מבוא לתורת האינפורמציה

עפ"י הרצאות של פרופ' מיכאל וורמן

סמסטר א', תש"ע

רשם: שיר פלד, באמצעות LyX גרסה 1.6.1
תיקונים יתקבלו בברכה במהלך ההפסקות או בכתובת מייל *shirpeled@cs*
ספרים:

Elements of Information Theory / Cover, Thomas

Information Theory Inference and Learning Theory / Mackay

1 שיעור 1

חידה: מיכאל בוחר שני מספרים שונים, מציג לנו את אחד מהם ובהסתברות גדולה מחצי עלינו לנחש אם זהו הגדול מבין השנים או הקטן.
איך עושים זאת? נשאיר חידה זו פתוחה ונענה עליה בהמשך הקורס (אולי).
נענה על שתי שאלות מרכזיות במהלך הקורס:

1. כמה אפשר לדחוס? והתשובה היא H - אנטרופיה.
2. כמה אפשר להעביר על ערוץ רועש? C - קיבולת.

במאמר של $Shannon$ משנות הארבעים יש כבר חסמים עליונים ותחתונים לשתי השאלות דלעיל. הסיבה היחידה עדיין לחקור את התחום (שכן השאלות העקרוניות פתורות) היא שהאלגוריתמים היו די כבדים, ואפילו אלו שופרו בשנים האחרונות.

הערה 1.1 בשלב זה נדבר על התפלגויות בדידות בלבד, בהמשך (עוד חודש וחצי) נגיע להתפלגויות רציפות.

אנחנו רוצים לכמת, עבור התפלגות נתונה, את מידת אי הודאות. במקרה הכי קל - כאשר ערך אחד מקבל הסתברות 1 וכל השאר 0 - הודאות גבוהה מאד. כאשר יש ערכים רבים ולכולם אותה הסתברות - אי הודאות גבוהה.

הגדרה 1.2 אנטרופיה $(Entropy)$- $H(X) = -\Sigma p(x)\log(p(x))$ כלומר האנטרופיה של המשתנה המקרי היא התוחלת של $-\log$ ההסתברות של x. המינוס בהגדרה הוא כדי לאזן את העובדה שלוג של מספר בין 0 ל 1 הוא שלילי. כמו כן לצרכי הגדרה נאמר ש $0\cdot\log 0 = 0$ כיוון שבגבול זה נכון, ואחרת זה יעשה לנו בעיות בסכימה. היחידות של אנטרופיה הן ביטים.

משפט 1.3

$$H(X) \geq 0$$

הוכחה: ברור מההגדרה ■

ניזכר בהתפלגות הבינומית:

$$X = \begin{cases} 1 & p \\ 0 & 1-p \end{cases}$$

ובמקרה כזה נקבל:

$$H(X) = -p \cdot \log(p) - (1-p)\log(1-p) = 1bit$$

שנסמנה ע"י $H(p)$.
במקרה יותר מורכב, של ההתפלגות:

$$X = \begin{cases} a & \frac{1}{2} \\ b & \frac{1}{4} \\ c & \frac{1}{8} \\ d & \frac{1}{8} \end{cases}$$

נקבל:

$$\begin{aligned} H(X) &= -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{2}{8}\log\left(\frac{1}{8}\right) \\ &= \frac{1}{2} + \frac{1}{2} + \frac{3}{4} = \frac{7}{4}bit \end{aligned}$$

(נזכור שאנחנו עובדים עם לוגריתם בבסיס 2)
במקרה כזה נרצה לקודד את a ב $-\log\left(\frac{1}{2}\right)$ ביטים, את b ב $-\log\left(\frac{1}{4}\right)$ ביטים וכן הלאה. למשל נוכל לבחור:

a	0
b	10
c	110
d	111

וזה קוד אופטימלי מבחינת נפח הקידוד, שכן נוכל להבטיח שהוא יהיה באורך של $\frac{7}{4}$ כפול מספר האותיות.
לכן אם נרצה, למשל, לשלוח משהו בשפה האנגלית, נקודד אותו לפי לוח תפוצת האותיות בטקסט, כיוון שזו בערך ההסתברות להתקל באות הטקסט אקראי.

אנטרופיה של שני משתנים

נניח שיש לנו שני משתנים מקריים X, Y ונגדיר אנטרופיה לפי האנטרופיה של ההתפלגות המשותפת:

$$H(X,Y) = -E[\log(p(X,Y))]$$

הגדרה 1.4 אנטרופיה מותנית:

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

זה אומר בעצם מה רמת האי וודאות של Y בהנתן שידוע X.

משפט 1.5 כלל השרשרת - $Chain\ Rule$

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ & \quad H(Y) + H(X|Y) \end{aligned}$$

הוכחה:

$$\log(p(x,y)) = \log(p(x)) + \log(p(y|x))$$

■

הגדרה 1.6 אנטרופיה יחסית $(KL\ Divergence)$:

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

יש בעיית הגדרה כאשר $q(x) = 0$ ו $p(x) \neq 0$, ואז נאמר שזה אינסוף. צורה אחרת לכתוב זאת היא $E_p\left[\log\left(\frac{p(x)}{q(x)}\right)\right]$.

דוגמא: אם אנחנו הולכים לקודד לפי התפלגות q, וההתפלגות האמיתית היא p - הקידוד שלנו יעלה בסופו של דבר $D(p||q)$ יותר ממה שחישבנו, כלומר $H(p) + D(p||q)$.
אינפורמציה משותפת $I(X;Y)$

$$I(X;Y) = \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

זה נותן לנו בעצם את רמת הקשר בין המשתנים הללו, כך למשל כאשר הם בלתי תלויים נקבל 0.

הערה 1.7 זה גם שווה ל $D(p(x,y)\,||p(x)\,p(y))$ וגם ל $E_{p(x,y)}\left[\log\left(\frac{p(X,Y)}{p(X)p(Y)}\right)\right]$.

מה הקשר בין I ובין H?

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

$$= \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

מכיוון ש $p(x,y) = p(y)\,p(x|y)$ נוכל לקבל:

$$= \sum p(x,y) \log \frac{p(y)\,p(x|y)}{p(x)\,p(y)} = \sum p(x,y) \log\left(\frac{p(x|y)}{p(x)}\right)$$

לפי כלל השרשרת:

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

ולכן:

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

משפט 1.8 כלל השרשרת ל n-יות:

$$\begin{aligned} H(X_1, X_2, ..., X_n) &= \sum_{i=1}^{n} H(X_i | X_{i-1}, ..., X_1) \\ &= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) ... \end{aligned}$$

הוכחה: באינדוקציה באופן דומה למקרה של שני משתנים. ■

מסקנה 1.9 כלל השרשרת לגבי I:

$$\begin{aligned} &I(X_1, X_2, ..., X_n; Y) \\ &= \sum I(X_i; Y | X_{i-1}, X_{i-2}, ..., X_1) \end{aligned}$$

2 שיעור 2

נזכיר - פונקציה קמורה $(Convex)$ היא פונקציה המקיימת $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$

משפט 2.1 (מהסתברות) אי שויון ינסן $(Jensen)$:

$$E[f(x)] \geq f(E[X])$$

עבור f קמורה.
עבור fקעורה זה נכון כאשר אי השוויון בכיוון השני.

הוכחה: עבור משתנה בינומי זה נובע מההגדרה. נראה באינדוקציה על משתנה (בדיד) שמקבל n ערכים.
עבור n משתנים נגדיר $p_i' = \frac{p_i}{1-p_n}$

$$\sum p_i f(x_i) = p_n f(x_n) + (1-p_n) \sum p_i' f(x_i)$$

ועבור הביטוי הימני ביותר - מתקיימת הנחת האינדוקציה. ■

טענה 2.2 $D(p||q) \geq 0$

הוכחה: כמובן כאשר $p = q$ נקבל שזה 0, לכן מספיק להראות שזה מינימום של הפונקציה כדי לקבל את הטענה.

$$D(p||q) = \sum p_i \log \frac{p_i}{q_i}$$

כאשר נסמן $p_i = p(x_i)$.
צריך להתקיים התנאי שההסתברויות על וקטור המ"מ $(x_1, x_2, ...)$מסתכמות ל 1, כלומר $\sum p_i = \sum q_i = 1$. נגזור את הפונקציה $G(p) = \sum p_i \log \frac{p_i}{q_i} + \lambda(\sum p_1 - 1)$ כדי לקבל את נקודת המינימום.

$$\begin{aligned} \frac{\partial G(p)}{\partial p_i} &= \log \frac{p_i}{q_i} + \frac{q_i}{p_i} \cdot p_i + \lambda = 0 \\ &= \log(p_i) - \log(q_i) + 1 + \lambda \end{aligned}$$

ולכן

$$\log(p_i) = \log(q_i) + 1 + \lambda$$

וכיוון שהלוגים משתווים עד כדי קבוע, מהסתכמות ההסתברויות ל 1 נובע שהקבוע הוא בדיוק 1 ולכן לכל i מובטח

$$p_i = q_i$$

ואכן הפונקציה מקבלת מינימום כאשר $p = q$ ומתקיימת הטענה.
ניתן להוכיח גם עם אי שוויון ינסן:

$$\begin{aligned} -D(p||q) &= -\sum p_i \log \frac{p_i}{q_i} = \sum p_i \log \frac{q_i}{p_i} \\ &\leq \log \sum p_i \cdot \frac{q_i}{p_i} = \log\left(\sum q_i\right) = \log 1 = 0 \end{aligned}$$

■

הערה 2.3 אי שוויון ינסן הוא חלש אם"ם $p_i = q_i$ ולכן באמת מתקיים הדרוש.

מתקבל באופן ישיר $I(X,Y) \geq 0$ כיוון שכפי שצוין בעבר מתקיים $D(p(x,y)||p(x)p(y)) = I(X,Y)$.

הערה 2.4 $H(X) \leq \log(|X|)$ (עבור משתנה דיסקרטי) כי ההתפלגות עם רמת האנטרופיה הגבוהה ביותר היא ההתפלגות האחידה, שבה ההסתברות לכל איבר היא $\frac{1}{|X|}$ ומהצבה בנוסחה של H מתקבל המבוקש.
$u = \frac{1}{|X|}$ משתנה קבוע.
נקבל $D(p||u) = \sum p(x) \log \frac{p(x)}{u(x)} = \log|X| - H(X)$ וזה מוכיח את ההערה.
כמו כן מתקיים:

$$H(X|Y) \leq H(X)$$

לפי ההגדרה של I באמצעות H.
ניישם את אי השוויון שקיבלנו לכלל השרשרת ונקבל:

$$H(X_1, X_2, ..., X_n) = \sum H(X_i|X_{i-1}, ..., X_1) \leq \sum H(X_i)$$

נזכיר שרשראות מרקוב. לפי תכונת האי־תלות של השרשראות מתקיים:

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y) \cdot p(z|y)$$

טענה 2.5 אי שוויון עיבוד המידע:
נניח שבשלב הראשון של תהליך נתבונן בעולם X ואז נרשום את המידע במחשב ונקבל את Y, כלומר $f(X) = Y$ ואז נעבד את המידע כדי לקבל Z. נראה ש $I(X;Y) \geq I(X;Z)$ כלומר - האינפורמציה המשותפת של המידע X, Y גדולה יותר מהאינפורמציה X, Z, במילים אחרות - המידע המעובד לא מספק לנו יותר מידע על העולם מהמידע שאינו מעובד, בסך הכולל.

הוכחה:

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0} \\ &= I(X;Y) + \underbrace{I(X;Z|Y)}_{=0} \end{aligned}$$

ולכן

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

מדוע האחרון הוא 0? מפני שאמרנו קודם שבשרשרת מרקוב מתקיים $p(x,z|y) = p(x|y)\,p(z|y)$, כלומר בהנתן y מתקיים ש x ו z הם בלתי תלויים, ולכן האינפורמציה המשותפת שלהם היא אפס. ■

יש לנו מטבע מוטה כלשהו, ויש תוצאות של הטלות (בלתי תלויות) $X_1,...,X_n$ ונרצה ליצור סדרה אחרת באורך k: $Z_1,...,Z_k$, של "תוצאות של הטלות" שכן תתפלג כאילו ההטלות הן של מטבע הוגן. כמובן מידת ההטיה של המטבע הראשון תקבע חסם על אורך הסדרה השניה שאנחנו יכולים ליצור. למשל - אם המטבע מוטה לקבל תמיד 1 (בהסתברות 1) אז לא נוכל ליצור סדרה כנ"ל, ואם הוא הוגן - נוכל ליצר סדרה כך ש $k = n$. מידת האנטרופיה לכל הטלה היא שווה והאנטרופיה המשותפת היא מכפלת האנטרופיות ולכן:

$$\begin{aligned} n \cdot H(p) = H(X_1,...,X_n) &\geq H(Z_1,Z_2,...,Z_k,k) \\ &= H(k) + H(Z_1,...,Z_k|k) \end{aligned}$$

כאן אנחנו אומרים ש k והסדרה Z_i הם תוצאה של עיבוד המיידע ואי השוויון שהסתמכנו עליו הוא אי שוויון עיבוד המידע. נשים לב כי $H(Z_1,...,Z_k|k)$ זוהי האנטרופיה של k הטלות של מטבע הוגן, כלומר של k ביטים בלתי תלויים, וזוהי בדיוק התוחלת של k ולכן:

$$n \cdot H(p) = H(k) + E(k)$$

אי שוויון פנו $(Fano)$

יש לנו שרשרת מרקוב $X \to Y \to \hat{X}$, כלומר יש מידע X שאנחנו מקבלים בצורת Y ואנחנו צריכים לנחש מה היה X במקור. הניחוש שלנו הוא $\hat{X}$, וננסה להעריך את ההסתברות לשגיאה (זהו משתנה ברנולי):

$$P_e = Pr\left(\hat{X} \neq X\right)$$

נקבל

$$H(P_e) + P_e \log|X| \geq H\left(X|\hat{X}\right) \geq H(X|Y)$$

אי השוויון הימני הוא פשוט יישום של אי־שוויון עיבוד המידע. וכיוון ש $0 \leq H(P_e) \leq 1$ מתקיים

$$1 + P_e \log|X| \geq H(X|Y)$$

ולכן

$$P_e \geq \frac{H(X|Y) - 1}{\log|X|}$$

הוכחה: נגדיר מ"מ $H(P_e) + P_e \log|X| \geq H\left(X|\hat{X}\right)$ ראשית לכל נוכיח את החלק הראשון שאומר

$$E = \begin{cases} 1 & \hat{X} \neq X \\ 0 & \hat{X} = X \end{cases}$$

ואז:

$$\begin{aligned} H\left(E,X,\hat{X}\right) &= H\left(X|\hat{X}\right) + \underbrace{H\left(E|X,\hat{X}\right)}_{=0} \\ &= \underbrace{H\left(E|\hat{X}\right)}_{\leq H(P_e)} + \underbrace{H\left(X|E,\hat{X}\right)}_{\leq Pr(E=1)\cdot\log|X|} \end{aligned}$$

הסברים:
הביטוי $H\left(E|X,\hat{X}\right)$ הוא 0 מכיוון שמרגע שנודע לנו X ו $\hat{X}$ברור לנו מה יהיה E.

