

פרק ו'

תורת האינפורמציה

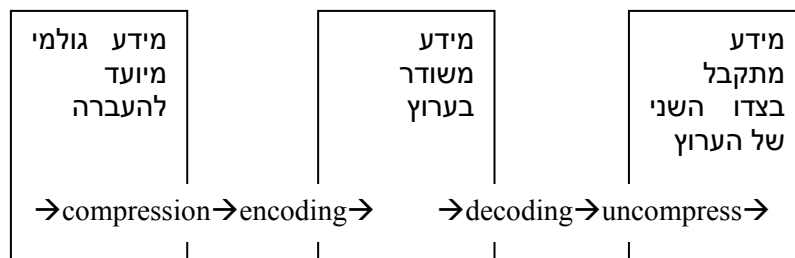
6.1 מבוא

6.1.1 תורת האינפורמציה

תורת האינפורמציה שהבסיס לה פותח כמעט בלעדית על ידי שנון (C. Shannon, 1949), תופסת בשנים האחרונות מקום מרכזי בניתוח של הקידוד והייצוג העצבי. התורה פותחה במקור על מנת לטפל בהעברת אינפורמציה במערכות תקשורת, אך כמעט מיד (למשל, Miller, 1953) הופיעו שימושים שלה לתיאור מערכות סנסוריות כמערכות המעבירות ומטפלות באינפורמציה. בפרק זה נסקור את שני המרכיבים הבסיסיים של תורת האינפורמציה **קידוד מקור** ו- **קידוד ערוץ** עבור משתנים מקריים בדידים.

6.1.2 מודל למערכות העברת אינפורמציה

המסגרת בה אנחנו עובדים היא הצורך להעביר באופן יעיל ככל האפשר מידע דיגיטלי נתון דרך ערוץ רועש ומשבש. שנון הראה כיצד ניתן לפרק את תהליך העברת המידע לשני שלבים נפרדים: בשלב הראשון נדחוס את האינפורמציה על ידי כך שנצל את התלויות הסטטיסטיות שבה ונקודד אותה בקצרה (**קידוד מקור**). בשלב השני, לצורך העברת המידע ללא שיבושים בערוץ הרועש, נוסיף למידע המועבר אינפורמציה יתירה, שתאפשר למקבל המידע בצדו השני של הערוץ הרועש לנקות את הרעשים (**קידוד ערוץ**). באופן ציורי, נוכל לתאר את מבנה מערכת העברת המידע באופן הבא



שנון הציג חסמים תיאורטיים על יעילות שני השלבים האלו, ובמשך השנים פותחו אלגוריתמים יעילים המתקרבים או משיגים חסמים אלו.

הפרק הנוכחי מורכב משלושה סעיפים עיקריים: ראשית נתאר את המרכיבים הבסיסיים של תורת האינפורמציה, ואז נשתמש בהם לתיאור קידוד מקור וקידוד ערוץ עבור משתנים מקריים בדידים.

6.2 מושגים בסיסיים: אנטרופיה ואינפורמציה

6.2.1 אנטרופיה

הגדרה

האנטרופיה של משתנה מקרי בדיד X בעל פונקציה התפלגות $p(x)$ מוגדרת כ-

$$(6.1) \quad H(X) = -\sum_x p(x) \log[p(x)]$$

כאשר מקובל להשתמש בבסיס הטבעי של הלוגריתם (אז תמדד האנטרופיה ביחידות הנקראות nats) או בבסיס 2 אז תמדד האנטרופיה בביטים (bits).

דוגמא

יהי X משתנה מקרי המקבל 1 בהסתברות p ואפס אחרת. אזי האנטרופיה שלו היא $H(X) = -[p \log(p) + (1-p) \log(1-p)]$. במקרה בו $p=1/2$ ובסיס הלוג הוא 2, נקבל

$$\begin{aligned} H(X) &= -1/2 \log_2(1/2) - 1/2 (\log_2(1/2)) = \\ &= -\log_2(1/2) = 1 \text{ bit} \end{aligned}$$

אנטרופיה כמדד לאי ודאות

האנטרופיה נחשבת מדד לאי הודאות על קבוצת המצבים האפשריים $\{x\}$. נשים לב כי בניגוד למדדי פיזור בהם עסקנו עד כה, כגון השונות של משתנה מקרי, האנטרופיה אינה תלויה כלל בערכים שהמשתנה מקבל אלא בפונקציית ההתפלגות שלו בלבד. נראה כעת כי האנטרופיה של משתנה מקרי בדיד מקבלת ערך מינימלי כאשר אי הוודאות מינימלית (יש רק ערך אפשרי אחד), וערך מקסימלי כאשר כל הערכים מתקבלים בהסתברויות שוות (אי ודאות מקסימלית).

האנטרופיה מקבלת ערכים חיוביים בלבד (היות וכל איבר בסכום הוא לוג של מספר קטן מאחד). קל לראות כי היא מקבלת ערך אפס כאשר קיים אחד בלבד שאיננו אפס, היות ומגדירים משיקולי רציפות $\lim_{x \rightarrow 0} x \log(x) = 0$. על מנת למצוא את הערך המקסימלי שהאנטרופיה יכולה לקבל עבור משתנה מקרי בדיד המקבל n ערכים, נרשום את הלגרנג'יאן

$$J = -\sum_{i=1}^n p_i \log(p_i) - \lambda \left(\sum_{i=1}^n p_i - 1 \right)$$

נגזור ביחס להסתברות של הערך x_i ונשווה לאפס

$$\frac{\partial J}{\partial p_i} = -(\log(p_i) + 1) - \lambda = 0$$

ונקבל כי

$$p_i = \exp(-\lambda - 1) \quad \text{for all } i$$

ומכיון ש- λ הוא קבוע, המקסימום מתקבל בהתפלגות האחידה $p_i = 1/n$. בהתפלגות זו האנטרופיה שווה ל-

$$(6.2) \quad H(X) = -\sum_{i=1}^n p_i \log(p_i) = +\frac{1}{n} \sum_{i=1}^n \log(n) = \log(n)$$

ניתן לראות את האנטרופיה של התפלגות כלשהי כמדד למרחק הסטטיסטי בין התפלגות נתונה לבין ההתפלגות האחידה

$$\begin{aligned} D\left[P \mid \frac{1}{m}\right] &= \sum_{\{x\}} p(x) \log\left(\frac{p(x)}{1/m}\right) \\ &= \sum_{\{x\}} p(x) \log p(x) - \sum_{\{x\}} p(x) \log\left(\frac{1}{m}\right) \\ &= -H[p] + \log m \end{aligned}$$

דוגמא

תא עצב מגיב לגירויים ראייתיים בירי של מספר פוטנציאלי פעולה, על פי ההתפלגות הבאה

מספר פ"פ	הסתברות ההופעה
0	0.3
1	0.5
2	0.15
3	0.04
4	0.01

האנטרופיה של התפלגות מספר פוטנציאלי הפעול בתגובה לגירוי היא

$$H(X) = -0.3 \log_2 0.3 - 0.5 \log_2 0.5 - 0.15 \log_2 0.15 - 0.04 \log_2 0.04 - 0.01 \log_2 0.01 = 1.6838 \text{ bits}$$

הגדרה: אנטרופיה משותפת

האנטרופיה של זוג משתנים מקריים, X ו- Y בעלי התפלגות משותפת $p(x, y)$ מוגדרת על ידי

$$(6.3) \quad H(X, Y) = - \sum_{\{x\}} \sum_{\{y\}} p(x, y) \log p(x, y) = -E_{p(x, y)} [\log p(X, Y)]$$

הגדרה: אנטרופיה מותנית

האנטרופיה המותנית (conditional entropy) $H(Y|X)$ תסומן $H(Y|X)$ והיא מוגדרת כ-

$$(6.4) \quad \begin{aligned} H(Y|X) &= \sum_{\{x\}} p(x) H(Y|X=x) = \\ &= - \sum_{\{x\}} p(x) \sum_{\{y\}} p(y|x) \log p(y|x) = \\ &= - \sum_{\{x\}} \sum_{\{y\}} p(x, y) \log p(y|x) \\ &= -E_{p(x, y)} [\log p(Y|X)] \end{aligned}$$

משפט: כלל השרשרת לאנטרופיה: $H(X, Y) = H(X) + H(Y|X)$

כלל זה מבטא את האדיטיביות של האנטרופיה, כלומר את היכולת שלנו לצבור אינפורמציה על ההתפלגות המשותפת, מידיעת האנטרופיה המותנית של כל אחד מהמשתנים.

