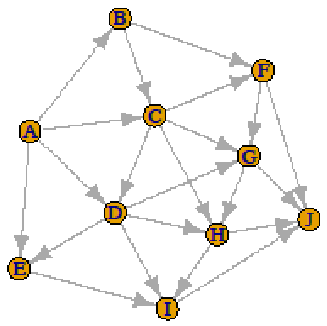
מתרגל: ניר קרת
סמסטר ב' תשפ"ג

חישוב סטטיסטי - פרוייקט 2

- יש לענות על כל השאלות.
- יש להגיש את הפרוייקט אל תיבת ההגשה במודל.
- את הפרוייקטים יש להגיש בזוגות.
- על גבי הקודים שאתם מגישים אתם מתבקשים להוסיף הערות (**באנגלית**) המסבירות מה כל קטע קוד אמור לבצע. המטרה היא לעזור לבודקת להבין איך התכוונתם לפתור את השאלה ומה הקוד אמור לעשות, צעד אחרי צעד. אם הפרוייקט יוגש ללא הסברים, הציון ייפגע. כתבו את הקוד וההערות בצורה מסודרת שתקל על הקריאה.
- יש להגיש הכל כקובץ PDF אחד, למשל באמצעות Rmarkdown.

שאלה 1: Importance Sampling



מערכות שונות (מכניות, אלקטרוניות, אנושיות וכו') ניתנות לתיאור ע"י גרף כגון הגרף לעיל. בגרף זה יש להעביר אות (=סיגנל) מנקודה A לנקודה J. ניתן לעשות זאת רק אם קיים מסלול רציף ותקין בין הנקודות. אות יכול לעבור מנקודה לנקודה רק לפי החיצים וכיוונם. לדוגמא, אות יכול לעבור מ-B ל-C, אך לא מ-C ל- B. הזמן עד לתקלה בכל קו המחבר בין שתי נקודות הוא מ"מ המתפלג $\exp(\theta)$, הפרמטר הוא בסקאלה של ימים, והזמנים עד לתקלות בקווים הנם בלתי תלויים.

השאלה עליה נרצה לענות: מהי ההסתברות שהמערכת מקולקלת לאחר 10 ימים (ז"א, לא ניתן לשלוח סיגנל מ-A ל-J)?

לשם כך נגדיר:

$$\underline{X} = \left(X_1, \ldots, X_{22} \right)$$

$$X_i = \begin{cases} 1 & \text{line } i \text{ is NOT working after 10 days} \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, \ldots, 22$$

$$h\left(\underline{X} \right) = \begin{cases} 1 & \text{the system is NOT working after 10 days} \\ 0 & \text{otherwise} \end{cases}$$

אזי, למעשה, יש לחשב את $E\{h(\underline{X})\}$. חישוב מדוייק של תוחלת זו היא משימה מסובכת קומבינטורית. במקום זאת נערוך סימולציה למערכת ומצבה לאחר 10 ימים. בכל אחד מהסעיפים הבאים עליכם לשרטט גרף המנטר את התכנסות האומד לתוחלת (העזרו בשקפים 11 ו-12 במצגת 2).

א. כתבו תוכנית סימולציה למערכת. מומלץ להיעזר בחבילה igraph (ובפונקציות השונות של חבילה זו, כגון: graph, plot, all_simple_paths, וכו'). שימו לב כי ניתן לדגום את הזמן בו כל קו מתקלקל ישירות מההתפלגות המעריכית המתאימה, או לחילופין לדגום מהתפלגות ברנולי (עם הפרמטר המתאים) האם הקו מקולקל או תקין. בשאלה זו (על כל סעיפיה) יש לדגום מהתפלגות **ברנולי** ולא ממעריכית.

עבור $\theta = 0.02$ הריצו סימולציה עם 15000 חזרות. מהו האומד לתוחלת? מהו אומד לשונות האומד?

ב. עבור $\theta = 0.005$ הריצו סימולצייה של 50000 חזרות. מהו האומד לתוחלת? מהו אומד לשונות האומד? בכמה מקרים של כשלון המערכת הצלחתם לחזות?

ג. בסעיף הקודם, בו ההסתברות לתקלה בכל קו נמוכה, ההסתברות שהמערכת אינה תקינה הנה קרובה מאוד לאפס, ולכן קשה לחזות בתקלות. שיטת importance sampling (IS) יכולה לסייע. עדכנו את תוכנית הסימולציה לשיטת IS עם דגימה מהתפלגות ברנולי אחרת.

