מבוא לתורת הקודים לתיקון שגיאות

שיעור ראשון – 21/10/2015

בירוקרטיה וכאלה

מרצה: אמיר שפילקה בנבנישתי.
אין תרגול, יהיו תרגילים. יהיו בערך פעם בשבוע-שבועיים, התרגיל הראשון יהיה השבוע. הוא יכלול בעיקר חישובים.
כל התרגילים הם להגשה. הציון הוא 80% בחינה 20% תרגילים.
שעת קבלה בשני ב־11:00 בחדר 118 בשרייבר.
מייל: shpilka@post.tau.ac.il

חומר

הקדמה

במאמר מ־1948 Shannonכתב על התרחיש הבא:
Aliceרוצה לשלוח הודעה ל־Bob. מה קורה כשבתווך יש רעש?
מודל: Binary Symmetric Channel: כשרוצים לשלוח ביט מקבלים בהסתברות $1-p$ את אותו הביט, ובהסתברות p את הביט השונה.
מודל נוסף: noiseless: אם נרצה לשלוח פקס, רוב הפיקסלים שנרצה לשלוח הם לבנים. כלומר מרחב ההודעות שלנו הוא **לא אחיד**. ולכן נרצה לחסוך, כלומר נרצה **לדחוס**.
במצב שבו אנחנו noiseless, יתקיימו מצבים שבהם נרצה לחסוך, ולכן נרצה לבצע compression.
נתמקד במודל noiseless בדוגמא של פקס.
פיקסל לבן: 99%, פיקסל שחור: 1%.
נחשוב על לבן כעל 0 ועל שחור כעל 1.
מסכימים מראש על הקידוד הבא:
נחתוך את הביטים לקבוצות של 10 ביטים. כשהבלוק הוא 10 אפסים, נשלח 0 יחיד. כשיש בו רק אחדות, נשלח 1 ואז את הבלוק.
נשאלת השאלה – **האם חסכנו בקידוד הזה?**
נניח שהפקס המקורי הכיל n פיקסלים. ננסה להבין כמה פיקסלים אנחנו מעבירים באמצעות הקידוד הזה בממוצע.

$$P(\text{block with only zeros}) = 0.99^{10} > 0.9$$

$$P(\text{block with at least one in it}) < 0.1$$

לכן כמות הביטים שנשלח יהיה:

$$\frac{n}{10}(0.9 \cdot 1 + 0.1 \cdot 11) = \frac{n}{5}$$

כלומר, חסכנו **פי 5** בערך.
Shannon שאל: מה האופטימום שאפשר להשיג מבחינת דחיסה?
הוא אפיין **בדיוק** כמה אפשר לחסוך.

אנטרופיה

Ω מרחב סופי. D התפלגות על Ω: $D : \Omega \to [0,1]$. $\sum_{x\in\Omega} D(X) = 1$.
נגדיר את האנטרופיה:

$$h(x) = \log \frac{1}{D(x)}$$

$$H(D) = Eh(x) = \sum_{x\in\Omega} D(X) \log \frac{1}{D(X)} = \sum_{x\in\Omega} -D(x) \cdot logD(X)$$

