ביג דאטה – תרגיל 1

אסף עזרא

1)

a. נחשב את $\hat{n}_{MLE}(x)$ לכל $x = 1, \dots, 10$:

$$\hat{n}_{MLE}(x = i) = \underset{n}{argmax} F(x, n) = \underset{n}{argmax}(\Pr[x = i|n])$$

כאשר:

$$\Pr[x = i|n] = 2^{-(i-1)} \Pr[x = i - 1|n - 1] + \left(1 - 2^{-i}\right) \Pr[x = i|n - 1]$$

זהו למעשה תנאי רקורסיבי אשר מקיים עבור $x = 1$:

$\Pr[x = 1|n] = 0.5^{n-1}$ – ההסתברות להגדיל את Morris counter בהופעה הראשונה היא 1 והיא כופלת את ההסתברות שלא להגדילו שוב באף צעד נוסף. זוהי כמובן פונקציה מונוטונית יורדת ממש עבור $-\infty < n < \infty$ ולכן $\hat{n}_{MLE}(x = 1) = 1$.

עבור $x = 2$ אם נסתכל על $a_i = \Pr[x = 2|i]$ נקבל סדרה הנדסית עם מקדם $\frac{0.75}{0.5}$ ו-$n - 2$ איברים:

$$\Pr[x = 2|n] = 0.5 \cdot 0.5^{n-2} + 0.75 \cdot \Pr[x = 2|n - 1] \Rightarrow a_n = 0.5^{n-1} + 0.75 \cdot a_{n-1}$$

$$a_n = 0.5^{n-1} + 0.75 \cdot 0.5^{n-2} + 0.75^2 \cdot 0.5^{n-3} + \cdots + 0.75^{n-2}0.5$$

$$= (0.5)^{n-1} \cdot \frac{\left(\left(\frac{0.75}{0.5}\right)^{n-1} - 1\right)}{\frac{0.75}{0.5} - 1} = 2(0.75^{n-1} - 0.5^{n-1})$$

פונקציה זו מקבלת מקסימום מקומי על פני החיוביים (ב-$-\infty$ היא שואפת ל-$-\infty$) ב:

$$0.75^{n-1} log 0.75 = 0.5^{n-1} log 0.5 \Rightarrow (n - 1)(log 0.75 - log 0.5) = \log\left(\frac{log 0.5}{log 0.75}\right) \Rightarrow$$

$$n \approx 3.16$$

ומחישוב פשוט נקבל ש:

$$\Pr[x = 2|n = 4] = 0.59375 < 0.625 = \Pr[x = 2|n = 3] \Rightarrow \hat{n}_{MLE}(x = 2) = 3$$

כעת באמצעות הרקורסיה נחשב עבור כל x את ה-$\hat{n}_{MLE}$ שלו (בעזרת מטלב):

חישבתי לכל x את ההסתברויות עבור $n = 1, \dots, 10000$. זה כמובן צעד מיותר כיוון שניתן היה לחשב את $\Pr[x = x'|n]$, רק עבור $n < \hat{n}_{MLE} + 2$, כיוון ש-$F(x, n)$ היא פונקציה יורדת ב-n עבור $n \geq \hat{n}_{MLE}$, ולכן (ומעצם ההגדרה מקבלת מקסימום מקומי ב-$\hat{n}_{MLE}$), ברגע ש-$F(x, n)$ תפסיק לעלות, סיימנו.

נוכיח כי $\forall x: \forall n \geq \hat{n}_{MLE} \in \mathbb{N}: F(x, n) \geq F(x, n + 1)$, באינדוקציה על x ובאינדוקציה על n:

עבור $x = 1$, הטענה נכונה כיוון ש- $F(1, n) = 0.5^{n-1}$, והיא כמובן יורדת ממש ב-n. כעת נניח נכונות הטענה עבור $x < x'$ ונוכיח עבור $x = x'$:

יהי $\hat{n}'_{MLE}$ המתאים ל-x', אזי נוכיח באינדוקציה כי

$\forall n \geq \hat{n}'_{MLE} \in \mathbb{N}: F(x', n) \geq F(x', n + 1)$:

השלב הראשון הוא כמובן טריוויאלי כיוון שמעצם ההגדרה $\forall n \in \mathbb{N}: F(x', \hat{n}'_{MLE}) \geq F(x', n)$.

