = j$ samt alle de andre.
I (5) trækker vi vores sum fra på begge sider og dividerer herefter med $\mathbf{D}_{ij}$.
I (6) benytter vi at alle indgange i $\mathbf{r}$ er uafhængige, så vi kan antage at $\mathbf{r}_j$ vælges til sidst. Siden det vælges uniformt fra $\{0, 1\}$ er der to unikke værdier det kan være, og højest én af dem vil opfylde ligningen i sandsynligheden.

Vi ser, at vi kan få vores fejlsandsynlighed ned på $\leq 1/2^t$ ved at lave t uafhængige verifikationer i alt med en køretid på $O(tn^2)$.

Streng-lighed 1: (mod p) fingerprint

Antag Alice har en n-bit streng $\mathbf{a} = (a_0, \ldots, a_{n-1})$ og Bob har en n-bit streng $(b_0, \ldots, b_{n-1}) = \mathbf{b}$. De ønsker nu at tjekke $\mathbf{a} \stackrel{?}{=} \mathbf{b}$ med høj sandsynlighed, men det skal foregå ved at sende relativt få bits (meget færre end n bits).
Lad

$$a = \sum_{i \in [n]} \mathbf{a}_i 2^i \qquad\qquad b = \sum_{i \in [n]} \mathbf{b}_i 2^i$$

Vælg et uniformt tilfældigt primtal $p < n^2$ og tjek derefter om:

$$a \bmod p \stackrel{?}{=} b \bmod p$$

Vi ser, at det højest bruger $2 \lg n$ bits kommunikation.

Analyse af sandsynlighed for false positive

Vi har en false positive (FP) når

$$a \bmod p = b \bmod p \quad | \quad a \neq b$$

Vi kan omskrive første udsagn til:

$$a \bmod p = b \bmod p \iff |a - b| \bmod p = 0 \tag{7}$$
$$\iff p \mid |a - b| \tag{8}$$

hvor $c = |a - b| < 2^n$.
Vi skriver nu vores c via primtalsfaktorisering:

$$c = \prod_i p_i^{d_i} \qquad \text{alle } p_i \geq 2$$

Så vi kigger på hvor mange primtal p_i der optræder hvor graden $d_i \geq 1$. Det kan højest være $\lg n$ fordi de alle sammen skal ganges sammen og de alle er mindst 2. F.eks.:

$$2 \cdot 3 \cdot 7^2 = \mathbb{Z} \Longrightarrow \mathbb{Z} > 2^3$$

Da må det betyde, at antal primtal $p \mid c$ er $\leq \lg c = n$.

$$\begin{aligned}
\mathbb{P}[\text{FP}] &= \mathbb{P}[p \mid c] \\
&\leq \frac{\#(p \mid c)}{\#(p < n^2)} && (9) \\
&\approx \frac{n}{n^2 / \ln n^2} && (10) \\
&= \frac{2 \ln n}{n}
\end{aligned}$$

I (9) har vi antallet af primtal p der dividerer c ud af alle de mulige primtal $p < n^2$ vi kunne have valgt.
I (10) benytter vi primtalssætningen der siger, at antallet af primtal mindre end tallet x konvergerer mod $x / \ln x$.

Streng-lighed 2: Polynomial identitet med fast p

Vi har igen to bitstrenge $\mathbf{a} = (a_0, \ldots, a_{n-1})$ og $\mathbf{b} = (b_0, \ldots, b_{n-1})$. Nu definerer vi:

$$A(x) = \sum_{i \in [n]} a_i x^i \qquad\qquad B(x) = \sum_{i \in [n]} b_i x^i$$

Vi starter med deterministisk at vælge et primtal $p \geq n^2$ og et r uniformt tilfældigt fra $[p]$. Derefter beregner vi:

$$A(r) \stackrel{?}{=} B(r) \pmod{p}$$

Fidusen er, at så kan vi tjekke om bitstrenge er ens ved kun at sende lidt over $2 \lg n$ bits.

Vi får at graden af A og B er $d = n - 1$, og der kan derfor højest være $n - 1$ forskellige steder hvor de er ens. Således får vi sandsynligheden for en false positive til:

$$\mathbb{P}[\text{FP}] = \mathbb{P}[A(r) = B(r) \mid A \neq B] \leq \frac{n-1}{p} \leq \frac{n}{n^2} = \frac{1}{n}$$

Sæt-lighed med produkter

Vi kan modificere en smule på ovenstående algoritme for at tjekke om to sæt er ens:

$$a = \{a_0, \ldots, a_{m-1}\} \stackrel{?}{=} \{b_0, \ldots, b_{m-1}\} = b$$

ved at vælge et r uniformt i en finit mængde $\mathbb{S}$ og derefter beregne:

$$\prod_{i \in [m]} (r - a_i) \stackrel{?}{=} \prod_{i \in [m]} (r - b_i)$$

Fidusen er at vi finder produktet, så derved bliver rækkefølgen ligegyldig.

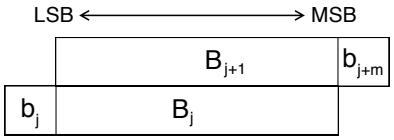
Pattern matching: Efficient Fingerprints

Givet $a = (a_0, \ldots, a_{m-1})$ og $b = (b_0, \ldots, b_{n-1})$ hvor $m \leq n$ ønsker vi at tjekke om der findes et j så:

$$a \stackrel{?}{=} (b_j, \ldots, b_{j+m-1}) = B_j$$

Vi vælger uniformt tilfældigt et $p < n^2$. Herefter beregner vi så $a \bmod p$. Derudover ønsker vi at beregne alle $B_j \bmod p$ i $O(n)$ tid, hvilket vi både udregner i baglæns rækkefølge og hvor vi aflæser deres bitstreng baglæns.

