האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

אלגוריתמים בביו' חישובית

76558

שערוך פרמטרים ושיטות היוריסטיות

תומי קפלן

14/1/2024

ציון העימוד ככלל החלטה אופטימלי

- שתי ההיפותיזות שלנו:

H_0: אין קשר בין שני הרצפים
H_1: לשני הרצפים יש אב קדמון משותף

- תחת הנחת אי-תלות בין העמדות:

$$P_1(S,T) = \prod_i P_1(S_i,T_i), \quad P_0(S) = \prod_i P_0(S_i)$$

- לוג יחס הנראות מתפרק לסכום ציוני העמודות:

$$LLR = \log \frac{P_1(S,T)}{P_0(S)\cdot P_0(T)} = \log \frac{\prod P_1(S_i,T_i)}{\prod P_0(S_i)\cdot \prod P_0(T_i)} = \sum_i \log \frac{P_1(S_i,T_i)}{P_0(S_i)\cdot P_0(T_i)} \triangleq \sum_i \sigma(S_i,T_i)$$

ביצד נשערך את P_1 ואת P_0?

- ובפרט, תחת אילוצי סימטריה: $P_1(a,b) = P_1(b,a)$
- ונוסחת ההסת׳ השלמה: $P_0(a) = \sum_b P_1(a,b)$

שערוך פרמטרים פשוט

- נניח מטבע עם התפלגות ברנולי: $\Omega = \{H, T\}$

$$P_\theta = \Omega \rightarrow [0,1]$$

- וסדרת תצפיות $D = \{a_1, a_2, a_3, \ldots, a_n\}$

- הגדרת הנראות:

$$L(\theta : D) = P_\theta(D) = \prod_i^n P_\theta(a_i)$$

שערוך פרמטרים פשוט

- ספציפית, אם $D = \{H, H, T, H\}$
- ונגדיר $P_\theta(H) = \theta, \quad P_\theta(T) = 1 - \theta$
- נקבל $L(\theta : D) = \theta^3(1 - \theta)$
- ובאופן כללי $L(\theta : D) = \theta^{N_H(D)}(1 - \theta)^{N_T(D)}$

$$N_H(D) = \sum_i 1\{a_i = H\}$$

$$N_T(D) = \sum_i 1\{a_i = T\}$$

סטטיסטיים מספיקים

- במקרה שלנו, N_H ו- N_T הם **נתונים מספיקים** [Sufficient Statistics] לחישוב הניראות $\vec{S}(D)$

- בלומר, בהינתן וקטור הססט' המספיקים הניראות כבר אינה תלויה בסדרה D:

$$\forall_{D_1}, \forall_{D_2}, \forall_{\theta}:$$

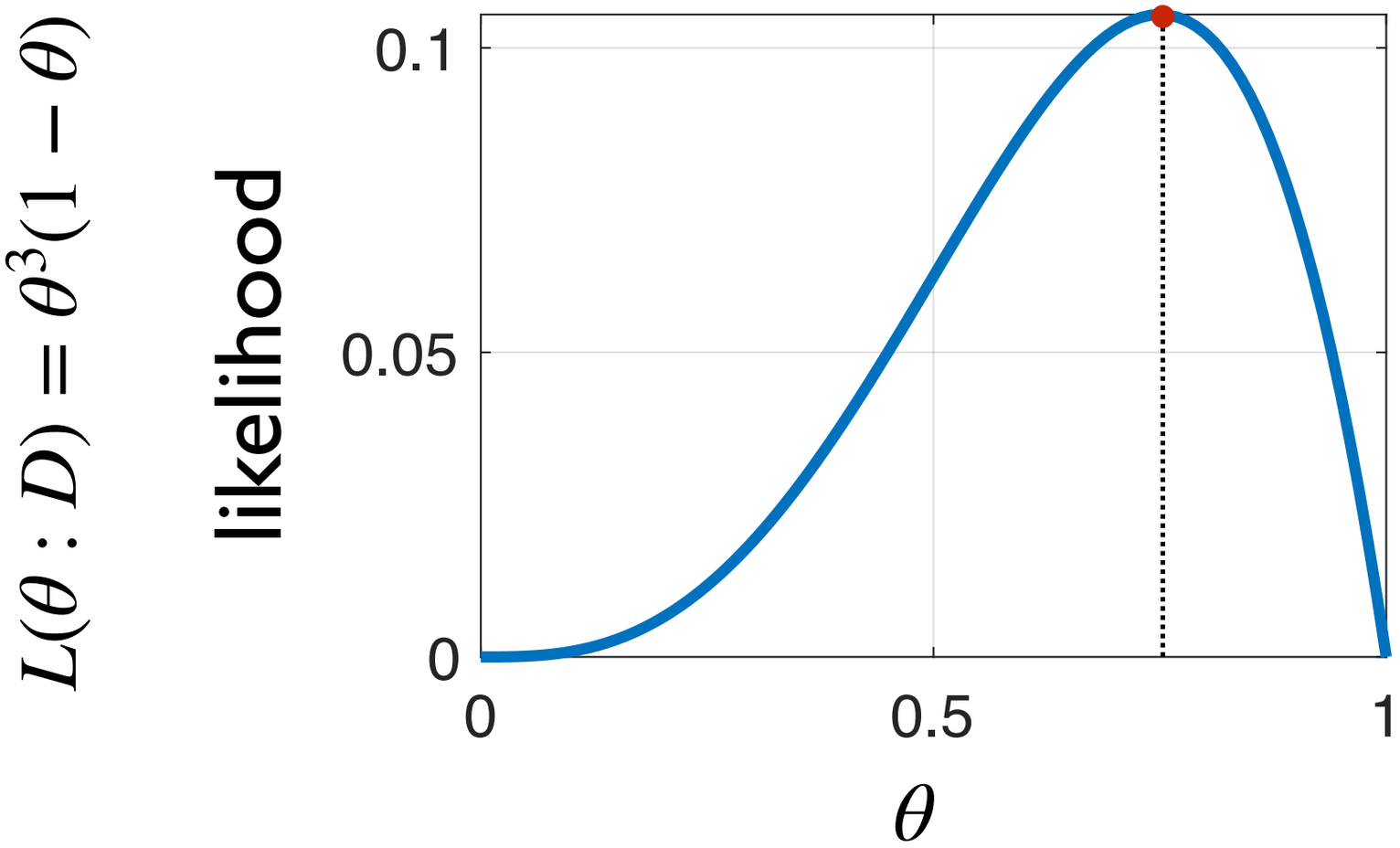
$$\vec{S}(D_1) = \vec{S}(D_2) \;\Rightarrow\; L(\theta : D_1) = L(\theta : D_2)$$

Maximum Likelihood Estimator

$$L(\theta : D) = \theta^3(1-\theta)$$

$$\hat{\theta}_{ML} = \arg\max_{\theta} L(\theta : D)$$

Maximum Likelihood Estimator



$$\hat{\theta}_{ML} = \arg\max_{\theta} L(\theta : D)$$

- איך נמצא? נגזור ונשווה לאפס. קללל.

מציאת ערך הניראות המקסימלי

$$L(\theta : D) = \theta^{N_H} \cdot (1-\theta)^{N_T}$$

- אהה! לוג היא פונ' מונוטונית.
- המקסימום של L מושג באותה הנקודה [=ארגמקס] בה מושג המקסימום של logL

$$\log L(\theta : D) = N_H \cdot \log\theta + N_T \cdot \log(1-\theta)$$

- נגזור ונקבל:

$$LL' = \frac{N_H}{\theta} + \frac{N_T}{1-\theta} \cdot (-1)$$

מציאת ערך הניראות המקסימלי

$$LL' = \frac{N_H}{\theta} - \frac{N_T}{1-\theta}$$

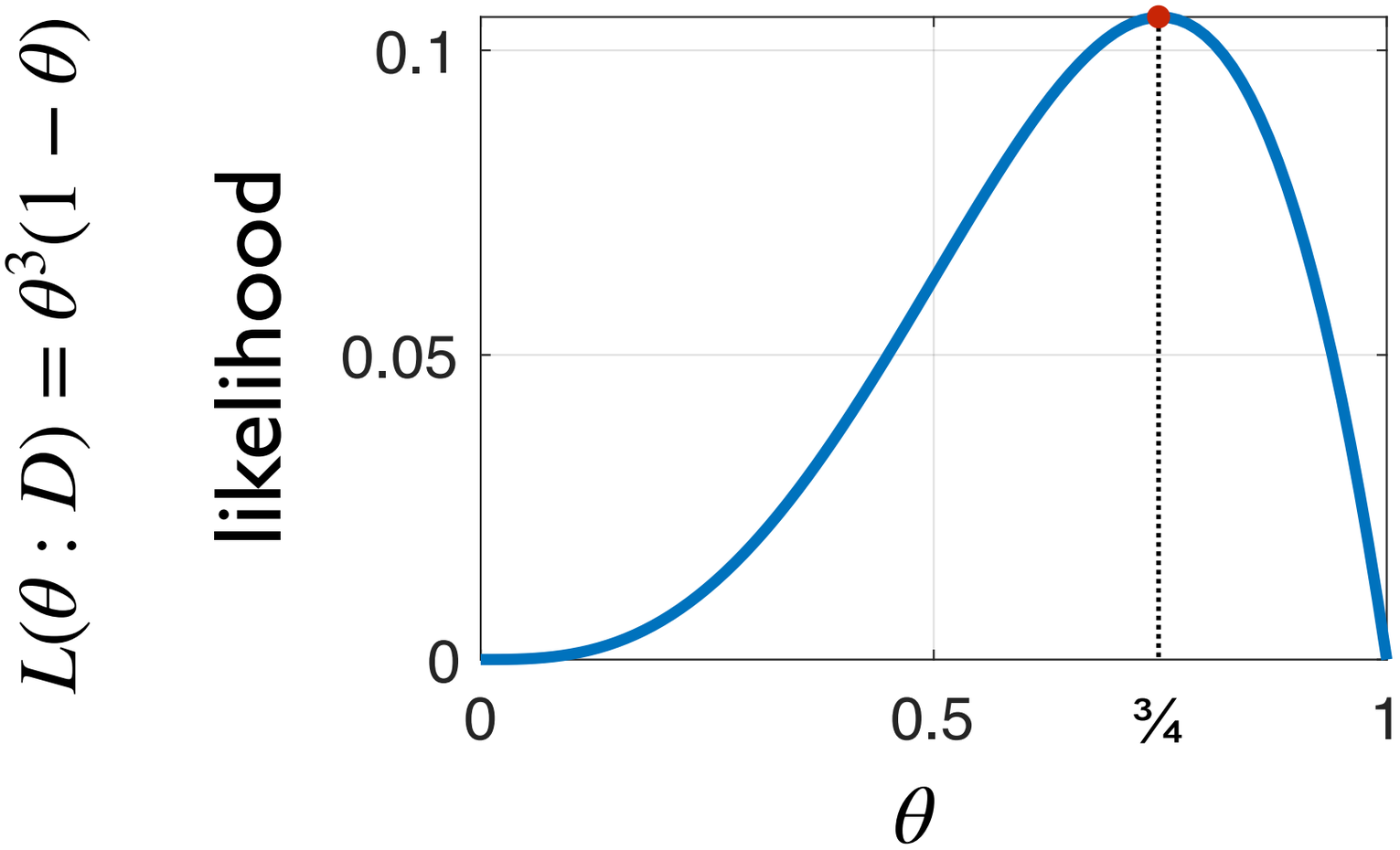
- נשווה לאפס

$$\frac{N_H}{\theta} = \frac{N_T}{1-\theta} \qquad (1-\theta)\cdot N_H = \theta \cdot N_T$$

$$N_H - \theta \cdot N_H = \theta \cdot N_T \qquad N_H = \theta \cdot (N_T + N_H)$$

- בלומר, אומד הניראות המקסימלי: $\hat{\theta} = \frac{N_H}{N_T + N_H}$

אומד נראות מקסימלי



- ועבור המטבע נקבל:

$$\hat{\theta}_{ML} = \frac{N_H}{N_H + N_T} = \frac{3}{4}$$

נכליל להתפלגות מולטינומיאלית

$$\vec{\theta} = < \theta_1, \theta_2, \ldots, \theta_k > \in \mathbb{R}^k \qquad \theta_i \geq 0, \quad \sum_i \theta_i = 1$$

- והניראות:

$$L(\theta : D) = \prod_i^n \theta_{a_i} = \prod_i^k \theta_i^{N_i}$$

- לוג הניראות:

$$\log L(\theta : D) = \sum_i^k N_i \log \theta_i$$

- מיהו אומד הניראות המירבית?

