Kodowanie, Kompresja, Kryptografia

Konspekt V, 15.V.2023

Kodowanie w obecności szumu — twierdzenie Shannona

Nieformalne oszacowanie efektywności transmisji:

$$\frac{\text{liczba bitów użytkowych}}{\text{długość słowa kodowego}} \;=\; \frac{m}{n}$$

Jeśli poziom szumu w kanale wynosi $0 < p < 0.5$, oczekiwana liczba przekłamanych bitów to np. Odległość słów kodowych d musi więc być dostatecznie duża, by zapewnić odpowiednią do warunków zdolność detekcyjną lub korekcyjną kodu,

$$d-1 \;>\; np \qquad \text{lub} \qquad t = \left\lfloor \frac{d-1}{2} \right\rfloor \;>\; np\,.$$

Mamy oszacowanie

$$n - m \;\geq\; \log_2 V_n(t)\,, \qquad V_n(t) = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{t}$$

czyli

$$\frac{m}{n} \;\leq\; \frac{n - \log V_n(t)}{n} \;=\; 1 - \frac{\log V_n(t)}{n}\,.$$

Korzystając z nierówności

$$\frac{1}{n+1} 2^{nh(q)} \;\leq\; \binom{n}{t} \;\leq\; 2^{nh(q)}\,,$$

gdzie $q = \frac{t}{n}$, można pokazać, że $\lim_{n\to\infty} \frac{\log V_n(t)}{n} \;=\; h(q)$, stąd jeśli przyjąć, że $t \approx np$, otrzymamy asymptotyczne ograniczenie

$$\frac{m}{n} \;\leq\; 1 - h(p).$$

dla kodu z odpowiednią do szumu zdolnością korekcyjną.

Dla $p = 0.01$, $h(p) = 0.08$, a więc $m \simeq 0.92n$,
dla $p = 0.05$, $h(p) = 0.29$, skąd $m \simeq 0.71n$.

Strategie dekodowania zniekształconego komunikatu

$B = \boldsymbol{x} \in \mathcal{B}^n$: zmienna losowa odpowiadająca nadanemu komunikatowi,
$B' = \boldsymbol{y} \in \mathcal{B}^n$: komunikat otrzymany.

1. Strategia *idealnego obserwatora*: po otrzymaniu $\boldsymbol{y}$ dekodujemy go jako $\boldsymbol{x}$ wg. kryterium
$$P(B = \boldsymbol{x} \mid B' = \boldsymbol{y}) \;=\; \max$$
2. Strategia *największego prawdopodobieństwa*: szukamy $\boldsymbol{x}$, dla którego
$$P(B' = \boldsymbol{y} \mid B = \boldsymbol{x}) \;=\; \max$$
3. Strategia *najmniejszej odległości*: wybieramy $\boldsymbol{x}$ wg. kryterium
$$d_H(\boldsymbol{x}, \boldsymbol{y}) \;=\; \min\,.$$

Wniosek 1. Jeśli wszystkie słowa kodowe $\boldsymbol{x}$ są jednakowo prawdopodobne, strategia idealnego obserwatora i strategia maksimum prawdopodobieństwa są równoważne.

$$P(B' = \boldsymbol{y} \mid B = \boldsymbol{x}) = \frac{P(B' = \boldsymbol{y}, B = \boldsymbol{x})}{P(B = \boldsymbol{x})} = P(B = \boldsymbol{x} \mid B' = \boldsymbol{y}) \frac{P(B' = \boldsymbol{y})}{P(B = \boldsymbol{x})}\,.$$

Wniosek 2. Jeśli kanał transmisyjny jest binarnym i symetrycznym kanałem bez pamięci z prawdopodobieństwem błędu $p < 0.5$, wówczas strategie maksimum prawdopodobieństwa i minimalnej odległości są równoważne.

$$P(B' = \boldsymbol{y} \mid B = \boldsymbol{x}) \;=\; \prod_{i=1}^{n} P(b'_j = y_j \mid b_j = x_j) \;=\; p^d (1-p)^{n-d} \;=\; (1-p)^n \left(\frac{p}{1-p}\right)^d\,,$$

gdzie d jest liczbą przekłamanych bitów $y_j \neq x_j$, a więc $d = d_H(\boldsymbol{x}, \boldsymbol{y})$. Ponieważ $\frac{p}{1-p} < 1$, maksimum $P(B' = \boldsymbol{y} \mid B = \boldsymbol{x})$ osiągane jest wtedy, gdy d jest najmniejsze.