כעת נניח נכונות הטענה לכל $n \leq m$ ונוכיח עבור $m + 1$:

$$\Delta(x', m + 1) = \Pr[x = x'|m + 1] - \Pr[x = x'|m]$$

$$= \left(1 - 2^{-x'}\right) \Pr[x = x'|m] + \left(2^{-x'+1}\right) \Pr[x = x' - 1|m]$$

$$- \left(\left(1 - 2^{-x'}\right) \Pr[x = x', m - 1] + (2^{-x'+1}) \Pr[x = x' - 1|m - 1]\right)$$

מכיוון ש- $\hat{n}_{MLE}(x)$ היא פונקציה לא יורדת נוכל להשתמש בטענה, עבור $x' - 1$ ונקבל $\Delta(x' - 1, m) = (\Pr[x = x' - 1|m] - \Pr[x = x' - 1|m - 1]) < 0$ וכן מהנחת האינדוקציה נקבל ש: $\Delta(x', m) = (\Pr[x = x'|m] - \Pr[x = x'|m - 1]) < 0$ לכן:

$$\Delta(x', m + 1) = \left(1 - 2^{-x'}\right)\Delta(x', m) + \left(2^{-x'+1}\right)\Delta(x' - 1, m) < 0$$

מש"ל האינדוקציה.

התוצאות (הקוד ליצירת מטריצת ההסתברויות מצורף בנספח A):

x	$\hat{n}_{MLE}$	Maximum Likelihood
1	1	1.0000
2	3	0.6250
3	8	0.5237
4	19	0.4867
5	39	0.4704
6	80	0.4625
7	162	0.4587
8	325	0.4568
9	652	0.4559
10	1306	0.4554

b. נראה את צעדי החישוב עבור כל סעיף ולאחר מכן נציג טבלה עם הנתונים:

(a) ה-bias מוגדר:

$$bias(\hat{n}_{MLE}, n) = E[\hat{n}_{MLE}, n] - n$$

כאשר עבור n מסוים:

$$E[\hat{n}_{MLE}, n] = \sum_{i=1}^{n} \Pr[x = i|n] \cdot \hat{n}_{MLE}(x)$$

את אלו כמובן חישבנו בסעיף הקודם לכל $i = 1, \dots, 10$ ולכל $n \leq 10000$.

(b) מלינאריות התוחלת ומכיוון שבכל שלב n הוא מספר ה-MSE:

$$MSE(\hat{n}_{MLE}, n) = E[(\hat{n}_{MLE} - n)^2] = E[\hat{n}_{MLE}^2] + n^2 - 2 \cdot n \cdot E[\hat{n}_{MLE}]$$

(*c*) בסעיף זה נחלק את שורש התוצאה מהסעיף הקודם ב-*n*

(*d*) כפי שראינו בשיעור, ה-*variance* של ה-*unbiased estimator* הוא:

$$V[\hat{n}] = \frac{n^2 - n}{2}, \mu[\hat{n}] = n \Rightarrow CV[\hat{n}] = \sqrt{\frac{n - 1}{2n}}$$

לבסוף נקבל:

n	$bias$	MSE	$NMSRE$	$CV\ of\ \hat{n}$
1	0	0	0	0
2	0	1	0.5	0.5
3	0.125	4.125	0.677	0.5774
4	0.3281	9.4844	0.7699	0.6124
5	0.5801	16.9668	0.8238	0.6325
6	0.8621	26.4124	0.8566	0.6455
7	1.1619	37.6808	0.8769	0.6547
8	1.4715	50.6707	0.8898	0.6614

