Kody polaryzujące

*Artur JEŻ**

* Wydział Matematyki i Informatyki, Uniwersytet Wrocławski

Z artykułu *O sztuce zadawania pytań* Δ_{22}^{12} dowiedzieliśmy się między innymi, jak można mierzyć „ilość informacji": taką miarą jest *entropia*, która intuicyjnie mówi nam, ile średnio bitów potrzeba, aby opisać wynik zmiennej losowej. Matematycznie, entropia zmiennej losowej A zadana jest jako $H(A) = \sum_a p_a \log_2 \frac{1}{p_a}$, gdzie p_a jest prawdopodobieństwem, że $A = a$, a $\log_2 \frac{1}{p_a}$ to intuicyjnie liczba bitów, którymi opisujemy to zdarzenie (liczba ta nie musi być oczywiście całkowita). Sama zaś entropia to wartość oczekiwana tejże liczby bitów; dlatego w dalszej części będę czasem pisał o bitach informacji zamiast o entropii. Ze wspomnianego artykułu dowiedzieliśmy się też, że wartość zmiennej losowej A można zakodować, używając średnio niewiele więcej niż $H(A)$ bitów. Dla uproszczenia notacji oznaczmy przez $H(p)$ entropię zmiennej losowej przyjmującej dwie wartości: $0, 1$, z prawdopodobieństwami $p, 1-p$. Proste rachunki pokazują, że $H(p)$ rośnie na przedziale $[0, \frac{1}{2}]$ i maleje na $[\frac{1}{2}, 1]$, a największą osiąganą wartością jest $H(\frac{1}{2}) = 1$.

Zmienna losowa to funkcja, która przypisuje zdarzeniom elementarnym liczby. Przykładowo moglibyśmy liczyć, ile razy wypadł orzeł przy rzucie monetą. Jeżeli rzucimy monetą raz, to zmienna losowa zdefiniowana będzie tak: $A(O) = 1$, $A(R) = 0$. Entropia w tym przypadku wynosi 1 (wynik opisuje 1 bit). Jeżeli rzucimy monetą dwa razy z rzędu, to zmienna losowa będzie następująca: $A(OO) = 2$, $A(OR) = A(RO) = 1$, $A(RR) = 0$. Prawdopodobieństwa kolejnych wyników wynoszą $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$, więc entropia jest równa 1,5: można powiedzieć, że wystarczy jeden bit, by opisać, że wynik jest równy 1, w przeciwnym razie potrzebujemy drugiego bitu, aby opisać, czy jest to 0, czy 2.

W tym artykule spróbujemy znaleźć rozwiązanie problemu błędów transmisji, pokażemy, co zrobić w przypadku, gdy między zapisem a odczytem mogą pojawiać się błędy. Niech $\mathbb{Z}_2$ będzie zbiorem $\{0, 1\}$ z dodawaniem modulo 2, które oznaczymy symbolem $\oplus$. Nadawany komunikat oznaczmy przez $X \in \mathbb{Z}_2$, a „szum" potraktujmy jako zmienną losową o wartościach w $\mathbb{Z}_2$, przy czym $\mathbb{P}(Z = 1) = p < 1/2$ to prawdopodobieństwo błędu. Odbiorca odczytuje $Y = X \oplus Z$. Ile bitów informacji można w tej sytuacji przesłać?

Podkreślmy, że entropia i kodowanie nie zależą od konkretnych wartości, jakie może przyjąć zmienna losowa, a jedynie od prawdopodobieństw ich przyjęcia. Wiemy, jakie wartości może przyjąć zmienna, i opisujemy jedynie, która z możliwości zaszła. Przykładowo, entropia zmiennej przyjmującej wartości 0, 1 z prawdopodobieństwem p, $1-p$ jest taka sama jak zmiennej przyjmującej wartości π, e z tymi prawdopodobieństwami.

Skoro Y ma dwie wartości, to $H(Y) \leqslant 1$. Ale $H(Z)$ informacji pochodzi z szumu, i wydaje się rozsądne stwierdzenie, że możemy przekazać najwyżej $H(Y) - H(Z) \leqslant 1 - H(p)$ bitów informacji. Powyższy model to uproszczenie pojęcia *kanału informacyjnego*, które wprowadził Claude Shannon; model ten jest, obok entropii, klasycznym pojęciem i narzędziem teorii informacji. Shannon sformalizował i udowodnił również powyższy argument o niemożności przesłania więcej niż $1 - H(p)$ bitów informacji.