הוכחה

$$\begin{aligned}
 H(X, Y) &= -\sum_{\{x\}} \sum_{\{y\}} p(x, y) \log p(x, y) \\
 &= -\sum_{\{x\}} \sum_{\{y\}} p(y|x) p(x) \log [p(y|x) p(x)] \\
 &= -\sum_{\{x\}} \sum_{\{y\}} p(y|x) p(x) \log [p(x)] \\
 (6.5) \quad &\quad - \sum_{\{x\}} \sum_{\{y\}} p(y|x) p(x) \log [p(y|x)] \\
 &= -\sum_{\{x\}} p(x) \log [p(x)] \\
 &\quad - \sum_{\{x\}} \sum_{\{y\}} p(x, y) \log [p(y|x)] \\
 &= H(X) + H(Y|X)
 \end{aligned}$$

על ידי הפעלה חוזרת של כלל השרשרת ניתן להרחיב אותו ל- n משתנים

$$(6.6) \quad H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1)$$

מסקנה: אנטרופיה של משתנים בלתי תלויים

$$\text{כאשר המשתנים בלתי תלויים אז } H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i)$$

מסקנה: אנטרופיה של תהליך מרקובי מסדר k

(6.7)

$$\begin{aligned}
 H(X_1, \dots, X_n) &= \\
 &= H(X_1) + H(X_2 | X_1) + \dots + H(X_k | X_1, \dots, X_{k-1}) + \sum_{i=k+1}^n H(X_i | X_{i-k+1}, \dots, X_{i-1})
 \end{aligned}$$

6.2.2 אינפורמציה משותפת

הגדרה: אנטרופיה יחסית

האנטרופיה היחסית (relative entropy) בין שתי התפלגויות, $p(x)$ ו- $q(x)$ היא שם אחר למרחק הסטטיסטי בין התפלגויות בו עסקנו בפרק 2.

$$(6.8) \quad D[p \parallel q] = \sum_{\{x\}} p(x) \log \frac{p(x)}{q(x)} = E_p \left[\log \frac{p(x)}{q(x)} \right]$$

הגדרה: אינפורמציה משותפת

האינפורמציה המשותפת (mutual information) בין שני משתנים מקריים X ו- Y , בעלי צפיפות משותפת: $p(x, y)$ והתפלגויות שוליות $p(x)$ ו- $p(y)$ הינה האנטרופיה היחסית (המרחק הסטטיסטי) בין ההתפלגות המשותפת ומכפלת ההתפלגויות השוליות

$$(6.9) \quad \begin{aligned} I(X; Y) &= D[p(x, y) \parallel p(x)p(y)] \\ &= \sum_x \sum_y p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \end{aligned}$$

$$I(X; Y) = H(X) - H(X|Y) \quad \text{טענה:}$$

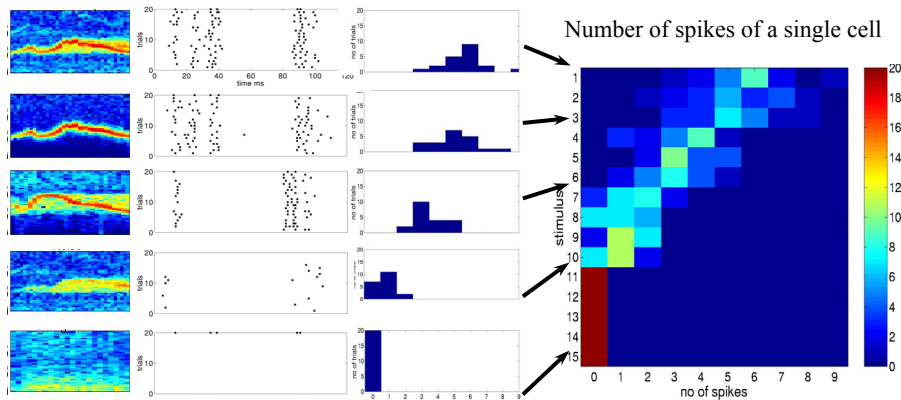
הוכחה

$$(6.10) \quad \begin{aligned} I(X; Y) &= D[p(x, y) \parallel p(x) \cdot p(y)] \\ &= \sum_{\{x\}} \sum_{\{y\}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= \sum_{\{x, y\}} p(x, y) \log \frac{p(x|y)}{p(x)} \\ &= - \sum_{\{x, y\}} p(x, y) \log p(x) + \sum_{\{x, y\}} p(x, y) \log p(x|y) \\ &= H(X) - H(X|Y) \end{aligned}$$

האינפורמציה המשותפת בין X ו- Y מבטאת את המידה שבה קטנה אי-הודאות בדבר ערכו של המשתנה X עקב הידיעה של המשתנה Y . נשים לב כי זהו מדד סימטרי ולכן מתקיים גם $I(X; Y) = H(Y) - H(Y|X)$.

דוגמא

האיור הבא מראה פעילות של תא במערכת השמיעה של חתול, בתגובה להשמעת גירויים אקוסטיים. חמישה גירויים כאלו מוצגים לדוגמה בעמודה משמאל. כל אחד מהגירויים הוצג 20 פעם, והתגובות מוצגות כאיור שבו כל נקודה מסמנת פוטנציאל פעולה (עמודה שניה משמאל). חזרות אלו מאפשרות לאמוד את ההתפלגות של סטטיסטיים שונים של שרשרת הספייקים. לדוגמא, התפלגות מספר הספייקים מוצגת בתגובה לחמשת הגירויים בעמודה השלישית משמאל. ניתן לחשב התפלגות זו עבור כל אחד מהגירויים שהוצגו (חמשה עשר בסך הכל). המטריצה המתקבלת היא אומד להתפלגות המשותפת של גירוי אל מול מספר פוטנציאלי פעולה. האינפורמציה המשותפת בהתפלגות שיצרה את המטריצה הזו היא חסם תחתון לאינפורמציה שמספק התא על הזהות של הגירוי שהוצג.



איור וניתוח הנתונים מתוך Chechik 2003. פירוט הניסוי ב- Bar Yosef et al 2003.

דוגמא

כשבוחנים את התפלגות מספר פוטנציאלי הפעולה של תא העצב מהדוגמא שבסעיף הקודם מגלים כי ההתפלגויות שונות מגירוי לגירוי

מספר פ"פ	בגירוי 1	בגירוי 2	סך הכל
0	0	0.30	0.30
1	0.35	0.15	0.50
2	0.10	0.05	0.15
3	0.04	0	0.04
4	0.01	0	0.01
סך הכל	0.50	0.50	1.00

כדי לחשב את כמות האינפורמציה שמספר פוטנציאלי הפעולה מספק על הגירוי, נסמן ב- X את מספר הספייקים וב- Y את הגירוי, ונרשום

$$\begin{aligned}
 H(X) &= 1.6838 \\
 H(X|Y) &= \sum_{i=1}^2 p(y_i) H(X|Y=y_i) \\
 &= -0.5[0.6 \log_2(0.6) + 0.3 \log_2(0.3) + 0.1 \log_2(0.1)] \\
 &\quad - 0.5[0.7 \log_2(0.7) + 0.2 \log_2(0.2) + 0.08 \log_2(0.08) \\
 &\quad + 0.02 \log_2(0.02)] \\
 &= 1.2622 \text{ bits} \\
 I(X;Y) &= H(X) - H(X|Y) = 0.4216 \text{ bits}
 \end{aligned}$$

כלל השרשרת לאינפורמציה משותפת

האינפורמציה המשותפת בין משתנה מקרי Y לאוסף משתנים X_1, \dots, X_n מוגדרת באופן טבעי

$$(6.11) \quad I(X_1, X_2, \dots, X_n; Y) = H(X_1, X_2, \dots, X_n) - H(X_1, X_2, \dots, X_n | Y)$$