9	1.7856	65.3201	0.898	0.6667
10	2.1008	81.599	0.9033	0.6708

c. נוכיח כי לכל $\hat{\theta}\ estimator\ of\ \theta$ מתקיים:

$$MSE(\hat{\theta}) = Var[\hat{\theta}] + Bias(\hat{\theta}, \theta)^2$$

נשים לב כי מתכונות ה-*Variance* כיוון ש-θ הוא מספר:

$$V[\hat{\theta}] = V[(\hat{\theta} - \theta)]$$

$$V[(\hat{\theta} - \theta)] = E\left[(\hat{\theta} - \theta)^2\right] - (E[(\hat{\theta} - \theta)])^2$$

לפי ההגדרה $MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$, ומלינאריות התוחלת נקבל ש: $E[(\hat{\theta} - \theta)] = E[\hat{\theta}] - \theta$,
כאשר לפי ההגדרה $Bias(\hat{\theta}, \theta) = E[\hat{\theta}] - \theta$.

נציב חזרה עבור $V[(\hat{\theta} - \theta)]$ ונקבל:

$$V[\hat{\theta}] = V[(\hat{\theta} - \theta)] = MSE(\hat{\theta}) - \left(Bias(\hat{\theta}, \theta)\right)^2$$

$$\Rightarrow MSE(\hat{\theta}) = V[\hat{\theta}] + \left(Bias(\hat{\theta}, \theta)\right)^2$$

מש"ל.

d. בהנחה שה-x_i הם *pairwise independent*:

i. עבור $\hat{n}' = \frac{1}{k}\sum_{i=1}^{k} \hat{n}_U(x_i)$:

כפי שראינו בשיעור עבור ה-*Average Estimator* עם $k\ counters$:

$$\hat{n}' = \frac{\sum_{i=1}^{k} Z_i}{k} : E[\hat{n}'] = E[\hat{n}], Var[\hat{n}'] = \frac{1}{k} Var[\hat{n}] \Rightarrow CV[\hat{n}'] = \frac{1}{\sqrt{k}} CV[\hat{n}]$$

לכן ה-*NRMSE* עבור $n = 8$ (שהוא ה-*CV* עבור *Unbiased estimators* כפי שנתון),
בהתאם לכך, מסעיף *b*:

$$CV[n = 8] = 0.6614 \Rightarrow C[\hat{n}', n = 8] = \frac{0.6614}{\sqrt{k}} = NRMSE\left(\frac{1}{k}\sum_{i=1}^{k} \hat{n}_U(x_i)\right)$$

ii. עבור $\hat{n}'' = \frac{1}{k}\sum_{i=1}^{k} \hat{n}_{MLE}(x_i)$ נבצע את אותו התהליך אשר ביצענו בשיעור עבור ה-
Average Estimator עם $k\ counters$ כאשר נניח כי $E[\hat{n}_{MLE}] = \mu$ וכן $Var[\hat{n}_{MLE}] = \sigma^2$:

$$E[\hat{n}''] = \frac{1}{k}\sum_{i=1}^{k} E[\hat{n}_{MLE}(x_i)] = \frac{1}{k} kE[\hat{n}_{MLE}] = E[\hat{n}_{MLE}] = \mu$$

בנוסף:

$$Var[\hat{n}''] = \frac{1}{k^2}\sum_{i=1}^{k} Var[\hat{n}_{MLE}(x_i)] = \frac{1}{k^2} kVar[\hat{n}_{MLE}] = \frac{\sigma^2}{k}$$

כעת נחשב את $Bias(\hat{n}'', n)$ אשר נעזר בו ובסעיף *c*:

$$Bias(\hat{n}'', n) = E[\hat{n}''] - n = \mu - n = E[\hat{n}_{MLE}] - n = Bias(\hat{n}_{MLE}, n)$$

