Lecture 13

Sampling and Random Algorithms for Matrix Multiplication

Bao-Jun Cai, 5/27/2026

Introduction to Algorithms for Data Science and Physics IMP@Fudan, 2026

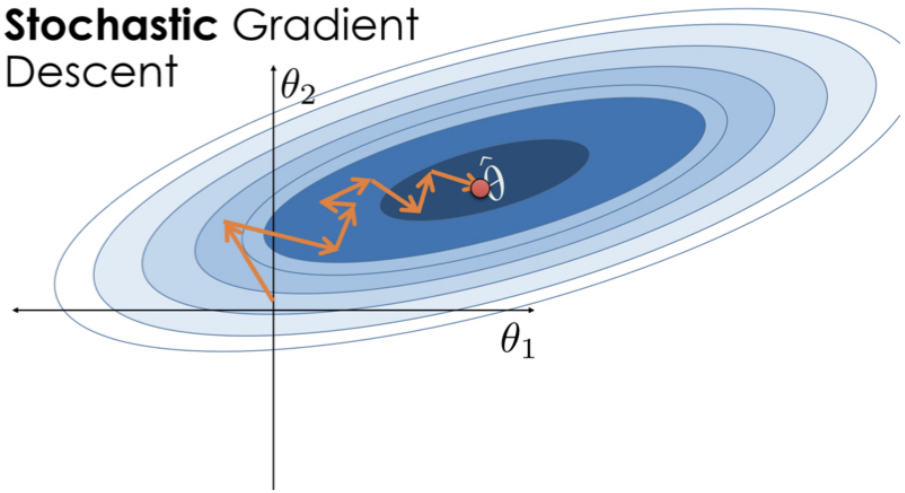
Topics of this lecture:

- reducing the accuracy to speed up the calculation(s)
- sampling a matrix with probability p_k
- length squared sampling scheme $p_k \sim \|\mathbf{A}(;,k)\| \cdot \|\mathbf{B}(k,;)\|$
- LR decomposition of AB $\mathcal{O}(mn) \to \mathcal{O}(mdn)$
- CUR or LUR decomposition of A error bound $\lesssim 2\|\mathbf{A}\|_F^2 \left(\frac{1}{\sqrt{r}} + \frac{r}{t}\right)$
- birthday paradox $K \sim \Theta(\sqrt{n})$

Motivation and examples

1. issue related to fast store in random access memory (RAM)

2. machine learning: we have a large population and want to take a small sample to perform some optimization and then argue that the optimal solution on the sample will be approximately optimal over the all population, example: stochastic gradient descent



Matrix multiplication revisited

$$\mathbf{A} \in \mathbf{R}^{m \times d},\ \mathbf{B} \in \mathbf{R}^{d \times n} \rightarrow \mathbf{AB} \in \mathbf{R}^{m \times n} \sim \mathcal{O}(mdn)$$

divide-and-conquer

Straussen's algorithm: $\sim \mathcal{O}(n^{\log_2 7})$

Can we do better?

approximated algorithm by paying the price that some accuracy is lost

Ex.: Show that $\mathbf{AB} = \sum_{k=1}^{d} \mathbf{A}(;,k)\mathbf{B}(k,;)$

Idea of sampling

$\mathbf{A} \in \mathbf{R}^{m \times d}, \mathbf{B} \in \mathbf{R}^{d \times n}$

$$\mathbf{M} = \frac{1}{p_k} \overbrace{\mathbf{A}(;,k)}^{k\text{th column}} \underbrace{\mathbf{B}(k,;)}_{k\text{th row}} \leftrightarrow p_k\text{: some probability}$$

$$\mathrm{E}[\mathbf{M}] = \sum_{k=1}^{d} p_k \frac{1}{p_k} \mathbf{A}(;,k)\mathbf{B}(k,;) = \sum_{k=1}^{d} \mathbf{A}(;,k)\mathbf{B}(k,;) = \mathbf{AB}$$

How can one select the probability p_k?