כמו כן $H\left(E|\hat{X}\right) \leq H\left(E\right) = H\left(P_e\right)$ (אי השוויון הוא מכיוון שללא מידע על $\hat{X}$ בוודאי האנטרופיה עולה, ולכל היותר שווה).

$$H\left(X|E,\hat{X}\right) = \underbrace{Pr\left(E=0\right)H\left(X|\hat{X},E=0\right)}_{=0} + Pr\left(E=1\right)H\left(X|\hat{X},E=1\right)$$

הביטוי הראשון הוא כאשר ידוע שצדקנו - ואז ברור לנו מה יהיה X (זהו בדיוק $\hat{X}$).
הביטוי השני הוא כאשר ניחשנו פעם אחת וטעינו, כלומר יש לנו ביד $\hat{X}$ כלשהו שידוע שאינו X, אז כרגע כדי לנחש את X יש רק $|X|-1$ אפשרויות. החסם על אנטרופיה הוא כזכור $\log$ של גודל הקבוצה ולכן בהנתן ניחוש אחד שגוי, גודל הקבוצה האפשרית הוא $|X|-1$ ואז האנטרופיה היא לכל היותר $\log\left(|X|-1\right)$ ומסיכום אלו נובע:

$$H\left(X|E,\hat{X}\right) \leq Pr\left(E=1\right)\cdot\log\left(|X|-1\right) \leq Pr\left(E=1\right)\log|X| = P_e\log|X|$$

■

דוגמא 1:

נניח שיש לנו שני משתנים בלתי תלויים X, X' בעלי אותה התפלגות ונרצה להראות חסם להסתברות שהם שווים ע"י:

$$Pr\left(X=X'\right) \geq 2^{-H(X)}$$

כמובן:

$$Pr\left(X=X'\right) = \sum_x p^2\left(x\right)$$

ולכן:

$$\underbrace{2^{E[\log(p(x))]}}_{2^{-H(X)}} \leq \underbrace{E\left[2^{\log(p(x))}\right]}_{\sum p^2(x)}$$

זהו אי שוויון ינסן כי 2^x היא פונקציה קמורה.

דוגמא 2:

נניח שיש שני משתנים מקריים $X \sim p$ ו $X' \sim r$ ואז:

$$p\left(X=X'\right) \geq 2^{-H(p)-D(p||r)}$$

מדוע?

$$\begin{aligned} 2^{-H(p)-D(p||r)} &= 2^{\sum p\cdot\log(p)+\sum p\cdot\log\frac{r}{p}} \\ &= 2^{\sum p\cdot\log(r)} \\ &\leq \sum p\cdot 2^{\log(r)} = \sum p\cdot r \end{aligned}$$

3 שיעור 3

חידה

נזכיר את החידה מתחילת הקורס: מיכאל בוחר שני מספרים שונים, מציג לנו את אחד מהם ובהסתברות גדולה מחצי עלינו לנחש אם זהו הגדול מבין השנים או הקטן.

איך עושים זאת?

בהנתן x - המספר שמיכאל נתן לנו, נגריל מספר נוסף, אם הוא קטן מ x - נכריז ש x הוא הגדול, ואחרת - נכריז ש x הוא הקטן. הרעיון הוא שאם המספר שהגרלנו z יצא גדול יותר משני המספרים שמיכאל הגריל - אין לנו אינפורמציה נוספת ונטעה בהסתברות חצי, כנ"ל אם z קטן משני המספרים. אבל אם במקרה z יצא באמצע - נקבל בוודאות את התשובה המלאה, ואם ההסתברות ש z יפול באמצע היא חיובית - הבטחנו שההסתברות שלא נטעה תהיה גדולה ממש מחצי.

עוד חידה

נתונים 100 מטבעות, כאשר 1 או 2 מהם עם משקל שונה מכל האחרים (אבל עם אותו משקל לשניהם). צריך להוכיח שיש בדיוק 2 (ולא 1), מבלי לגלות מיהם המטבעות הללו.

איך עושים זאת? מחלקים לשתי קבוצות של חמישים ושמים מטבע דפוק בכל קבוצה, מהשוויון בין שתי הקבוצות ברור שאין רק מטבע אחד דפוק.

נניח שחילקנו לשתי קבוצות ובחרנו מטבע מכאן ומטבע מכאן, מה ההסתברות שקלענו למטבעות המקולקלים? $\frac{1}{2500}$

בהנתן קבוצה של 100 מטבעות, נבחר שניים אקראית, מה ההסתברות שקלענו למטבעות המקולקלים? $\frac{1}{\binom{100}{2}}$, שזה הרבה יותר.

וחידה שלישית

יש n אנשים, שלכל אחד יש פיסת מידע, וכולם מדברים עם כולם בטלפון (רק שניים על כל קו) עד שכולם יודעים את הכל. מה מספר השיחות הטלפון המינימלי?

ניחוש: צריך לפחות $n-1$ ובוודאי אפשר בפחות מ $2n-1$.

דחיסת מידע - $Asymptotic\,Equipartition\,Property$

ההתפלגות שהכי קל לנו לקודד היא התפלגות אחידה, מדוע? נניח שיש התפלגות שנותנת 1 בהסתברות 0.999 ו 0 בהסתברות 0.001, אז אם נשתמש בביט אחד, בעצם "נבזבז" את הביט שהקצינו ל-0 מכיוון שההסתברות שנקבל אותו קטנה מאד.

הרעיון - ניצור n-יות מהקידוד, שהן שוות הסתברות ומכילות את רוב מרחב ההסתברות, ולכן הקידוד יהיה חסכוני ככל האפשר.

הגדרה 3.1 התכנסות בהסתברות:

$$\forall \varepsilon > 0 : Pr\left(|X_n - X| > \varepsilon\right) \to 0$$

טענה 3.2 אם $X_1, X_2, ..., X_n$ הם משתנים $iid \sim p\left(X\right)$ אז"ז (אז ורק אז)

$$-\frac{1}{n}\log\left(p\left(X_1, ..., X_n\right)\right) \to H\left(X\right)$$

הוכחה: מאי תלות נובע:

$$-\frac{1}{n}\log\left(p\left(X_1, ..., X_n\right)\right) = -\frac{1}{n}\sum \log p\left(X_i\right)$$

ומחוק המספרים הגדולים:

$$\rightarrow -E\left[\log p\left(X\right)\right]=H\left(X\right)$$

■

הגדרה 3.3 קבוצה טיפוסית $A_{\varepsilon}^{(n)}=\left\{\left(x_1,...,x_n\right):2^{-n(H(X)+\varepsilon)}\leq p\left(x_1,...,x_n\right)\leq 2^{-n(H(X)-\varepsilon)}\right\}$

טענה 3.4 :

1. ע"י לקיחת לוג משני האגפים של ההגדרה נקבל:

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

2. עבור n מספיק גדול:

$$Pr\left(A_{\varepsilon}^{(n)}\right)>1-\varepsilon$$

3.

$$\left|A_{\varepsilon}^{(n)}\right|\leq 2^{n(H(X)+\varepsilon)}$$

4.

$$\left|A_{\varepsilon}^{(n)}\right|\geq\left(1-\varepsilon\right)2^{n(H(X)-\varepsilon)}$$

הוכחה: .
לטענה 2 חוק המספרים הגדולים:

$$Pr\left(\left|-\frac{1}{n}\log\left(p\left(x_1,...,x_n\right)-H\left(X\right)\right)\right|<\varepsilon\right)\geq 1-\varepsilon$$

לטענה 3: אם z היא n־יה אז:

$$1=\sum p\left(z\right)\geq\sum_{z\in A_{\varepsilon}^{(n)}}p\left(z\right)\geq\sum_{z\in A_{\varepsilon}^{(n)}}2^{-n(H(X)+\varepsilon)}$$

כיוון שהסתברות היחידונים בקבוצה $A_{\varepsilon}^{(n)}$ היא גדולה מ $2^{-n(H(X)+\varepsilon)}$ ולכן מתקבל:

$$1=\left|A_{\varepsilon}^{(n)}\right|2^{-n(H(X)+\varepsilon)}$$

לטענה 4: ידוע לנו שעבור n מספיק גדול

$$Pr\left(A_{\varepsilon}^{(n)}\right)>1-\varepsilon$$

ומכיוון שההסתברות לכל יחידון קטנה מ $2^{-n(H(X)-\varepsilon)}$ אז האגף השמאלי קטן או שווה ל:

$$\sum_{z\in A_\varepsilon^{(n)}} 2^{-n(H(X)-\varepsilon)} = \left|A_\varepsilon^{(n)}\right| \cdot 2^{-n(H(X)-\varepsilon)}$$

■

כיוון שהקבוצה של ה n־יות שעלינו לקודד היא "בערך" בגודל $2^{n\cdot H(X)}$, אז אפשר לקודד אותה עם $n\cdot(H(X)+1)$ ביטים. את מה שמחוץ לקבוצה $A_\varepsilon^{(n)}$ נכנה X, ונקודד את חלק זה של המרחב ע"י $n\cdot(\log X+1)$ ביטים. נוסיף ביט אחד בתחילת המילה שמספר לנו האם המילה בתוך הקבוצה $A_\varepsilon^{(n)}$ או מחוצה לה (היינו בתוך X).
נחשב את תוחלת מספר הביטים בקידוד שלנו :

$$\sum_{z\in A_\varepsilon^{(n)}} p(z)\left[n(H(X)+\varepsilon)+2\right] + \sum_{z\notin A_\varepsilon^{(n)}} p(z)\left(n\log|X|+2\right)$$
$$\leq n(H(X)+\varepsilon)+2+\varepsilon\left(n\cdot\log|X|+2\right)$$

כי כאמור ההסתברות לקבל משהו מחוץ לקבוצה היא לכל היותר ε, וההסתברות לקבל משהו בתוך הקבוצה היא לכל היותר 1.
ואם נבחר ε' מספיק קטן כדי שהביטוי $\varepsilon'\cdot\log|X|$ יהיה זניח נקבל שהנ"ל נותן בערך

$$n(H(X)+\varepsilon)$$

קצבי אנטרופיה בתהליכים סטוכסטיים

הגדרה 3.5 תהליך סטוכסטי סטאציונרי $(Stationary)$ אם

$$P(X_i=a, X_j=b, ...) = P(X_{i+k}=a, X_{j+k}=b, ...)$$

כלומר אם ההסתברות לשרשרת היא קבועה, בלי משמעות להזזה.

.

הגדרה 3.6 קצב אנטרופיה $Entropy-Rate$

$$H(X) := \lim_{n\to\infty}\frac{1}{n}H(x_1,...,x_n)$$

אם הגבול קיים. מצד ימין זהו H לפי הגדרת האנטרופיה שאנחנו מכירים, ומצד שמאל אנחנו מגדירים משמעות חדשה ל H ביחס לתהליך סטוכסטי X (להבדיל ממשתנה מקרי).

דוגמא:

ניקח משתנים בלתי תלויים שווי התפלגות (iid): $X_1, X_2, ..., X_n$ ואז

$$H(X) = \lim_{n\to\infty}\frac{1}{n}H(x_1,...,x_n)$$
$$= \lim_{n\to\infty}\frac{1}{n}nH(X_1) = H(X_1)$$

(מצד ימין זהו ה H הישן)

דוגמא 2:

יהיו $X_1, ..., X_n$ משתנים בלתי תלויים (id) אבל לא בהכרח בעלי אותה התפלגות, ואז:

$$H(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, ..., X_n)$$
$$= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i)$$

אבל לא בטוח שזה מתכנס, זה תלוי בהתפלגויות השונות וכיצד הן מסתדרות אלו עם אלו...

הגדרה 3.7

$$H'(X) := \lim_{n\to\infty} H(X_n|X_1, ..., X_{n-1})$$

באופן אינטואיטיבי זו הערכה למידה שבה ידיעת העבר מסייעת לחזות את העתיד. נראה עוד מעט שזה שקול ל H שהגדרנו קודם.