נקודת המקסימום בכפוף לאילוצים!

שיטת כופלי לגראנז׳

מציאת אופטימום בכפוף לאילוצים:

$$\arg\max_{\vec{X}\in\mathbb{R}^k} f(\vec{X}) \quad \text{s.t.} \begin{cases} c_1(\vec{X}) = 0 \\ \vdots \\ c_l(\vec{X}) = 0 \end{cases}$$

- נוסיף **λ** כופלי לגראנז׳:

$$J(\vec{X},\vec{\lambda}) = f(\vec{X}) - \sum_i c_i(X)\cdot\lambda_i$$

- כאשר X עומד בתנאים:

$$\forall_i\ c_i(X) = 0 \ \Rightarrow\ J(\vec{X},\vec{\lambda}) = f(\vec{X})$$

כופלי לגראנז׳ להתפלגות מולטינומית

$$\arg\max_{\vec{\theta}\in\mathbb{R}^k} f(\vec{\theta})$$

אילוצים $\sum_i \theta_i = 1$

- λ כופלי לגראנז׳:

$$J = \sum_i^k N_i \log\theta_i - \lambda\left(\sum_i \theta_i - 1\right)$$

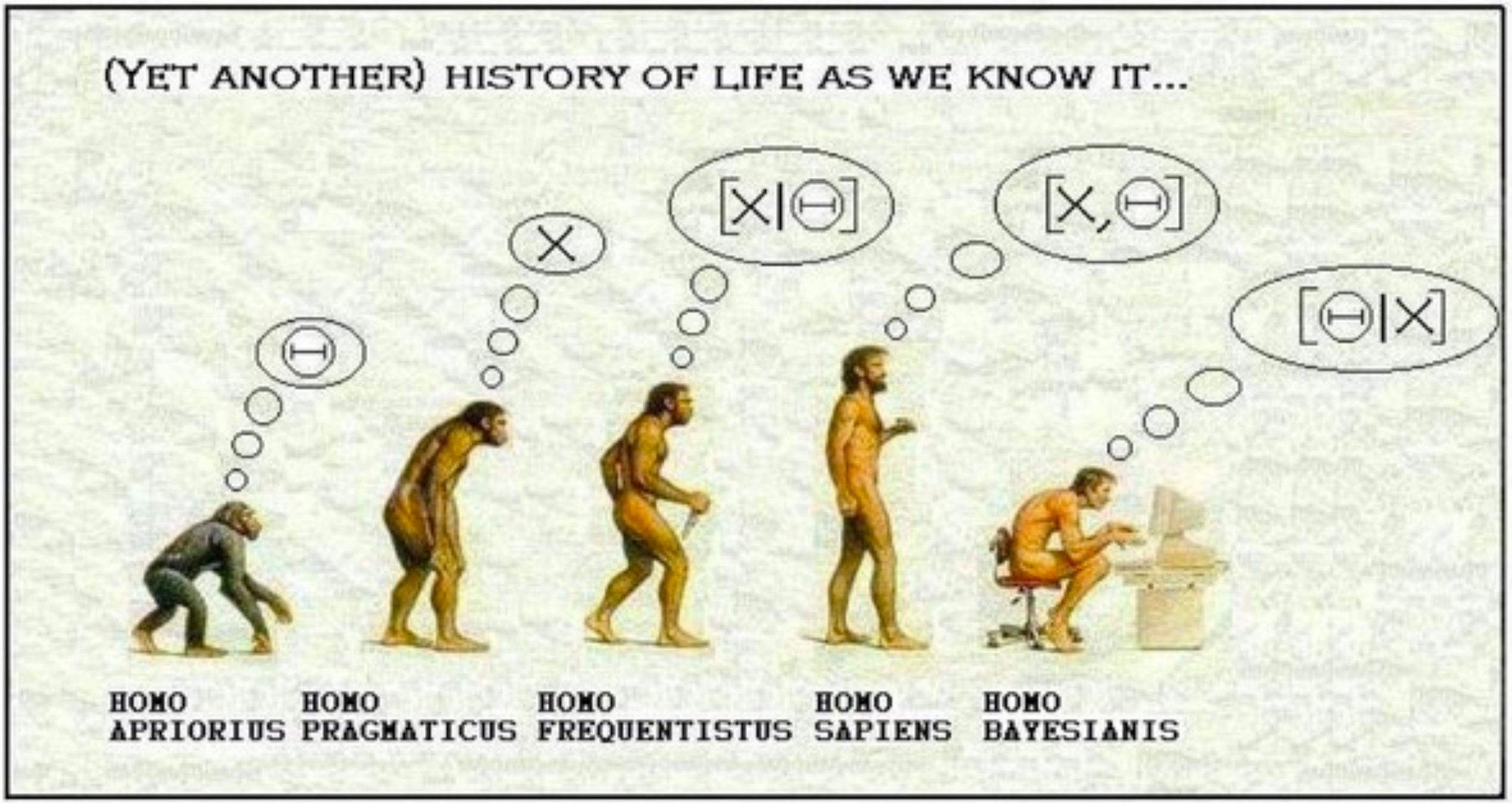
$$\frac{\partial J}{\partial \theta_i} = \frac{N_i}{\theta_i} - \lambda \qquad \frac{\partial J}{\partial \theta_i} = 0 \quad \Rightarrow \quad \hat{\theta}_i = \frac{N_i}{\lambda}$$