Error produced by the sampling

$$\mathrm{var}[\mathbf{M}] \equiv \mathrm{E}\left[\|\mathbf{AB}-\mathbf{M}\|_F^2\right]$$

Frobenious norm: $\|\mathbf{A}\|_F \equiv \sqrt{\sum_{i,j} a_{ij}^2}$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{var}[m_{ij}] = \sum_{i,j}\left(\mathrm{E}[m_{ij}^2] - \mathrm{E}^2[m_{ij}]\right)$$

$\mathbf{AB} \in \mathrm{R}^{m\times n}$

$$= \left(\sum_{i,j}\sum_{k} p_k \frac{1}{p_k^2} a_{ik}^2 b_{kj}^2\right) - \|\mathbf{AB}\|_F^2$$

Ex.: for $f = \sum_k c_k/p_k$ with $\sum_k p_k = 1$, what is the optimal p_k for minimizing f ?

$$= \sum_k \frac{1}{p_k}\left(\sum_i a_{ik}^2\right)\left(\sum_j b_{kj}^2\right) = \sum_k \frac{1}{p_k}\|\mathbf{A}(:,k)\|^2 \cdot \|\mathbf{B}(k,:)\|^2$$

Length squared sampling

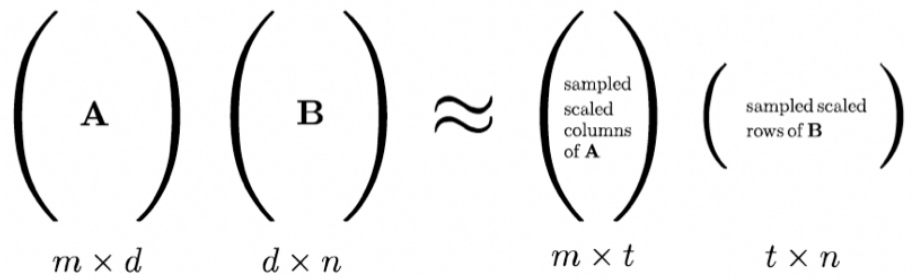
probability $p_k \sim \|\mathbf{A}(;,k)\| \cdot \|\mathbf{B}(k,;)\|$

$$\mathbf{B} = \mathbf{A}^\top \rightarrow p_k \sim \|\mathbf{A}(;,k)\|^2 \rightarrow p_k = \frac{\|\mathbf{A}(;,k)\|^2}{\|\mathbf{A}\|_F^2}$$

$$\mathrm{var}[\mathbf{M}] = \sum_k \frac{1}{p_k} \|\mathbf{A}(;,k)\|^2 \cdot \|\mathbf{B}(k,;)\|^2$$

$$\mathrm{var}[\mathbf{M}] \leq \|\mathbf{A}\|_F^2 \sum_k \|\mathbf{B}(k,;)\|^2 = \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2$$

More trials



$$\langle \mathbf{M} \rangle_t = t^{-1} \sum_{i=1}^{t} \mathbf{M}_i$$

$$= \frac{1}{t} \left(\frac{\mathbf{A}(:,k_1)\mathbf{B}(k_1,:)}{p_{k_1}} + \frac{\mathbf{A}(:,k_2)\mathbf{B}(k_2,:)}{p_{k_2}} + \cdots + \frac{\mathbf{A}(:,k_t)\mathbf{B}(k_t,:)}{p_{k_t}} \right)$$

$$= \underbrace{\left(\frac{\mathbf{A}(:,k_1)}{\sqrt{t p_{k_1}}}, \frac{\mathbf{A}(:,k_2)}{\sqrt{t p_{k_2}}}, \cdots, \frac{\mathbf{A}(:,k_t)}{\sqrt{t p_{k_t}}} \right)}_{\mathbf{L} \in \mathbb{R}^{m \times t}, \text{ columns}} \cdot \underbrace{\left(\frac{\mathbf{B}(k_1,:)}{\sqrt{t p_{k_1}}}, \frac{\mathbf{B}(k_2,:)}{\sqrt{t p_{k_2}}}, \cdots, \frac{\mathbf{B}(k_t,:)}{\sqrt{t p_{k_t}}} \right)}_{\mathbf{R} \in \mathbb{R}^{t \times n}, \text{ rows}}$$

$$= \mathbf{LR}$$

$$\mathbf{AB} \approx \mathbf{LR}$$

Ex.: Show that $\mathrm{E}[\mathbf{L}\mathbf{L}^\top] = \mathbf{A}\mathbf{A}^\top$, $\mathrm{E}[\mathbf{R}^\top\mathbf{R}] = \mathbf{B}^\top\mathbf{B}$.