תחת הגדרה זו מתקיים כלל השרשרת הבא עבור האינפורמציה המשותפת

$$(6.12) \quad I(X_1, X_2, \dots, X_n; Y) = I(X_1; Y) + \sum_{i=2}^n I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1)$$

הוכחה

מתוך כלל השרשרת לאנטרופיה נובע

$$\begin{aligned}
 I(X_1, X_2, \dots, X_n; Y) &= H(X_1, X_2, \dots, X_n) - H(X_1, X_2, \dots, X_n | Y) \\
 &= \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) - \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1, Y) \\
 &= \sum_{i=1}^n I(X_i; Y | X_1, X_2, \dots, X_{i-1})
 \end{aligned}$$

ערך אינפורמטיבי של מאורע בודד

נניח כי נתונים לנו שני משתנים מקריים X ו- Y וההתפלגות המשותפת שלהם $P(X, Y)$, ובצענו ניסוי יחיד בו קיבלנו כי ערכו של המשתנה המקרי Y הוא y . האינפורמציה שהניסוי הבודד מספק לנו על X תהיה

$$(6.13) \quad I(X; y) = H(X) - H(X | y).$$

כלומר, האינפורמציה תהיה ההפרש בין האנטרופיה של X לפני שידענו כי $Y=y$, לאנטרופיה של X בהינתן הערך של y (לסקירה השוואתית של מדדים לאינפורמציה מתצפית בודדת ראה (DeWeese and Meister 1999)).

סיכום

נסכם את התכונות היסודיות של האינפורמציה המשותפת

$$I(X;Y) = D[p(x,y) \| p(x)p(y)]$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y;X)$$

$$I(X;X) = H(X)$$

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1)$$

6.2.3 סדרה אופיינית

חוק המספרים הגדולים קובע כי עבור משתנים מקריים שמתפלגים i.i.d. הממוצע $\frac{1}{n} \sum_{i=1}^n x_i$ קרוב לתוחלת של x , עבור n גדול. המשפט האנלוגי בתורת האינפורמציה הוא AEP (Asymptotic Equipartition Property), והוא קובע כי הגודל $-\frac{1}{n} \log(p(x_1, x_2, \dots, x_n))$ קרוב לאנטרופיה כאשר $\{x_1, \dots, x_n\}$ הם בלתי תלויים, ו- $p(x_1, \dots, x_n)$ הוא ההסתברות לראות את הסדרה $\{x_1, \dots, x_n\}$. העברת אגפים תתן לנו

$$(6.14) \quad p(x_1, \dots, x_n) \xrightarrow{n \rightarrow \infty} 2^{-nH(X)}$$

תוצאה זו מאפשרת לנו לחלק את אוסף הסדרות האפשריות באורך n $\{x_1, \dots, x_n\}$ לשתי קבוצות: קבוצה של סדרות אופייניות (typical sets) שתסומן $A_{\mathcal{E}}^{(n)}$ שבהן האנטרופיה של הסדרה קרובה לאנטרופיה האמיתית עד כדי ε , וקבוצת הסדרות הלא אופייניות. ובאופן יותר פורמלי, ההסתברות לקבל סדרה שההסתברות שלה קרובה ל- $2^{-nH(X)}$ מקיימת

$$\Pr(A_{\mathcal{E}}^{(n)}) = \Pr\left(\{x_1, \dots, x_n\} \mid 2^{-n(H(X)+\varepsilon)} \leq p(x_1, \dots, x_n) \leq 2^{-n(H(X)-\varepsilon)}\right) > 1 - \varepsilon$$

משפט AEP (שנון-מקמילן):

אם X_1, \dots, X_n הם i.i.d. אז מתקיימת התכנסות בהסתברות

$$-\frac{1}{n} \log p(x_1, \dots, x_n) \rightarrow H(X)$$

הוכחה

הפעלת פונקציה על סדרת משתנים מקריים בלתי תלויים נותנת אף היא סדרת משתנים מקריים בלתי תלויים ולכן

$$\begin{aligned} -\frac{1}{n} \log p(X_1, \dots, X_n) &= -\frac{1}{n} \sum_{i=1}^n \log p(X_i) \\ &\xrightarrow{n \rightarrow \infty} -E[\log(p(X))] \text{ in probability} \\ &= H(X) \end{aligned}$$

הגדרה: קבוצה אופיינית (typical set)

הקבוצה האופיינית $A_\varepsilon^{(n)}$ ביחס ל- $p(x)$ היא קבוצת הסדרות $\{x_1, x_2, \dots, x_n\}$ המקיימות

$$(6.15) \quad 2^{-n(H(X)+\varepsilon)} \leq p(x_1, \dots, x_n) \leq 2^{-n(H(X)-\varepsilon)}$$

משפט: תכונות סדרה אופיינית

(1) אם סדרה $\{x_1, x_2, \dots, x_n\}$ היא אופיינית $A_\varepsilon^{(n)}$ אז

$$H(X) - \varepsilon \leq -\frac{1}{n} \log(p(x_1, \dots, x_n)) \leq H(X) + \varepsilon$$

$$p_r \{A_\varepsilon^{(n)}\} \geq 1 - \varepsilon \text{ for sufficiently large } n \quad (2)$$

$$2^{n(H(X)+\varepsilon)} \text{ מספר האיברים ב- } A_\varepsilon^{(n)} \text{ קטן מ-} \quad (3)$$

$$(1 - \varepsilon) \cdot 2^{n(H(X)-\varepsilon)} \text{ מספר האיברים ב- } A_\varepsilon^{(n)} \text{ גדול מ-} \quad (4)$$

למרות פשטות ההוכחה נדלג עליה כאן (ניתן למצוא אותה למשל בפרק 3 ב- Cover and Thomas) ונדגיש כי משמעות המשפט היא כי למרות שכמות הסדרות ב- $A_\varepsilon^{(n)}$ אינה בהכרח גדולה, הרי שההסתברות לפגוש סדרה ב- $A_\varepsilon^{(n)}$ היא כמעט אחת.

6.2.4 עיבוד ואיבוד ואינפורמציה

טיפול במידע כולל העברה שלו ממקום למקום (למשל העברת אינפורמציה מאיברי החושים אל מערכת העצבים המרכזית, העברת קבצים ממחשב למחשב, העברת קול בטלפון וכדומה), ועיבוד של המידע, כלומר מיצוי של החלקים החשובים או הרלוונטים של האינפורמציה. שני תהליכים אלו גורמים בדרך כלל לאיבוד אינפורמציה: העברת אינפורמציה דרך מערכת פיסיקאלית כלשהי גורמת להוספה של "רעש" לאינפורמציה שנשלחה, ועיבוד אינפורמציה (כלומר ביצוע של פונקציה כלשהי על האינפורמציה הראשונית) גורם לסינון של חלק ("לא חשוב" אם העיבוד הוא טוב) מהאינפורמציה הראשונית.

פורמלית, נייצג את תהליך ההעברה או העיבוד על ידי שרשרת מרקובית

$$(6.16) \quad X \rightarrow Y \rightarrow Z.$$

דוגמא

נניח ש- X הוא משתנה מקרי המתאר את כמות הפוטונים הפוגעים ברצפטור יחיד ברשתית בפרק זמן נתון, ו- Y הוא כמות הטרנסמיטור שהרצפטור מפריש בתגובה לגירוי האור. לדוגמא, Z הוא מספר פוטנציאלי הפעולה שתא גנגליון ברשתית יורה כתגובה לאותו הגירוי. הפעילות העצבית של הרצפטור Y מספקת אינפורמציה על העולם החיצון $I(X;Y)$, אינפורמציה זו עוברת עיבוד כך שהפעילות העצבית בתא הגנגליון מספקת גם היא אינפורמציה על העולם החיצון אך זוהי אינפורמציה שונה $I(X;Z)$. ומתקיים הקשר המרקובי $X \rightarrow Y \rightarrow Z$. המשפט הבא קובע קשר בין שני מדדי אינפורמציה אלו.