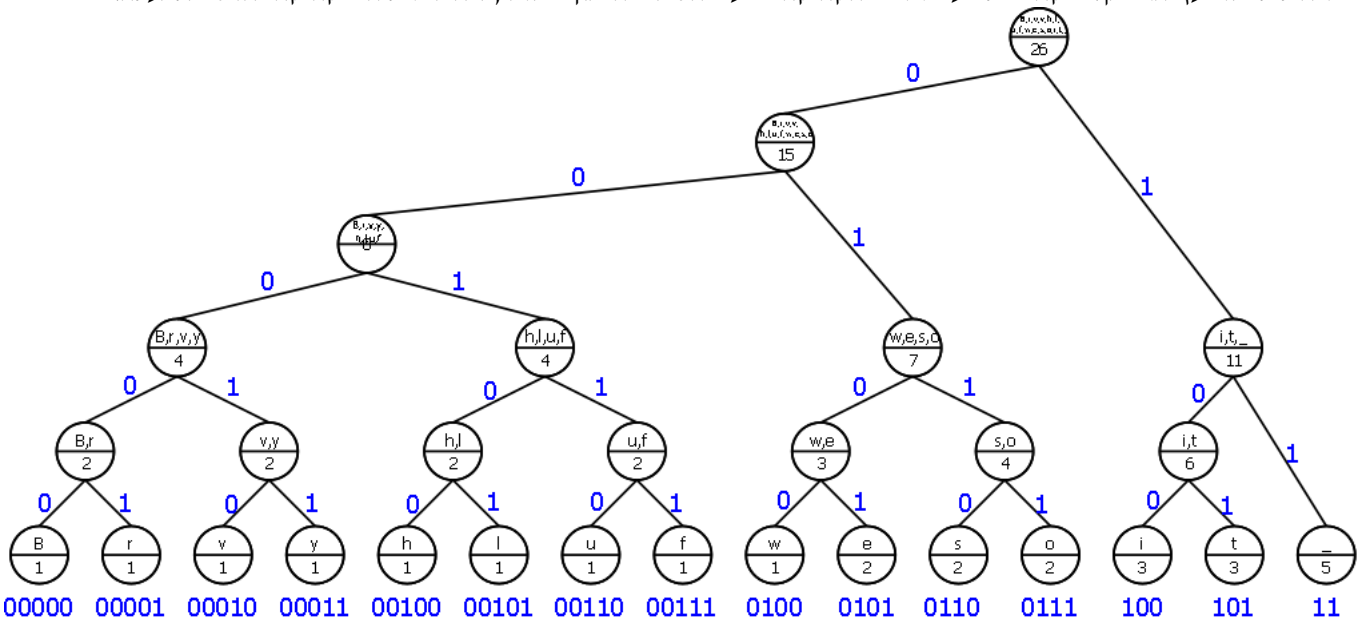
<u>משפט שאנון לערוץ ללא רעש:</u>
לכל D, Ω כנ"ל יש פונקציית $Enc : \Omega \to \{0,1\}^\star$, $Dec : \{0,1\}^\star \to \Omega$ כך שלכל $x \in \Omega$ מתקיים:

$$Dec(Enc(x)) = x$$

וגם

$$\mathbb{E}_{X\sim D}[|Enc(X)|] \in [H(D), H(D)+1]$$

ואי אפשר טוב יותר.
הוכחה: <u>Huffman tree</u>
לשם פשטות, נניח שכל ההסתברויות מהצורה $\frac{1}{2^i}$, $i \in \mathbb{N}$.
נתחיל לבנות עץ האפמן. ניקח בכל פעם את 2 הקודקודים עם הסכום הנמוך ביותר, ונחבר אליהם קודקוד חדש מלמעלה.



נשים לב שאף מחרוזת בקידוד אינה רישא של מחרוזת אחרת.
תכונה מרכזית של הקידוד: איבר שההסתברות שלו היא $\frac{1}{2^i}$ יקבל מחרוזת באורך i.
ניתן להוכיח תכונה זו באינדוקציה על כמות האיברים ב-Ω.
נראה שהבנייה הזו משיגה **בדיוק** את האנטרופיה.

$$\mathbb{E}_{X\sim D} = \sum_x \frac{1}{2^{i_x}} i_x = \sum_x D(x) \log \frac{1}{D(x)} = H(D)$$

במקרה שבו החזקות הן לא כאלה, ניתן לקחת את המספר הכי קרוב אליו מלמעלה מהצורה $\frac{1}{2^i}$, ואז ניתן לראות שאנחנו "מבזבזים" לכל היותר 1+.

Noisy channel

אלפבית הודעות מקור Σ.

אלפבית שהערוץ מציג Γ.

לכל הודעה ב-Σ קיימת הסתברות מסוימת לקבל הודעה מ-Γ.

דוגמא: BSC (Binary Symmetric Channel) עם פרמטר p.

עוד דוגמא: BEC: Binary Erasure Channel עם פרמטר p, בו יש הסתברות שהודעה מגיעה מחוקה במקום לקבל את הביט ההפוך. ואז ניתן לדעת שהייתה בעיה.