יהי X תהליך סטציונרי, ואז ברור ש:

$$H(X_n|X_{n-1}, ..., X_1) \leq H(X_n|X_{n-1}, ..., X_2) = H(X_n|X_{n-2}, ..., X_1)$$

כי בעצם אנחנו מקטינים את הידע שלנו, ולכן האנטרופיה עולה. באופן דומה זה נכון אם נמשיך להאריך את הסדרה (נמשיך להקטין את האנטרופיה) ומכיוון ש H חסומה מלרע ע"י 0 - הסדרה מתכנסת.

הגדרה 3.8 מפוצץ צ'זארו
אם יש $a_n \to a$ אם נגדיר $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$ אז נקבל $b_n \to a$

נסתמך על ממוצע צ'זארו ואז מתקבל מכלל השרשרת:

$$\frac{1}{n} H(X_1, ..., X_n) = \frac{1}{n} \sum H(X_i|X_{i+1}, ..., X_1)$$

וצד ימין שווה ל H' בגבול ע"י ממוצע צ'זארו, ולכן בתהליך סטציונרי מתקבל ש $H = H'$.

הערה 3.9 בתהליך מרקובי סטציונרי מתקבל:

$$H(X) = H'(X) = \lim_{n\to\infty} H(X_n|X_{n-1}) = H(X_2|X_1)$$

אם התהליך הוא אי־פריק ולא־מחזורי (לא נתעכב על ההגדרה הזו כרגע), אחרי מספיק "זמן" התהליך מגיע להתפלגות קבועה μ. ואז

$$H(X) = -\sum_{i,j} \mu_i p_{i,j} \log(p_{i,j})$$

כלומר אנחנו סוכמים את ההסתברות להיות במצב i כפול ההסתברות לעבור ל j כפול לוג של ההסתברות הזו, וכיוון שעברנו על כל הזוגות, עברנו על כל מרחב ההסתברות וזה באמת מתאים להגדרה של H.

4 שיעור 4

נתבונן בהילוך מקרי על גרף לא מכוון. נניח שמקודקוד i יוצאת צלע לקודקוד j, אז משקלה w_{ij}, ונגדיר שההסתברות ללכת על הצלע הזו מ i ל j היא w_{ij} חלקי משקל כל הצלעות היוצאות מ i, שנסמנו w_i. נאמר ש W הוא סך כל המשקלות, ואז $\sum_i w_i = 2W$ כיוון שסופרים כל צלע פעמיים. אז $\mu_i = \frac{w_i}{2W}$ התפלגות סטציונרית.
אם נגדיר את מטריצת המעברים מבעיית הגרף הקודמת ע"י P, אז:

$$\sum_i \mu_i P_{ij} = \sum \frac{w_i}{2W} \cdot \frac{w_{ij}}{w_i} = \sum_i \frac{w_{ij}}{2W} = \frac{w_j}{2W} = \mu_j$$

ולכן זו אכן התפלגות סטציונרית.

מה האנטרופיה של ההילוך המקרי?

מכיוון שזה תהליך מרקובי, אז:

$$H(X) = H(X_2|X_1) = -\sum_i \mu_i \sum_j P_{ij} \log P_{ij}$$

נציב את הידוע לנו על μ_i (המעבר השני הוא על ידי הכפלה ב $\frac{2W}{2W}$):

$$= -\sum_i \frac{w_i}{2W} \sum_j \frac{w_{ij}}{w_i} \log \frac{w_{ij}}{w_i} = -\sum_i \sum_j \frac{w_{ij}}{2W} \log \frac{w_{ij}}{2W} + \sum_i \sum_j \frac{w_{ij}}{2W} \log \frac{w_i}{2W}$$

$$= H\left(..., \frac{w_{ij}}{2W}, ...\right) - H\left(..., \frac{w_i}{2W}, ...\right)$$

דוגמה 1:

אם כל הצלעות בעלות אותו משקל $w_{ij} = 1$ אז נקבל:

$$\log(2E) - H\left(\frac{E_1}{2E}, \frac{E_2}{2E}, ...\right)$$

וכן הלאה, כאשר E_i=דרגת הקודקוד ה i.

דוגמה 2:

מה האנטרופיה של הילוך מקרי של מלך על לוח שחמט אינסופי, ההסתברות לצעד בכל כיוון היא $\frac{1}{8}$, ולכן מספר הביטים שצריך כדי לקודד כל צעד הוא 3 (לוג 8).

דוגמה 3:

על לוח סופי האנטרופיה קצת יותר קטנה, כיוון שבהנתן שהמלך בפינה (או בצד) - יש פחות אפשרויות, וכשסוכמים את כל המשבצות והאפשרויות זה נותן משהו כמו $0.92 \cdot \log 8$.

החוק השני של התרמודינמיקה

אם יש לנו שני תהליכים μ_n, μ'_n (אלו סדרות של התפלגויות) עם אותו חוק מעבר, אז מה קורה ל $D(\mu_n||\mu'_n)$ כאשר n גדל?

התשובה - זה הולך וקטן. מדוע? אם P ההסתברות של μ_n ו Q ההסתברות של μ'_n אז:

$$p(X_n, X_{n+1}) = p(X_n) \cdot r(X_{n+1}|X_n)$$

$$q(X_n, X_{n+1}) = q(X_n) \cdot r(X_{n+1}|X_n)$$

ואז:

$$D(p(X_n, X_{n+1})||q(X_n, X_{n+1})) = D(p(X_n)||q(X_n)) + \underbrace{D\left(\underbrace{p(X_{n+1}|X_n)}_{=r}||\underbrace{q(X_{n+1}|X_n)}_{=r}\right)}_{=0}$$

$$= D(p(X_{n+1})||q(X_{n+1})) + \underbrace{D(p(X_n|X_{n+1})||q(X_n|X_{n+1}))}_{\geq 0}$$

ולכן:

$$D\left(p\left(X_{n+1}\right)||q\left(X_{n+1}\right)\right) \geq D\left(p\left(X_n\right)||q\left(X_n\right)\right)$$

ואכן האנטרופיה היחסית אינה גדלה.

$$D\left(\mu_n||\mu'_n\right) \geq D\left(\mu_{n+1}||\mu'_{n+1}\right)$$

ואם μ'_n סטציונרית μ כלשהי, אז:

$$D\left(\mu_n||\mu\right) \geq D\left(\mu_{n+1}||\mu\right)$$

ובאופן אינטואיטיבי, אי אפשר להתרחק יותר מדי (במובן של D ובממוצע לאורך זמן) מהתפלגות סטציונרית, וזה נכון גם אם יש יותר מהתפלגות סטציונרית אחת.
אם μ אחיד, אז

$$D\left(\mu_n||\mu\right) = \log|X| - H\left(X_n\right)$$

ומכיוון שהאנטרופיה היחסית קטנה או קבועה - נובע מכך שהאנטרופיה גדלה או נותרת בעינה, וזהו החוק השני של התרמודינמיקה בניסוח היותר פיסיקלי שלו, שכן אנחנו מניחים התפלגות אחידה.

עירבוב

T- עירבוב, כלומר זוהי איזושהי פרמוטציה שפועלת על 52 קלפים.
X-פרמוטציה של הקלפים, כלומר סידור כלשהו.
(בעצם שני הדברים הללו הם תמורות, אבל משום מה אנחנו מגדירים אותם אחרת...)
כמובן מתקיים

$$H\left(TX\right) \geq H\left(TX|T\right)$$

כאשר TX מצב הקלפים כשהתחלנו עם X וערבבנו ע"י T. מדוע אי השוויון נכון? כי תמיד התניה מקטינה את האנטרופיה (לכל היותר).
ואז:

$$H\left(TX|T\right) = H\left(T^{-1}TX|T\right) = H\left(X|T\right) = H\left(X\right)$$

מדוע השוויון הראשון משמאל? כי כשמפעילים תמורה על איזושהי התפלגות, ולוקחים אנטרופיה (שבמהלכה סוכמים הסתברויות ולוגים בלי חשיבות לסדר) - האנטרופיה לא משתנה. המסקנה היא שלאחר העירבוב האנטרופיה יכולה רק לגדול, וזה די חשוב לקזינו.

$Channel\,Capacity$ - קיבולת

נניח שיש הודעה שרוצים לשלוח w, מקודדים אותה ל X^n (n-יה של אותיות), נשלח אותה בערוץ רועש כלשהו, יתקבל בצד השני Y^n, ובצד השני יש מערכת מפענחת שמייצרת על סמך זה את $\hat{w}$ ונקווה ש $w = \hat{w}$. נניח כרגע שאותיות לא מתווספות ולא הולכות לאיבוד, אבל אולי משתנות בדרך. האותיות השונות "מתקלקלות" לפי התפלגות $p\left(y|x\right)$, כלומר זו ההסתברות לקבל אות y בהנתן שהאות המקורית היתה x, וההתפלגות הזו מגדירה ערוץ תקשורת.

$$C = \max_{p(X)} I\left(X;Y\right) = H\left(X\right) - H\left(X|Y\right)$$

באופן אינטואיטיבי זה מגדיר - אחרי שראיתי את Y - מה רמת אי הוודאות שלי לגבי X?

דוגמה 1

ערוץ ללא רעש, כאשר מתקבל בדיוק הביט שנשלח, הוא בעל קיבולת 1 (כי בהנתן Y ידוע X בוודאות).

דוגמה 2 - $noisy\,typewriter$

אם שולחים א - מקבלים א' או ב' בהסתברות חצי.
אם שולחים ב - מקבלים ב' או ג' בהסתברות חצי.
...
...
אם שולחים ת - מקבלים ת' או א' בהסתברות חצי.

$$C=\max_{p(x)}H\left(Y\right)-\underbrace{H\left(Y|X\right)}_{=1}=\max_{p(X)}H\left(Y\right)-1=\underbrace{\log 22}_{uniform\,probability}-1$$

פתרון: אפשר לשלוח רק אותיות אי זוגיות (כלומר א', ג', ה' וכן הלאה), וכך תמיד בהנתן Y נדע מה היתה X, נקודד באמצעות זה את כל 22 האותיות.

5 שיעור 5

דוגמה 3

אם אנחנו משדרים n-יה, כלומר X^n, כמה n-יות טיפוסיות (Y^n) עשויות להתקבל? $2^{n\cdot H(Y|X)}$, נכנה קבוצה זאת $C\left(X^n\right)$, זה נכון לכל n-יה קבועה, וסך הכל ה n-יות הטיפוסיות ב Y מספרן $2^{n\cdot H(Y)}$. אם אנחנו רוצים שכל ה $C\left(X^n\right)$ תהיינה זרות, אז בהנתן איבר מ Y^n נדע בוודאות איזה X^n היה המקור שלו. אם זה נכון וקבוצות אלו הן זרות, אז מספרן הוא בדיוק $\frac{2^{n\cdot H(Y)}}{2^{n\cdot H(Y|X)}}=2^{n\cdot(H(Y)-H(Y|X))}=2^{nI(X;Y)}$.

הגדרה 5.1 ערוץ בדיד ($Discrete\,Channel$) מוגדר ע"י שלישיה $(X,p\left(y|x\right),Y)$, כלומר המשתנה המקרי שנשלח, המשתנה המקרי שמתקבל וההתפלגות המותנית שלהם (כלומר מה ההסתברות ל y מסויים בהנתן ששלחנו x על הערוץ).
ועבור n-יות: $\left(X^n,p\left(y|x\right),Y^n\right)$
נאמר שהערוץ חסר זיכרון ($Memoryless$), ובקיצור DMC כאשר

$$p\left(y_k|x^k,y^{k-1}\right)=p\left(y_k|x_k\right)$$

נאמר שהערוץ ללא משוב $(Without\,feedback)$ כאשר:

$$p\left(x_k|x^{k-1},y^{k-1}\right)=p\left(x_k|x^{k-1}\right)$$

כלומר פונקציית ההסתברות מוגדרת כך שאין מתחשבים ב Y, במילים אחרות - לא יודעים מה יצא בצד השני, ולכן אי אפשר לתקן או להתייחס.

בערוץ חסר זיכרון ללא משוב מתקיים (למרות שהיעדר המשוב לא חשוב כאן):

$$p\left(y^n|x^n\right)=\prod p\left(y_i|x_i\right)$$

קוד נגדיר על ידי (M,n) כאשר:

1. ההודעות האפשריות הן $\{1,...,M\}$
2. יש פונקציית קידוד $\chi^n:\{1,...,M\}\to X^n$ שמכל הודעה מייצרת מילה ב X^n, כלומר n-יה של האלף בית של X.

