תרגיל 4 | CBIO (76558)

שם: אלון מרקוביץ' וגיא טרוזמן | ת`ז: 313454902 ו־312517303

שאלה 1.1

סטטיסטי מספקים

הסטטיסטים המספקים הם $\#(a_i = b_i)$ ו־$\#(a_i \neq b_i)$. נגדיר $A = a_1, \dots, a_n\ ,\ B = b_1, \dots, b_n$. אם כן =

$$L\left(A \overbrace{\rightarrow}^{t} B\right) = \prod_{i=1}^{n} P\left(a_i \overbrace{\rightarrow}^{t} b_i\right) = P\left(a_i \overbrace{\rightarrow}^{t} b_i\right)^{\#(a_i=b_i)} \cdot P\left(a_i \overbrace{\rightarrow}^{t} b_i\right)^{\#(a_i \neq b_i)} =$$

$$= \left(\frac{1}{4} \cdot (1 + 3e^{-4\alpha t})\right)^{\#(a_i=b_i)} \cdot \left(\frac{1}{4} \cdot (1 - e^{-4\alpha t})\right)^{\#(a_i \neq b_i)}$$

$$\arg\max_t \left(\pi_A \cdot L\left(A \overbrace{\rightarrow}^{t} B\right)\right) \underset{\star}{=} \log(\pi_A) + \#(a_i = b_i) \cdot \log\left(\frac{1}{4} \cdot (1 + 3e^{-4\alpha t})\right) + \#(a_i \neq b_i) \cdot \log\left(\frac{1}{4} \cdot (1 - e^{-4\alpha t})\right)$$

$\star$ ־ $\log$ היא פונקציה מונוטונית עולה.

כדי למצוא את t אשר ממקסם את הביטוי, נגזור לפי t נשווה ל־0 ונקבל :

$$\#(a_i = b_i) \cdot \log\left(\frac{1}{4} \cdot (1 + 3e^{-4\alpha t})\right) + \#(a_i \neq b_i) \cdot \log\left(\frac{1}{4} \cdot (1 - e^{-4\alpha t})\right)' = \tag{'}$$

$$= \#(a_i = b_i) \cdot \frac{-3\alpha e^{-4\alpha t}}{\frac{1}{4} \cdot (1 + 3e^{-4\alpha t})} + \#(a_i \neq b_i) \cdot \frac{\alpha e^{-4\alpha t}}{\frac{1}{4} \cdot (1 - e^{-4\alpha t})} = 0$$

$$\Downarrow / \cdot \frac{4}{\alpha e^{-4\alpha t}}$$

$$\#(a_i = b_i) \cdot \frac{-3}{1 + 3e^{-4\alpha t}} + \#(a_i \neq b_i) \cdot \frac{1}{1 - e^{-4\alpha t}} = 0$$

$$\Downarrow$$

$$\#(a_i \neq b_i) \cdot (1 + 3e^{-4\alpha t}) = \#(a_i = b_i) \cdot 3 \cdot (1 - e^{-4\alpha t})$$

$$\Downarrow$$

$$e^{-4\alpha t} \cdot 3 \cdot (\#(a_i = b_i) + \#(a_i \neq b_i)) = 3 \cdot \#(a_i = b_i) - \#(a_i \neq b_i)$$

$$\Downarrow$$

$$t = \frac{\ln\left(\frac{3\cdot\#(a_i=b_i)-\#(a_i\neq b_i)}{3\cdot(\#(a_i=b_i)+\#(a_i\neq b_i))}\right)}{-4\alpha}$$

■

שאלה 1.2 ־ בניית דוגם לענף

(a) בהינתן מרחק t ואות a, בנו הליך שיגדום את b מ־ $P_{JC}\left(a \overset{\frown}{\rightarrow} b\right)$.

נדגום את b להיות a בהסתברות $\frac{1}{4}\cdot(1+3e^{-4\alpha t})$ ולהיות אות שונה מ־a בהסתברות $\frac{1}{4}\cdot(1-e^{-4\alpha t})$.