הגדרה אינפורמציה משותפת מותנית

האינפורמציה המשותפת המותנית של X ו- Z בהינתן Y

$$(6.17) \quad \begin{aligned} I(X;Z|Y) &= H(X|Y) - H(X|Z,Y) = \\ &= D[p(X,Z|Y) \| p(X|Y)p(Z|Y)] \end{aligned}$$

האינפורמציה המשותפת המותנית מבטאת את צמצום אי-הודאות ב- X הנובע מידעת Z כאשר Y נתון. מתוך העובדה ש- $D[p \| q] = 0$ אם ורק אם $p(x) = q(x)$ כמעט בכל מקום נובע כי

$$(6.18) \quad I(X;Z|Y) = 0 \Leftrightarrow p(X,Z|Y) = p(X|Y)p(Z|Y)$$

מההגדרה של התפלגות מותנית

$$P(x,y,z) = p(x)p(y|x)p(z|x,y)$$

נזכיר כי הסדרה $X \rightarrow Y \rightarrow Z$ היא מרקובית מסדר ראשון אם מתקיים

$$P(x, y, z) = p(x)p(y|x)p(z|y)$$

משפט: אי-שוויון עיבוד האינפורמציה

אם $X \rightarrow Y \rightarrow Z$ מהווים שרשרת מרקובית מסדר ראשון, אז

$$(6.19) \quad I(X;Y) \geq I(X;Z)$$

הוכחה: מכלל השרשרת נובע כי אנו יכולים לפתח את האינפורמציה המשותפת בשתי דרכים

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y) \end{aligned}$$

מאחר ו- X ו- Z בלתי תלויים בהנתן Y , נקבל כי $I(X;Z|Y) = 0$ ומאחר ו- $I(X;Y|Z) \geq 0$ נובע

$$I(X;Y) \geq I(X;Z)$$

ושוויון מתקיים אם ורק אם $I(X;Y|Z) = 0$ כלומר אם: $X \rightarrow Z \rightarrow Y$ מהווים גם-כן שרשרת מרקובית מסדר ראשון. באופן דומה ניתן להראות כי $I(Z;Y) \geq I(X;Z)$. בפרט נובע מאי שוויון האינפורמציה כי אם $Z = g(Y)$ אז $I(X;Y) \geq I(X;g(Y))$.

דוגמא

כדי להבין את המשמעות של אי שוויון זה לגבי עיבוד מידע במוח, נחזור לדוגמא ממערכת הראיה. נסמן ב- X את התפלגות הקלטים על הרשתית, ב- Y את התפלגות התגובות של תאי הרצפטורים ברשתית, וב- Z את התפלגות התגובות של תאי הגנגליון ברשתית. נניח שתגובת תאי הגנגליון תלויה אך ורק בפעילות תאי הרצפטור ברשתית ונקבל את השרשרת המרקובית $X \rightarrow Y \rightarrow Z$. מתוך אי שוויון עיבוד האינפורמציה נובע כי תאי הגנגליון מספקים פחות אינפורמציה על הגירויים. ירידה זו בכמות האינפורמציה תלך ותחמיר ככל שנוסיף עוד ועוד שכבות עיבוד (תלמוס, קורטקס...). לכאורה כל רמת עיבוד כזו מזיקה היות והיא גוררת הפסד של אינפורמציה על הגירוי. האבחנה הקריטית כאן היא שמטרת תהליך העיבוד המוחי איננה לשמור אינפורמציה על הקלט הגולמי שהתקבל ברשתית, אלא דווקא לזרוק אינפורמציה לא חשובה בו, ולהשאיר רק את המבנים הסטטיסטיים

החשובים מבחינה התנהגותית. כך מתאפשר לנו למשל לזהות את אותו הפרצוף מזוויות שונות בהבעות שונות ובתנאי תאורה שונים.

טענה: סטטיסטים מספיקים ואינפורמציה

בהינתן מדגם X מתוך התפלגות פרמטרית $f_\theta(x)$ הסטטיסטי $S(X)$ מספיק עבור θ אם ורק אם

$$(6.20) \quad I(\theta; X) = I(\theta; S(X))$$

הוכחה

← כיוון ראשון:

- נניח כי $S(X)$ הוא סטטיסטי מספיק ונוכיח שוויון האינפורמציות.
- א. ראשית נשים לב כי לכל פונקציה של X , ובפרט ל- $S(X)$ מתקיים הקשר המרקובי $\theta \rightarrow X \rightarrow S(X)$, ולכן בהכרח $I(\theta; X) \geq I(\theta; S(X))$.
- ב. בנוסף לכך עבור סטטיסטי מספיק מתקיים על פי ההגדרה כי $p(\theta | X, S(X)) = p(\theta | S(X))$, כלומר θ בלתי תלוי ב- X בהינתן $S(X)$, ולכן מתקיים גם קשר מרקובי נוסף $\theta \rightarrow S(X) \rightarrow X$. כתוצאה מהקשר המרקובי הזה מתקיים $I(\theta; S(X)) \geq I(\theta; X)$.
- ג. מצירוף שני אי השוויונים לעיל מתקבל שוויון האינפורמציות.

→ כיוון שני:

- נניח שוויון האינפורמציות ונוכיח כי $S(X)$ הוא סטטיסטי מספיק.
- א. ראשית נשים לב כי היות והתניה מפחיתה אנטרופיה, מתקיים לכל שלושה משתנים A, B, C כי $H(A|B) \geq H(A|B, C)$ ושוויון מתקיים אם ורק אם A אינו תלוי ב- C בהינתן B . (נעיר כי אם C פונקציה של B אז השוויון מתקיים מיידית)
- ב. נביט על הפרש האינפורמציות

$$\begin{aligned} & I(\theta; X) - I(\theta; S(X)) \\ &= H(\theta) - H(\theta | X) - H(\theta) + H(\theta | S(X)) \\ &= -H(\theta | X) + H(\theta | S(X)) \\ &= -H(\theta | X, S(X)) + H(\theta | S(X)) \\ &= H(\theta | S(X)) - H(\theta | S(X), X) \end{aligned}$$

- ג. על פי א. הפרש זה מתאפס אם ורק אם θ בלתי תלוי ב- X בהינתן $S(X)$.

6.3 קידוד מקור

לאחר שביססנו את המושגים היסודיים בתורת האינפורמציה נעבור לתאר את החלק הראשון בתאוריה של שנון, העוסק בקידוד אינפורמציה וייצוגה באופן קומפקטי.

6.3.1 קידוד

נטפל בתרחיש המוכר בו נתונה לנו סדרת סימנים (למשל סדרת ערכים של משתנה מקרי שהוגרלו באופן בלתי תלוי, או רצף אותיות וסימני פיסוק המהווה שיר באנגלית). קידוד של הסדרה הוא תהליך בו ממפים את סימני הסדרה לסימנים ורצפי סימנים אחרים. המטרות העיקריות של קידוד הן דחיסת אינפורמציה (לצורך הקטנת המקום הנדרש לאחסון או הקטנת משאבי התקשורת הנדרשים), חסינות מפני רעשים (Error Correction Codes), הצפנה, ותרגום האינפורמציה ל"שפה המובנת לצרכן האינפורמציה" כגון שפת מחשב (computer code) או לקוד עצבי המובן לאזורים מסויימים במוח. בפרק הנוכחי נדון בשני הנושאים הראשונים בלבד. בחרנו להשתמש בפרק זה בסימונים של (Cover,1991).

הגדרות: קידוד מקור

קידוד מקור של משתנה מקרי X הוא מיפוי כל אחד מהערכים x למחרוזת מתוך אלפא-בית Σ בן d סימנים $C(x) \in \Sigma^*$.

הגדרה: קוד לא סינגולרי

קוד נקרא לא סינגולרי אם אין סימן קוד ב- d שמתאימים לו שני סימנים שונים בשפה

$$x_i \neq x_j \Rightarrow C(x_i) \neq C(x_j)$$

הגדרה: הרחבה של קוד

הרחבה של קוד מתבצעת על ידי שרשור של מילות קוד

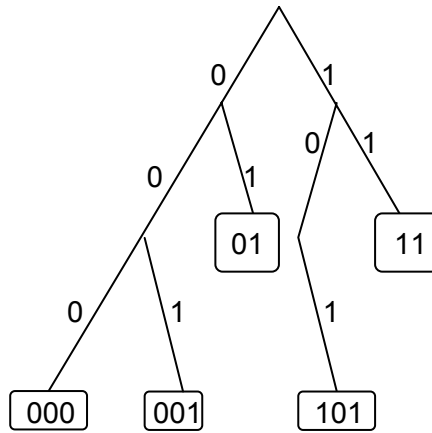
$$C(X_1, \dots, X_n) = C(x_1)C(x_2), \dots, C(x_n)$$

הגדרה

קוד נקרא ניתן לפענוח יחיד (uniquely decodable) אם ההרחבה שלו אינה סינגולרית.