3. יש פונקציית פיענוח $g : Y^n \to \{1, ..., M\}$

הגדרה 5.2 השגיאה של i היא

$$\lambda_i = Pr\left(g\left(Y^n\right) \neq i | X^n\left(i\right)\right)$$

כלומר מה ההסתברות שפונקציית הפיענוח שגתה בפיענוח המילה.

$$\lambda^n = \max_i \lambda_i$$

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i$$

נגדיר את הקצב ($Rate$) של M, n ע"י:

$$R = \frac{\log M}{n}$$

קצב R הוא ישיג אם יש סדרת קודים $\left(\underbrace{2^{n \cdot R}}_{M}, n\right)$ כך ש:

$$\lambda^n \to_{n\to\infty} 0$$

נגדיר JTS- $Joint Typical Sequences$ (סדרות טיפוסיות משותפות), כיוון שידוע לנו $p(x, y)$:

$$A_\varepsilon^n = \{(x^n, y^n)\}$$

כל הזוגות של n-יות המקיימים:

$$\left| -\frac{1}{n} \log\left(p\left(x^n\right)\right) - H\left(X\right) \right| < \varepsilon$$

$$\left| -\frac{1}{n} \log\left(p\left(y^n\right)\right) - H\left(Y\right) \right| < \varepsilon$$

$$\left| -\frac{1}{n} \log\left(p\left(x^n, y^n\right)\right) - H\left(X, Y\right) \right| < \varepsilon$$

תכונות של JTS:

1.

$$Pr\left(\left(x^n, y^n\right) \in A_\varepsilon^n\right)_{n\to\infty} \longrightarrow 1$$

2.

$$|A_\varepsilon^n| \leq 2^{n \cdot (H(X,Y)+\varepsilon)}$$

3. נבחר n-יה של Xים וn-יה של Y-ים באופן בלתי תלויה, ולכן ההסתברות לקבלת כל זוג ספציפי $x\tilde{^n}, y\tilde{^n}$ תהיה $p\left(x\tilde{^n}\right) \cdot p\left(y\tilde{^n}\right)$ מאי תלות. אזי

$$(1-\varepsilon)\, 2^{-n(I(X;Y))} \leq Pr\left[\left(x\tilde{^n}, y\tilde{^n}\right) \in A_\varepsilon^n\right] \leq 2^{-n(I(X;Y))}$$

נוכיח את 2:

$$\begin{aligned} 1 &= \sum p\left(x^n, y^n\right) \geq \sum_{A_\varepsilon^n} p\left(x^n, y^n\right) \\ &\geq \left|A_\varepsilon^n\right| 2^{-n(H(X,Y)-\varepsilon)} \\ &\Rightarrow \left|A_\varepsilon^n\right| \leq 2^{n\cdot(H(X,Y)+\varepsilon)} \end{aligned}$$

נוכיח את 3:

$$\begin{aligned} Pr\left(\left(x\tilde{^n}, y\tilde{^n}\right) \in A_\varepsilon^n\right) &= \sum_{x^n, y^n \in A_\varepsilon^n} p\left(x^n\right) \cdot p\left(y^n\right) \\ &\leq \underbrace{2^{n\cdot(H(X,Y)+\varepsilon)}}_{\left|A_\varepsilon^n\right|} \cdot \underbrace{2^{-n(H(X)-\varepsilon)}}_{p(x^n)} \cdot \underbrace{2^{-n\cdot(H(Y)-\varepsilon)}}_{p(y^n)} \\ &= 2^{-n(I(X;Y)-3\varepsilon)} \end{aligned}$$

החסם מהצד השני מוכח באופן דומה.

טענה 5.3 אם $R < C$ אזי R ישיג.

הוכחה: בהנתן $M = 2^{nR}$ הודעות אפשריות לשליחה, נגריל מטריצה $2^{nR} \times n$, שהיא תהיה ספר הקוד שלנו במובן שההודעה ה c תקודד ע"י השורה ה c במטריצה (n ביטים). ההסתברות לקבלת מטריצה כלשהי K היא מכפלת ההסתברויות של איבריה (כיוון שבחרנו כל תא במטריצה באופן בלתי תלוי) כלומר:

$$Pr\left(K\right) = \prod_{j=1}^{2^{nR}} \prod_{i=1}^{n} p\left(x_i\left(w\right)\right)$$

כאשר w היא הודעה, וההודעות מתפלגות אחיד אז

$$Pr\left(W = w\right) = 2^{-nR}$$

כמובן ואז

$$Pr\left(Y^n | X^n\right) = \prod \left(p\left(y_i | x_i\left(w\right)\right)\right)$$

לכאורה היינו צריכים לבדוק בהנתן השורה שהתקבלה - מה השורה שהכי סביר שנשלחה, אבל לעשות את זה ישירות זה קשה ומסובך, ולכן נמצא דרך אחרת לעשות זאת שאסימפטוטית תיתן את אותה תוצאה.
המפענח שלנו מקבל y^n ושואל האם קיים ויחיד x^n כך ש $\left(x^n, y^n\right) \in A_\varepsilon^n$, אם כן - אזי x^n היא השורה המתאימה (ההודעה שנשלחה), אחרת - טעינו.

$$\begin{aligned} Pr\left(E\right) &= \sum_c Pr\left(c\right) \cdot P_\varepsilon^n\left(c\right) \\ &= \sum_c Pr\left(c\right) \left(\underbrace{\frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w\left(c\right)}_{P_e^n(c)} \right) \end{aligned}$$

הקטע בסוגריים מודד את ממוצע הטעויות בפיענוח לפי הספר c.
מכיוון שספרי הקודים הם אקראיים לחלוטין והגרלנו אותם באופן בלתי תלוי - ההסתברות לטעות בפיענוח של המילה שקידודה בשורה הראשונה - שווה להסתברות לטעות בפיענוח של המילה שבשורה ה-i <הוכחה לא הושלמה>
■

6 שיעור 6

המשך ההוכחה מהשיעור הקודם: **הוכחה:**

$$\begin{aligned}&=\frac{1}{2^{nR}}\sum_i\sum_c Pr\left(c\right)\cdot\lambda_i\left(c\right)\\&=\sum_c Pr\left(c\right)\cdot\lambda_i\left(c\right)=Pr\left(E|w=1\right)\end{aligned}$$

נגדיר $E_i=\left(X^n\left(i\right),Y^n\right)\in A_\varepsilon^n$, זהו המאורע שהזוג $X^n\left(i\right)$ (קידוד i) ו Y^n, מה שהתקבל בצד השני של הערוץ, אכן שייכים ל A_ε^n.
ואז נקבל שהסכום הנ"ל שווה ל:

$$Pr\left(\overline{E_1}\cup E_2\cup...\cup E_{2^{nR}}\right)=P\left(\overline{E}_1\right)+\sum_{i=2}^{2^{nR}}P\left(E_i\right)$$

(אנחנו בודקים את ההסתברות לטעות בשליחת ההודעה 1, וזה יכול לקרות כאשר Y^n מופיע עם עוד X^nים אחרים ב A_ε^n, או שהוא לא מופיע כלל).

$$\begin{aligned}&=\varepsilon+2^{nR}\cdot2^{-n(I(X;Y)-3\varepsilon)}\\&=\varepsilon+\underbrace{2^{3n\varepsilon}\cdot2^{-n(I(X;Y)-R)}}_{\to_{n\to\infty}\varepsilon}\end{aligned}$$

ואם $R<I\left(X;Y\right)-3\varepsilon$ אז הביטוי הנ"ל שואף ל 2ε ככל ש n שואף לאינסוף.
קיבלנו מצב שבו בספר קודים אקראי ההסתברות הממוצעת לטעות היא 2ε, כעת אם נזרוק מחצית מהשורות בטבלה, ודוק - את מחצית השורות שבן ההסתברות לטעות גבוהה יותר. נשארנו עם מחצית מספר הקודים, אבל הטעות המקסימלית היא קטנה מ 4ε (אחרת נובע שכל אלו שזרקנו, ההסתברות לטעות בהם גדולה מ 4ε, ואז הממוצע גדול מ 2ε בסתירה).
הקצב החדש שהושג על ידי ספר הקודים הזה הוא $R'=\frac{\log2^{nR-1}}{n}=R-\frac{1}{n}$.
ולכן ניתן לקרב את R כרצוננו ע"י בחירת n גדול. ואכן בנינו קוד שהקצב שלו קרוב ל R כרצוננו. ■

למה 6.1 $I\left(X^n;Y^n\right)\leq n\cdot C$

הוכחה:

$$\begin{aligned}I\left(X^n;Y^n\right)&=H\left(Y^n\right)-H\left(Y^n|X^n\right)\\&=H\left(Y^n\right)-\sum H\left(Y_i|Y_1,Y_2,...,Y_{i-1},X^n\right)\\&=H\left(Y^n\right)-\sum H\left(Y_i|X_i\right)\\&\leq\sum H\left(Y_i\right)-\sum H\left(Y_i|X_1\right)\leq n\cdot C\end{aligned}$$

■

משפט 6.2 אם $C<R$ אז R אינו ישיג.

הוכחה: נניח שההודעות ב W מתפלגות אחיד, ולכן האנטרופיה של W היא לוג גודל הקבוצה ונקבל:

$$\log\left(\left|\{w\,is\,a\,message\}\right|\right)=nR=H\left(W\right)$$

$$\begin{aligned}&=H\left(W|\hat{W}\right)+I\left(W;\hat{W}\right)\\&\underbrace{\leq}_{Fano}1+P_e^nnR+I\left(W;\hat{W}\right)\\&\underbrace{\leq}_{Information\,Loss}1+P_e^nnR+I\left(X^n;Y^n\right)\\&\underbrace{\leq}_{Lemma}1+P_e^nnR+nC\end{aligned}$$

מקבלים

$$R\leq\frac{1}{n}+P_e^nR+C$$

$$P_e^n\geq1-\frac{c}{R}-\frac{1}{nR}$$

■

7 שיעור 7

הערוץ הגאוסי הדיסקרטי

$$Y_i=X_i+Z_i$$

כלומר מכניסים מספר ממשי ומקבלים בצד השני מספר ממשי + רעש שהוא מספר ממשי שמתפלג נורמלי. $Z_i\sim G\left(0,N\right)$

אם אין רעש כלל - הקיבולת היא אינסוף - כי כל מספר שנכניס נקבל בדיוק אותו בצד השני.

נחסום את X_i על ידי:

$$\frac{1}{n}\sum X_i^2<p$$

אם נרצה לשלוח ביט אחד, אז בהתאם להגבלה נקודד את 0 על ידי $-\sqrt{p}$ ואת 1 על ידי $\sqrt{p}$, ונכריע בצד המקבל לפי "האם מה שקיבלנו גדול או קטן מאפס".

אם האינפורמציה שיש לשלוח מתפלגת חצי-חצי על 0 ו 1, נקבל כי ההסתברות לטעות היא (ההסתברות ששלחנו חיובי וקיבלנו קטן מאפס וכו'):

$$\begin{aligned}P_e&=\frac{1}{2}Pr\left(Y<0|X=\sqrt{p}\right)+\frac{1}{2}Pr\left(Y>0|X=-\sqrt{p}\right)\\&=\frac{1}{2}Pr\left(Z<-\sqrt{p}\right)+\frac{1}{2}Pr\left(Z>\sqrt{p}\right)=Pr\left(Z>\sqrt{p}\right)\\&=1-erf\left(\frac{\sqrt{p}}{n}\right)\end{aligned}$$

כאשר $erf=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$

כרגיל מגדירים את הקיבולת כ

$$C=\max_{\underbrace{f(X)}_{X\ density}:E[X^2]<p} I(X;Y)$$

כעת לפי הגדרה:

$$\begin{aligned}I(X;Y)&=h(Y)-h(Y|X)=h(Y)-h(X+Z|X)\\&=h(Y)-h(Z|X)=h(Y)-\underbrace{h(Z)}_{\frac{1}{2}\log_2 2\pi eN}\end{aligned}$$

(Z בלתי תלוי ב X)

כעת נחשב את השונות של Y כדי לקבל חסם על האנטרופיה (ידוע לנו שבהנתן השונות האנטרופיה המקסימלית מושגת על ידי המשתנה הנורמלי בעל אותה שונות, מהתרגיל):

$$E\left[Y^2\right]=E\left[(X+Z)^2\right]=\underbrace{E\left[X^2\right]}_{\leq P}+2\underbrace{E[X]}_{=0}E[Z]+\underbrace{E\left[Z^2\right]}_{N}$$

ולכן השונות חסומה מלעיל ע"י $P+N$, ומכאן שהאנטרופיה של הערוץ חסומה על ידי $\frac{1}{2}\log_2 2\pi e(P+N)$ ומכך נחזור לחישוב הקיבולת של הערוץ ונקבל את החסם לקיבולת הערוץ על ידי:

$$C=\underbrace{\frac{1}{2}\log_2 2\pi e(P+N)}_{h(Y)}-\underbrace{\frac{1}{2}\log_2 2\pi eN}_{h(Z)}=\frac{1}{2}\log\left(1+\frac{P}{N}\right)$$

נתבונן כעת ב n־יות של X_i־ים, האילוץ הוא $\frac{1}{n}\sum X_i^2<P$.

בעצם אנחנו מחפשים וקטורים n מימדיים בממשיים, כך שהנורמה שלהם קטנה מ $\sqrt{nP}$, במילים אחרות אנחנו מתבוננים בספירה שהרדיוס שלה הוא $\sqrt{nP}$. כל הודעה שנשלח נקודד על ידי נקודה בספירה, אם יהיו רעשים ־ ההודעה שתתקבל בצד השני היא וקטור שעשוי להיות מרוחק במידה מסויימת מהוקטור ששלחנו, המרחק של הוקטור שנשלח מהוקטור שהתקבל חסום ע"י $\sqrt{nN}$, מדוע? כיוון שזו סטיית התקן של n משתנים בלתי תלויים עם שונות N.

כלומר ־ נרצה לבחור את הנקודות שלנו כך שתמיד נוכל לתקן את השגיאות. יש ספירה ברדיוס $\sqrt{nN}$ סביב לכל קידוד של הודעה, שהיא טווח הטעות שלנו, אם יתקבל משהו בספירה הזו ־ נדע שההודעה המקודדת היתה מרכז הספירה. נבחר את הספירות שלנו להיות זרות, וכך כאשר יתקבל וקטור בצד המקבל ־ נדע מה היה מקורו לפי הספירה שבה הוא נמצא.

כאשר נשלח ־ נבחר את הקידודים מתוך ספירה ברדיוס $\sqrt{nP}$ כאמור, מאידך, בצד המקבל יתקבלו הודעות בתוך הספירה ברדיוס $\sqrt{n(P+N)}$, מדוע? כי זה התחום של סטיית תקן אחת.