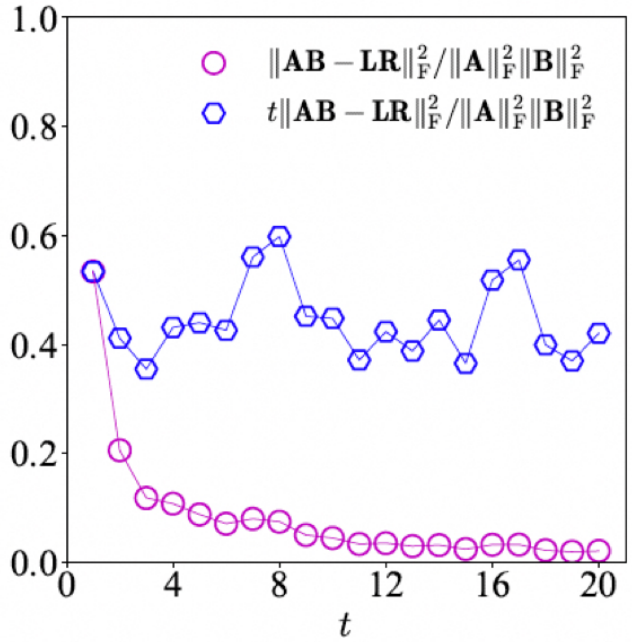
$$\mathrm{E}\left[\|\langle \mathbf{M} \rangle_t - \mathbf{AB}\|_{\mathrm{F}}^2 \right] \leq \frac{\|\mathbf{A}\|_{\mathrm{F}}^2 \|\mathbf{B}\|_{\mathrm{F}}^2}{t}$$

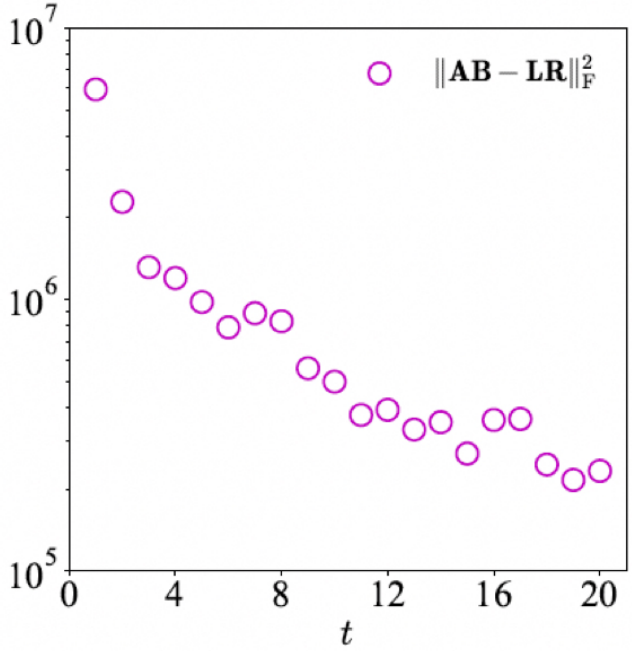
$$\mathrm{E}\left[\|\langle \mathbf{M} \rangle_t - \mathbf{AB}\|_{\mathrm{F}}^2 \right] \leq \epsilon^2 \|\mathbf{A}\|_{\mathrm{F}}^2 \|\mathbf{B}\|_{\mathrm{F}}^2$$

→if $\epsilon \sim \mathcal{O}(1)$, then $t \sim \mathcal{O}(1) \rightarrow \mathcal{O}(mn)$

better than $\mathcal{O}(mnd)$ ↑

Simulation example





Ex.: How smooth the pattern of the curve in the right panel?