Jak obliczyć $P(B = \boldsymbol{x}|B' = \boldsymbol{y})$ itp?

Nasze dane:

– $P(B = x_i)$ rozkład częstości znaków wejściowych $\mathcal{A} = \{x_1, x_2, ..., x_k\}$
– $P(B' = y_j \,|\, B = x_i) = p_{ij}$ charakterystyka kanału

Stąd możemy wyznaczyć:

– $P(B' = y_j,\ B = x_i) \ = \ P(B' = b_j|B = x_i) \cdot P(B = x_i)$
– $P(B' = y_j) \ = \ \sum_{x_i \in \mathcal{A}} P(B' = y_j|B = x_i)P(B = x_i)$,

a dalej można wyliczyć $H(B)$, $H(B')$, $H(B', B)$, ...

Przypomnienie: $H(Y|X) = H(Y,X) - H(X)$,

X, Y niezależne $\Rightarrow$ $H(Y,X) = H(Y) + H(X)$, $H(Y|X) = H(Y)$.

Gdy $Y = f(X)$, to jest gdy $P(Y = y_i|X = x_i) = \begin{cases} 0 & \text{gdy } f(x_i) \neq y_i \\ 1 & \text{gdy } f(x_i) = y_i \end{cases}$

wtedy $H(Y,X) = H(X)$ i $H(Y|X) = 0$.

Niech teraz $X = B$ (komunikat nadawany) i $Y = B'$ (komunikat odbierany). W tym kontekście

$$\begin{aligned} H(B'|B) &= \text{niepewność co do } B' \text{ gdy znane jest } B \\ &= \text{niepewność wynikająca z przekłamań kanału} \end{aligned}$$

Zatem pozostała część niepewności

$$I(B', B) := H(B') - H(B'|B)$$

wynika wyłącznie z niewiedzy o tym jaka litera została nadana. Innymi słowy, to ta część niepewności $H(B')$, która znika, gdy wiemy wszystko o B. Jest to tzw. *wzajemna informacja między B i B'*.

Własności

a) $I(B', B) = H(B') - H(B'|B) = H(B') + H(B) - H(B', B)$
$= H(B) - H(B|B') = I(B, B')$

b) $I(B', B) \geq 0$, z równością jedynie gdy B i B' są niezależne

c) $I(B', B) \leq H(B')$, z równością jedynie gdy $B' = f(B)$

d) $I(B', B) \leq H(B)$, z równością jedynie gdy $B = f(B')$.

b) Gdy $B \perp B'$, $I(B', B) = H(B') - H(B'|B) = H(B') - H(B') = 0$

c) Gdy $B' = f(B)$, $I(B', B) = H(B') - H(B'|B) = H(B') - 0 = H(B')$

Definicja: Pojemność informacyjna kanału Φ (information capacity)

$$\mathcal{C}(\Phi) \;=\; \sup_{P(B)} I(B', B)$$

Przykład: Binarny kanał symetryczny

$$\Phi \quad \longrightarrow \quad \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}.$$

Niech $\mathcal{A} = \{0, 1\}$, $P(B=0) = t$, $P(B=1) = 1-t$. Stąd

$$H(B) \;=\; -t\log t - (1-t)\log(1-t) \;=\; h(t)$$

$$\begin{aligned} H(B'|B) &= H(B'|B=0)P(B=0) \;+\; H(B'|B=1)P(B=1) \\ &= h(1-p)\cdot t \;+\; h(p)\cdot(1-t) \\ &= h(p) \end{aligned}$$

$$\begin{aligned} P(B'=0) &= P(B'=0|B=0)P(B=0) \;+\; P(B'=0|B=1)P(B=1) \\ &= (1-p)\cdot t \;+\; p\cdot(1-t) \\ &= t+p-2tp \;=\; q \end{aligned}$$

$$\begin{aligned} P(B'=1) &= P(B'=1|B=0)P(B=0) \;+\; P(B'=1|B=1)P(B=1) \\ &= p\cdot t \;+\; (1-p)\cdot(1-t) \\ &= 1-(t+p-2tp) \;=\; 1-q \end{aligned}$$

Stąd $H(B') = h(q)$. Dalej

$$I(B', B) \;=\; H(B') - H(B'|B) \;=\; h(q) - h(p) \;=\; h(t+p-2tp) - h(p)\,.$$

Musimy wybrać t tak, aby $I(B', B) = \max$, a więc aby $h(t+p-2tp) = \max$. Zachodzi to gdy

$$t+p-2tp = \frac{1}{2} \quad \Rightarrow \quad t = \frac{\frac{1}{2}-p}{1-2p} = \frac{\frac{1-2p}{2}}{1-2p} = \frac{1}{2}\,.$$

Wtedy $\mathcal{C}(\Phi) \;=\; \max I(B', B) \;=\; h(1/2) - h(p) \;=\; 1 - h(p)$.