נשאל כמה כדורים זרים אפשר להכניס בתוך הכדור ברדיו $\sqrt{nP}$, שמשמעו ־ כמה הודעות אפשר לקודד. זה כמובן חסום מלעיל על ידי גודל הכדור הגדול לחלק לכדור הקטן, יותר מזה אי אפשר. הנפח של ספירה n מימדית ברדיוס r, היא קבוע כלשהו C כפול r^n ולכן היחס בין נפחי הכדורים דלעיל הוא:

$$\frac{C\cdot(n\cdot(P+N))^{\frac{n}{2}}}{C\cdot(nN)^{\frac{n}{2}}}=\frac{(P+N)^{\frac{n}{2}}}{N^{\frac{n}{2}}}$$

נגריל ספר קודים (כלומר מטריצת $2^{nR}\times n$) שבכל תא יש לו מספר ממשי מוגרל מהתפלגות גאוסיאנית $G(0,P-\varepsilon)$.

נזכיר שכך יש לנו 2^{nR} קידודים להודעות, בני n אותיות כ"א. נשלח את שורה מספר 1.

טעות מספר 1 ־ בשורה ה i קיבלנו סדרה שמפירה את אילוץ ההספק, כלומר $\frac{1}{n}\sum X_i<P$, נסמנה E_0

טעות מספר 2 ־ המפענח חשב שהשורה ששלחנו שייכת לשורה j

טעות מספר 3 ־ המפענח קיבל שורה שלא מצא למי היא שייכת

כבעבר יש לנו:

$$\underbrace{W}_{1\ldots 2^{nR}}\rightarrow X^n\rightarrow_{+Z^n}\rightarrow Y^n\rightarrow\hat{W}$$

על פי פאנו:

$$H\left(W|\hat{W}\right) \leq 1 + \underbrace{nR}_{H(W)} \cdot P_e^n = n\varepsilon_n$$

(כאשר $\varepsilon_n = \left(RP_e^n + \frac{1}{n}\right)$, לצרכי נוחות הסימון)
נראה שכאשר הטעות קטנה מאד - הקיבולת חסומה על ידי C, כלומר נתעלם מהספר האקראי שבנינו קודם לכן, וכעת נתבונן בכל ספר נתון ונראה שהקיבולת שלו חסומה כך.

$$\begin{aligned} nR = H(W) &= I\left(W;\hat{W}\right) + H\left(W|\hat{W}\right) \leq I\left(W;\hat{W}\right) + n\varepsilon_n \\ \underbrace{\leq}_{loss\ of\ information} & I\left(X^n,Y^n\right) + n\varepsilon_n = h\left(Y^n\right) - h\left(Y^n|X^n\right) + n\varepsilon_n \\ &= h\left(Y^n\right) - h\left(Z^n|X^n\right) + n\varepsilon_n \\ &\leq \sum h\left(Y_i\right) - \sum h\left(Z_i\right) + n\varepsilon_n \end{aligned}$$

(*)
המעבר האחרון מכלל השרשרת ומאי־תלות של Z־ים זה בזה וב X, זהו רעש אקראי לחלוטין (גאוסיאני).

$$= \sum I\left(X_i;Y_i\right) + n\varepsilon_n$$

הביטוי $I\left(X_i;Y_i\right)$ הוא האינפורמציה המשותפת בין האותיות ה i במילה הנשלחת והמתקבלת.
נגדיר

$$P_i = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} X_i^2\left(w\right)$$

$$Y_i = X_i + Z_i$$

ואז

$$E\left[Y_i^2\right] = P_i + N$$

ואז כזכור ההתפלגות על Y_i שתמקסם את האנטרופיה בהנתן השונות הנ"ל היא ההתפלגות הנורמלית עם שונות זו ולכן:

$$\max h\left(Y_i\right) = \frac{1}{2}\log 2\pi e\left(P_i + N\right)$$

נחזור ל (*):

$$\begin{aligned} &\leq \sum h\left(Y_i\right) - \sum h\left(Z_i\right) + n\varepsilon_n \leq \sum \frac{1}{2}\log 2\pi e\left(P_i+N\right) - \sum \underbrace{\frac{1}{2}\log 2\pi eN}_{h(gaussian)} + n\varepsilon_n \\ &= \sum\left[\frac{1}{2}\log\frac{\left(P_i+N\right)}{N}\right] + n\varepsilon_n = \sum\left[\frac{1}{2}\log 2\pi e\left(1+\frac{P_i}{N}\right)\right] + n\varepsilon_n \end{aligned}$$

כעת אם נסכום את ריבועי תאי המטריצה ונחלק במספר השורות, נקבל את סכום ריבועי התאים הממוצע בשורה, כלומר ריבוע הנורמה הממוצע של הוקטורים שהם שורות המטריצה, וזה לפי דרישתנו המקורית קטן מ P ועל כן:

$$\frac{1}{n}\sum P_i < P$$

כיוון ש $\frac{1}{2}\log(1+x)$ היא פונקציה קעורה, אזי מאי שוויון ינסן מתקבל:

$$\sum\left[\frac{1}{2}\log 2\pi e\left(1+\frac{P_i}{N}\right)\right]+n\varepsilon_n \leq n\cdot\frac{1}{2}\log\left(1+\frac{1}{n}\sum\frac{P_i}{N}\right)+n\varepsilon_n$$

$$\leq n\cdot\frac{1}{2}\log\left(1+\frac{P}{N}\right)+n\varepsilon_n = n\left(\underbrace{\frac{1}{2}\log\left(1+\frac{P}{N}\right)}_{C}+\varepsilon_n\right)$$

ולכן $R \leq C+\varepsilon_n$ וכאשר $\varepsilon_n \to 0$ (וזה קורה כאשר n גדל) מקבלים את הדרוש.

8 שיעור 8

פורייה

ניסה לחשב את זרימת החום על גבי ריבוע מתכת, כאשר מחממים צלע אחת - איך מתפשט החום על פני הריבוע.
אמר שבהנתן f מחזורית 2π אז מגדירים

$$\hat{f}(n) = \frac{1}{2}\pi\int_{-\pi}^{\pi} f(x)\,e^{-inx}dx$$

אז מקבלים:

$$f(x) = \sum_{n=-\infty}^{\infty}\hat{f}(n)\,e^{inx}$$

אפשר לחשוב על e^{inx} בתור e_n ואז מקבלים ש:

$$f = \sum \underbrace{(f,e_n)}_{inner\,product}\cdot e_n$$

וכן:

$$\sum \hat{f}^2 = \int |f|^2$$

$$i\cdot n\cdot\hat{f}(n) = \hat{f}'(n)$$

ולכן גזירה מגבירה את המשקל של החלקים בסדרה שבהם n גדול, כלומר של תדרים גבוהים.
משפט הקונוולוציה: מקדם הפורייה של $f\hat{*}g$ הוא מכפלת המקדמים $\hat{f}\cdot\hat{g}$. כאשר מגדירים:

$$f*g = \int f(x)\,g(t-x)\,dx$$

כעת לפונקציות לא מחזוריות:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} dx$$

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi) e^{2\pi i x \xi} dx$$

משפט הדגימה

הרעיון הוא שכאשר נתונה לנו פונקציה מקושקשת ומסובכת אך חסומה, מספיק לדגום אותה כדי להבין איך היא נראית.

השיטה:

ראשית ניקח את $\hat{f}$, ונכפול אותה בפונקציה קבועה בגובה 1, אשר , שמסומנת ע"י $rect$.

נציין שפונקצית $\delta(x)$ היא פונקציה המקיימת לכל f את השוויון

$$\int \delta(x) f(x) dx = f(0)$$

פונקציית מסרק היא בעצם צירוף של פונקציות דלתה שנותנות את ערכי הפונקציה לא רק ב 0 אלא גם במקומות אחרים. אינטואיטיבית פונקציית δ היא מין פונקציה שהאינטגרל שלה הוא 1, והתומך שלה מוכל בקטע שארכו שואף לאפס.

כאשר עושים קונוולוציה של $\hat{f} \cdot rect$ כלשהי עם פונקציית מסרק - מקבלים העתקים של הפונקציה במרחקים קבועים (לפי פונקציית המסרק).

נעיר כי הפונקציה $sinc(x) = \frac{sin\pi x}{\pi x}$ היא הפורייה של $rect$, ואז אם נסמן את פונקציית המסרק ע"י M נקבל כי:

$$f * sinc \cdot M$$

הוא הפורייה של הפונקציה המקורית.

במילים אחרות - מספיק לדגום את הפונקציה לפי צפיפות המסרק לאחר שהעברנו אותה קונוולוציה עם $sinc$, ואז נוכל לשחזר את הפונקציה.

כעת אם מתקיים:

$$\int |f|^2 = \int \left|\hat{f}\right|^2 = 1$$

ועיקרון אי הוודאות אומר:

$$\int x^2 f(x)^2 \int \xi^2 \hat{f}(\xi)^2 \geq \frac{\pi}{4}$$

כלומר מכפלת השונות של הפונקציה ושל הפורייה שלה חסומות מלמטה, אי אפשר להקטין אחד מהם מבלי להגדיל את האחר.

שליחת פונקציה רציפה על גבי ערוץ

כמובן אי אפשר ממש לשלוח פונקציה רציפה, אבל באמצעות פורייה ראינו שמספיק לשלוח דגימות של הפונקציה שהן מספיק צפופות כדי לספק אינפורמציה על הפונקציה.

נניח שהרעש בערוץ שלנו הוא $\frac{N_0}{2}$ וואט\הרץ.

רוחב הפס - w.

קצב השידור יהיה $\frac{1}{2w}$

ממוצע 0, ובזמן t נשדר $2wt$ דגימות.
השונות המשותפת היא $\frac{N_0}{2} \cdot I$
הקיבולת של ערוץ גאוסיאני היתה כזכור

$$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$$

קיבולת הערוץ לדגימה תהיה:

$$C = \frac{1}{2}\log\left(1 + \frac{P}{N_0 w}\right)$$

וכיוון שבכל שניה שולחים $2w$ דגימות, אז לשניה הקיבולת תהיה:

$$C = w\log\left(1 + \frac{P}{N_0 w}\right) \to \frac{P}{N_0}\log e$$

נניח שאנחנו שולחים בו זמנית בערוצים מקבילים בלתי תלויים, ונתבונן בשליחה בכולם יחד כבערוץ אחד.
בכל ערוץ שולחים X_j, מקבלים $Y_j = X_j + Z_j$ כך ש $Z_j \sim N(0, N_j)$, כלומר אלו ערוצים גאוסיאנים מקבילים.
יש לנו אילוץ על העצמה לכל הערוצים יחד ע"י

$$E\left[\sum X_i^2\right] \le P$$

נגדיר קיבולת הערוץ כרגיל:

$$\begin{aligned}
C =& \max_{f(X_1,...,X_k):E\left[\sum X_i^2\right]\le P} I(X_1,...,X_k;Y_1,...,Y_k) \\
=& h(Y_1,...,Y_k) - h(Y_1,...,Y_k|X_1,...,X_k) \\
=& h(Y_1,...,Y_k) - h(Z_1,...,Z_k)
\end{aligned}$$

מאי תלות מתקיים $h(Z_1,..,Z_k) = \sum h(Z_i)$ וכן $h(Y_1,...) \le \sum h(Y_i)$ ולפיכך:

$$\begin{aligned}
h(Y_1,...,Y_k) - h(Z_1,...,Z_k) &\le \sum h(Y_i) - \sum h(Z_i) \\
&\le \sum \frac{1}{2}\log\left(1 + \frac{P_i}{N_i}\right)
\end{aligned}$$

נרצה למצוא את P_i כדי להבין איך לחלק את עצמת הערוץ בין הערוצים המקבילים כדי למקסם את האינפורמציה, נמצא ע"י כופלי לגראנז':

$$\max_{\sum P_i = P} \sum \frac{1}{2}\log\left(1 + \frac{P_i}{N_i}\right)$$

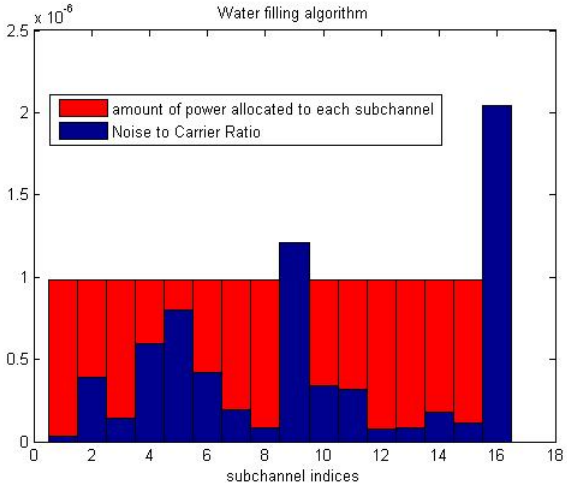
נקבל:

$$\frac{\partial}{\partial P_i} = \frac{1}{2}\cdot\frac{1}{P_i + N_i} + \lambda = 0$$

ואז:

$$P_i = \upsilon - N_i$$

ואז נוודא שמה שהתקבל הוא אכן מספר חיובי.
האלגוריתם נקרא $Water - Filling$, כביכול לכל ערוץ, שונות הרעש שלו N_i נותנת עמודה יותר גבוהה, שחוסמת את השימוש בערוץ זה, כל עוד יש ערוצים יותר "שקטים" שאפשר להשתמש בהם:



KKT

$f:\mathbb{R}^n\to\mathbb{R}$ ונניח שיש קבוצת אילוצים שהם שוויונים $h_i=...$ ואילוצים שהם אי שוויונים $g_i=...$, אזי אופטימום x^* יקיים:

$$\nabla f\left(x^*\right)+\sum\mu_i\nabla g_i\left(x^*\right)+\sum\lambda_j\nabla h_j\left(x^*\right)=0$$

כאשר $\mu_i\geq 0$, ו $\mu_i g_i\left(x^*\right)=0$

9 שיעור 9

התבוננו בעבר בערוץ שהוא בעצם כמה ערוצים גאוסים מקבילים ובלתי תלוים, נתבונן כעת במקרה שבו יש תלות בין הערוצים המקבילים.