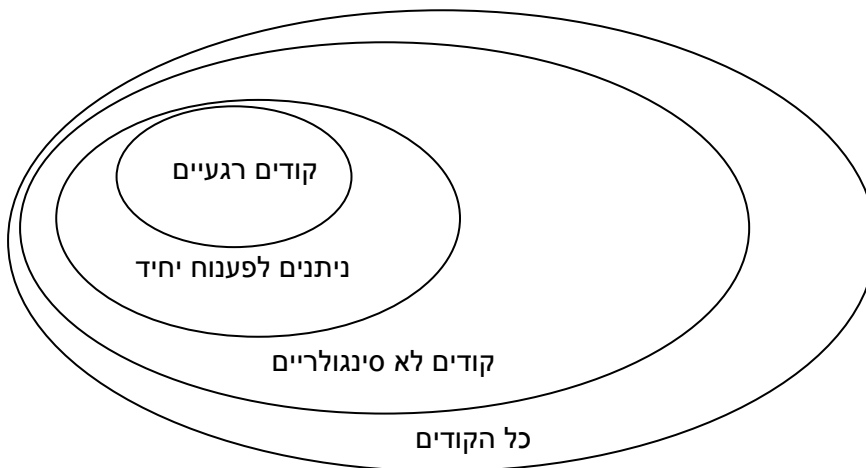
הגדרה: קוד רגעי , קוד רישא

קוד רגעי (instantaneous) או **קוד רישא** (prefix code) הוא קוד שבו אף מילת קוד אינה תחילית של אף מילת קוד אחרת. במקרה זה ניתן להציג את הקוד בצורת עץ



קוד רגעי ניתן לפיסוק (segmentation) תוך מעבר אחד על סדרת מילות הקוד.

את ההירארכיה של סוגי קידוד ניתן לתאר בדיאגרמה הבאה



דוגמא

X	סינגולרי	לא סינגולרי אך לא ניתן לפענוח יחיד	פענוח יחיד אך לא רגעי	קוד רגעי (קוד רישא)
A	0	1	10	0
B	0	101	00	10
C	1	01	11	110
D	1	10	110	111

6.3.2 דחיסת אינפורמציה באמצעות קידוד

על מנת לדחוס אינפורמציה נשאף לקודד את המקור באופן כזה שמחרוזות שכיחות תהינה קצרות, ומחרוזות נדירות – ארוכות. תהליך דומה קורה באופן טבעי בשפות טבעיות: מילים שכיחות הן בדרך כלל קצרות יותר (למשל, ברוב השפות "כן" ו"לא" הן מילים בנות הברה אחת). במקומות בהם השפה דינמית (כמו למשל בצבא) תהליך זה קורה באופן שוטף כך שנוצרים ר"ת, קיצורים וכד'. עבור סימן x_i ששכיחותו $p(x_i)=p_i$ והוא מקודד ל- $C(x_i)$ נסמן את אורך מילת הקוד שלו ב- $l(x_i)=l_i$. האורך הממוצע של הקוד C יהיה $L(C) = \sum_x p(x)l(x)$, והקידוד היעיל ביותר לדחיסת אינפורמציה יהיה הקידוד שעבורו $L(C)$ הוא מינימלי. יחס הדחיסה יהיה היחס בין האורך הממוצע של מילות הקוד במקור לבין האורך הממוצע של המילים המקודדות. יחס זה יבטא גם את היחס בין גודל הקובץ לפני הדחיסה לבין גודל הקובץ אחריה.

משפט: אי-שוויון Kraft-MacMillan

לכל קוד רישא על אלפא-בית בן d סימנים, אורכי מלות הקוד $l(x)$ מקיימים

$$(6.21) \quad \sum_{\{x\}} d^{-l(x)} \leq 1$$

הוכחה

קוד רישא מעל אלפא-בית בן d סימנים ניתן לתיאור באמצעות עץ d -נארי. (כלומר עץ שבו לכל קודקוד יש לכל היותר d בנים). כל ענף בעץ מציין סימן-קוד וכל עלה בעץ מייצג את מילת הקוד הנוצרת על ידי סימני-הקוד לאורך הנתיב שבין השורש לעלה. נסמן את העומק המקסימאלי של העץ ב- m . ונספור את כמות הקודקודים בעץ המתאר את הקוד. בעץ המלא בעומק m יש כמובן d^m קודקודים, אך בעץ המתאר את הקוד עשויים להיות עלים שאינם בעומק מלא. כדי להשלים את עץ

הקוד לעץ מלא יש להשלים לכל עלה שאינו בעומק m אלא עומק l , תת-עץ שבו d^{m-l} עלים.

היות וסך הקודקודים בעץ הוא לכל היותר d^m אז מתקיים

$$(6.22) \quad \sum_{\{x\}} d^{m-l(x)} \leq d^m$$

ומכאן נובע כנדרש כי

$$(6.23) \quad \sum_{\{x\}} d^{-l(x)} \leq 1.$$

ניתן להוכיח גם טענה משלימה: בהנתן אורכים של מלות קוד $l(x_1), \dots, l(x_n)$ המקיימים את אי שוויון קראפט, נוכל לבנות עץ המתאר את הקוד באופן הבא. סדר את האורכים בסדר עולה. כעת מצא את הקודקוד הראשון בעץ (בסדר לקסיקוגרפי) שאורכו l_1 , ומחק את תת העץ שלו. המשך כך עבור האורכים הבאים. משיקולים של ספירת כמות הקודקודים שאותם מנצלים ניתן לראות כי אכן העץ שיבנה יהיה עץ חוקי.

אורכי מלות הקוד של קוד אופטימלי

ראינו כי כדי לבנות קוד רישא, אורכי מילותיו צריכות לקיים את אי-שוויון קראפט. נרצה לכן למצוא אורכים העומדים בדרישה זו, ומביאים למינימום את האורך הממוצע של הקוד $L(C) = \sum_i p_i l_i$. זוהי בעיית אופטימיזציה סטנדרטית, וכדי לפתור אותה נתעלם כרגע מהעובדה ש l_i חייב להיות שלם, נניח שוויון באי-שוויון קראפט ונכתוב מחדש כבעיית מינימיזציה המשתמשת בכופלי לגרנג'

$$(6.24) \quad J = \sum_i p_i l_i + \lambda \left(\sum_i d^{-l_i} - 1 \right)$$

נגזור ביחס ל- l_i ,

$$\frac{\partial J}{\partial l_i} = p_i - \lambda d^{-l_i} \log(d)$$

נשווה לאפס ונקבל

$$(6.25) \quad d^{-l_i} = \frac{p_i}{\lambda \log(d)}$$

נציב שוויון זה באילוץ (קראפט) ונקבל כי הערך של λ הוא $1/\log(d)$, נציב אותו ונקבל כי האורכים (הלא שלמים) של קוד אופטימלי מקיימים

$$(6.26) \quad l_i^* = -\log_d(p_i)$$

לו ניתן היה לבנות קוד שהיה משתמש באורכים כאלו, אז האורך הממוצע של הקוד האופטימלי היה

$$(6.27) \quad L^*(C) = \sum_i p_i l_i^* = -\sum_i p_i \log_D(p_i) = H_D(X)$$

שהיא האנטרופיה של המשתנה המקרי שאותו אנחנו מקודדים. בפועל, היות ועלינו להגביל את אורכי מלות הקוד לערכים שלמים, הרי שלא תמיד נוכל להשיג ערך זה. כדי לראות מתי ניתן להשיג את האנטרופיה נוכיח כעת באופן ישיר כי אורך הקוד הממוצע גדול או שווה לאנטרופיה של המקור.