$$\frac{\partial J}{\partial \lambda} = -\sum_i \theta_i - 1 \qquad \frac{\partial J}{\partial \lambda} = 0 \quad \Rightarrow \quad \sum_i \hat{\theta}_i = 1$$

$$\hat{\theta}_i = \frac{N_i}{N}$$ אומד הניראות המירבית

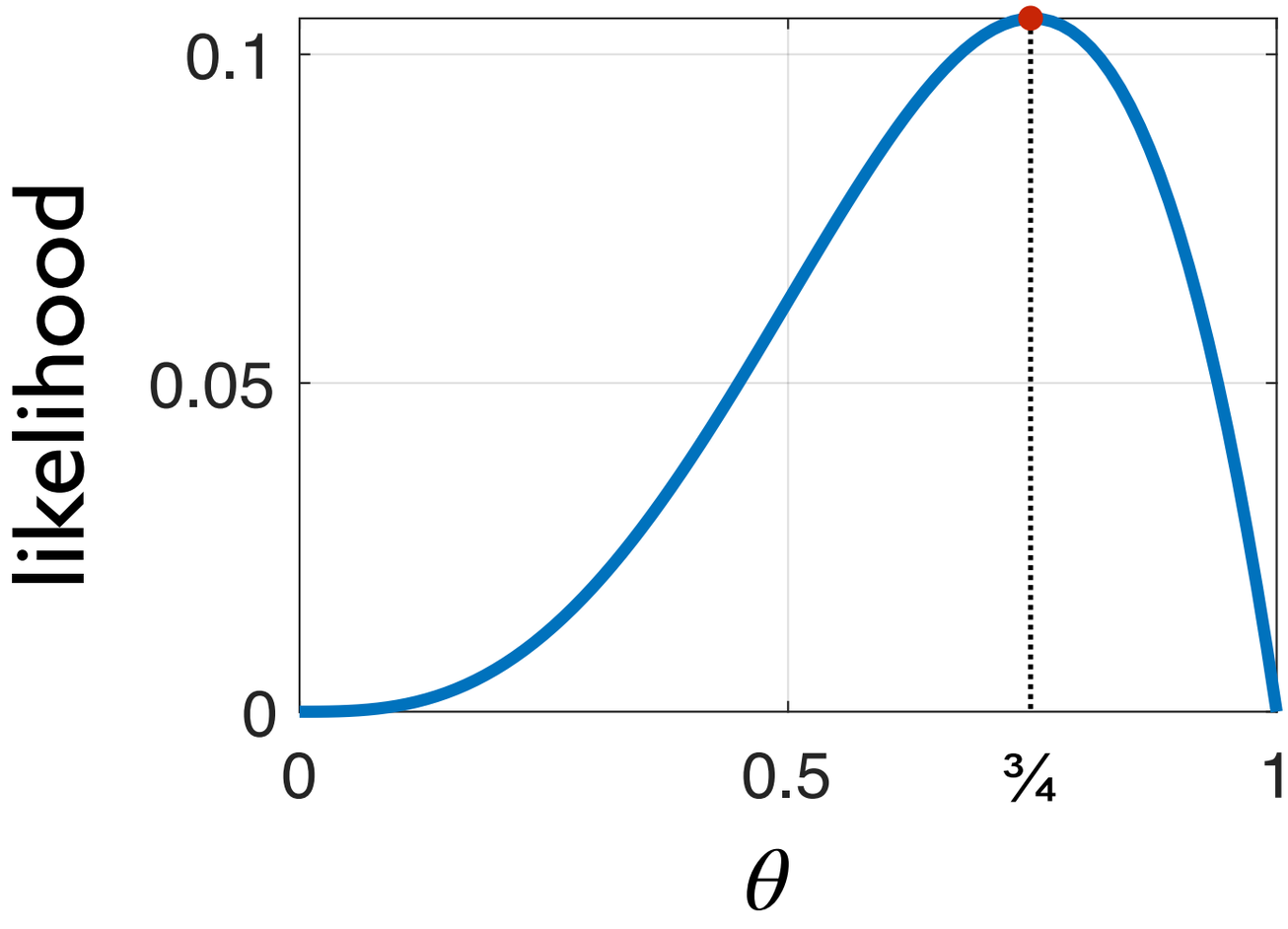
$$\sum_i \hat{\theta}_i = \frac{\sum_i N_i}{\lambda} = 1 \quad \Rightarrow \quad \lambda = \sum_i N_i = N$$