9.1 רעש צבוע $(Colored\,Noise)$

נניח שהרעשים הם $Z_1,...,Z_n$ שהממוצע שלהם 0, ומטריצת $n\times n$ השונויות המשותפות שלהם היא K_Z. (כמובן סימטרית, אי שלילית)

כמו כן נניח שיש אילוץ על העצמה ע"י:

$$\frac{1}{n}\sum\mathbb{E}\left[X_i^2\right]\leq P$$

ובאופן שקול (כיוון שהשונות המשותפת של כל X עם עצמו היא התוחלת):

$$\frac{1}{n}Tr\left(K_X\right)\leq P$$

(כאשר K_X היא מטריצת השונויות המשותפות של הקלט)

הערה 9.1 במקרה זה משוב יכול לעזור, מדוע? בעצם המטריצה K_Z מספקת מתאם מסויים בין הרעשים שבערוצים המקבילים, ולכן אם ידוע לנו הרעש שנוצר בערוץ מסויים - זה יכול לספק לנו אינפורמציה על הרעש שיווצר כעת בערוץ אחר, ולכן נוכל לפזר את השליחה בערוצים השונים בצורה יותר נבונה.

כעת:

$$\begin{aligned}I\left(X_1,...,X_n;Y_1,...,Y_n\right)&=h\left(Y_1,...,Y_n\right)-h\left(Y_1,...,Y_n|X_1,...,X_n\right)\\&=h\left(Y_1,...,Y_n\right)-h\left(X_1+Z_1,...,X_n+Z_n|X_1,...,X_n\right)\\&=h\left(Y_1,...,Y_n\right)-h\left(Z_1,...,Z_n\right)\end{aligned}$$

אם מניחים שגם X גאוסי, וזה הרי המקרה הגרוע ביותר, נזכיר שחישבנו סכום של אנטרופיות של משתנים גאוסים ונקבל:

$$\frac{1}{2}\log\left(2\pi e\right)^{n}\left|K_{X}+K_{Z}\right|$$

ונרצה למצוא מה קיבולת הערוץ, כלומר למקסם את הביטוי הזה, תחת האילוץ $\frac{1}{n}Tr\left(K_{X}\right)\leq P$
למטריצה חיובית תמיד יש פירוק כך ש $K_{Z}=Q\Lambda Q^{t}$ כך ש Λ אלכסונית ו $Q\cdot Q^{t}=I$.
לכן כעת:

$$\begin{aligned}\left|K_{X}+K_{Z}\right| &= \left|K_{X}+Q\Lambda Q^{t}\right| = \left|\underbrace{Q^{t}K_{X}Q}_{A:=}+\Lambda\right| \\ &= \left|A+\Lambda\right|\end{aligned}$$

למה 9.2 אי שוויון $Hadamard$:

$$h\left(X_{1},...,X_{n}\right)\leq\sum h\left(X_{i}\right)$$

ואם X_i גאוסי, אז:

$$\begin{aligned}\frac{1}{2}\log\left(\left(2\pi e\right)^{n}\left|K_{X}\right|\right) &\leq \sum\frac{1}{2}\log\left(2\pi e\ \underbrace{k_{i,i}}_{\sigma\ of\ X_{i}}\right) \\ \frac{1}{2}\sum\log\left(2\pi e\right)+\frac{1}{2}\log\left|K\right| &\leq \frac{1}{2}\sum\log\left(2\pi e\right)+\frac{1}{2}\sum\log k_{i,i} \\ \log\left|K\right| &\leq \sum\log k_{i,i} \\ \log\left|K\right| &\leq \log\left(\prod k_{i,i}\right) \\ \left|K\right| &\leq \prod k_{i,i}\end{aligned}$$

נעשה בזה שימוש:

$$\left|A+\Lambda\right|\leq\Pi\left(A_{i,i}+\Lambda_{i}\right)$$

וכיוון ש $Tr\left(K_{X}\right)=Tr\left(QK_{X}Q^{t}\right)$ אז נרצה למקסם את $\left|A+\Lambda\right|$ תחת האילוץ:

$$\frac{1}{n}\sum A_{i,i}\leq P$$

ונעשה זאת באמצעות כופלי לגראנז':

$$\prod_{j}\left(A_{j,j}+A_{j}\right)+\lambda\sum_{i}A_{i,i}$$

$$\frac{\partial}{\partial A_{i,i}}=\prod_{j\neq i}\left(A_{j,j}+\Lambda_{j}\right)+\lambda$$

$$A_{i,i} + \Lambda_i = \nu$$

$$A_{i,i} = (\nu - \Lambda_i)$$

מכאן שכדי לשלוח באופן שמנצל את הקיבולת המקסימלית, יש למצוא את המטריצה Λ כנ"ל, ואז אפשר להגדיר את המטריצה K_X בסופו של דבר כך שתהיה מקסימום אינפורמציה משותפת.

9.2 $Rate\,Distortion$

נניח שיש מקור שפולט מספרים ממשיים לפי התפלגות נורמלית, ונניח שיש לנו אפשרות להעביר רק נתונים דיסקרטיים, מה עושים?

אנחנו רוצים להעביר ע"י ביט אחד את המספר שהתקבל, אזי אם התקבל משהו אי־שלילי נשלח 1, ואחרת - נשלח 0.

בצד המקבל - נניח שקיבלנו 1, מה ננחש שנשלח? מה הכי כדאי?

אם אנחנו רוצים למזער את הטעות - ננחש את החציון של החלק החיובי. אם אנחנו רוצים למזער את הטעות בריבוע, ננחש את הממוצע של החלק החיובי (ובאופן סימטרי לחלק השלילי).

מה הממוצע?

$$\int_0^\infty x \cdot f(x)\, dx = \sqrt{\frac{2}{\pi}} \cdot \sigma$$

(עבור השלילי זה מטעמי סימטריה $-\sqrt{\frac{2}{\pi}} \cdot \sigma$)

למה הממוצע ממזער את סכום ההפרשים בריבוע? נחפש מהו Z כך ש:

$$\sum (X_i - Z)^2$$

מינימלי.

נגזור ונקבל:

$$\frac{\partial}{\partial Z} = 2\sum X_i - 2\sum Z = 0$$

$$Z = \frac{\sum X_i}{\sum 1} = \frac{\sum X_i}{n}$$

וזה אכן הממוצע.

נניח שאנחנו צריכים להעביר נקודה כלשהי במישור ע"י r ביטים. אזי בהנתן x, נחפש את $g(x)$, זהו קידוד של נקודה אחת מבין 2^r נקודות על המישור, כלומר עברנו מנתון רציף למשהו דיסקרטי, העברנו את הנקודה שהתקבלה לצד השני, וכעת המקבל צריך לנחש מהי הנקודה שנשלחה.

נתבונן בבעיה הזאת לגבי הישר, נניח שיש התפלגות על הממשיים, ואנחנו רוצים למצוא אוסף של נקודות על הישר (זה בעצם מגדיר את g, כי לכל $x \in \mathbb{R}$ נתאים את הנקודה הקרובה אליו ביותר) כך שהטעות תהיה מזערית.

9.2.1 אלגוריתם $Lloyd$

ניקח פיזור אקראי של 2^r הנקודות על הישר, כעת זה מגדיר חלוקה של הישר ל 2^r קטעים, כפי שראינו - לכל קטע המיקום האופטימלי עבור נקודה מסויימת הוא הממוצע בקטע, לכן נחשב אותו עבור כל אחד מ 2^r הקטעים, ו"נתקן" את מיקומי הנקודות להיות הממוצעים בקטעים הללו (ביחס להתפלגות על הממשיים).

מעניין לציין שהאלגוריתם הזה אמנם משפר בכל איטרציה ושואף לאופטימום מקומי - אבל לא בהכרח שואף לאופטימום גלובלי, ובכלל לא ידוע על אלגוריתם שנותן אופטימום גלובלי לבעיה זו, אפילו בישר, שלא לדבר על המישור ועל מרחבים n־מימדיים.

9.2.2 בעיית מיזעור טעויות בשליחת נתונים רציפים מעל ערוץ דיסקרטי (במטבנרמע"ד, בטרמינולוגיה המקובלת)

יש n־יה נתונה של מספרים, יש מקודד $f(x)$ שמעביר את זה ל 2^{nR} ביטים, יש מפענח שמעביר את הביטים הללו ל $\hat{X}^n$, ויש מטריקה על הטעויות:

$$d:\left(X\times\hat{X}\right)\to\mathbb{R}^{+}$$

מרחק האמינג ($Hamming$) הוא בעצם המטריקה הדיסקרטית ע"י:

$$d\left(x,\hat{x}\right)=\begin{cases}0 & x=\hat{x}\\ 1 & x\neq\hat{x}\end{cases}$$

ואז:

$$\mathbb{E}\left[d\left(X,\hat{X}\right)\right]=Pr\left(X\neq\hat{X}\right)$$

מטריקה אחרת היא הטעות הריבועית:

$$d\left(x,\hat{x}\right)=\left(x-\hat{x}\right)^{2}$$

ואז נגדיר אותה לכל ה n־יה ע"י ממוצע הטעויות:

$$d\left(x^{n},\hat{x}^{n}\right)=\frac{1}{n}\sum d\left(x,\hat{x}\right)$$

הגדרה 9.3 $\left(2^{nR},n\right)-rate\,Distortion\,Code$

$$f_n:X^n\to\left\{1,...,2^{nR}\right\}$$

$$g_n:\left\{1,...,2^{nR}\right\}\to\hat{X}^n$$

$$D\left(R\right)=\mathbb{E}\left[d\left(x^n,g_n\circ f_n\left(x^n\right)\right)\right]=\sum_{x^n}p\left(x^n\right)d\left(x^n,g_n\circ f_n\left(x^n\right)\right)$$

נשים לב כי ככל ש R גדל, $D\left(R\right)$ האופטימלי שאפשר לקבל דועך אקספוננציאלי, כלומר בקצב שידור גבוה נוכל למזער את הטעות, ולהיפך ־ אם נרשה הרבה טעויות, אפשר להסתפק בקצב שידור קטן מאד.

הגדרה 9.4

$$R^{I}\left(D\right)=\min_{p(\hat{x}|x):\sum_{(x,\hat{x})}p(x)p(\hat{x}|x)d(x,\hat{x})\leq D}I\left(X,\hat{X}\right)$$

למשל עבור משתנה ברנולי עם הסתברות p, אז:

$$R\left(D\right)=H\left(p\right)-H\left(D\right)$$

כאן נקבע את מטריקת השגיאה להיות מרחק המינג.

$$\begin{aligned}I\left(X;\hat{X}\right)=H\left(X\right)-H\left(X;\hat{X}\right)=H\left(p\right)-H\left(X\oplus_{\mathbb{Z}_2}\hat{X}|\hat{X}\right)\\ \geq H\left(p\right)-H\left(X\oplus\hat{X}\right)\geq H\left(p\right)-H\left(D\right)\end{aligned}$$

נשים לב ש $X\oplus\hat{X}$ מונה את הטעויות שקיבלנו, אבל כפי שציינו D הוא החסם המינימלי לטעות, ולכן אי השוויון האחרון. מכאן נובע:

$$R\left(D\right)\geq H\left(p\right)-H\left(D\right)$$

10 שיעור 10

במקרה ש X משתנה מקרי גאוסי, $X \sim N\left(0,\sigma^2\right)$ אז:

$$\begin{aligned} I\left(X;\hat{X}\right) &= h\left(X\right) - h\left(X|\hat{X}\right) = \frac{1}{2}\log\left(2\pi e\sigma^2\right) - h\left(X-\hat{X}|X\right) \\ &\geq \frac{1}{2}\log\left(2\pi e\sigma^2\right) - h\left(X-\hat{X}\right) \end{aligned}$$

מכיוון שמשתנה נורמלי עם שונות נתונה ממקסם אנטרופיה, מתקיים $h\left(Z\right) \geq H\left(X-\hat{X}\right)$ אם השונות של Z היא כמו השונות של $X-\hat{X}$ ולכן:

$$\begin{aligned} &\geq \frac{1}{2}\log\left(2\pi e\sigma^2\right) - h\left(N\left(0,\mathbb{E}\left[\left(X-\hat{X}\right)^2\right]\right)\right) \\ &\geq \frac{1}{2}\log\left(2\pi e\sigma^2\right) - \frac{1}{2}\log\left(2\pi eD\right) \end{aligned}$$

כיוון שהמרחק בריבוע בין X ל $\hat{X}$ מתפלג כמו D.
ואז:

$$= \frac{1}{2}\log\frac{\sigma^2}{D}$$

ואם ניקח $\hat{X} \sim N\left(0,\sigma^2-D\right)$ ונשדר אותו עם רעש $N\left(0,D\right)$ אכן נקבל $X \sim N\left(0,\sigma^2\right)$, ולכן זו ההתפלגות שחיפשנו, כי אי אפשר לקבל R קטן מ $\frac{1}{2}\log\frac{\sigma^2}{D}$ לפי אי השוויון שהראינו לעיל.

$$I\left(X;\hat{X}\right) = h\left(X\right) - h\left(X|\hat{X}\right)$$

וזה נותן את הגודל הדרוש.
וכעת נחלץ מ $R = \frac{1}{2}\log\frac{\sigma^2}{D}$ את D:

$$D = \sigma^2 2^{-2R}$$

ומכאן נובע שבכל ביט נוסף (כלומר כל תוספת ל R) הטעות תקטן בפקטור כפלי של 4.