$m = d = n = 10^2$, $a_{ij}, b_{ij} \sim \mathrm{Unif}[0, 1]$

roughly, $\dfrac{t\|\mathbf{LR}-\mathbf{AB}\|_{\mathrm{F}}^2}{\|\mathbf{A}\|_{\mathrm{F}}^2\|\mathbf{B}\|_{\mathrm{F}}^2} \lesssim 0.5$

Is the algorithm always useful?

$$\mathbf{B} = \mathbf{A}^\top,\ \mathbf{A} = \vec{\mathbf{l}}_m$$

$$\mathrm{E}\left[\|\mathbf{AB} - \mathbf{LR}\|_\mathrm{F}^2\right] \leq \frac{\|\mathbf{A}\|_\mathrm{F}^2 \|\mathbf{B}\|_\mathrm{F}^2}{t}$$

$$(\|\mathbf{A}\mathbf{A}^\top\|_\mathrm{F}^2 = m)$$

$$\leftrightarrow m \leq \frac{m^2}{t} \leftrightarrow t > m \text{ (no advantage)}$$

Spectrum of the matrices involved

if SV $\sigma_k \in \mathbf{A}$, then SV $\sigma_k^2 \in \mathbf{A}\mathbf{A}^\top$

therefore $\|\mathbf{A}\mathbf{A}^\top\|_F^2 = \sum_k \sigma_k^4,\ \|\mathbf{A}\|_F^2 = \sum_k \sigma_k^2$

so $\mathrm{E}\left[\|\mathbf{A}\mathbf{A}^\top - \mathbf{L}\mathbf{R}\|_F^2\right] \le \|\mathbf{A}\mathbf{A}^\top\|_F^2 \leftrightarrow t \ge \frac{(\sum_k \sigma_k^2)^2}{\sum_k \sigma_k^4}$

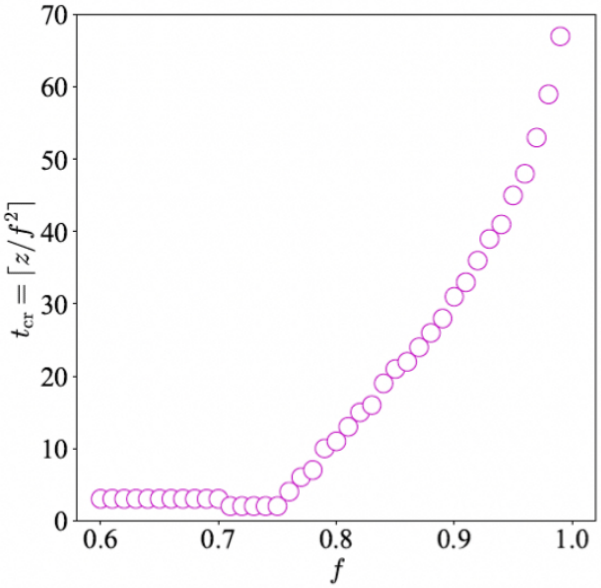
$\mathrm{rank}\mathbf{A} = r \rightarrow$ best upper bound of $\frac{(\sum_k \sigma_k^2)^2}{\sum_k \sigma_k^4} \approx r \rightarrow t \gtrsim r \rightarrow$ if $\mathbf{A}$ is full rank, no advantage

related case:

if $\sigma_1^2 + \cdots + \sigma_z^2 \ge f\left(\sigma_1^2 + \cdots + \sigma_r^2\right)$

then $\frac{(\sigma_1^2 + \cdots + \sigma_r^2)^2}{\sigma_1^4 + \cdots + \sigma_r^4} \le \frac{1}{f^2} \frac{(\sigma_1^2 + \cdots + \sigma_z^2)^2}{\sigma_1^4 + \cdots + \sigma_z^4} \le \frac{z}{f^2}$

$t_{cr} \approx 3 \sim 5$



Ex.: Try to approximate $\mathbf{A}\vec{1}$.

$$t \gtrsim t_{cr} \approx \lceil z/f^2 \rceil$$

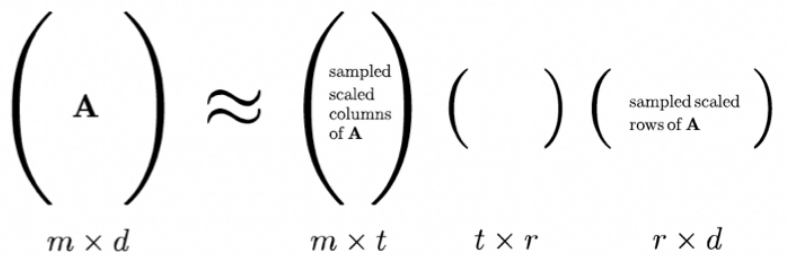
Better choice for the identity matrix

define $\mathbf{P} = \mathbf{R}^\top(\mathbf{R}\mathbf{R}^\top)^{-1}\mathbf{R}$ and $\mathbf{x} = \mathbf{R}^\top\mathbf{y}$ (row space of $\mathbf{R}$)