נושאים לדיון



שערוך בייזיאני

- במקום אומד MLE, נשערך אומד שמתחשב בכל ערכי התטא האפשריים [כל אחד לפי הפוסטריור שלו]



שערוך בייזיאני

- במקום אומד MLE, נשערך אומד שמתחשב בכל ערכי התטא האפשריים [כל אחד לפי הפוסטריור שלו]

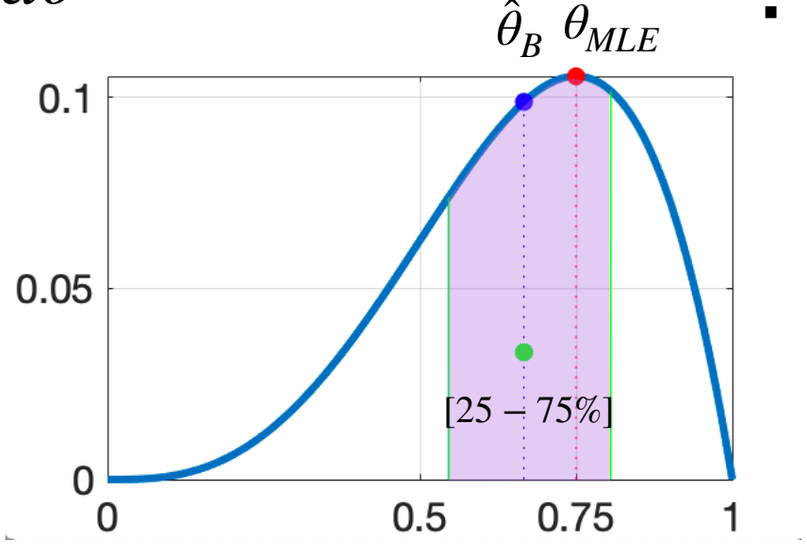
$$\hat{\theta} = \mathbb{E}(\theta \mid D) = \int \theta \cdot P(\theta \mid D)\ d\theta$$

- לפי חוק ביים:

$$P(\theta \mid D) = \frac{P(D \mid \theta) \cdot P(\theta)}{P(D)} \propto \underbrace{P(D \mid \theta)}_{\text{נראות}} \cdot \underbrace{P(\theta)}_{\text{פריור}}$$