(b) השתמשו בהליך שלעיל כדי לחולל N דגימות של b והשוו בין השכיחות הנראית לבין זו הצפויה.

$$a \neq b$$

actual, a!=b:

	0.15	0.40	0.90
10	0.2	0.1	0.4
100	0.11	0.13	0.2
1000	0.094	0.187	0.233
100000	0.11251	0.19903	0.24301

prediction, a!=b:

	0.15	0.40	0.90
10	0.112797	0.199526	0.243169
100	0.112797	0.199526	0.243169
1000	0.112797	0.199526	0.243169
100000	0.112797	0.199526	0.243169

$$a = b$$

actual, a=b:

	0.15	0.40	0.90
10	0.8	0.6	0.1
100	0.69	0.37	0.29
1000	0.683	0.411	0.279
100000	0.6604	0.40232	0.26887

prediction, a=b:

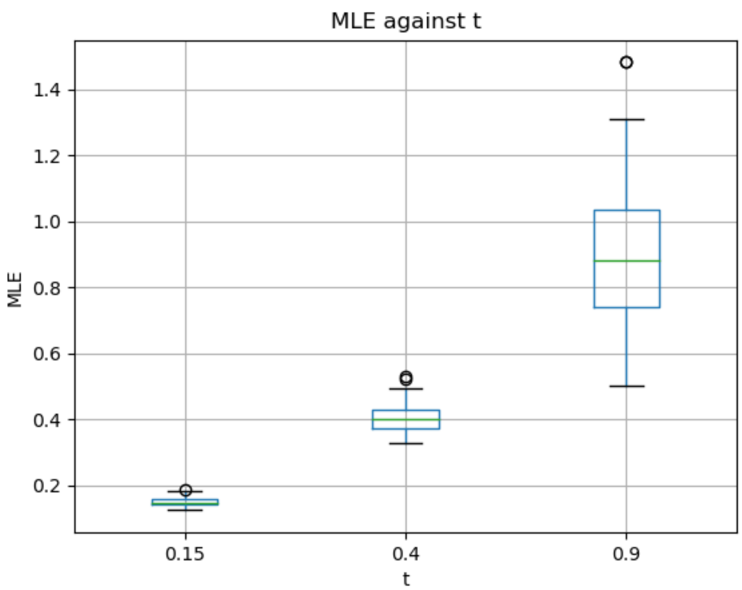
	0.15	0.40	0.90
10	0.661609	0.401422	0.270493
100	0.661609	0.401422	0.270493
1000	0.661609	0.401422	0.270493
100000	0.661609	0.401422	0.270493

(c) דונו בתוצאותיכם.

נצפה לראות שככל ש־t גדול יותר אז הסיכוי ש־$a=b$ אמור להיות קטן יותר (ובהתאמה הסיכוי של $a\neq b$ גדול יותר). כמו כן, עבור t גדול נצפה שנתקרב להתפלגות אחידה (קרי 0.25). התוצאות מתאימות לציפיות ולרוב מתקרבות לערך האמת ככל שהמדגם גדול יותר. אם כן, כדי להיות קרובים לתוצאות הנכונות נצטרך לדגום מדגם גדול יחסית (תלוי ברמת הדיוק שנרצה).

שאלה 1.3

(b) השתמשו בהליך מסעיף (a) כדי לדגום זוג רצפים באורך N במרחק t אחד מהשני. בדקו את היחס בין t ה"אמיתי" לבין ה-MLE.



(c) מה מסקנותיכם על אומדן אורך הענף? איך זה ישפיע על שיטות מבוססות מרחק לשחזור עץ?

תחילה, נציין שעבור $t = 0.9$ קיבלנו ערכים בעייתיים (מונה שלילי, לדוגמא) ולכן נרמלנו את הנקודות (הורדנו נקודות בעייתיות). מהגרף ניתן להסיק שככל ש-t גדול יותר כך ערכי הקיצון (ה-$outliers$) יהיו רחוקים יותר מערך ה-t האמיתי (הדגימות מפוזרות ואינן מרוכזות סביב החציון וכן החציון אינו מדויק עבור $t = 0.9$). אם כן, נסיק שעבור t גדול נקבל עץ פחות מדויק. הסבר ביולוגי לכך הוא ש-t גדול מעיד על מרחק אבולוציוני גדול, כלומר עבר הרבה זמן ולכן סביר שמדובר בעץ עם פיצולים רבים (ההסתברות לפיצול גדלה ככל שעובר הזמן, כמו שראינו לעיל). אם כן, העץ המקורי הוא עץ סבוך ומורכב ולכן הסיכוי לטעות במהלך השחזור שלו גדל עם t.

```
Median of t=0.15 is 0.15

Median of t=0.4 is 0.4

Median of t=0.9 is 0.88
```

שאלה 2

(1)

$$R = \begin{pmatrix} -4 & 2 & 1 & 1 \\ 1 & -4 & 2 & 1 \\ 1 & 1 & -4 & 2 \\ 2 & 1 & 1 & -4 \end{pmatrix}$$

מכיוון שנרצה התפלגות סטציונרית אחידה נקבל ש־$1 = \frac{\pi_b}{\pi_a}$. הוכחנו בשיעור שעל מנת שמטריצה אינה תהיה רווסיבילית צריך להתקיים

$$\exists a,b,t \ s.t \ \pi_a \cdot P(a \overbrace{\to}^{t} b) \neq \pi_b \cdot P(b \overbrace{\to}^{t} a)$$