then $\mathbf{P}\mathbf{x} = \mathbf{R}^\top(\mathbf{R}\mathbf{R}^\top)^{-1}\mathbf{R}\mathbf{R}^\top\mathbf{y} = \mathbf{x} \rightarrow \mathbf{P}$ plays the role of $\vec{I}$

consider decompose $\mathbf{AP}$

* sample t columns of $\mathbf{A}$ to form $\mathbf{L} \in \mathrm{R}^{m\times t}$
* sample r rows of $\mathbf{A}$ to form $\mathbf{R}$
* construct $\mathbf{P} = \mathbf{R}^\top(\mathbf{R}\mathbf{R}^\top)^{-1}\mathbf{R}$
  (orthogonal projection onto the row space of $\mathbf{R}$)
  ($\mathbf{AP}$: projection of $\mathbf{A}$ onto the sampled row subspace)
* so $\mathbf{AP} = \mathbf{BC}$ with $\mathbf{B} = \mathbf{A}\mathbf{R}^\top$, $\mathbf{C} = (\mathbf{R}\mathbf{R}^\top)^{-1}\mathbf{R}$
  (each column of $\mathbf{B}$ is from sampled row of $\mathbf{A}$)
* construct $\mathbf{C}$, these rows form $\mathbf{U}$



LUR or CUR decomposition

*Algorithm analysis

$\|\mathbf{P}\|_F^2 \leq r$

$$\mathrm{E}\left[\|\mathbf{AP}-\mathbf{LUR}\|_2^2\right] \leq \mathrm{E}\left[\|\mathbf{AP}-\mathbf{LUR}\|_F^2\right] \leq \frac{\|\mathbf{A}\|_F^2\|\mathbf{P}\|_F^2}{t} \leq \frac{r}{t}\|\mathbf{A}\|_F^2$$

*bound of $\|\mathbf{A}-\mathbf{AP}\|_2^2$

$$\|\mathbf{A}-\mathbf{AP}\|_2^2 = \max_{\|\mathbf{x}\|=1} \|(\mathbf{A}-\mathbf{AP})\mathbf{x}\|^2$$

$\mathbf{x}$ is in row space of $\mathbf{R} \rightarrow \mathbf{Px} = \vec{0} \rightarrow (\mathbf{A}-\mathbf{AP})\mathbf{x} = \mathbf{Ax}$

$$\|(\mathbf{A}-\mathbf{AP})\mathbf{x}\|^2 = \|\mathbf{Ax}\|^2 = \mathbf{x}^\top\mathbf{A}^\top\mathbf{Ax} = \mathbf{x}^\top\left(\mathbf{A}^\top\mathbf{A}-\mathbf{R}^\top\mathbf{R}\right)\mathbf{x} \leq \left\|\mathbf{A}^\top\mathbf{A}-\mathbf{R}^\top\mathbf{R}\right\|_2 \overbrace{\|\mathbf{x}\|^2}^{1}$$

$$\rightarrow \|\mathbf{A}-\mathbf{AP}\|_2^2 \leq \|\mathbf{A}^\top\mathbf{A}-\mathbf{R}^\top\mathbf{R}\|_2 \rightarrow \|\mathbf{A}-\mathbf{AP}\|_2^2 \leq \frac{\|\mathbf{A}\|_F^2}{\sqrt{r}}$$

$\frac{1}{\sqrt{r}}$: statistical error

*Algorithm analysis: continues

$$\|\mathbf{A}-\mathbf{LUR}\|_2 \le \|\mathbf{A}-\mathbf{AP}\|_2 + \|\mathbf{AP}-\mathbf{LUR}\|_2$$

$$\rightarrow \|\mathbf{A}-\mathbf{LUR}\|_2^2 \le 2\|\mathbf{A}-\mathbf{AP}\|_2^2 + 2\|\mathbf{AP}-\mathbf{LUR}\|_2^2$$

$$\|\mathbf{A}-\mathbf{AP}\|_2^2 \le \frac{\|\mathbf{A}\|_F^2}{\sqrt{r}}$$

$$\mathrm{E}\left[\|\mathbf{AP}-\mathbf{LUR}\|_2^2\right] \le \frac{r}{t}\|\mathbf{A}\|_F^2$$

$$\boxed{\mathrm{E}\left[\|\mathbf{A}-\mathbf{LUR}\|_2^2\right] \le 2\|\mathbf{A}\|_F^2\left(\frac{1}{\sqrt{r}}+\frac{r}{t}\right)}$$