Powyższe rozważania są trudno zrozumiałe w przypadku pojedynczego bitu: jak mamy przesłać $1 - H(p)$ bitu? Naturalniej jest rozważać transmisję *wielu* bitów: wysyłając n bitów, chcielibyśmy – pomimo szumu – przesłać $n(1 - H(p))$ bitów informacji. Ustalmy n i niech $m \approx H(p)n$ będzie liczbą bitów informacji pochodzących z szumu. Wysyłając n bitów, chcemy zatem przekazać $n - m$ bitów informacji. Formalnie: wiadomość $\vec{C} = (C_1, \ldots, C_{n-m}) \in \mathbb{Z}_2^{n-m}$ kodujemy (czyli ustalamy pewną funkcję różnowartościową $c : \mathbb{Z}_2^{n-m} \to \mathbb{Z}_2^n$) za pomocą wektora $\vec{X} = (X_1, \ldots, X_n) \in \mathbb{Z}_2^n$, który przesyłamy bit po bicie. Wektorem błędów jest $\vec{Z} = (Z_1, \ldots, Z_n) \in \mathbb{Z}_2^n$, gdzie $\mathbb{P}(Z_i = 1) = p < 1/2$ dla $1 \leqslant i \leqslant n$, czyli dla każdego bitu prawdopodobieństwo błędu wynosi p. Zakładamy, że zmienne $Z_1, \ldots, Z_n$ są niezależne. Odbiorca odczytuje $\vec{Y} = \vec{X} \oplus \vec{Z}$ i chce odtworzyć $\vec{C}$ przynajmniej z dużym prawdopodobieństwem (względem rozkładu $\vec{Z}$). O podobnym problemie pisałem już w artykule *Kody korekcyjne* Δ_{21}^3. Kodowanie c jest również kodem korekcyjnym, ale obecnie rozważamy inny model pojawiania się i poprawiania błędów. Poprzednio rozważaliśmy model pesymistyczny: błędy pojawiają się złośliwie i chcemy zawsze móc je poprawić, o ile nie jest ich za wiele; obecnie rozważamy model probabilistyczny: błędy pojawiają się losowo, ale chcemy z dużym prawdopodobieństwem bezbłędnie przekazać informację.

Na przykład, dla $n = 4$ i $m = 2$ naszą oryginalną wiadomość $\vec{C} = (0, 1)$ kodujemy w wektorze:

$$\vec{X} = (0, 1, 0, 1)$$

z użyciem pewnej funkcji c (opiszemy ją za chwilę). Ten wektor wysyłamy bit po bicie, ale ze względu na błędy transmisji bit 2 dochodzi zmieniony: wektorem błędu jest więc

$$\vec{Z} = (0, 1, 0, 0),$$

a odbiorca otrzymuje:

$$\vec{Y} = \vec{X} \oplus \vec{Z} = (0, 0, 0, 1).$$

Model pesymistyczny bywa nazywany *modelem Hamminga*, a probabilistyczny *modelem Shannona*.

Konstrukcja

Podejdziemy do problemu od strony szumu: skoro wiemy, że wektor błędów spełnia $H(\vec{Z}) \approx m$, to jak wspomnieliśmy na wstępie, możemy zakodować go, używając średnio m bitów. Pomysł polega na rozdzieleniu wysyłanych n bitów na dwa zbiory: m bitów przeznaczamy na zakodowanie $\vec{Z}$, a pozostałych $n - m$ bitów użyjemy do przekazania oryginalnej wiadomości. Pomysł ten może się w pierwszej chwili wydawać naiwny i niemożliwy do zrealizowania… a jednak,

jak pokażemy, jest to możliwe! W tym celu zastosujemy kompresję, która *zawsze*, a nie średnio, używa m bitów. Niech zatem P będzie oznaczeniem na funkcję kompresji ($P : \mathbb{Z}_2^n \to \mathbb{Z}_2^m$). Wprawdzie różne ciągi błędów mogą być teraz „zakodowane" w ten sam sposób, jednak dla pewnej funkcji dekompresującej Dec zachodzić będzie z dużym prawdopodobieństwem $\mathrm{Dec}(P(\vec{Z})) = \vec{Z}$. Ponadto P będzie *przekształceniem liniowym*, czyli $P(\vec{A} \oplus \vec{B}) = P(\vec{A}) \oplus P(\vec{B})$ oraz suriekcją (tzn. $P(\mathbb{Z}_2^n) = \mathbb{Z}_2^m$).

Formalnie równość $\mathrm{Dec}(P(\vec{Z})) = \vec{Z}$ zachodzi z prawdopodobieństwem, względem rozkładu $\vec{Z}$, dążącym do 1 przy $n \to \infty$. Dla uproszczenia będziemy jednak pisać równość.