עוד דוגמה

יש לנו nיה של משתנים $X_1,...,X_n$ שכולם מתפלגים נורמלי ע"י $X_i \sim N\left(0,\sigma_i^2\right)$ תחת האילוץ:

$$\sum_i \left(X_i - \hat{X}_i\right)^2 \leq D$$

נרצה לקבל מספר מינימלי של ביטים שצריך לשלוח כדי לא לעבור את הטעות D.

$$\begin{aligned} I\left(X^n;\hat{X^n}\right) &= \sum h\left(X_i\right) - \sum h\left(X_i|\underbrace{X^{i-1}}_{:=X_1,X_2,...,X_{i-1}},\hat{X^n}\right) \\ &\geq \sum h\left(X_i\right) - \sum h\left(X_i|\hat{X}_i\right) \\ &= \sum I\left(X_i;\hat{X}_i\right) \\ &\geq \sum R\left(D_i\right) \\ &= \sum \frac{1}{2}\log\left(\frac{\sigma^2}{D_i}\right) \end{aligned}$$

כיוון ש I גדול מביטוי זה, נרצה

$$\sum \frac{1}{2}\ln\frac{\sigma^2}{D_i}+\lambda\sum D_i$$

$$D_i=\lambda'$$

כדי לעמוד בתנאי ההתחלתי של "D טעות מקסימלית", נצטרך D_i כך שסכומם קטן מ D.
קיבלנו שה D_i-ים שווים, כלומר מבצעים משהו שנקרא $Inverse\,water\,filling$, שהוא בעצם לשדר קודם לכן בתדרים (ב X_k-ים) שבהם שונות הרעש היא גדולה יותר.

10.1 איך מחשבים את $\hat{X}$?

כזכור:

$$R\left(D\right)=\min_{q(x|\hat{x}),\sum p(x)q(\hat{x}|x)d(x,\hat{x})\leq D}I\left(X;\hat{X}\right)$$

נמצא באמצעות כופלי לגראנז' מינימום:

$$J=\sum_x\sum_{\hat{x}}p\left(x\right)q\left(\hat{x}|x\right)\log\left(\frac{q\left(\hat{x}|x\right)}{\sum_x p\left(x\right)q\left(\hat{x}|x\right)}\right)+\lambda\sum_x\sum_{\hat{x}}p\left(x\right)q\left(\hat{x}|x\right)d\left(x,\hat{x}\right)+\sum_x\underbrace{\nu\left(x\right)}_{lag.\,coefficients}\sum_{\hat{x}}q\left(\hat{x}|x\right)$$

זו בעיה קשה כי בסכום האחרון יש המון משתנים.
נצטרך איזושהי דרך לחשב נומרית את q.

11 שיעור 11

נסמן את המכנה שבתוך הלוגריתם ע"י $r\left(\hat{x}\right)=\sum_x p\left(x\right)q\left(\hat{x}|x\right)$ כדי לפשט מעט את הביטוי והוא קבוע, לכן לכל x ו $\hat{x}$ נקבל:

$$\frac{\partial J}{\partial q\left(\hat{x}|x\right)}=p\left(x\right)\log\left(\frac{q\left(\hat{x}|x\right)}{r\left(x\right)}\right)+p\left(x\right)+\lambda p\left(x\right)d\left(x,\hat{x}\right)+\nu\left(x\right)$$

ונגדיר את μ ע"י $\frac{\nu(x)}{p(x)}=\log\mu\left(x\right)$ ואז נקבל:

$$\frac{\partial J}{\partial q\left(\hat{x}|x\right)}=p\left(x\right)\left(\log\left(\frac{q\left(\hat{x}|x\right)}{r\left(x\right)}\right)+1+\lambda d\left(x,\hat{x}\right)+\log\mu\left(x\right)\right)$$

וכאשר ביטוי זה מתאפס:

$$q\left(\hat{x}|x\right)=\frac{r\left(\hat{x}\right)e^{-\lambda d(x,\hat{x})}}{\mu\left(x\right)}$$

אבל מכיוון ש q היא הסתברות אנחנו יודעים ש $\sum_{\hat{x}}q\left(\hat{x}|x\right)=1$ ולכן:

$$\frac{\sum_{\hat{x}}r\left(\hat{x}\right)e^{-\lambda d(x,\hat{x})}}{\mu\left(x\right)}=1$$

ולכן:

$$\sum_{\hat{x}}r\left(\hat{x}\right)e^{-\lambda d(x,\hat{x})}=\mu\left(x\right)$$

ונוכל לקבל מכך:

$$q\left(\hat{x}|x\right)=\frac{r\left(\hat{x}\right)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}}r\left(\hat{x}\right)e^{-\lambda d(x,\hat{x})}}$$

שנסמנו (*) ונחזור אליו בהמשך.

11.1 אלגוריתם $Blahut-Arimoto$

בהנתן ההתפלגויות $p(x)$ ו $p(y|x)$ (זו התפלגות אחת ועוד n התפלגויות) מתקיים:

$$D\left(p\left(x\right)p\left(y|x\right)||p\left(x\right)r^{*}\left(y\right)\right)$$

כאשר $r^{*}\left(y\right)$ היא ההתפלגות שממזערת את הנ"ל, נקבל:

$$r^{*}\left(y\right)=\sum_{x}p\left(x\right)p\left(y|x\right)$$

(לא נרשום פה את ההוכחה).

$$R\left(D\right)=\min_{r(\hat{x})}\qquad\min_{q(\hat{x},x),\,s.t.:\sum p(x)q(\hat{x}|x)\cdot d(x,\hat{x})\leq D}\sum_{x}\sum_{\hat{x}}p\left(x\right)q\left(\hat{x}|x\right)\log\left(\frac{q\left(\hat{x}|x\right)}{r\left(\hat{x}\right)}\right)$$

אלו תנאי העיוות.
נגדיר את A = ההתפלגויות המשותפות עם שוליים $p(x)$ שמקיימים את תנאי העיוות.
B=התפלגויות מכפלה $p\left(x\right)q\left(\hat{x}\right)$
והמרחק הוא D.

11.2 סיבוכיות קולמוגורוב

הגדרה 11.1 סיבוכיות קולמוגורוב של מחרוזת היא אורך תכנית המחשב הדרושה כדי לייצר אותה. (מונח שפותח גם כן ע"י צ'ייטין וסולומון).

$$K_{u}\left(x\right)=\min_{p:u(p)=x}l\left(p\right)$$

u-מכונה אוניברסלית (אולי טיורינג) כלשהי, משמעות ההגדרה היא שבהנתן תכנה p - u מייצרת את x.
נוסיף גם את האילוץ שאין תכנית שהיא רישא של תכנית אחרת.

נגדיר $K_{u}\left(x|l\left(x\right)\right)$ להיות אורך תכנית המחשב המינימלית שמייצרת את x, ביודעה מראש מה אורך המחרוזת המצופה ממנה.
נוכל להציג תכנית מחשב שארכה $2\log x+1$ אם נגדיר את קוד התכנית להיות המחרוזת כך שכל ביט מופיע פעמיים, וביט הסיום מסומן ע"י 01, למשל התכנית:

$$00110011111101$$

תייצר את המחרוזת

$$0101111$$

נוכל לשפר זאת אם קודם כל נשלח את האורך של הייצוג של x (זהו מספר מסדר גודל $\log x$) בשיטה הקודמת, אורך הייצוג של הייצוג הוא $\log\log x$ ולכן על פי השיטה הקודמת - השליחה תהיה $2\log\log x$ ועוד מספר הביטים בייצוג, המכונה תדע לעצור אחרי $\log x$ ביטים לפי המספר הראשון ששלחנו, ולכן כעת אורך התכנית הוא

$$2\log\log x+\log x$$

וכך נוכל להקטין זאת עוד ולקבל שבאמצעות $l\left(x\right)+\lg\left(l\left(x\right)\right)+\log\log\left(l\left(x\right)\right)+\log\log\log\left(l\left(x\right)\right)$ נוכל לעשות זאת וכן הלאה.
נתבונן בכל סדרות הביטים באורך k או קטן מזה - יש $2^{k+1}-1$ כאלו. כביכול היינו מעדיפים לשלוח תכנית מחשב שתכתוב מחרוזת במקום את המחרוזת עצמה, היינו מצפים לחסוך כך מקום, אבל ב k ביטים אפשר לתאר רק 2^{k-1} סדרות שהן לא רישא אלו של אלו, ורק הסדרות הללו מייצגות מכונות טיורינג כלשהן, ולכן את "רוב" הסדרות באורך k או קטן מכך - לא נוכל לתאר ע"י מכונת טיורינג שאורכה k לכל היותר, ולכן לא תהיה לנו ברירה אלא לשלוח את הסדרה עצמה.

הגדרה 11.2 נאמר שסדרה כנ"ל (שסיבוכיות קול' שלה היא לפחות אורכה) היא אקראית

נניח שנרצה לקודד מחרוזת שבה לכל יום גשום בעשר השנים האחרונות יופיע 1 במקום הנכון, ולכל יום לא־גשום יופיע 0.

אפשרות 1 - נרשום את המיקום במחרוזת של הימים שבהם ירד גשם. כיוון שאנחנו יודעים שמספר קטן בהרבה ממספר הימים שאינם גשומים - נקבל יעילות לא רעה של $k\log n$, כאשר k תוחלת מספר הימים הגשומים, ו n מספר הימים הכולל.

אפשרות 2 - להשתמש ברעיון של סדרות טיפוסיות. כלומר נכין רשימה של כל הסדרות שבהן יש בערך k אחדות, ופשוט נשלח את מספרה ברשימה של הסדרה הרלבנטית. כמה סדרות כאלו יש?

$$2^{nH\left(\frac{k}{n}\right)}$$

כמובן מספר הסדרות האמיתי הוא $\binom{n}{k}$, ויש לנו אי שוויון:

$$\sqrt{\frac{n}{8k\left(n-k\right)}}2^{nH\left(\frac{k}{n}\right)}\leq\binom{n}{k}\leq\sqrt{\frac{n}{\pi k\left(n-k\right)}}2^{nH\left(\frac{k}{n}\right)}$$

12 המשך סיבוכיות קולמוגורוב

מאי שוויון קראפט נובע שאם נסמן ב P את קבוצת התוכנות, אז:

$$\sum_{p\in P}2^{-l(p)}\leq1$$

ואם ניקח מתוכן רק את אלו שגם עוצרות אז נקבל:

$$\sum_{p\in P,\,and\,U(p)\,halts}2^{-l(p)}<1$$

דוגמה:
x_i משתנים iid המתפלגים עם $B\left(\Theta\right)$
אז ההסתברות

$$p\left(x^n\right)=\prod p\left(x_i\right)$$

ו:

$$H\left(x\right)\leq\frac{1}{n}\sum_{x^n}f\left(x^n\right)k\left(x^n|n\right)\leq H\left(x\right)+\frac{\log n}{n}+c$$

"נעשה כל מיני חישובים, למי שאין מושג כי הוא מהנדס או משהו, שישאל" (מיכאל)

12.0.1 נוכיח שיש אינסוף מספרים ראשוניים.

נניח בשלילה שיש רק k מספרים ראשוניים. כמובן את כל הטבעיים אפשר לקודד עם חזקות של ראשוניים. כידוע לנו משיקולי ספירה, קיים מספר x בן nביטים כך ש $k\left(x|n\right)\geq n$.

מספר הביטים במספר הוא n, ולכן בהנתן המספרים הראשוניים - לתאר את מעריכי חזקותיהם זה $\log n$ ביטים. ולכן ניתן לתאר כל מספר כנ"ל באמצעות $k\cdot2\log n<n$ ביטים בסתירה.

12.0.2 ניתן ההערכה להתפלגות הראשוניים

אם n מספר כלשהו, אפשר לרשום אותו על ידי מכפלת הראשוני הגדול ביותר שמחלק אותו במנת המספר הזה בראשוני זה. אם נסמן את הראשוני הגדול ביותר שמחלק את n ב p_i כוונתנו ש $n = p_i \cdot \frac{n}{p_i}$. נניח ש p_i הוא הראשוני ה i בסדרת הראשוניים (האינסופית מההוכחה הקודמת).

אז כדי לקודד את p_i מספיקים $\log i$ ביטים. וכדי לקודד את $\frac{n}{p_i}$ צריך $\log n - \log p_i$ ביטים.

נרצה שבקידוד יגולם מתי מסתיים הקידוד של p_i ומתחיל הקידוד של $\frac{n}{p_i}$ ולכן נרצה שקידוד של i יהיה חסר רישא, ולכן גדלו יהיה $\log i + 2\log\log i$, ואז גודל הקידוד יהיה:

$$\log i + 2\log\log i + \log n - \log p_i$$

כיוון שממילא מספיקים $\log n$ ביטים כדי לתאר את n אז נקבל:

$$\log n \leq \log i + 2\log\log i + \log n - \log p_i$$

כיוון שתיארנו את p_i ע"י $\log i + 2\log\log i$ ביטים, אז :

$$\log p_i \leq \log i + 2\log\log i$$

ולכן:

$$p_i \leq i \cdot \log^2 i$$

וזה נותן לנו בעצם חסם לא רע על התשובה לשאלה - כמה רחוק צריך לטפס במספרים הטבעיים כדי לקבל i ראשוניים בקבוצה?

12.0.3 אין מכונה רגולרית שמקבלת את הסדרות 0^k1^k

ניקח k שהוא לא דחיס (כלומר שסיבוכיות קולמוגורוב שלו גדולה מ k), כלומר אקראי, ונרצה לתאר אותו.

נניח שיש מכונה כזו T, אז נוכל לתאר את k על ידי הזוג (T, q), כאשר q הוא המצב באוטומט שבו נימצא אחרי קריאת k אפסים. קל להבין שזה מתאר את k, כיוון שאז אפשר לקחת את האוטומט, להתחיל מהמצב q ולהזין 1 עד שנגיע למצב מקבל. מספר ה 1-ים שהוזנו יהיה k.