- עם פריור יוניפורמי:

$$\hat{\theta} = \int \theta \cdot \frac{P(D \mid \theta)}{\cancel{P(D)}} \cancel{P(\theta)}\ d\theta$$



[אפשר להבין על זה סקרייב]

שערוך P_1 ו- P_0

- לפי למת ניימן–פירסון, הסקור האופטימלי לעימוד רצפים צריך להיות

$$\sigma(a,b) \triangleq \log \frac{P_1(a,b)}{P_0(a)\cdot P_0(b)} \overset{?}{(+c)}$$

- הסיכוי למצוא a בחלבון מקרי $P_0(a)$
- הסיכוי למצוא a מועמד מול b $P_1(a,b)$

[בעמדות מתאימות בחלבונים בעלי אב קדמון משותף]

- אומד MLE

$$\hat{P}_0(a) = \frac{N_a}{N}$$

עבור דאטה–ביים של חלבונים?

[אפשר להכין על זה סקרייב]

שערוך P_1 ו- P_0

- רגע רגע – אבל זה תלוי במרחק האבולוציוני ביניהם!

זהות מוחלטת — איזור הדמדומים

0% 15% 40% 62% 80% 100%

BLOSUM

σ אלכסונית — σ שואפת ל-P_0

Henikoff & Henikoff ['92]

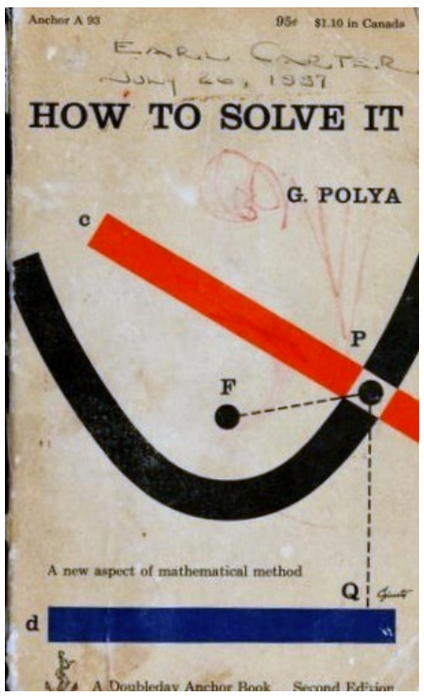
1. נעמד זוגות חלבונים [לפי מרחק עריכה פשוט]
2. נמצא איזורים בעלי % זהות עד 62%
3. נחשב עליהם סטטיסטיקות:

$$\text{BLOSUM}_{62} \triangleq \left\lfloor \log \frac{P_{62}(a,b)}{P_{62}(a) \cdot P_{62}(b)} \right\rfloor$$

4. והרי לנו σ לעימוד חלבונים

[אפשר להבין על זה סקרייב]

היוריסטיקות



- **פוליה 45'** – "איך לפתור את זה?"
- גישה מעשית ויעילה לפתרון בעיות, **ללא** הבטחה למציאת הפתרון האופטימלי
- ארבעה שלבים:
  1. הבנת הבעיה
  4. ניתוח פתרון והכללה
  2. גיבוש תכנית פעולה
  3. פתרו!

[אפשר להכין על זה סקרייב]

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

אלגוריתמים בביו' חישובית

76558

שערוך פרמטרים ושיטות היוריסטיות

תומי קפלן

16/1/2024

FASTA [Pearson & Lipman, 88']

- היוריסטיקה למציאת עימוד לוקאלי בין שני רצפים.
- מבוססת על מציאת תתי-מילים [seeds] זהות.