Vi starter med naivt at udregne den B_{j+1} længst til højre i bitstrengen, B_{n-m}. Derefter beregner vi B_j baglæns:



Figur 1: Pattern Matching bitstrenge. Læg mærke til at LSB og MSB står omvendt af hvordan vi normalt ville læse det, da vi udregner det baglæns.

Vi antager at vi kender B_{j+1} Da skal vi starte med at få vores mest signifikante bit b_{j+m} væk, og herefter gange med 2 for at forskyde vores streng med 1. Derefter skal vi blot lægge bit b_j til. Til sidst skal vi køre modulus med p. Altså bliver vores formel:

$$B_j \bmod p = \left(2\left(B_{j+1} - b_{j+m}2^{m-1}\right) + b_j\right) \bmod p$$

Således kan vi beregne B_{n-m} i $\mathcal{O}(m)$ tid og herefter alle de resterende B_j frem til B_0 i hver $\mathcal{O}(1)$ tid, som giver os en endelig køretid på $\mathcal{O}(n \cdot 1 + m) = \mathcal{O}(n)$.

(Hvis tid)

Freivalds teknik: Polynomial identitet

Givet to polynomier $P_1(x), P_2(x) \in \mathbb{F}[x]$ (legement af f.eks. réelle tal, primtal, etc) af grad $\leq d$ som black boxes, bestem da $P_1 \stackrel{?}{=} P_2$.

Lad $\mathbb{S} \subseteq \mathbb{F}$ være en finit mængde og vælg uniformt et $r \in \mathbb{S}$.
Da siger vi "$P_1 = P_2$" hvis $P_1(r) = P_2(r)$. Lad $Q = P_1 - P_2$. Da får vi at sandsynligheden for en false positive er:

$$\mathbb{P}\left[P_1(r) = P_2(r) \mid P_1 \neq P_2\right] = \mathbb{P}\left[Q(r) = 0 \mid Q \neq 0\right] \leq \frac{d}{|\mathbb{S}|}$$

Dette gælder da ligningen $Q(x) = 0$ højest har d løsninger x, men vi har et udfaldsrum der er $|\mathbb{S}|$ stort.

Schwartz-Zippel theoremet - Multivariable polynomier

Vi kan generalisere ovenstående til casen med flere variable. Da definerer vi graden af leddet $\alpha x_1^{d_1} x_2^{d_2} \dots x_n^{d_n}$ til at være $d_1 + d_2 + \cdots + d_n$ og den totale grad af polynomiet d til at være maksimum graden af alle dens led.

Da siger Schwartz-Zippel theoremet:
Lad polynomium $Q(x_1, \dots, x_n) \in \mathbb{F}[x_1, \dots, x_n]$ have en total grad $\leq d$. Lad igen $\mathbb{S} \subseteq \mathbb{F}$ være en finit mængde og vælg uniformt tilfældigt $r_1, \dots, r_n \in \mathbb{S}^n$.
Da får vi en generel formel for vores unikke case før:

$$\mathbb{P}\left[Q(r_1, \dots, r_n) = 0 \mid Q \neq 0\right] = \frac{d}{|\mathbb{S}|}$$

Induktionsbevis

Vi har allerede bevist casen når $n = 1$. Antag $n \geq 2$ og det holder for alle mindre n. Lad $Q \neq 0$ og lad $k > 0$ være den største eksponent af x_n.

Da vil der eksistere $Q_0, \dots, Q_k$ således at

$$Q(x_1, \dots, x_n) = \sum_{i=0}^{k} Q_i(x_1, \dots, x_{n-1}) x_n^i$$

(Det kan vi få på simpel vis ved bare at gruppere alle led der indeholder x_n^i og flytte x_n^i uden for en parentes.)
Vi har at $Q_k \neq 0$ da k er en eksponent af x_n som indgår i Q. $\deg(Q_k) \leq d-k$, da $Q_k(x_1, \dots, x_{n-1}) x_n^k$ er et led i Q, så graden af dette må være $\leq d$ og når vi fjerner x_n^k fjerner vi k fra graden. Vær opmærksom på dette kun gælder for Q_k, ikke nødvendigvis for de andre Q_i.

Nu vælger vi uniformt tilfældigt $r_1, \dots, r_{n-1} \in \mathbb{S}$. Lad $C_i = Q_i(r_1, \dots, r_{n-1})$. Da $Q_k \neq 0$ har vi pr. vores induktionsantagelse at $\mathbb{P}\left[C_k = 0\right] \leq \frac{d-k}{|\mathbb{S}|}$.

Hvis $C_k \neq 0$ kan vi definere $q(x) = \sum_{i=1}^{k} C_i x^i = Q(r_1, \dots, r_{n-1}, x)$. Hvis $q(x) \neq 0$ kan vi se det har graden k, så for uniform $r_n \in \mathbb{S}$ får vi:

$$\mathbb{P}\left[q(r_n) = 0 \mid C_k \neq 0\right] \leq \frac{k}{|\mathbb{S}|}$$

Endelig kan vi udregne:

$$\begin{aligned} \mathbb{P}\left[Q(r_1, \dots, r_n) = 0\right] &\leq \mathbb{P}\left[C_k = 0\right] + \mathbb{P}\left[q(r_n) = 0 \mid C_k \neq 0\right] && (11) \\ &\leq \frac{d-k}{|\mathbb{S}|} + \frac{k}{|\mathbb{S}|} = \frac{d}{|\mathbb{S}|} && (12) \end{aligned}$$

Hermed har vi altså bevist Schwartz-Zippel Theoremet, som er en generalisering for polynomial identitet for multivariable polynomier.