לכן בהצבה לסעיף c ומכיוון שלפיו מתקיים

$$Var[\hat{n}_{MLE}] = MSE(\hat{n}_{MLE}) - \left(Bias(\hat{n}_{MLE}, n)\right)^2$$:

$$MSE(\hat{n}'') = Var[\hat{n}''] + \left(Bias(\hat{n}'', n)\right)^2 = \frac{Var[\hat{n}_{MLE}]}{k} + \left(Bias(\hat{n}_{MLE}, n)\right)^2$$

$$= \frac{MSE(\hat{n}_{MLE}) - \left(Bias(\hat{n}_{MLE}, n)\right)^2}{k} + \left(Bias(\hat{n}_{MLE}, n)\right)^2$$

בהצבה ל-(c) מסעיף b כאשר ראינו כי עבור $n = 8$, $MSE(\hat{n}_{MLE}, n = 8) = 50.6707$,
$Bias(\hat{n}_{MLE}, n) = 1.4715$:

$$NMRSE(\hat{n}'', n = 8) = \frac{1}{8}\sqrt{\frac{50.6707 - (1.4715)^2}{\mathrm{k}} + (1.4715)^2}$$

$$= \sqrt{\frac{0.7579}{k} + 0.0338}$$

2)

a. נוכיח באינדוקציה כי בכל שלב $n \in \mathbb{N}$ מתקיים $\hat{c}_i \leq c_i$
   i. עבור $n = 1$:
      1. אם התקבל $j \neq i, (j, \Delta)$ אזי $c_i = \hat{c}_i = 0$ כיוון שעל פי ההגדרה לא נעשה שינוי ב-counterים של i.
      2. אם התקבל (i, +1), אזי לפי ההגדרה $c_i = 1$ ומכיוון שאין שום counter עדיין גם $\hat{c}_i = 1$.
      3. אם התקבל (i, -1), כיוון ש-$c_i = \hat{c}_i = 0$ לא נעשה שינוי, לפי ההגדרה של $c_i = \max(0, c_i + \Delta)$
   ii. כעת נניח כי הדבר נכון לצעד ה-n' ונוכיח עבור $n' + 1$, בסימון $c_i^{n'}$ הערך של c_i בצעד ה-n:
      1. אם התקבל $j \neq i, (j, \Delta)$ אזי $c_i^{n'+1} = c_i^{n'}$ - לא השתנה כיוון שעל פי ההגדרה הוא אינו מושפע ממידע שאינו נוגע ל-i. לעומת זאת ב-$MG \pm$:
         a. אם קיים counter $\hat{c}_j$ נוסיף לו 1. אם לא וישנו מקום להוסיף counter כזה, ניתן לו את הערך 1. כך או כך בסוף הצעד לא נשפיע על $\hat{c}_i^{n'}$ ולכן ולפי הנחת האינדוקציה מקיים: $\hat{c}_i^{n'+1} = \hat{c}_i^{n\prime} \leq c_i^{n'} = c_i^{n'+1}$
         b. אם כל ה-counterים מלאים, וקיים counter ל-i נוריד ממנו 1 כמו משאר ה-counterים ונקבל לפי הנחת האינדוקציה: $\hat{c}_i^{n'+1} < \hat{c}_i^{n\prime} \leq c_i^{n'} = c_i^{n'+1}$.
         c. אם כל ה-counterים מלאים ולא קיים counter ל-i אזי לפי ההגדרה והנחת האינדוקציה $0 = \hat{c}_i^{n'+1} = \hat{c}_i^{n\prime} \leq c_i^{n'} = c_i^{n'+1}$
      2. אם התקבל (i, +1), אזי לפי ההגדרה $c_i^{n'+1} = c_i^{n'} + 1$ כיוון שלפי ההגדרה, לכל היותר $\hat{c}_i^{n'+1} = \hat{c}_i^{n'} + 1$ אם קיים counter או אם יש מקום להוסיף אותו, תמיד נקבל לפי הנחת האינדוקציה ש:

$$\hat{c}_i^{n'+1} \le \hat{c}_i^{n'} + 1 \le c_i^{n'} + 1 = \hat{c}_i^{n'+1}$$

3. אם התקבל (i, -1) אזי לפי ההגדרה $c_i^{n'+1} = \max(0, c_i^{n'} - 1)$, ולכן לפי ההגדרה גם $\hat{c}_i^{n'+1} = \max(0, \hat{c}_i^{n'} - 1)$ כאשר במידה ולא קיים counter ל-i, $\hat{c}_i^{n'}$, אז אין גם ממה להחסיר ולכן $\hat{c}_i^{n'+1}$ לכל צורך מעשי יהיה 0, ולכן לפי הנחת האינדוקציה:

$$\hat{c}_i^{n'+1} \le \max(0, \hat{c}_i^{n'} - 1) \le \max(0, c_i^{n'} - 1) = \hat{c}_i^{n'+1}$$

כלומר בכל מקרה $\hat{c}_i^{n'+1} \le c_i^{n'+1}$, מש"ל האינדוקציה.

b. נשים לב כי הפער בין ה-counter $\hat{c}_i$ לבין c_i גדל רק:

- כאשר מתבצע decrement ל-$\hat{c}_i$ שלא נובע מהופעה של (i, -1)
- כאשר מופיע (i,+1) ו-$\hat{c}_i$ לא קיים ואין מקום להוסיף אותו (אחרת שניהם היו מושפעים מההופעה באותה מידה). זהו כמובן צעד decrement!
- בשאר המקרים, הפער קטן (אם רק c_i קטן כי $\hat{c}_i$ לא קיים או שווה ל-0) או נשאר כפי שהיה אם שני ה-counterים מושפעים מהופעה.

כפי שראינו בהרצאה, $\forall i: c_i - \hat{c}_i \le \#decrements$ כאשר פה הכוונה היא כמובן לצעדי ה-$decrement$ הנגרמים בשל הופעת "+" ללא ממקום להוספת counter עבור ה-i אליו הוא שייך.

כעת נשתמש בחסם עליון על ה-$decrements$ של כל ה-counterים, בדומה לזה אשר ראינו בהרצאה:

לא נוכל להשתמש בחסם אותו ראינו בהרצאה כיוון שכעת קשה לנו לספור את כל המקרים בהם הגיע "–" ל-i כאשר עוד לא הופיע i כלל – כלומר, מקרים שבהם ה-"-" לא משפיע על גודל ה-stream הכולל. מצד שני, נדרוש כי במקרה המנוון בו מגיעים אך ורק +ים יתכנס החסם לזהו אשר הופיע בהרצאה.

- יהי m' סכום ה-counterים בסוף הריצה, $\hat{c}_i$.
- נסמן ב-m את סכום ה-counterים האמיתיים, c_i, נשתמש בסימון לצורך נוחות ונראה בהמשך שהוא למעשה מתחלק לגדלים מדידים וגדלים שאינם מדידים אך אינם שייכים לתוצאה הסופית.
- נסמן ב-$\mathbb{P}$ את ה-counter של כלל ה-"+" שהופיעו ב-stream.
- נסמן ב-$\mathbb{M}$ את ה-counter של כלל ה-"-" אשר השפיעו על ה-counterים, כלומר כאשר הופיעו, ה-counter האמיתי, c, המתאים לערך איתו הופיעו היה שונה מ-0. למשל אם הופיע, $(i, -)$ אך $c_i = 0$ לפי ההגדרה, הוא לא משפיע על הערך של c_i ולכן לא משפיע על m ולא נכלל ב-$\mathbb{M}$. נשים לי ש-$\mathbb{M}$ אינו ניתן לחישוב כיוון שאיננו יודעים את מספר הפעמים בהם הופיע "-" וה-counter האמיתי המתאים לו היה 0.
בסימונים הנ"ל מתקיים:

$$(*) \quad m = \mathbb{P} - \mathbb{M}$$

- לכן, נסמן ב-$\mathbb{M}_{\hat{c}}$ את מספר ה-"-" אשר השפיעו על אחד מה-counterים שאנחנו מחזיקים $\hat{c}$. בפרט, "-"ים אלו השפיעו גם על ה-count האמיתי.