Można łatwo pokazać, że jeśli przekazujemy n bitów niezależnie przez równoległe kanały Φ^n, to

$$\mathcal{C}(\Phi^n) \;=\; n\mathcal{C}(\Phi)\,.$$

Reguły łańcucha

$$H(X_1, X_2, ..., X_n) \;=\; H(X_1)+H(X_2|X_1)+H(X_3|X_2, X_1)+\cdots+H(X_n|X_{n-1}, ..., X_1)$$

$$I(X_1, X_2, ..., X_n,\, Y) \;=\; I(X_1, Y)+I(X_2, Y \,|\, X_1)+\cdots+I(X_n, Y \,|\, X_n-1, ..., X_1)$$

przy czym warunkowa wzajemna informacja zdefiniowana jest tak:

$$I(X, Y \,|\, Z) \;=\; H(X \,|\, Z) \;-\; H(X \,|\, Y, Z)\,.$$

Data processing inequality

Dla sekwencji niezależnych kanałów $X \xrightarrow{\Phi} Y \xrightarrow{\Psi} Z$ zachodzi

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

Niezależność kanałów oznacza, że prawdopodobieństwo odbioru $Z = z_k$ gdy nadano $X = x_i$ obliczamy z iloczynu macierzy prawdopodobieństw $P(\Psi) \cdot P(\Phi)$:

$$P(Z = z_k \,|\, X = x_i) \;=\; \sum_j P(Z = z_k \,|\, Y = y_j) P(Y = y_j \,|\, X = X_i)$$

Inaczej można to wyrazić następująco:

$$\begin{aligned} &P(X = x_i,\, Y = y_j,\, Z = z_k) \\ &\qquad =\; P(Z = z_k \,|\, Y = y_j) \;\cdot\; P(Y = y_j \,|\, X = x_i) \;\cdot\; P(X = x_i)\,. \end{aligned}$$

Oznacza to w szczególności, że rozkłady zmiennych warunkowych $P(X|Y = y)$ i $P(Z|Y = y)$ są niezależne dla wszystkich y, $P(X, Z|Y) \;=\; P(X|Y) \cdot P(Z|Y)$, a stąd

$$H(X, Z \,|\, Y) \;=\; H(X \,|\, Y) \;+\; H(Z \,|\, Y)\,.$$

Dowód nierówności $I(X, Z) \leq I(X, Y)$.

Stosując regułę łańcucha na dwa sposoby otrzymamy alternatywne rozwinięcia dla $I(X, Y, Z)$:

$$I(X, Y, Z) \;=\; I(X, Z) + I(X, Y \,|\, Z) \;=\; I(X, Y) + I(X, Z \,|\, Y)$$

Na podstawie niezależności $X|Y$ i $Z|Y$, $I(X,Z\,|\,Y)=0$:

$$\begin{aligned} I(X,Z\,|\,Y) &= H(X\,|\,Y) - H(X\,|\,Y,Z) \;=\; H(X\,|\,Y) - (H(X,Y,Z) - H(Y,Z)) \\ &= H(X\,|\,Y) \;-\; \Big(H(X,Y,Z) \;-\; (H(Z\,|Y\,) + H(Y))\Big) \\ &= H(X\,|\,Y) \;+\; H(Z\,|Y\,) \;-\; H(X,Z\,|\,Y) \;=\; 0\,. \end{aligned}$$

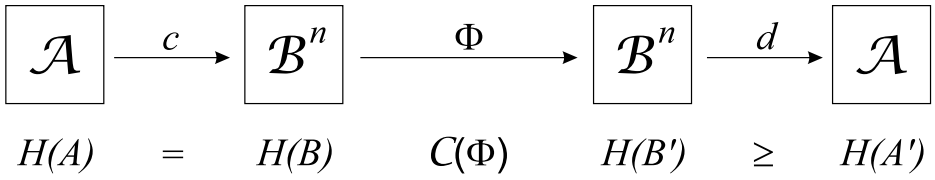
Zatem $I(X,Z) + I(X,Y\,|\,Z) \;=\; I(X,Y)$, a ponieważ $I(X,Y\,|\,Z) \;\geq\; 0$, otrzymujemy nierówność data processing.