השקול לכך שקיימים a,b כך ש־$\frac{R_{a,b}}{R_{b,a}} \neq \frac{\pi_b}{\pi_a}$ אך מכיוון ש־$1 = \frac{\pi_b}{\pi_a}$ נצטרך שיהיו קיימים a,b כך ש־$1 \neq \frac{R_{a,b}}{R_{b,a}}$ כמופיע בטבלה.
ניתן לראות שעבור $t=4$ המטריצה מתכנסת להתפלגות סטציונרית אחידה.

```
t=1
[[0.25243995 0.25221521 0.24879942 0.24654542]
 [0.24654542 0.25243995 0.25221521 0.24879942]
 [0.24879942 0.24654542 0.25243995 0.25221521]
 [0.25221521 0.24879942 0.24654542 0.25243995]]
t=2
[[0.24999209 0.2500191  0.25001098 0.24997782]
 [0.24997782 0.24999209 0.2500191  0.25001098]
 [0.25001098 0.24997782 0.24999209 0.2500191 ]
 [0.2500191  0.25001098 0.24997782 0.24999209]]
t=3
[[0.24999985 0.25000002 0.25000016 0.24999997]
 [0.24999997 0.24999985 0.25000002 0.25000016]
 [0.25000016 0.24999997 0.24999985 0.25000002]
 [0.25000002 0.25000016 0.24999997 0.24999985]]
t=4
[[0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]
 [0.25 0.25 0.25 0.25]]
```

(2)

$$R = \begin{pmatrix} -3 & 1 & 1 & 1 \\ 2 & -4 & 1 & 1 \\ 2 & 1 & -4 & 1 \\ 2 & 1 & 1 & -4 \end{pmatrix}$$

תחילה, מהתמונה לקמן ניתן לראות שהמטריצה מתכנסת להתפלגות סטוציונרית שאינה אחידה.

```
t=1
[[0.40404277 0.19865241 0.19865241 0.19865241]
 [0.39730482 0.20539036 0.19865241 0.19865241]
 [0.39730482 0.19865241 0.20539036 0.19865241]
 [0.39730482 0.19865241 0.19865241 0.20539036]]
t=2
[[0.40002724 0.19999092 0.19999092 0.19999092]
 [0.39998184 0.20003632 0.19999092 0.19999092]
 [0.39998184 0.19999092 0.20003632 0.19999092]
 [0.39998184 0.19999092 0.19999092 0.20003632]]
t=3
[[0.40000018 0.19999994 0.19999994 0.19999994]
 [0.39999988 0.20000024 0.19999994 0.19999994]
 [0.39999988 0.19999994 0.20000024 0.19999994]
 [0.39999988 0.19999994 0.19999994 0.20000024]]
t=4
[[0.4 0.2 0.2 0.2]
 [0.4 0.2 0.2 0.2]
 [0.4 0.2 0.2 0.2]
 [0.4 0.2 0.2 0.2]]
```

נשים לב, שעבור $a=A$, לכל $b\neq A$ מתקיים ש-$\frac{\pi_b}{\pi_a}=\frac{1}{2}$ וכן $\frac{R_{a,b}}{R_{b,a}}=\frac{1}{2}$. כמו כן, עבור $b=A$, לכל $a\neq A$ מתקיים ש-$\frac{\pi_b}{\pi_a}=2$ וכן $\frac{R_{a,b}}{R_{b,a}}=2$. לבסוף, לכל $a,b\neq A$ מתקיים $\frac{\pi_b}{\pi_a}=1$ וכן $\frac{R_{a,b}}{R_{b,a}}=1$. אם כן, לכל a,b מתקיים $\frac{\pi_b}{\pi_a}=\frac{R_{a,b}}{R_{b,a}}$ וממה שראינו בהרצאה זה שקול לכך ש-R היא מטריצה רוורסיבלית.

שאלה 3

(1) הראו ש- $P\left(X_1,\ldots,X_{2n-1}\right)=\left[\prod_i \pi_{X_i}\right]\cdot\prod_{(i\to j)\in T}\frac{\left[e^{t_{ij}R}\right]_{X_i,X_j}}{\pi_{X_j}}$.