חיזרו על הסימולציה של סעיף א' עם IS ודגימה מהתפלגות עם $\theta = 0.05$. האם מתקבל אומד דומה לזה שקיבלתם בסעיף א'? השוו גם את האומדים לשונויות האומדים.

ד. חיזרו על סעיף ב' עם שיטת IS ודגימה עם $\theta = 0.02$ ומצאו אומד לתוחלת ואומד לשונות האומד.

סכמו בקצרה את תוצאות כל הסעיפים ומסקנותיכם.

שאלה 2: Antithetic sampling

א. יש לאמוד את הפרמטר μ וברשותינו שני אומדים חסרי הטיה, $\hat{\mu}_1$ ו- $\hat{\mu}_2$, כאשר כל אחד מהם מבוסס על m תצפיות ב"ת, ומתקיים:

$$Var\left(\hat{\mu}_1\right) = \frac{\sigma^2}{m} \qquad Var\left(\hat{\mu}_2\right) = \frac{\sigma^2}{m} \qquad Corr\left(\hat{\mu}_1, \hat{\mu}_2\right) = \rho$$

נציע את האומד הבא:

$$\tilde{\mu} = \frac{1}{2}\left(\hat{\mu}_1 + \hat{\mu}_2\right)$$

מתרגל: ניר קרת
סמסטר ב' תשפ"ג

הראו כי

$$.Var\left(\tilde{\mu}\right)=\frac{\left(1+\rho\right)\sigma^2}{2m}$$

<u>**מסקנה**</u>: אם נמצא שני אומדים אשר ביניהם **מתאם שלילי**, אזי, מלבד הגדלת המכנה של השונות פי 2, ניתן גם להקטין את המונה בעזרת הגורם $1+\rho$.

ב. כעת עולה השאלה הבאה: אם יש לנו $\hat{\mu}_1$, כיצד ניתן "בקלות" לייצר את $\hat{\mu}_2$ כך שיהיו עם מתאם שלילי ביניהם? על שאלה זו נענה בעזרת הטענה הבאה והמסקנה שבסעיף ג' להלן.

<u>**טענה**</u>: יהי $\underline{X}=\left(X_1,\dots,X_n\right)$, כאשר $X_1,\dots,X_n$ מ"מ ב"ת, וכן h_1 ו- h_2 שתי פונקציות כלשהן $R^n\to R$ העולות/יורדות בכל רכיביהן (שתיהן באותו כיוון). הוכיחו כי

$$Cov\left(h_1\left(\underline{X}\right),h_2\left(\underline{X}\right)\right)=E\left\{h_1\left(\underline{X}\right)h_2\left(\underline{X}\right)\right\}-E\left\{h_1\left(\underline{X}\right)\right\}E\left\{h_2\left(\underline{X}\right)\right\}\ge 0$$

הוכיחו טענה זאת באינדוקציה בעזרת השלבים הבאים:

1. עבור $n=1$, הסבירו מדוע לכל ערכים x ו- y כלשהם
$$.\left\{h_1\left(x\right)-h_1\left(y\right)\right\}\left\{h_2\left(x\right)-h_2\left(y\right)\right\}\ge 0$$
2. מה ניתן להסיק מכך על $E\left[\left\{h_1\left(X\right)-h_1\left(Y\right)\right\}\left\{h_2\left(X\right)-h_2\left(Y\right)\right\}\right]$ כאשר X ו- Y הם מ"מ ב"ת ושווי התפלגות (ש"ה)?
3. מה ניתן להסיק מכך על $Cov\left(h_1\left(X\right),h_2\left(X\right)\right)$? רמז: העזרו בעובדה כי X ו- Y ש"ה.
4. נניח כי הטענה נכונה עבור $n-1$. הוכיחו כי
$$.E\left\{h_1\left(\underline{X}\right)h_2\left(\underline{X}\right)\mid X_n=x_n\right\}\ge E\left\{h_1\left(\underline{X}\right)\mid X_n=x_n\right\}E\left\{h_2\left(\underline{X}\right)\mid X_n=x_n\right\}$$
5. **(**סעיף רשות**)** הראו כי
$$\begin{aligned}E\left\{h_1\left(\underline{X}\right)h_2\left(\underline{X}\right)\right\}&\ge E\left[E\left\{h_1\left(\underline{X}\right)\mid X_n=x_n\right\}E\left\{h_2\left(\underline{X}\right)\mid X_n=x_n\right\}\right]\\&\ge E\left\{h_1\left(\underline{X}\right)\right\}E\left\{h_2\left(\underline{X}\right)\right\}\end{aligned}$$

ג. על-סמך הטענה בסעיף קודם, נסיק את התוצאה הבאה.