W ogólności mamy także

$$I(X,Y) \;\geq\; I(X,f(Y))\,,$$

gdy drugi kanał jest bezszumowy $Z=f(Y)$ (jednak f nie musi być różnowartościowa i stąd dodatkowa utrata informacji).

Nierówność Fano



$H(A) \;=\; H(B)$, bo kodowanie c jest odwracalne. Dla prostoty załóżmy, że $H(A) = \log|\mathcal{A}|$.
$A' = d(B')$ niekoniecznie 1-1, więc $H(B') \geq H(A')$.
Prawdopodobieństwo błędnego odczytania a' gdy wysłano a podlegające minimalizacji

$$P(\text{błąd}) \;=\; \sum_{a\in\mathcal{A}} P(\text{błąd}\,|\,a \text{ nadane})\cdot P(A=a)\,.$$

To prawdopodobieństwo jest zależne od sposobu kodowania c, więc minimalizujemy

$$\epsilon(c) \;=\; \max\{P(\text{błąd}\,|\,a \text{ nadane}):\quad a\in\mathcal{A}\}$$

Kosztem tej minimalizacji jest spadek efektywności transmisji, którą chcielibyśmy — przeciwnie — maksymalizować,

$$\rho(c) \;=\; \frac{H(A)}{n} \;=\; \frac{\log|\mathcal{A}|}{n}\,.$$

Nierówność pomocnicza: X przyjmuje wartości w k-elementowym zbiorze $\mathcal{A}$. Dla ustalonego $a \in \mathcal{A}$ oznaczmy przez $p = P(X \neq a) = 1 - P(X = a)$. Wtedy

$$H(X) \leq p \log(k-1) + h(p)\,. \qquad (*)$$

DOWÓD: Określamy pomocniczą zmienną losową (tzw. indykator)

$$J = \begin{cases} 1 & \text{gdy } X \neq a \\ 0 & \text{gdy } X = a \end{cases}$$

Mamy $P(J=1) = p$, $P(J=0) = 1-p$, $H(J) = h(p)$ oraz

$$H(X|J) = H(X,J) - H(J)\,.$$

Ale $J = f(X)$, zatem $H(X,J) = H(X)$, czyli $H(X) = H(X|J) + H(J)$.

$$\begin{aligned} H(X|J) &= H(X|J=0)P(J=0) + H(X|J=1)P(J=1) \\ &= 0 \cdot (1-p) + pH(X|J=1) \leq p\log(k-1)\,, \end{aligned}$$

a stąd

$$H(X) \leq p \log(k-1) + h(p)\,.$$

Nierówność Fano: Niech X i Y przyjmują wartości w k-elementowym zbiorze $\mathcal{A}$, $p = P(X \neq Y)$. Wówczas

$$H(X|Y) \leq p \log(k-1) + h(p)\,.$$

DOWÓD. Weźmy $(*)$ z warunkiem $Y = y$, przy czym J jest indykatorem relacji $X \neq Y$:

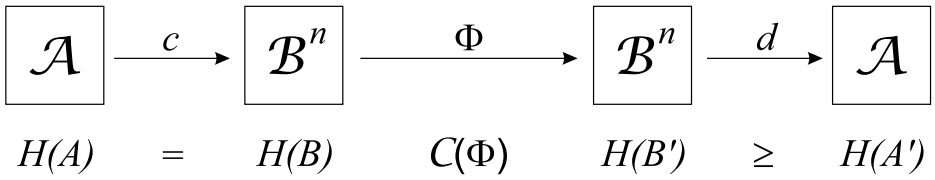
$$H(X \,|\, Y = y) \leq P(X \neq Y \,|\, Y = y) \log(k-1) + H(J \,|\, Y = y)\,.$$

Mnożymy obustronnie przez $P(Y = y)$ i sumujemy po y, otrzymując

$$H(X \,|\, Y) \leq P(X \neq Y) \log(k-1) + H(J \,|\, Y) \leq p \log(k-1) + h(p)\,.$$

Interpretacja: $H(X|Y)$ mierzy niepewność co do poprawnego odczytania nadanego symbolu X jeśli odebrano Y. Ta niepewność składa się z 2 części: $h(p)$ niepewność tego czy odczyt jest błędny i $p\log(k-1)$ niepewność jakie przekłamanie powstało, gdy wiadomo, że błąd odczytu rzeczywiście powstał.