Załóżmy na razie, że rzeczywiście uda nam się taką funkcję P skonstruować. Niech $\mathcal{C}_n = \{\vec{X} \in \mathbb{Z}_2^n \,:\, P(\vec{X}) = \vec{0}\}$. Zbiór $\mathcal{C}_n$ ma 2^{n-m} elementów: można bowiem pokazać, że warunek z definicji $\mathcal{C}_n$ odpowiada układowi równań liniowych, który ma 2^{n-m} rozwiązań (patrz margines). Ustalamy dowolną bijekcję między $\mathbb{Z}_2^{n-m}$ a $\mathcal{C}_n$, która pozwoli nam zakodować wiadomość $\vec{C}$ jako wektor $\vec{X}$ z $\mathcal{C}_n$. Korzystając z liniowości P, możemy teraz z $\vec{Y}$ odzyskać $P(\vec{Z})$:

$$P(\vec{Y}) = P(\vec{X} \oplus \vec{Z}) = P(\vec{X}) \oplus P(\vec{Z}) = P(\vec{Z}).$$

Po zdekodowaniu (z dużym prawdopodobieństwem) dostaniemy $\vec{Z}$, który po dodaniu do $\vec{Y}$ da szukany $\vec{X}$:

$$\vec{Y} \oplus \mathrm{Dec}(P(\vec{Y})) = (\vec{X} \oplus \vec{Z}) \oplus \mathrm{Dec}(P(\vec{Z})) = (\vec{X} \oplus \vec{Z}) \oplus \vec{Z} = \vec{X}.$$

Łatwo udowodnić, że wektor $(W_1, \ldots, W_m) = P(X_1, \ldots, X_n)$ jest postaci $W_i = c_{i,1}X_1 \oplus \cdots \oplus c_{i,n}X_n$ dla pewnych $c_{i,j} \in \mathbb{Z}_2$. Wtedy warunek $P(X_1, \ldots, X_n) = (0, \ldots, 0)$ przekłada się na układ m równań liniowych o n niewiadomych (z konstrukcji będzie wynikało, że równania te są niezależne). Można dowolnie ustalić $n - m$ niewiadomych (na 2^{n-m} sposobów) i jednoznacznie rozwiązać powstały układ, co daje 2^{n-m} wektorów w $\mathcal{C}_n$.

Użycie zbioru $\{\vec{X} \,:\, P(\vec{X}) = 0\}$ (dla odpowiednio dobranego P) do kodowania wiadomości jest standardowym podejściem w teorii kodów korekcyjnych, również kolejne kroki są standardowe: P nazywana jest macierzą parzystości, a $P(\vec{Z})$ to tak zwany syndrom. Trudność jest w konstrukcji odpowiedniego P.

Wróćmy więc do pytania, jak skonstruować P. Ograniczymy się do przypadku, gdy $n = 2^k$. Aby osiągnąć nasz cel, skonstruujemy najpierw odwracalną funkcję liniową $P'_n : \mathbb{Z}_2^n \to \mathbb{Z}_2^n$ i za P weźmiemy odpowiednio wybrane m z n współrzędnych P'_n. Konstrukcja P'_{2^k} jest rekurencyjna: definiujemy P'_1 jako identyczność, a dla $k \geqslant 1$

$$P'_{2^k}(\vec{Z}) = (P'_{2^{k-1}}(\vec{Z}_{\leqslant 2^{k-1}}) \oplus P'_{2^{k-1}}(\vec{Z}_{>2^{k-1}}), P'_{2^{k-1}}(\vec{Z}_{>2^{k-1}})).$$

Mamy więc $P'_2(Z_1, Z_2) = (Z_1 \oplus Z_2, Z_2)$, a konstrukcja P'_4 wygląda następująco:

Korzystamy z notacji dla wektorów $\vec{A}_{\leqslant i} = (A_1, \ldots, A_i)$, analogicznie definiujemy $\vec{A}_{\geqslant i}, \vec{A}_{<i}, \vec{A}_{>i}$.

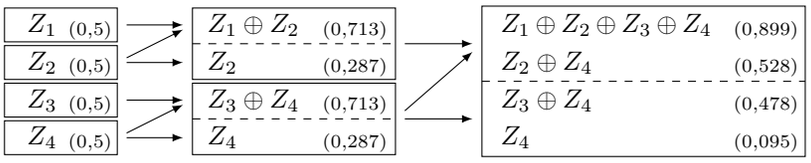

Diagram przedstawia konstrukcję P'_4: $P'_4(Z_1, Z_2, Z_3, Z_4)$ skonstruowane jest z $P'_2(Z_1, Z_2)$ oraz $P'_2(Z_3, Z_4)$. One z kolei skonstruowane są z $P'_1(Z_1), \ldots, P'_1(Z_4)$. W nawiasie podane są entropie warunkowe, które wyjaśnimy później.

Łatwo sprawdzić, że P' jest odwracalne. Co ciekawe, konstrukcja P' *nie zależy* od parametru p, ale P zależy: wybieramy $m \approx nH(p)$ współrzędnych z P'. Przykładowo dla $n = 4$ i $m = 2$, jeśli do P weźmiemy pierwsze dwie współrzędne P', otrzymamy

$$P(Z_1, Z_2, Z_3, Z_4) = (Z_1 \oplus Z_2 \oplus Z_3 \oplus Z_4, Z_2 \oplus Z_4).$$

Do pytania, które współrzędne należy wybrać, powrócimy za chwilę.