<u>**תוצאה**</u>: תהי h פונקציה עולה/יורדת בכל רכיביה. יהיו $U_1,\dots,U_n$ מ"מ ב"ת מהתפלגות $U\left(0,1\right)$. אזי

$$Cov\left(h\left(U_1,\dots,U_n\right),h\left(1-U_1,\dots,1-U_n\right)\right)\le 0$$

הוכיחו תוצאה זו על סמך המשפט מהסעיף הקודם.

<u>רמז</u>: אם $h\left(u_1,\dots,u_n\right)$ עולה/יורדת בכל רכיביה, מה ניתן לומר על $-h\left(1-u_1,\dots,1-u_n\right)$?

ד. עתה נשתמש במה שהוכחנו על מנת ליצור את האומד $\hat{\mu}_2$. נניח כי

$$\hat{\mu}_1=\hat{\mu}_1\left(\underline{U}\right)=\frac{1}{m}\sum_{j=1}^{m}h\left(F_1^{-}\left(U_{j1}\right),\ldots,F_n^{-}\left(U_{jn}\right)\right)$$

מכיוון שפונקציית ההתפלגות המצטברת מונוטונית עולה, גם ההופכית לה מונוטונית עולה. לכן, אם h מונוטונית בכל אחד מרכיביה $F^{-}(U_j)$, היא גם מונוטונית ב- U_j. לפיכך, נגדיר:

$$\hat{\mu}_2=\hat{\mu}_2\left(\underline{U}\right)=\frac{1}{m}\sum_{j=1}^{m}h\left(F_1^{-}\left(1-U_{j1}\right),\ldots,F_n^{-}\left(1-U_{jn}\right)\right)$$

וקיבלנו כי $Cov\left(\hat{\mu}_1\left(\underline{U}\right),\hat{\mu}_2\left(\underline{U}\right)\right)\le 0$. לכן, האומד $\tilde{\mu}=\frac{1}{2}\left(\hat{\mu}_1+\hat{\mu}_2\right)$ הוא עם שונות קטנה יותר מהשונות של $\hat{\mu}_1$ וכן של $\hat{\mu}_2$ (הרווחנו כאן פעמיים – הגדלת המדגם פי 2 מבלי לדגום שוב פעם מהתפלגות אחידה, וגם המתאם השלילי שמביא להקטנת השונות).

ד1. נחזור לבעיית המערכת בשאלה 1. שנו את התוכנית משאלה 1 סעיף א, כך שבמקום לדגום מהתפלגות ברנולי, הגרילו את זמן הכשלון של כל קו מהתפלגות **מעריכית**. הסבירו מדוע הפונקציה $h\left(\underline{X}\right)$ מונוטונית בכל אחד מרכיבי $\underline{U}=\left(U_1,\ldots,U_{22}\right)$, כאשר $U_1,\ldots,U_{22}$ מ"מ ב"ת מהתפלגות $U\left(0,1\right)$ ובעזרתם נדגמים 22 המ"מ מההתפלגות המעריכית.

ד2. בעזרת שיטת antithetic sampling, אימדו את $E\{h(\underline{X})\}$ עבור $\theta=0.08$ ו- $m=15000$ חזרות. מהו אומד לשונות האומד? השוו תוצאות אלו לאומד מונטה-קרלו "נאיבי".

ד3. (סעיף זה אינו קשור לבעיית המערכת מהשאלה הקודמת והינו סעיף עצמאי):

יהי $X\sim N(0,1)$ ונגדיר $h(x)=(e^{0.5x}+7)^{\frac{x}{3}}$ אמדו את $E\left\{h\left(X\right)\right\}$ בעזרת שיטת antithetic sampling על סמך 10000 דגימות מהתפלגות נורמלית סטנדרטית (חישבו מהו ה- *antithetic variate* במקרה זה). מהו אומד לשונות האומד? מהו הרווח בשונות האומד בשיטת antithetic sampling לעומת אומד מונטה-קרלו רגיל עם 10000 דגימות מהתפלגות נורמלית סטנדרטית?