כעת ניקח k דחיס שהוא גדול ממש מהתיאור של T ועוד המרחב הדרוש כדי לתאר מצב כלשהו ב T, ואז נקבל שמצאנו דרך להציג את k על ידי תיאור שהוא קטן מ k, בסתירה.

12.0.4 אין מכונה רגולרית שפולטת מספרים ראשוניים

נניח שיש מכונה רגולרית ששפתה L. נגדיר

$$L_x = \{y : xy \in L\}$$

נרשום את כל ה L_x בסדר מסויים, אלו כל הסיפות שמכניסות את x לשפה. כדי לתאר את L_x נמצא דרך לתאר את y_n, כלומר האיבר ה n-י ב L_x. אז נוכל לשלוח את המכונה T, ששפתה L, את המספר n, ואת המצב במכונה של L שמגיעים אליו אחרי קריאת x. ולכן:

$$k(y_n) \leq k(n) + \underbrace{c}_{the\ length\ of\ T's\ description\ +\ one\ of\ its\ states}$$

השפה : $L = \{0^p : p\ is\ prime\}$ אינה רגולרית.

נניח בשלילה שכן, אז $p_k - p_{k-1}$ הוא האיבר הראשון ב $L_{p_{k-1}}$, וראינו שאפשר לתאר זאת בגודל קבוע. עם זאת - לכל מרווח שנבחר - יש שני ראשוניים עוקבים (בסדרת הראשוניים) שההפרש ביניהם גדול ממרווח זה, ולכן עצם המסקנה שהגענו אליה, שניתן לתאר אינסוף מספרים באמצעות גודל קבוע, היא סתירה ולכן הנחתנו שהשפה רגולרית היא שגויה.

12.0.5 מה אורך הסגמנט הקבוע המקסימלי בסדרה אקראית?

משמעות השאלה - בהנתן מספר רנדומלי, כלומר סדרה אקראית, מה הסדרה הרצופה הקבועה הכי ארוכה שיש בתוכו? $\log n$, נוכיח זאת.

נניח בשלילה ש n רנדומלי וזה לא מתקיים בתוכו, אזי:

$$k\left(u0^{2\log n}v|n\right)\geq n$$

נתאר את n ע"י תיאור u ותיאור v. כזכור, במצב כזה התיאור של u צריך להיות חסר רישא, כדי שנדע מתי מתחיל התיאור של v, ואז נקבל שגודל התיאור של u,v הוא

$$|u|+|v|+\log|v|+2\log\log|v|\geq n$$

וכשנפענח - נדע לכתוב את u, אחריו $2\log n$ פעמים 0 ואחריו את v. כמובן $|u|+|v|$ זה בדיוק אורך המחרוזת בלי $2\log n$ אפסים ולכן זה $n-2\log n$ ואז נקבל מהאמור לעיל:

$$n-2\log n+\log n+2\log\log n\geq n$$

ולכן:

$$\log n+2\log\log n\geq 2\log n$$

כלומר:

$$2\log\log n\geq\log n$$

בסתירה.

12.0.6 בתוך סדרה אקראית בהכרח יש סדרה קבועה של אפסים באורך $\frac{1}{2}\log n$

נניח שאין בתור סדרה נתונה סדרה קבועה של אפסים באורך הנ"ל, אז נוכל להתבונן על הסדרה כולה בתור $\frac{n}{\frac{1}{2}\log n}$ רצפים שאורך כל אחד מהם $\frac{1}{2}\log n$. מהנחתנו נובע שאף אחד מהרצפים הללו אינו רצף של אפסים, ולכן מייצג מספר בין 1 ל $2^{\frac{1}{2}\log n}=\sqrt{n}$, אבל לא אפס. יש $\sqrt{n}-1$ מספרים בטווח זה ולכן בסך הכל יש $(\sqrt{n}-1)^{\frac{n}{\frac{1}{2}\log n}}$ סדרות באורך n מהסוג המתואר. :

$$\left(\sqrt{n}-1\right)^{\frac{2n}{\log n}}=\sqrt{n}^{\frac{2n}{\log n}}\left(1-\frac{1}{\sqrt{n}}\right)^{\frac{2n}{\log n}}\approx 2^n e^{\frac{-2\sqrt{n}}{\log n}}$$

ואז לוג של מספר הסדרות הוא:

$$\frac{n}{\log n}\cdot\log n-\frac{2\sqrt{n}}{\log n}<n$$

ולכן ניתן לייצג את כל הסדרות מסוג זה בפחות מ n ביטים.

12.1 התער של אוקאם

נגדיר:

$$p_U(x)=\sum_{p:U(p)=x}2^{-l(p)}$$

כלומר ההסתברות של סדרה היא סכימה על כל התכניות שבמכונה U מפיקות את הסדרה x, כאשר המשקל של כל תכנית הוא 1 חלקי מינוס אורך התכנית, כלומר במובן מסויים תכניות קצרות יותר הן אמינות יותר, במובן זה שמחרוזות שניתן לייצר על ידי תכנית קצרה יקבלו הסתברות גבוהה יחסית.
נגדיר:

$$\Omega = \sum_{p:U(p)\ halts} 2^{-l(p)}$$

נראה שאי אפשר לחשב מספר זה, וכן שבהנתן מחשב זה ניתן לתת מענה לשאלה - האם תכנה כלשהי עוצרת, בסתירה לבעיית העצירה שאנחנו מכירים מחישוביות.
אם המכונה שלנו ניתנת לתיאור ע"י n ביטים, וידועים לי n הביטים הראשונים של Ω. ניקח את כל המכונות שאורך תיאורן הוא n ביטים או פחות. מכיוון שידוע לנו Ω עד הביט ה m, נובע מכך שכל המכונות שגדולות יותר מהמכונות שלקחנו, הן לא תורמות דבר ל n הביטים הראשוניים (לפי האופן שבו מוגדרת Ω). בכל פעם שמכונה עוצרת - נניח שאורך תיאורה הוא m - נוסיף את 2^{-m} למונה קטן שנחזיק בצד, וכל עוד לא הגענו ל Ω - נמשיך. ברגע שהגענו ל Ω נבדוק האם p ספציפית עצרה (שהרי היא נכללה באוסף התוכניות שאנחנו מריצים במקביל) וכך נקבל תשובה.

13 שימושים בסיבוכיות קולמוגורוב

13.1 מספרי רמזי

הגדרה 13.1 מספר רמזי $R(3,3)$ הוא מספר הקודקודים בגרף השלם המינימלי שצביעתו בשני צבעים תבטיח משולש בצבע אחד או משולש מהצבע השני. $R(3,3) = 6$.

בעיקרון למצוא אפילו חסמים על מספר רמזי זה לא דבר פשוט, נשתמש בסיבוכיות קולמגורוב כדי לתאר את הגרף.
מסיבוכיות קולמגורוב אנחנו יודעים שיש גרף בן n קודקודים כך שצריך לפחות $\binom{n}{2}$ ביטים כדי לתארו (זה מספר הצלעות, ואורך התיאור הנאיבי).
אם נתון לנו שיש k-קליקה בתוך הגרף - אז לא צריך לתאר את הצלעות בקליקה, אלא מספיק לומר מיהם אותם קודקודים - ולשם כך צריך $k\log n$ ביטים (כל קודקוד מתוך ה n מתואר ע"י $\log n$ ביטים). מכך נובע שקודם תיארנו את הקליקה ע"י $\binom{k}{2}$ ביטים וכעת ע"י $k\log n$ ביטים, ולכן חסכנו את ההפרש. יש גרפים שעבורם אי אפשר לחסוך ולכן בהכרח $R(k,k) > 2^k$.

13.2 .

נתונה תת קבוצה לא ידועה של $\{1,...,n\}$ ויש $D_1,...,D_k$ קבוצות ידועות, כאשר נותנים לנו רק מה גודל החיתוכים של הקבוצה העלומה עם הקבוצות הידועות. נרצה למצוא בודאות מהי הקבוצה העלומה M. ליתר דיוק - נרצה למצוא את ה k המינימלי שנותן זאת.
נניח שהתיאור של M אינו דחיס, מכיוון שיש 2^n תת קבוצות אז גודל התיאור יהיה $\log_2 2^n = n$.
בהנתן $D_1,...,D_k$ נוכל לומר לכל אחת מהן מה גודל חיתוכה עם M באמצעות $\log n$ ביטים, ולכן סה"כ $k\log n$ ביטים. על פי האילוץ שהתיאור של M אינו דחיס נובע ש

$$k\log n > n \Rightarrow k > \frac{n}{\log n}$$

13.3 אומגה

נשוב לבעיית Ω מהשיעור הקודם. כזכור מספר שזה הוא בלתי חשיב שכן אחרת זה היה סותר את בעיית העצירה.
נוכיח שלא ניתן לדחוס שום רישא של Ω. **הוכחה:** נגדיר את x להיות המספר הקטן ביותר כך שאין מכונה שתיאורה הוא n ביטים (לכל היותר) שפולטת אותו. סיבוכיות קולמוגורוב של x היא לפחות $n+1$, שכן אין מכונה שפולטת אותו. נתבונן באלגוריתם לחישוב x בהנתן n הביטים הראשונים של Ω: נריץ את כל המכונות שתיאורן קטן מ n עד שיעצרו (זמן סופי בהנתן Ω), ונבדוק מהו המספר הקטן שלא יצא באף אחת מהמכונות. כיוון שהחישוב האחרון די פשוט - סיבוכיות קולמוגורוב שלו קבועה, ולכן סיבוכיות קולמוגורוב של האלגוריתם כולו היא (אורך תיאור הרישא של Ω באורך n ביטים) + (קבוע). אם ניתן לתאר את הרישא בגודל קטן מ n - אז קיבלנו את x. ■

13.4 הימורים - $Universal\ gambling$

משחק הטלות מטבע - בכל שלב אנחנו מהמרים מה יצא, ומיכאל אומר לנו מה התוצאה האמיתית, עד שלמיכאל נמאס. אם אנחנו צודקים - הרווחנו כסף, אחרת - מיכאל הרוויח. אנחנו רוצים לקבוע מראש מהי הסדרה שאנחנו מהמרים עליה (כי בעצם אנחנו מהמרים על סדרה כלשהי).
נגדיר עבור סדרה x :

$$p\left(x\right)=2^{-K(x)}$$

כלומר ההסתברות בעינינו לסדרה היא הופכית לסיבוכיות קולמוגורוב שלה. כיוון שתכניות המחשב בהגדרה שלנו הן חסרות רישא - אז זה מסתכם לפחות מ 1, ואנחנו נהמר פרופורציוניות להסתברות.

טענה 13.2

$$\log\left(\underbrace{S\left(x\right)}_{revenue\ for\ successful\ gambling\ on\ x}\right)+K\left(x\right)\geq l\left(x\right)$$

הוכחה:

$$S\left(x\right)=\sum_{x\ is\ a\ prefix\ of\ x'}2^{l(x)}p\left(x'\right)$$

מכיוון שאם הימרנו על x ויצאה רישא כלשהי של x - עדיין הרווחנו.
ואז מכיוון ש $p\left(x'\right)<p\left(x\right)$ מתקבל מהצבה:

$$S\left(x\right)\geq 2^{l(x)}2^{-K(x)}$$

(מדוע ויתרנו על הסכימה אחרי ההצבה? שכן כשמוסיפים ביט אחד באורך - ההסתברות יורדת בחצי ולכן ניתן להתעלם מכל ההסתברויות החל מאורך מסויים ולא נכפיל את התוצאה ביותר מ 2).
ואז מתקבלת הטענה אחרי שנוציא לוג:

$$\log\left(S\left(x\right)\right)\geq l\left(x\right)-K\left(x\right)$$

ומהעברת אגפים הטענה נובעת. ■

13.5 האם מחר תזרח השמש?

לפלס הניח שהשמש זורחת בהתפלגות ברנולי כלשהי, כלומר $B\left(\Theta\right)\sim$ אבל Θ לא ידוע. לכן הוא הניח ש Θ מתפלג אחיד בין 0 ל 1 (כלומר הקבוע שקובע את תהליך ברנולי הוא עצמו נקבע לפי התפלגות רציפה אחידה בין 0 ל 1).
ואז:

$$\begin{aligned}P\left(X_{n+1}=1|X_n=1,X_{n-1}=1,...,X_1=1\right)&=\frac{P\left(X_{n+1}=1,X_n=1,...\right)}{P\left(X_n=1,X_{n-1}=1,...\right)}\\&=\frac{\int_0^1\Theta^{n+1}d\Theta}{\int_0^1\Theta^n d\Theta}\\&=\frac{n+1}{n+2}\end{aligned}$$

כאשר n מספר ימי השמש עד כה.
נחשב זאת באמצעות סיבוכיות קולמגורוב.

מה סיבוכיות הסדרה 1^{n+1}? היא קרובה לסיבוכיות סדרה אינסופית של 1, ויש תכנה פשוטה בגודל קבוע שעושה זאת. ולכן ההסתברות שלה קבועה.

מה סיבוכיות הסדרה 1^n0? (כלומר n ימי שמש, וביום האחרון השמש לא תזרח) זה יהיה $\log n$ ואז הסיבוכיות תהיה בערך $2^{-\log n}$.

כיוון שרק אחת מהן תתממש מחר, ולכן ההסתברות לכך היא:

$$\frac{p\left(1^{n+1}\right)}{p\left(1^{n+1}\right)+p\left(1^n0\right)}=\frac{c}{\frac{1}{n}+c}$$

וזה שואף ל 1 כל עוד n דחיס.