משפט שאנון עבור BSC(p)

לכל $0 \leq p < \frac{1}{2}$ קיים קבוע $0 < c < \infty$ וזוג פונקציות Enc, Dec

$$Enc : \{0,1\}^k \to \{0,1\}^n$$

$$Dec : \{0,1\}^n \to \{0,1\}^k$$

$$n = c \cdot k$$

כך שאם בוחרים באקראי הסתברות אחידה $x \in \{0,1\}^k$, מקודדים ל $Enc(x)$ ו"שולחים בערוץ הרועש". מסתכלים על המחרוזת $Enc(x) \oplus \eta$ כאשר $Pr(\eta_i = 1) = p$, $\eta \in \{0,1\}^n$ אז בהסתברות גבוהה

$$Dec(Enc(x) + \eta) = x$$

עבור BSC מספיק לקחת $c > \frac{1}{1-H(p)}$. (כאשר $H(p) = -p \cdot \log p - (1-p) \cdot \log(1-p)$).

הוכחה:

לכל $x \in \{0,1\}^k$ נבחר את $Enc(x)$ באקראי מתוך $\{0,1\}^n$.

נגדיר Hamming distance:

$$u, v \in \{0,1\}^n, dist(u,v) = |\{i | u_i \neq v_i\}|$$

$$Dec(y) = x : dist(y, Enc(x)) \text{ is minimal}$$

(כשיש כמה מחרוזות מתאימות, נבחר אחת מהן).

נקבע איזשהו $x \in \{0,1\}^k$ ואת $Enc(x)$.

נגדיר $y = Enc(x) \oplus \eta$

1. בהסתברות גבוהה המשקל של η הוא לכל היותר $(p+\varepsilon)n$.

$$wt(\eta) = |\{i | \eta_i = 1\}| = dist(\eta, 0)$$

2. נראה שההסתברות שאיזשהו $x' \neq x$ מקיים $dist(Enc(x'), y) \leq (p+\varepsilon)n$ היא "קטנה".

נוכיח את 1:

1.

$$Pr\left(wt(\eta) > (p+\varepsilon)n\right) \overset{chernoff}{<} e^{-\frac{\varepsilon^2}{3}n}$$

2. נקבע x' כלשהו.

$$Pr\left(dist\left(Enc\left(x\right), y\right) \leq \left(p+\varepsilon\right)n\right) = \frac{Vol\left(Ball\left(y, \left(p+\varepsilon\right)n\right)\right)}{2^n}$$

ולכן ההסתברות שאיזשהו x' יפול ב"כדור" היא לכל היותר $\frac{2^k \cdot Vol(Ball(y,(p+\varepsilon)n))}{2^n}$.

$$Vol(Ball(0,r)) = \binom{n}{0} + \binom{n}{1} + ... + \binom{n}{r} = 2^{H(\frac{r}{n})n + O(\log n)}$$

אפשר להראות את זה עם קירוב סטירלינג.
ולכן

$$Vol(Ball(y,(p+\varepsilon)n)) = 2^{n(\frac{k}{n} + H(p+\varepsilon) - 1 + O(1))}$$

אם אנחנו בוחרים את n בצורה הבאה:

$$n > \frac{k}{1 - H(p+\varepsilon)}$$

אנחנו מקבלים את התוצאה הרצויה.

$$Pr_{E,\eta}\left(Dec\left(Enc\left(x\right) \oplus \eta\right) \neq x\right) < e^{-\frac{\varepsilon^2}{3}n} + 2^{-n(1-\frac{k}{n} - H(p+\varepsilon) - o(1))}$$

לכל $c > \frac{1}{1-H(p)}$ אם $n = c \cdot k$ אז ההסתברות שניכשל בפענוח של x היא קטנה מהחלק הימני של האי שוויון האחרון עבור איזשהו $\varepsilon' > 0$.
היחס $\frac{k}{n}$ נקרא **קצב**. $1 - H(p)$ נקרא הקיבול של $BSC(p)$.

<u>משפט:</u>
לכל $0 < p < \frac{1}{2}$ ולכל $0 < \delta$ יש n_0 כך שאם $n \geq n_0$ וגם $k \geq (1 - H(p)) + \varepsilon)n$ אז לכל זוג פונקציות

$$E : \{0,1\}^k \to \{0,1\}^n, D : \{0,1\}^n \to \{0,1\}^k$$

מתקיים:

$$Pr_{\eta, x \in \{0,1\}^k}\left(D\left(E(x) \oplus \eta\right) = x\right) < \delta$$