משפט

אורך הקוד הממוצע $L(C)$ גדול או שווה מהאנטרופיה של המקור $H_d(X)$

$$(6.28) \quad L(C) = \sum_i p_i l_i \geq H_d(X) .$$

הוכחה

$$\begin{aligned} L(C) - H_d(X) &= \sum_i p_i l_i - \sum_i p_i \log_d \left(\frac{1}{p_i} \right) \\ &= -\sum_i p_i \log_d (d^{-l_i}) + \sum_i p_i \log_d (p_i) \end{aligned}$$

נסמן

$$(6.29) \quad q_i = \frac{d^{-l_i}}{c}, \quad c = \sum_i d^{-l_i}$$

ונקבל

$$\begin{aligned} L(C) - H_d(X) &= -\sum_i p_i \log_d d^{-l_i} + \sum_i p_i \log_d (p_i) \\ &= \sum_i p_i \log_d \left(\frac{p_i/c}{d^{-l_i}/c} \right) \\ &= \sum_i p_i \log_d \left(\frac{p_i}{q_i} \right) - \log_d (c) \\ &= D[p \parallel q] - \log_d (c) \geq 0 \end{aligned}$$

ואי השוויון האחרון נובע מחיוביות האנטרופיה היחסית (המרחק הסטטיסטי – ראה פרק 2) ואי שוויון קראפט על פיו $c \leq 1$. מכאן ששוויון יתקיים רק אם יש שוויון באי שוויון קראפט וגם המרחק הסטטיסטי מתאפס, שאז

$$D[p \parallel q] = 0 \quad \Rightarrow \quad p(x) = q(x) = \frac{d^{-l(x)}}{\sum_{\{x\}} d^{-l(x)}} = d^{-l(x)}$$

וקיבלנו שוב כי אורכי המילים צריכים לקיים

$$(6.30) \quad l(x) = \log_d \frac{1}{p(x)} .$$

כאמור, מאחר ומילים כוללות מספר שלם של סימני-קוד אין בדרך כלל אפשרות לקודד את המילים באופן שארכן יהיה שווה בדיוק ללוג של שכיחותן. אולם תמיד ניתן לקודד מילים כך שיתקיים

$$(6.31) \quad \log_d \frac{1}{p(x)} \leq l(x) \leq \log_d \frac{1}{p(x)} + 1$$

אם נמצע אי-שוויון זה על פי השכיחות של המילים $p(x)$, נקבל

$$(6.32) \quad H_d(x) \leq L(C) \leq H_d(x) + 1$$

דהיינו אורך מילת הקוד הממוצעת חסום בין האנטרופיה ל- "אנטרופיה + 1". כאשר האנטרופיה קטנה, חסם זה יכול להיות גרוע, ונרצה לכן לבנות קוד בו האורך הממוצע של מילת קוד שואף ממש לאנטרופיה.

6.3.3 משפט קידוד המקור של שנון

משפט הקידוד הראשון של שנון קובע כי ניתן תמיד להתקרב אסימפטוטית לקידוד שבו $L(C) = H_d(x)$. הרעיון הבסיסי של ההוכחה הינו קידוד של "בלוקים" הכוללים n סימני מקור. כלומר, אנו מפסקים את המקור באופן כזה שכל מילת מקור הינה רצף של n סימני מקור: $X^{(n)} = (x_1, x_2, \dots, x_n)$ ונבנה קוד כך ש

$$\log_d \frac{1}{p(X^{(n)})} \leq l(X^{(n)}) \leq \log_d \frac{1}{p(X^{(n)})} + 1$$

ואם אנו ממצעים את המשוואה אנו מקבלים, עבור n גדול מספיק, על-פי משפט ה-AEP (ראה סעיף 6.2.3)

$$(6.33) \quad n \cdot H_d(x) \leq \langle l(X^{(n)}) \rangle \leq n \cdot H_d(x) + 1$$

והאורך הממוצע (עבור סימן מקור) $\langle l \rangle = \frac{1}{n} \langle l(X^{(n)}) \rangle$ מקיים:

$$(6.34) \quad H_d(x) \leq \langle l \rangle \leq H_d(x) + \frac{1}{n}$$

כלומר: האורך הממוצע (עבור סימן מקור x) ישאף לאנטרופיה $H_d(x)$. דחיסה מקסימלית של אינפורמציה מתבצעת כאשר אנו "משתמשים בכל התלות הסטטיסטית". לאחר שגמרנו להשתמש בה, הסדרה המקודדת תראה אקראית לחלוטין כיון שמיצינו את כל התלויות הסטטיסטיות שקיימות בסדרה. פעולה זו נקראת לעתים "הלבנה של הערוץ".

מה קורה אם אנו מקדדים מקור שהתפלגותו היא $p(x)$ כאילו היה בעל התפלגות אחרת $q(x)$?

אם בצענו קידוד אופטימלי לפי $q(x)$ ההפרש בין $\langle l \rangle$ לאנטרופיה יהיה

$$\begin{aligned}
 \langle l \rangle - H_d &= \sum_{\{x\}} p(x) \log_d \frac{1}{q(x)} - \sum_{\{x\}} p(x) \log_d \frac{1}{p(x)} \\
 (6.35) \quad &= \sum_{\{x\}} p(x) \log_d \frac{p(x)}{q(x)} \\
 &= D_d[p \parallel q] \geq 0
 \end{aligned}$$

6.4 העברת אינפורמציה בערוץ רועש

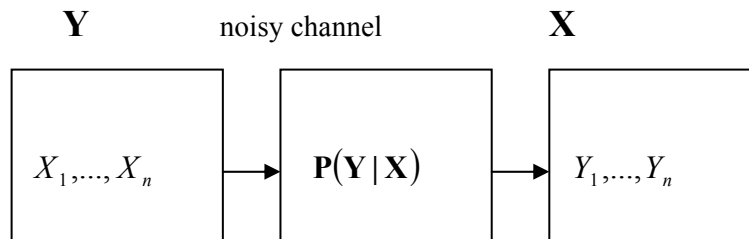
כל תהליך של העברת אינפורמציה בין שתי נקודות חשוף לרעשים: העברת אינפורמציה מנקודה לנקודה היא תהליך פיסיקלי, ותהליכים פיסיקלים אחרים שמתרחשים באותו זמן מתווספים אל האות המועבר כרעש לזוּאי: למשל רעשים אלקטרומגנטיים יכולים להופיע כשלג על מסך הטלוויזיה, פעילות ספונטנית של נירונים יוצרת רעש המתווסף לאות המועבר מהחושים לקליפת המוח וכדומה. אנו נהנים היום מתקשורת כמעט חסינה לרעשים (בעיקר בתקשורת דיגיטלית, כמו בקשר בין מחשבים), ואיננו מרגישים, בדרך כלל, את הרעש הנובע מפעילות ספונטנית של נירונים. עובדה זו נובעת מכך שניתן להעביר אינפורמציה באופן שיהיה חסין לרעשים גם בסביבה רועשת. החסינות לרעשים נוצרת על ידי הוספת אינפורמציה "יתירה" ("Redundant") שתאפשר אחר כך שחזור של ההודעה המקורית למרות ה"רעשים" שנלוו אליה. השפה הטבעית, שנועדה בעיקר להעברת אינפורמציה, התפתחה כך שקיימת בה יתירות של כ-50%. עובדה זו מאפשרת לנו לשחזר טקסט שנכתב באופן בלתי קריא, להשלים את המילה החסרה בסופו של לקרוא עברית בלי ניקוד, לפתור תשבצים וכו'.

הוספת האינפורמציה היתירה מאיטה את קצב העברת האינפורמציה. על מנת להביא לאופטימום את קצב ההעברה עלינו להתאים את כמות האינפורמציה היתירה לרעש. לשם כך נגדיר מדדים כמותיים.