BLAST [Altschul et al, 90']

- היוריסטיקה דומה:
- תתי-רצפים ארוכים יותר
- לא נדרשת התאמה מלאה

[אפשר להכין על זה סקרייב]

FASTA

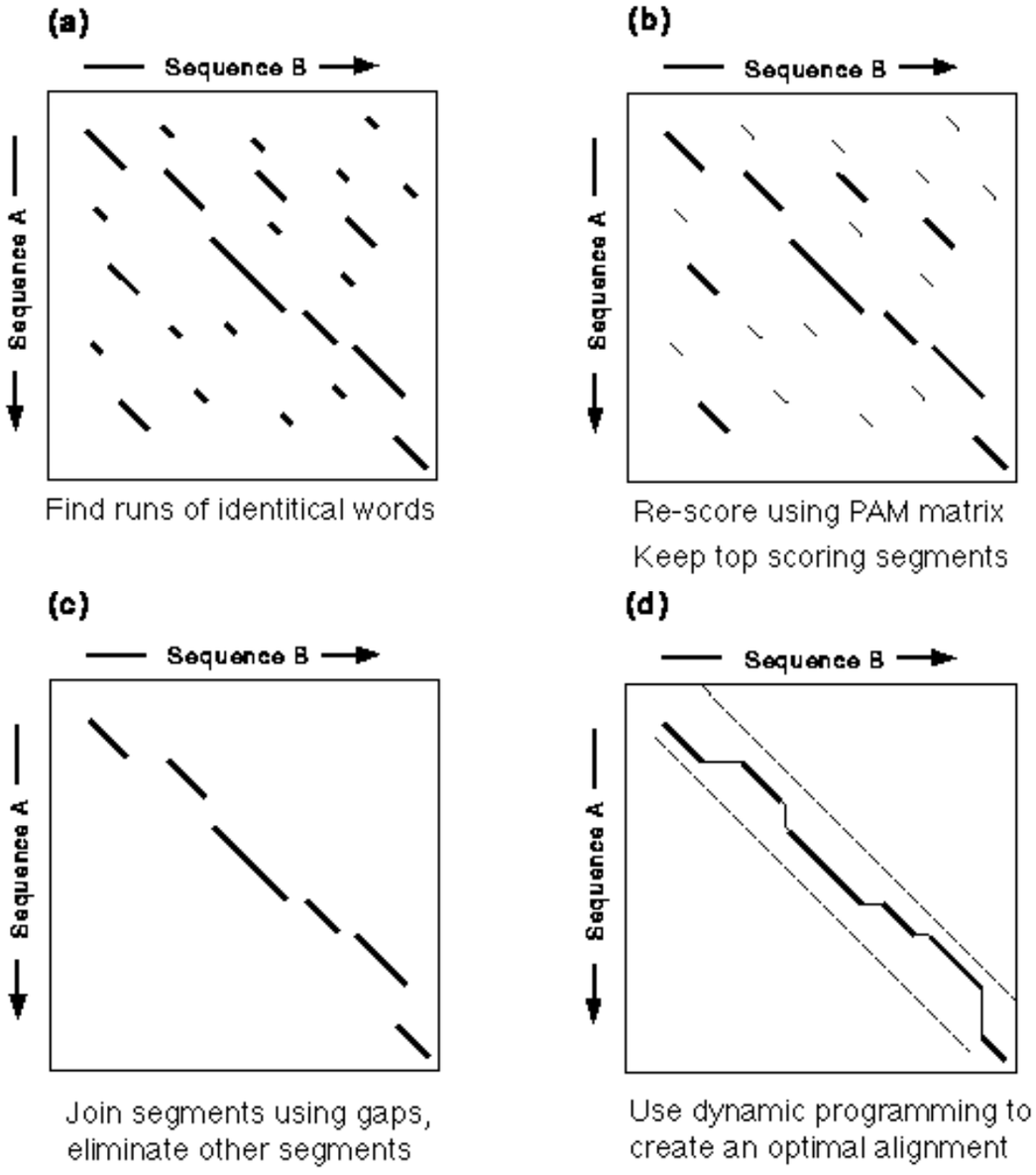
- Heuristic: local alignments of query vs. database
- "Good alignments have exact small matches"

[a] Find promising matches
[focus on top diagonals]

[b] Re-score [using PAM]

[c] Join gapped segments

[d] Finalize [local alignment]



BLAST

- Heuristic: local alignments of query vs. database
- "Good alignments have multiple small hits"

[a] Break query into words

[b] Filter low-scoring words

[c] Store in hash table

[d] Run other sequences through hash table
[allowing for mismatches]

[e] Find high-scoring pairs [HSPs]

[f] Extend