$$P\left(X_1,\ldots,X_{2n-1}\right)=P\left(X_{2n-1}\right)\cdot\prod_{(i\to j)\in T}P\left(X_i\overset{\frown}{\rightarrow}X_j\right)\underset{(*)}{=}\pi_{X_{2n-1}}\cdot\prod_{(i\to j)\in T}\left[e^{t_{ij}R}\right]_{X_i,X_j}=$$

$$=\pi_{X_{2n-1}}\cdot\prod_{(i\to j)\in T}\left[e^{t_{ij}R}\right]_{X_i,X_j}\cdot\prod_{k=1}^{2n-2}\frac{\pi_{X_k}}{\pi_{X_k}}=\pi_{X_{2n-1}}\cdot\prod_{(i\to j)\in T}\left[e^{t_{ij}R}\right]_{X_i,X_j}\cdot\prod_{k=1}^{2n-2}\pi_{X_k}\cdot\prod_{k'=1}^{2n-2}\frac{1}{\pi_{X_{k'}}}=$$

$$\underset{(**)}{=}\prod_{(i\to j)\in T}\left[e^{t_{ij}R}\right]_{X_i,X_j}\cdot\prod_{k=1}^{2n-1}\pi_{X_k}\cdot\prod_{k'=1}^{2n-2}\frac{1}{\pi_{X_{k'}}}\underset{(***)}{=}\prod_{k=1}^{2n-1}\pi_{X_k}\cdot\prod_{(i\to j)\in T}\frac{\left[e^{t_{ij}R}\right]_{X_i,X_j}}{\pi_{X_j}}$$

כאשר המעבר $(*)$ הוא הצבת הנתונים, המעבר $(**)$ הוא תוצאה של $\pi_{X_{2n-1}}\cdot\prod_{k=1}^{2n-2}\pi_{X_k}$ והמעבר $(***)$ נובע מכך שיש מעבר אל על אחד מהקודקודים למעט השורש.

אם כן, הוכחנו ש-$P\left(X_1,\ldots,X_{2n-1}\right)=\left[\prod_i \pi_{X_i}\right]\cdot\prod_{(i\to j)\in T}\frac{\left[e^{t_{ij}R}\right]_{X_i,X_j}}{\pi_{X_j}}$, כנדרש. ■

(2) הראו שאם התהליך הוא רוורסיבלי, שינוי השורש לא ישנה את ההתפלגות המשותפת.

כפי שאמרנו, השורש הוא X_{2n-1}. אם כן, שינוי השורש שקול לשינוי כיוון הצלעות (חלק או כל). ניזכר שאם התהליך רוורסיבלי, אזי מתקיים

$$\forall i,j,t\ \pi_{X_i} P\left(X_i \overbrace{\rightarrow}^{t} X_j\right) = \pi_{X_j} P\left(X_j \overbrace{\rightarrow}^{t} X_i\right) \Rightarrow P\left(X_i \overbrace{\rightarrow}^{t} X_j\right) = \frac{\pi_{X_j}}{\pi_{X_i}} P\left(X_j \overbrace{\rightarrow}^{t} X_i\right)$$

אם כן,

$$P\left(X_1,\ldots,X_{2n-1}\right) = \left[\prod_i \pi_{X_i}\right]\cdot \prod_{(i\to j)\in T} \frac{\left[e^{t_{ij}R}\right]_{X_i,X_j}}{\pi_{X_j}} \underset{(*)}{=} \left[\prod_i \pi_{X_i}\right]\cdot \prod_{(i\to j)\in T} \frac{\frac{\pi_{X_j}}{\pi_{X_i}} P\left(X_j \overbrace{\rightarrow}^{t} X_i\right)}{\pi_{X_j}} =$$

$$= \left[\prod_i \pi_{X_i}\right]\cdot \prod_{(i\to j)\in T} \frac{P\left(X_j \overbrace{\rightarrow}^{t} X_i\right)}{\pi_{X_i}} \underset{(*)}{=} \left[\prod_i \pi_{X_i}\right]\cdot \prod_{(i\to j)\in T} \frac{\left[e^{t_{ji}R}\right]_{X_j,X_i}}{\pi_{X_i}}$$

כאשר המעבר $(*)$ מתבסס על כך שבשל הרוורסיבליות, לכל i,j מתקיים $\left[e^{t_{ij}R}\right]_{X_i,X_j} = P\left(X_i \overbrace{\rightarrow}^{t} X_j\right) = \frac{\pi_{X_j}}{\pi_{X_i}} P\left(X_j \overbrace{\rightarrow}^{t} X_i\right)$.

אם כן, הוכחנו שהפיכה של כל כיווני הצלעות שומרת על ההתפלגות המשותפת ולכן בפרט שינוי כיווני חלק מהצלעות גם הוא ישמור על ההתפלגות המשותפת (כלומר נקבל שתי מכפלות, אחת לצלעות שלא השתנו והשניה לצלעות שהשתנו, כאשר עבור אלו שהשתנו ראינו שההפיכה נשארת זהה). אם כן, קיבלנו ששינוי השורש לא ישנה את ההתפלגות המשותפת, כנדרש. ■