6.4.1 קיבולת של ערוץ Channel Capacity

נתאר ערוץ רועש במודל הסתברותי. מצידו האחד של הערוץ אנו שולחים מילה או אות כלשהו x_j , ובצידו השני של הערוץ מתקבלת המילה או האות y_j , בסיכוי

$$P(y_j | x_j)$$

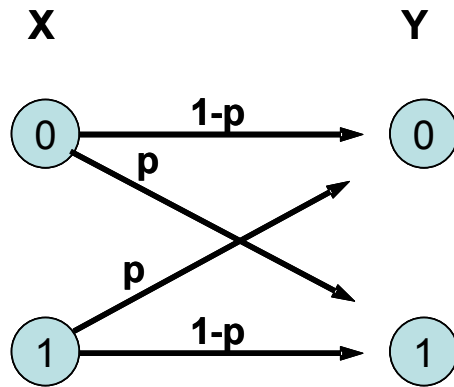


אנו מגדירים את קיבולת האינפורמציה של הערוץ (channel capacity) באופן הבא

$$(6.36) \quad C = \max_{p(x)} I_2(X; Y)$$

דוגמא: ערוץ בינארי סימטרי

ערוץ שבו בסיכוי p מועבר ביט רועש ובסיכוי $q=1-p$ מעבירים את הביט הנכון



נגדיר $p_{00} \equiv P(Y=0|X=0)$ ובאופן דומה עבור p_{01}, p_{10}, p_{11} . ואז לפי ההגדרה נרשום

$$\begin{aligned}
 I_2(X;Y) &= H_2(Y) - H_2(Y|X) \\
 &= H_2(Y) - \sum_{\{x=0,1\}} p(x) \cdot H_2(Y|X=x) \\
 &= H(Y) - \Pr(x=1)(-p_{11} \log p_{11} - p_{01} \log p_{01}) \\
 &\quad - \Pr(x=0)(-p_{10} \log p_{10} - p_{00} \log p_{00}) \\
 &= H(Y) - \Pr(x=1) \cdot (-q \log q - p \log p) - \\
 &\quad - \Pr(x=0) \cdot (-p \log p - q \log q) \\
 &= H_2(Y) - H_2([p]) [\Pr(x=1) + \Pr(x=0)] \\
 &= H_2(Y) - H_2([p]) \\
 &\leq 1 - H_2([p])
 \end{aligned}$$

כאשר אנו מסמנים ב- $H_2([p])$ את האנטרופיה בבסיס 2 של משתנה ברנולי עם הסתברות p .

כעת עלינו למצוא את ההתפלגות $p(x)$ שעבורה יתקבל מקסימום בביטוי לעיל, והמקסימום מתקבל כאשר

$$H_2(Y) = -p(y=1)\log_2 p(y=1) - p(y=0)\log_2 p(y=0) = 1$$

כלומר כאשר $p(y=0) = p(y=1) = \frac{1}{2}$. מכך שהערוץ סימטרי נובע כי גם $p(x=0) = p(x=1) = \frac{1}{2}$, ולכן

$$(6.38) \quad C = 1 - H_2([p]) = 1 - (-p\log_2 p - (1-p)\log_2(1-p))$$

אם $p=0$ או $p=1$ אז $C=1$ ואם $p=0.5$ אז $C=0$ (ואז אנחנו לא יכולים להעביר שום אינפורמציה - תמיד בסיכוי $\frac{1}{2}$ נקבל 1, ובסיכוי $\frac{1}{2}$ נקבל 0).

6.4.2 ערוץ רועש וקוד לתיקון שגיאות

הגדרה: ערוץ רועש חסר זיכרון

ערוץ רועש חסר זכרון הוא ערוץ המקיים

$$(6.39) \quad P(Y^{(n)} | X^{(n)}) = \prod_{i=1}^n P(y_i | x_i)$$

כאשר $X^{(n)} = x_1, x_2, \dots, x_n$ היא מילת מקור המורכבת מ- n סימני מקור. $Y^{(n)} = y_1, y_2, \dots, y_n$ היא המילה המתקבלת בקצה הערוץ כאשר נשלחה המילה $X^{(n)}$. אנו נגביל את הדיון לערוצים רועשים חסרי זיכרון. כאמור, על מנת לאפשר גילוי ותיקון שגיאות עלינו להוסיף אינפורמציה יתרה. במערכת העצבים יתרות מבוטאת לעתים על ידי שיגור מספר גדול של פוטנציאלי-פעולה והעברת אותה אינפורמציה במספר ערוצים במקביל. בתקשורת הדיגיטלית אנו מוסיפים ביטים המאפשרים תיקון השגיאות. דוגמא מוכרת היא המנגנון של "Parity Check" – לכל שביעיית ביטים מוסיפים ביט נוסף כך שמספר ה"1"ים בשמינייה שנוצרה יהיה תמיד זוגי. אם בצדו השני של הערוץ מתקבל מספר אי זוגי של ביטים אז יודעים שנפלה טעות ויכולים לבקש שישלחו לנו שוב את השמינייה. קוד כזה מאפשר גילוי שגיאה, אך לא ניתן להסיק מתוך השמינייה שהתקבלה איזה מהביטים הוא הביט המשובש.

עקרונית, ככל שרמת הרעש בערוץ גדולה יותר - נצטרך להוסיף יותר ביטים שאינם מעבירים אינפורמציה אלא משמשים רק לתיקון השגיאות. אם המספר הכולל של ביטים לשניה (BPS) שניתן להעביר בערוץ הוא חסום (כפי שבד"כ המצב) – אזי הקצב האפקטיבי של העברת האינפורמציה יורד.

דוגמא

נשתמש בקידוד שבו חוזרים על כל ביט שלוש פעמים (שלוש פעמים? שלוש פעמים!), ומחליטים מה היה הביט המקורי על פי הצבעת רוב

$$101 \rightarrow 111,000,111 \rightarrow 101,001,111 \rightarrow 101$$

אז הורדנו את קצב האינפורמציה בערוץ פי שלוש: העברנו תשעה ביטים בעוד שבערוץ שקט היה מספיק להעביר שלושה ביטים.

הגדרה: הקצב של קוד

אם יש לנו קוד שמכיל M מילים שכל אחת מהן היא בת n ביטים – אז הקצב (Rate) של הקוד יהיה

$$(6.40) \quad R = \frac{\log_2 M}{n}$$

דוגמא

בקוד הקודם שהגדרנו יש רק שתי "מילים": 000 (אם הביט המועבר הוא "0") ו-111 (אם הביט המועבר הוא "1") ולכן: $M=2, n=3$ ו-

$$(6.41) \quad R = \frac{\log_2 2}{3} = \frac{1}{3}$$

6.4.3 קודי המינג Hamming Codes

על מנת להבטיח שנדע להבחין בין מילות הקוד השונות לאחר שעברו שיבושים – נרצה שמילות הקוד תהיינה מספיק שונות זו מזו, כך שגם אם ישתבשו כמה ביטים נדע עדיין לאיזו מבין המילים האפשריות התכוון השולח, לפי מילת הקוד הדומה ביותר למילה שנשלחה. נגדיר מרחק בין שתי מילים בינאריות על ידי מספר הביטים השונים בין שתי המילים, מרחק זה נקרא מרחק Hamming. עבור ערוץ בינארי סימטרי בעל הסתברות p לשיבוש ביט יהיו בממוצע pn ביטים משובשים. כשמקודדים בלוקים גדולים (n גדול) התפלגות מספר הביטים המשובשים חדה סביב pn (עד כדי $\pm\sqrt{p \cdot n}$) ולכן אם נוודא שהמרחק המינימלי בין כל שתי מילות קוד יהיה גדול יותר מ $2 \cdot p \cdot n + \alpha\sqrt{p \cdot n}$, (α הוא "מקדם בטחון") נוכל להבטיח כי בסבירות גבוהה כרצוננו נוכל להבחין בין מילות קוד.

ואמנם, ניתן לבנות משפחה של מלות קוד שבה המרחק המינימלי בין המילים גדול כרצוננו. לדוגמא אנו מביאים קבוצה של מלים בנות שבעה ביטים בהן המרחק המינימלי בין כל זוג מלים הוא שלושה ביטים

0000000, 0100101, 1000011, 1100110.
 0001111, 0101010, 1001100, 1101001.
 0010110, 0110011, 1010101, 1110000.
 0011001, 0111100, 1011010, 1111111.

משפט הקידוד השני של שנון

משפט קידוד הערוץ קובע כי בהנתן ערוץ שלו קיבולת C , ניתן לבנות קוד בעל קצב הקרוב ל- C כרצוננו, שיאפשר תיקון כל השגיאות המתקבלות בהעברת האינפורמציה. הקוד האופטימלי מבוסס על קידוד בלוקים גדולים של אינפורמציה.