כעת, מכיוון שבכל צעד $decrement$, אנחנו מאבדים $k+1$ איברים מהספירה, נקבל (נסמן $\#decrements = d$):

$\mathbb{M} - \mathbb{M}_{\hat{c}}$ זהו בדיוק מספר הפעמים שבהם הופיע "-" שלא השפיע על *counter* שאנחנו מחזיקים אך השפיע על ה-*counter* האמיתי, מכאן נובע:

$$m - m' = d(k+1) - (\mathbb{M} - \mathbb{M}_{\hat{c}})$$

כעת נציב את $(*)$:

$$\mathbb{P} - \mathbb{M} - m' = d(k+1) - \mathbb{M} + \mathbb{M}_{\hat{c}}$$

$$\Rightarrow d = \frac{\mathbb{P} - \mathbb{M}_{\hat{c}} - m'}{k+1}$$

כעת נותר להציב חזרה בחסם העליון על $c_i - \hat{c}_i$ את החסם על $\#decrements$ ולקבל:

$$c_i - \hat{c}_i \leq \#decrements = \frac{\mathbb{P} - \mathbb{M}_{\hat{c}} - m'}{k+1}$$

בדיקת שפיות על המקרה בו אין הופעות של "-" (אז מתקיים $m = \mathbb{P}$) מאשרת שהחסם במקרה כזה מתכנס לזה אשר ראינו בשיעור. לכן נשמור counterים עבור מספר ה-"+" ומספר הפעמים בהם הופיע "-" עם ערך j אשר עבור $\hat{c}_j \neq 0$. כפי שראינו בשיעור, עבור MG החסם הנ"ל הוא הדוק, לכן גם במקרה זה יהיה הוא הדוק. קל מאוד לבנות מקרה בו כל ה-$decrements$ משפיעים על ערך למשל i, ובו אין כלל צעדי "-" (או לחילופין ישנם כאלה אך הם אינם בקבוצה $\mathbb{M}\backslash\mathbb{M}_{\hat{c}}$), במקרה כזה ההפרש עבור i הוא בדיוק $\#decrements$ ולכן נקבל שזהו חסם, בפרט ישנו מקרה אשר בו בדיוק ההפרש ולכן הוא חסם הדוק.

3)

a. במידה וקיים ה-counter נחזיר $\frac{x}{p}$ במידה וה-counter לא קיים נחזיר 0. יותר פורמלית:

$$\hat{c} = \begin{cases} \frac{x}{p}, & if\ there\ exists\ a\ counter \\ 0, & otherwise \end{cases}$$

נוכיח באינדוקציה כי estimator הוא unbiased כלומר עבור c פקטות ב-flow, $E[\hat{c}] = c$.
עבור $c = 1$:
אם נדגמה הפקטה, אזי $x = 1$ אחרת $\hat{c} = 0$ ולכן:

$$E[\hat{c}] = \sum_{i=0}^{c-1} \frac{i}{p} \Pr[x = i] = \frac{0}{p} \cdot (1-p) + \frac{1}{p}(p) = 1$$

כעת נניח נכונות הטענה לכל $c < c'$ ונוכיח עבור c', בסימון x_c יהיה ה-counter כשעברו c פקטות, והוא שווה ל-0 במידה ואין counter:

$$E[\hat{c}] = E\left[\frac{x_c}{p}\right] = \sum_{i=1}^{c-1} \Pr[x_{c-1} = i]\, E\left[\frac{x_c}{p} \middle| x_{c-1} = i\right]$$

נבחין כי $E\left[\frac{x_c}{p} \middle| x_{c-1} = i\right]$ מקיים:

- בהסתברות p: $x_c = i+1 \Rightarrow \frac{x_c}{p} = \frac{i+1}{p}$
- בהסתברות $1-p$: $x_c = i \Rightarrow \frac{x_c}{p} = \frac{i}{p}$