Cel — określenie relacji między pojemnością kanału a maksymalną możliwą efektywnością transmisji.



Jak poprzednio zakładamy, że $P(A = a) = \frac{1}{k}$. Na podstawie nierówności Fano

$$H(A \,|\, A') \;\leq\; h(p) + p \log(k-1) \;<\; h(p) + p \log k$$

Stąd

$$I(A, A') \;=\; H(A) - H(A|A') \;\geq\; \log k - (h(p) + p \log k) \;=\; (1-p) \log k - h(p)$$

Ponadto $I(A', A) \;\leq\; I(A', B) \;\leq\; I(B', B) \;\leq\; n\mathcal{C}(\Phi)$, a więc

$$n\mathcal{C}(\Phi) \;\geq\; (1-p) \log k \;-\; h(p) \qquad \Rightarrow \qquad \mathcal{C}(\Phi) \;+\; \frac{h(p)}{n} \;\geq\; (1-p)\rho(c)\,.$$

Twierdzenie: Jeśli $|\mathcal{A}| = k$ i słowa kodowe $c : \mathcal{A} \to \mathcal{B}^n$ są przesyłane przez bezpamięciowy stacjonarny kanał o przepustowości $\mathcal{C}(\Phi)$, przy czym prawdopodobieństwo błędnego dekodowania otrzymanego komunikatu wynosi $p = \epsilon(c)$, wówczas efektywność transmisji spełnia nierówność

$$\rho(c) \;\leq\; \frac{\log k}{n} \;\leq\; \frac{\mathcal{C}(\Phi)}{1-p} \;+\; \frac{h(p)}{n(1-p)} \;\underset{p \to 0}{\longrightarrow}\; \mathcal{C}(\Phi)\,.$$

Twierdzenie Shannona o kodowaniu w obecności szumu

Dla binarnego symetrycznego kanału Φ i dowolnego $\varepsilon > 0$ istnieje dostatecznie duże n i kod $c_n : \mathcal{A}_n \to \{0,1\}^n$, dla którego efektywność transmisji spełnia nierówność

$$\rho(c_n) \;=\; \frac{\log |\mathcal{A}_n|}{n} \;\geq\; \mathcal{C}(\Phi) - \varepsilon\,,$$

z gwarancją, że $\epsilon(c_n) < \varepsilon$.

PRZYKŁAD. Zamierzamy nadać komunikat $\alpha \in \mathcal{A}^*$, $|\mathcal{A}| = k$, przez binarny kanał symetryczny o poziomie szumu $p = 0.1$. Jaką maksymalną efektywność transmisji możemy osiągnąć przy możliwie małym prawdopodobieństwie przekłamań $\epsilon(c)$?

$\mathcal{C}(\Phi) = 1 - h(0.1) = 0.53$, więc nie można liczyć na większą efektywność transmisji, niezależnie od sposobu kodowania. Ale tw. Shannona gwarantuje istnienie kodów, dla których efektywność może być dowolnie bliska 0.53. Niestety nie mówi nic o tym jak je konstruować.

Powiedzmy, że chcemy osiągnąć efektywność $\rho(c) = 0.5$, gwarantując jakość transmisji na poziomie $\epsilon(c) < 0.01$. Twierdzenie Shannona orzeka o istnieniu kodu $c_n : \mathcal{A}_n \to \{0,1\}^n$ o takich parametrach transmisji, za cenę być może dużego n (oznacza to, że należy kodować bloki liter z $\mathcal{A}$). Powiedzmy, że znamy ten kod i stosowne n. Jak zakodować napis α?

Wyznaczmy właściwą długość bloków liter

$$r \;=\; \left\lfloor \frac{0.5n}{\log k} \right\rfloor , \qquad \text{a więc} \quad |\mathcal{A}^r| \;=\; k^r \;\leq\; 2^{0.5n}$$

To gwarantuje, że można zakodować odwracalnie bloki liter z $\mathcal{A}^r$ w alfabecie $\mathcal{A}_n$, bo $|\mathcal{A}_n| \geq 2^{0.5n}$ skoro $\rho(c_n) \geq 0.5$. Efektywność transmisji wynosi wówczas

$$\frac{\log |\mathcal{A}^r|}{n} \;\simeq\; 0.5\,.$$