*if t is fixed $\rightarrow r = t^{2/3}$, choosing $t \sim \mathcal{O}(1/\epsilon^3)$ and $r \sim \mathcal{O}(1/\epsilon^2) \rightarrow$ bound $\sim \mathcal{O}(\epsilon)\|\mathbf{A}\|_F^2$

$\sigma_1^2 = 1$ (scaling)$\rightarrow \mathrm{E}[\|\mathbf{A}-\mathbf{LUR}\|_2^2] \le \epsilon \sum_i \sigma_i^2(\mathbf{A})$

if top k SVs of $\mathbf{A} \sim \mathcal{O}(1)$ for $k \gg m^{1/3} \rightarrow \sum_i \sigma_i^2(\mathbf{A}) \gg m^{1/3}$, then $\epsilon \sim \mathcal{O}(m^{-1/3}) \rightarrow t > m$

*if just the few SVs of $\mathbf{A}$ are large: quite good

Quiz 4: 5/27/2026

Quiz 4.1:
Explain the role(s) of ℓ-1 loss, ℓ-2 loss, ℓ-1 penalty and ℓ-2 penalty.
What is LASSO? What is Huber Loss? Explain/describe them.

Quiz 4.2:
What is difference between the *k*-median and *k*-center clusterings? Write down the definitions.

Quiz 4.3:
What is the motivation for doing spectral clustering? Give an example.

Quiz 4.4:
Show that two random vectors in *d*D are probably orthogonal.

Quiz 4.5:
Under which condition that a best-fit subspace for $\mathbf{A} \in \mathrm{R}^{d \times d}$ exists? Use SVD to write down it.

Birthday problem: randomized design/analysis

One asks how many people must there be in a room before there is a 50% chance that two of them were born on the same day of the year?

$B_K = \cap_{i=1}^{K} A_i,\ B_K = B_{K-1} \cap A_K$ $\quad A_i =$ person i's birthday is different from person j's for all $j < i$

$$\begin{aligned}
\text{Prob}(B_K) =& \text{Prob}(B_{K-1})\text{Prob}(A_K|B_{K-1}) \\
=& \text{Prob}(B_{K-2})\text{Prob}(A_{K-1}|B_{K-2})\text{Prob}(A_K|B_{K-1}) = \cdots \\
=& \text{Prob}(B_1)\text{Prob}(A_2|B_1)\text{Prob}(A_3|B_2)\cdots\text{Prob}(A_K|B_{K-1}) \\
=& 1 \cdot \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{n-K+1}{n} \\
=& 1 \cdot \left(1-\frac{1}{n}\right) \cdot \left(1-\frac{2}{n}\right) \cdots \left(1-\frac{K-1}{n}\right) \\
\le& \exp\left(-\frac{1}{n}\right)\exp\left(-\frac{2}{n}\right)\exp\left(-\frac{K-1}{n}\right) \\
=& \exp\left[-\left(\frac{1}{n}+\frac{2}{n}+\cdots+\frac{K-1}{n}\right)\right] = \exp\left(-\frac{K(K-1)}{2n}\right)
\end{aligned}$$

$$\text{Prob}(A_1) = \text{Prob}(B_1) = 1$$

$$\text{Prob}(b_i = d) = \frac{1}{n},\ \text{Prob}(b_i = d, b_j = d) = \frac{1}{n^2}$$

$$\text{Prob}(b_i = b_j) = \sum_{d=1}^{n} \text{Prob}(b_i = d, b_j = d) = \frac{1}{n}$$

$$\text{Prob}(B_K) \le 50\% \rightarrow K \ge \frac{1}{2}(1+\sqrt{1+8n\log 2}) \approx 23$$

$$K \sim \Theta(\sqrt{n})$$