לכן:

$$E\left[\frac{x_c}{p}\middle|x_{c-1}=i\right]=p\left(\frac{i+1}{p}\right)+(1-p)\left(\frac{i}{p}\right)=i+1+\frac{i}{p}-i=1+\frac{i}{p}$$

נציב חזרה עבור $E[\hat{c}]$:

$$E[\hat{c}]=\sum_{i=1}^{c-1}\Pr[x_{c-1}=i]\,E\left[\frac{x_c}{p}\middle|x_{c-1}=i\right]=\sum_{i=1}^{c-1}\Pr[x_{c-1}=i]\left(1+\frac{i}{p}\right)=$$

$$\underbrace{\sum_{i=1}^{c-1}\Pr[x_{c-1}=i]\left(\frac{i}{p}\right)}_{E(\hat{c}-1)=c-1\text{ לפי הנחת האינדוקציה}}+\underbrace{\sum_{i=1}^{c-1}\Pr[x_{c-1}=i]}_{1}$$

ולכן:

$$E[\hat{c}]=c-1+1=c$$

מש"ל האינדוקציה.

כמובן שבהינתן הרמז לסעיף b. ניתן היה להסתכל על x כמתפלג בינומית עם הסתברות p - לכל פקטה יש התסברות p להידגם ולכן $x\sim B(c,p)$. כיוון שכך ניתן היה לתת הוכחה פשוטה הרבה יותר, שהרי ידוע כי שאם $X\sim B(n,\lambda)$ ולכן $E[x]=n\lambda$ ולכן $E[\hat{c}]=E\left[\frac{x}{p}\right]=\frac{E[x]}{p}=\frac{cp}{p}=c$, כאשר $E\left[\frac{x}{p}\right]=\frac{E[x]}{p}$ מלינאריות התוחלת.

b. נחשב את $Var[\hat{c}]$:

$$Var[\hat{c}] = Var\left[\frac{x}{p}\right]$$

לפי תכונות ה-variance מכיוון ש-pזה מספר:

$$Var\left[\frac{x}{p}\right] = \frac{1}{p^2}Var(x)$$

כיוון ש-$x \sim B(c,p)$:

$$Var(x) = cp(1-p)$$

ולכן נקבל:

$$Var[\hat{c}] = \frac{1}{p^2}cp(1-p) = c\left(\frac{1-p}{p}\right)$$

c. כיוון שהוכחנו שמתקיים $E[\hat{c}] = c$

$$CV[\hat{c}] = \frac{\sqrt{Var[\hat{c}]}}{E[\hat{c}]} = \sqrt{\left(\frac{1-p}{p}\right)\left(\frac{1}{c}\right)}$$

d. בהנחה שההסתברות לדגום פקטה היא p בשניהם, הם ניתנים למיזוג בקלות. נניח שעבור *NetFlow* X קיבלנו *counter* x ועבור *NetFlow* Y קיבלנו *counter* y. אזי *counter* המיזוג יהיה $z = x + y$. נוכיח כי $E[z] = X + Y$:
מלינאריות התוחלת, ומכיוון שה-*estimator*ים שלנו הם *Unbiased*:

$$E[z] = E\left[\frac{x+y}{p}\right] = E\left[\frac{x}{p}\right] + E\left[\frac{y}{p}\right] = X + Y$$