פורמלית: לכל קצב $R < C$ ולכל $\varepsilon \geq 0$ קיים $n(\varepsilon)$ כך שניתן לממש קוד הכולל $2^{n(\varepsilon)R}$ מילות קוד באורך $n(\varepsilon)$ והסתברות השגיאה (כלומר ההסתברות שכאשר נקבל פלט $Y^{(n)}$ נפענח אותו בתור המילה \hat{w} השונה מהמילה המקורית w) היא קטנה מ- ε

$$(6.42) \quad \Pr \left[g \left(Y^{(n)} \right) = \hat{w} \neq w \right] \leq \varepsilon$$

אם אנו מנסים להעביר אינפורמציה בקצב הגדול מקיבולת הערוץ $R > C$ אז בהכרח תיפולנה שגיאות בפיענוח, וניתן להראות כי

$$(6.43) \quad P_{err} \geq 1 - \frac{C}{R} - \frac{1}{nR}$$

נתאר באופן לא פורמלי את הרעיון הכללי של ההוכחה: כאשר שולחים מילה בת n ביטים, $X_i^{(n)}$, בערוץ היא יכולה כתוצאה מהשיבושים להפוך ל- $2^{nH(Y|X)}$ מילים אפשריות זאת על פי משפט ה-AEP. לדוגמא: אם נשלח את המילה 1...1111111 בערוץ בינארי סימטרי בעל הסתברות שגיאה η יופיעו לנו כ- $\eta \cdot n$ אפסים (אם n מספיק גדול).

על מנת שנוכל להבחין בין מילים שונות שיועברו - נרצה לחלק את מרחב המילים המועברות ל- "כדורים" נבדלים, כך שמרחק Hamming בין המרכזים של שני כדורים גדול מרדיוס הכדורים (במונחים של מרחק Hamming) - כלומר אין חפיפה בין הכדורים. מילות הקוד תהיינה הקואורדינטות של מרכזי ה"כדורים" ומילה שהיא תולדה של שיבוש בערוץ של מילת המקור תיכלל בתוך הכדור המתאים. מספר מילות הקוד המרוחקות מספיק זו מזו יהיה, לפיכך, מספר הכדורים.

ניתן לקבל חסם על מספר הכדורים על ידי חלוקת נפח המרחב בנפח הכדור. נפח המרחב (מספר הסדרות בנות n ביטים האפשריות במרחב Y , משוקללות בהתאם להסתברותן) הוא $2^{nH(Y)}$, ולכן היחס בין נפח המרחב לנפח הכדור הוא

$$(6.44) \quad \frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{nH(Y)-nH(Y|X)} = 2^{nI(Y|X)} \leq 2^{nC}$$

ולכן M , מספר המילים בקוד, קטן או שווה מ- 2^{nC}

$$(6.45) \quad \Rightarrow R = \frac{\log_2 M}{n} \leq C$$

שנן הצליח להראות, כאמור, כי ניתן להתקרב לחסם זה כרצוננו.

בבניית קוד תיקון שגיאות צריך גם להביא בחשבון את זמן הקידוד והפענוח. קוד טוב יהיה בעל זמן קידוד ופענוח לינאריים באורך המילה וקצב קרוב ככל האפשר ל- C . כיום ניתן לקבל קודים בעלי קצב שקטן רק עד כדי קבוע מ- C (עד תחילת שנות ה-70 לא הצליחו לבנות קודים שהקצב שלהם איננו שואף אסימפטוטית לאפס).

תרגילים

1. פונקציות של משתנים מקריים

א. הראו כי אם $H(Y|X) = 0$ אזי Y היא פונקציה של X , כלומר, לכל x עבורו $p(x) \geq 0$, קיים רק ערך יחיד של y עבורו $p(x, y) \geq 0$.

ב. יהי X משתנה מקרי בדיד. הראו כי עבור כל פונקציה $g(X)$ מתקיים כי $H(X) \geq H(g(X))$.

2. יהיו X, Y ו- Z שלושה משתנים מקריים. נטפל בביטוי $Q = I(X; Y) - I(X; Y|Z)$.

א. מצא דוגמא כך ש- Q חיובי וכן דוגמא כך ש- Q שלילי.

ב. הוכח כי מתקיים $Q = I(X, Y; Z) - [I(Y; Z) + I(X; Z)]$.

3. נתונים 12 מטבעות זהים למראה שאחד מהם מזויף (קל או כבד מן השאר).

א. מהי האנטרופיה של המצב המתואר בשאלה ומהו החסם על מספר השקילות באמצעות מאזניים לצורך מציאת המטבע המזויף?

ב. מצאו אלגוריתם למציאת המטבע במספר מינימלי של שקילות.

ג. תארו את האלגוריתם כעץ טרינרי וכקוד רישא.

ד. אם המטבעות צבועים כך שלארבעת המטבעות הירוקים סיכוי אפריורי של 0.6 להיות מזויפים, לארבעת המטבעות האדומים סיכוי 0.4 ולארבעת המטבעות הכחולים סיכוי 0.2 – מה יהיה עכשיו האלגוריתם האופטימלי ומהי תוחלת מספר השקילות הנדרשות.

ה. אותה שאלה, רק שהפעם נסו לגלות גם אם המטבע המזויף כבד או קל מן השאר.

ו. מה עבור 13 מטבעות?

4. הוכיחו את חוק השרשרת עבור אינפורמציה משותפת

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1)$$

5. "ערוץ מוחק סימטרי" "Binary erasure channel" מתאר מצב בו חלק מהביטים

נמחקים: לכל ביט מקור יש סיכוי α להמחק (ואז מגיע בקצה הערוץ סיגנל

"מוחק", אותו נסמן ב- e) ובסיכוי $1 - \alpha$ מגיע הביט המקורי שנשלח (0 או 1)

מהי הקיבולת של ערוץ כזה?

6. נניח כי קודדנו מקור X בעל התפלגות p אך קבענו את הקוד לפי התפלגות אחרת q , דהיינו בחרנו $l(x) = \text{ceil}(\log(1/q(x)))$. הראו כי תוחלת אורך הקוד מקיימת

$$H(p) + D(p \| q) E_p [l(x)] < H(p) + D(p \| q) + 1$$

7. השאלה הנוכחית עוסקת בנושא אמידה של אינפורמציה מתוך מדגם אמפירי. נתון מדגם $n(r, s)$ של פעילויות עצביות אפשריות r_1, \dots, r_l (למשל בניית ירי של תאי עצב) בתגובה להצגה של גירויים מתוך סט סופי s_1, \dots, s_k . בשאלה זו נניח כי המדגם נוצר על ידי התהליך הבא: ראשית מגרילים בהתפלגות אחידה אחד מבין k הגירויים, ואז מגרילים את אחת התגובות לגירוי על פי הסתברות קבועה מראש $p(r|s)$.

א. מהו אומד נראות מירבית $\hat{p}(r, s)$ להתפלגות $p(r, s)$ בהנתן המדגם $n(r, s)$?

ב. יהי $\hat{I}(R; S)$ אומד של האינפורמציה המשותפת $I(R; S)$ המוגדר על ידי

$$\hat{I}(R; S) = D_{KL}[\hat{p}(r, s) \| \hat{p}(r)\hat{p}(s)]$$

$$\hat{p}(s) = \sum_r \hat{p}(r, s) \quad \text{ו-} \quad \hat{p}(r) = \sum_s \hat{p}(r, s) \quad \text{ו-} \quad \hat{p}(r, s) = n(r, s) / n$$

הם ההתפלגות השוליות, ו- n הוא גודל המדגם. הוכח כי אם R ו- S הם בלתי תלויים אזי אומד זה הוא מוטה.

ג. תן הערכה לגודל ההטיה של האומד מהסעיף הקודם, על ידי שימוש בעובדה שניתן לקרב את המרחק הסטטיסטי על ידי פונקציה לינארית של הסטטיסטי χ^2 (ראה תרגיל בפרק 2), המוגדר באופן הבא

$$\chi^2 = \sum_{s=1}^k \sum_{r=1}^l \frac{(n(r, s) - n(r)n(s)/n)^2}{n(r)n(s)/n}$$

ושימוש בכך שהתוחלת של סטטיסטי זה היא $(k-1)(l-1)$.