אוניברסיטת בן גוריון
המחלקה להנדסת מערכות תקשורת
מבוא לתורת המידע
סמסטר ב' 2016

תרגיל תכנות 2- קידוד מקור אריתמטי

מועד הגשה: 23:55 8/06/2016

בעבודה זו נעסוק בקידוד אריתמטי. בפרט נחקור את השפעת המודל (שערוך ההסתברויות הסימבולים של המקור) על טיב הדחיסה בארבע מקרים שונים. בניגוד לתרגיל הקודם, בו קיבלתם קובץ לדחיסה, בתרגיל זה אתם תייצרו קובץ בעל תוכן אקראי ותדחסו אותו.

אפיון מקור

בחלק זה נממש מקור של אינפורמציה. המקור יהיה מקור מרקובי בעל 4 מצבים (סמבולים): $X \in \{0,1,2,3\}$.
להלן מטריצת הסתברויות מעברים בין המצבים (הסימבולים) של המקור:

$$\Pr(X_i \mid X_{i-1}) = \begin{bmatrix} 1-\alpha & \frac{\alpha}{5} & 0 & \frac{4\alpha}{5} \\ \frac{4\beta}{5} & 1-\beta & \frac{\beta}{5} & 0 \\ 0 & \frac{4\gamma}{5} & 1-\gamma & \frac{\gamma}{5} \\ \frac{\delta}{5} & 0 & \frac{4\delta}{5} & 1-\delta \end{bmatrix}$$

עבור $\alpha = \beta = 0.8, \quad \gamma = \delta = 0.9$, הגרילו 10^6 סמבולים של המקור (הסימבול הראשון יוגרל באופן שרירותי) ותכתבו אותם לקובץ בשם "orig". אין לרשום סמבולים כטקסט ascii אלא בצורה בינארית. הקובץ "orig" אמור להיות בגודל 250KBytes. בכל הסעיפים יש להשתמש באותו קובץ "orig".

קידוד אריתמטי

1. יישמו קידוד אריתמטי עבור קובץ "orig" באמצעות המודל של לפלס שנלמד בכיתה, ותיצרו קובץ דחוס "arithmLaplace". **שימו לב**: ה-input לדוחס (בכל הסעיפים בעבודה זו) יהיו סמבולים (0,1,2,3) ולא ביטים. לכן, עליכם לתרגם את הביטים מתוך קובץ "orig" לסמבולים. שימו לב: זמן ריצה של האלגוריתם לא צריך לעלות על דקות ספורות.

כפי שלמדנו בכיתה, במודל לפלס, ההסתברות לכל סימבול משתנה במהלך הקידוד (המודל מתעדכן באופן אדפטיבי למקור). במהלך הריצה של הדוחס, תשמרו את הסתברויות הסימבולים של המקור בזמנים הבאים:

i. בתחילת הדחיסה
ii. לאחר דחיסה של רבע הראשון של הקובץ
iii. לאחר דחיסה של רבע השני של הקובץ
iv. לאחר דחיסה של רבע השלישי של הקובץ
v. בסיום הדחיסה

2. תרשמו קטע קוד העובר על כל הקובץ ומחשב את ההסתברויות להופעה של כל אחד מהסימנים בקובץ. לסט הסתברויות אלו נקרא הסתברויות ה"אמפיריות" של המקור. <u>ענו: האם הסתברויות האמפיריות שוות להסתברויות הסופיות מסעיף הקודם? הסבירו את התשובה</u>. יישמו קידוד אריתמטי עבור קובץ "orig" תחת ההנחה שהמקור מתפלג לפי הסתברויות האמפיריות שחישבתם ותיצרו קובץ דחוס "arithmAmp".

3. יישמו קידוד אריתמטי עבור קובץ "orig" תחת ההנחה שהסמבולים הם בלתי תלויים ומתפלגים אותו הדבר, כלומר כאשר מניחים $p(0)=p(1)=p(2)=p(3)=\frac{1}{4}$, ותיצרו קובץ דחוס "arithmIid".

4. הפעם נשתמש במודל מתוחכם יותר, אשר יניח שהמקור שיצר את הקובץ הוא מרקובי וינסה לגלות במהלך הקידוד את $(\alpha,\beta,\gamma,\delta)$. תיצרו קובץ דחוס "arithmLaplaceMarkov".

נממש את המודל כך:

נחזיק מטריצה 4×4 (F_{ji}) עבור כל המעברים אפשריים של המקור. ובדומה למודל לפלס, נספור מספר הופעות של כל מעבר עד לסמבול הנוכחי. ועל סמך המידע שאספנו, נקבע את ההתפלגות של הסמבול הנוכחי. כלומר,

$P(X_n=i \mid X_{n-1}=j)=\frac{1+F_{ji}}{|X|+A}$, כאשר F_{ji} - זה מספר <u>המעברים</u> מ- j ל- i שהיו עד לסמבול ה- n -י,

$|X|$ - זהו הגודל של האל"ב של המקור, ו- A - זה סה"כ מספר המעברים שהיו מ- j (לא משנה למי) עד לסמבול ה- n -י. שימו לב: זמן ריצה של האלגוריתם לא צריך לעלות על 3 דקות.

5. ענו על השאלות הבאות:

a. מהי הדחיסה (מספר הביטים לסמבול) בכל אחד מארבעת המקרים (מודל לפלס, מודל אמפירי, מודל יוניפורמי, ומודל מרקובי)?
b. מה היא הדחיסה היעילה ביותר? הסבירו את האינטואיציה לתשובה שלכם.
c. מה המרחק בין **מספר הביטים לסמבול** בכל שיטה, לגובל התיאורטי? איך ניתן לקצר מרחק זה?

הוראות הגשה :

1. ניתן לעשות את העבודה ביחידם או זוגות.
2. העבודה תבוצע ב- Mathematica או בmatlab.
3. תוודאו שהקוד שלכם רץ ללא באגים.
4. יש להגיש :
   a. דוח מסכם בפורמט PDF.
   b. קבצי קוד בפורמט שניתן להריץ (כאשר בתוכם מופיעים הסבר על הקוד) .

בהצלחה!