במידה וההסתברויות שונות:
נניח שעבור *NetFlow* X והסתברות לדגימה p קיבלנו *counter* x ועבור *NetFlow* Y והסתברות לדגימה q קיבלנו *counter* y. אזי, במקרה כזה, תמיד נמזג את ה-*counter* עם הסתברות הדגימה הגדולה לתוך ה-*counter* עם הסתברות הדגימה הקטנה. נניח בלי הגבלת הכלליות כי $p \geq q$ *counter* המיזוג יהיה $z = x + \frac{p}{q}y$ ניתן לבצע מעבר על הפקטות שדגמנו ולדגום אותן בהסתברות $\frac{p}{q}$ או פשוט להכפיל את ה-*count* הנוכחי. כעת נוכיח כי $E[z] = X + Y$:

$$E[z] = E\left[\frac{x + \frac{p}{q}y}{p}\right] = E\left[\frac{x}{p}\right] + E\left[\frac{\frac{p}{q}y}{p}\right] = E\left[\frac{x}{p}\right] + E\left[\frac{y}{q}\right] = X + Y$$

כמובן שמההסתכלות על $x, y \sim B$ אפשר לראות את זה מאוד בקלות.

e. בצורה די מתבקשת חלוקה בה כל הפקטות עוברות ב-*flow* אחר תניב מקסימום *counter*ים, כמובן שהיא לא בהכרח היחידה, אבל לכל *stream* היא נוכל לקבל יותר *counter*ים בחלוקה אחרת.
נוכיח בשני אופנים:

i. נסתכל על הפקטות ב-*stream* אשר נדגמו S, ברור שאם כל $s \in S$ הייתה ב-*flow* אחר. היינו מקבלים $|S|$ *counter*ים. שהוא כמובן מקסימלי כיוון שלא נדגמו יותר פקטות לכן בפרט לא יכולים להיות יותר *counter*ים. למעשה ברור שכל מקרה שבו אין *counter* עם יותר מ-1 בערכו הוא מקרה שבו כל פקטה שנדגמה יצרה *counter* חדש וגם הוא סידור אפשרי למקסום ה-*counter*ים. לכן, תמיד בסידור כל פקטה ב-*NetFlow* שונה ייתן מקסימום על ה-*counter*ים.

ii. נסתכל על כל פקטה בנפרד לפי מיקומה ב-*NetFlow* שלה. עבור פקטה w_i^j (פקטה שהיא ה-*i*ית ב-*j flow*) מתקיים שהסיכוי שלה לפתוח *counter* חדש הוא:

$$\Pr\left[w_i^j\right] = p(1-p)^{i-1}$$

זהו בדיוק הסיכוי שלא נדגמה אף פקטה ב-*flow* לפניה ושהיא אכן נדגמה. ומספר ה-*counter*ים יהיה, בהנחה של k *flows*:

$$S = \sum_{j=1}^{k} \sum_{i=1}^{m_j} p(1-p)^{i-1}$$

כאשר m_j הוא מספר הפקטות ב-*j flow*.

פוקנציה זו מקבלת מקסימום עבור $i = 1$ לכל פקטה ואינה תלויה ב-j, לכן קל לראות שעל מנת למקסם את ההסתברות של כל פקטה ליצור *counter* עלינו לשים אותה ראשונה ב-*flow* שלה, כך נקבל עבור m פקטות ב-*stream* (ומכאן ש-m *flows*):

$$S = \sum_{j=1}^{m} p = pm$$

זהו מספר ה-*counter*ים הצפוי להיפתח.

נספח *A* – קוד המקור של מציאת $\hat{n}_{MLE}$:

Probability.m:

```
function [prob] = Probability(i,n,AllResults)
   if i==1
      prob = (0.5)^(n-1);
      return;
   end
   if i==0 || n<i
      prob = 0;
      return;
   end
   if AllResults(i-1,n-1)==-1
      [AllResults(i-1,n-1)]=Probability(i-1,n-1,AllResults);
   End
   if AllResults(i,n-1)==-1
      [AllResults(i-1,n-1)]=Probability(i,n-1,AllResults);
   end
   prob = (2^(-(i-1)))*AllResults(i-1,n-1) + (1-(2^(-i)))*AllResults(i,n-1);
end
```

Main.m:

```
I=10;
N=10000;
AllResults = zeros(I,N)-1;
for i = 1:I
   for j=1:N
      [AllResults(i,j)]=Probablity(i,j,AllResults);
   end
end
```