Dla tak zdefiniowanej macierzy P mamy: $\mathcal{C}_n = \{(0,0,0,0), (0,1,0,1), (1,0,1,0), (1,1,1,1)\}$, bo są to wszystkie wektory, które po zastosowaniu P przechodzą na $(0,0)$. Jeżeli teraz określimy bijekcję c w następujący sposób: $c(0,0) = (0,0,0,0)$, $c(0,1) = (0,1,0,1)$, $c(1,0) = (1,0,1,0)$, $c(1,1) = (1,1,1,1)$, to wspomniana przykładowa wiadomość $\vec{C} = (0,1)$ rzeczywiście zostanie zakodowana w $\vec{X} = (0,1,0,1)$. Kontynuując poprzedni przykład, jeżeli wektorem błędu będzie $\vec{Z} = (0,1,0,0)$, to stosując P na wektorze $\vec{Y} = \vec{X} \oplus \vec{Z}$, otrzymamy $P(\vec{Z})$:

$$P(\vec{Y}) = P(0,0,0,1) = (1,0) = P(\vec{Z}).$$

Analiza

Aby zrozumieć tę konstrukcję, użyjemy *entropii warunkowej*: $H(A|B = b)$ to entropia A, jeśli wiemy, że $B = b$, zaś $H(A|B) = \sum_b p_b \cdot H(A|B = b)$ to średnia liczb $H(A|B = b)$ ważona prawdopodobieństwami p_b.

Prześledźmy działanie P'_4. Weźmy $p = 0{,}11$, co daje $H(p) \approx 0{,}5$. Przyjmiemy więc $m = \frac{n}{2}$, czyli do P wybieramy połowę zmiennych z P'. Niech $W^{(j)} = (W_1^{(j)}, \ldots, W_4^{(j)})$ dla $j = 1, 2, 4$ będą kolejnymi wektorami uzyskanymi w konstrukcji P'_4, tzn.: $W^{(1)} = (P'_1(Z_1), \ldots, P'_1(Z_4))$, $W^{(2)} = (P'_2(Z_1, Z_2), P'_2(Z_3, Z_4))$, $W^{(4)} = P'_4(Z_1, \ldots, Z_4)$. Przeanalizujemy $H(W_i^{(j)}|W_{<i}^{(j)})$, czyli entropię warunkową, która mówi nam, ile bitów informacji niesie $W_i^{(j)}$, jeżeli poznaliśmy już wartości na poprzednich współrzędnych (wartości te podane są w nawiasach na diagramie).

Będziemy używać następujących własności entropii:

1. $H(A|B) \leqslant H(A)$: dodatkowa informacja nie może zaszkodzić; równość zachodzi dla A, B niezależnych.
2. $H(A,B) \leqslant H(A) + H(B)$ i równość zachodzi, gdy A, B są niezależne.
3. $H(A_1, \ldots, A_n) = H(A_1) + H(A_2|A_1) + \ldots + H(A_n|A_{<n})$: informacja całości to suma przyrostów informacji.
4. $H(A) \geqslant H(f(A))$ dla dowolnej funkcji f i równość zachodzi dla funkcji odwracalnych: przekształcanie może jedynie „stracić" informację.

$n = 1$: P'_1 to identyczność, więc $W_1^{(1)}, \ldots, W_4^{(1)}$ to $Z_1, \ldots, Z_4$ i są one niezależne, czyli $H(W_i^{(1)}|W_{<i}^{(1)}) = H(W_i^{(1)}) = H(p) \approx 0{,}5$.

$n = 2$: $P'_2(Z_1, Z_2) = (Z_1 \oplus Z_2, Z_2)$. Wtedy Ogólnie $p < 2p(1-p) < 1-p$ i z własności entropii mamy $H(W_1^{(2)}) = H(2p(1-p)) > H(p) = H(Z_1)$. Jednocześnie

$$H(Z_1 \oplus Z_2) + H(Z_2|Z_1 \oplus Z_2) = H(Z_1 \oplus Z_2, Z_2) = H(P'_2(Z_1, Z_2)) = \\ = H(Z_1, Z_2) = H(Z_1) + H(Z_2) \approx 1,$$

ponieważ P'_2 jest odwracalne, a Z_1, Z_2 niezależne. Czyli $H(Z_2|Z_1 \oplus Z_2) \approx 1 - 0{,}713 = 0{,}287$. Analogiczne obliczenia przeprowadzamy dla $Z_3 \oplus Z_4$ i Z_4.

$n = 4$: Dla $W_1^{(4)} = Z_1 \oplus Z_2 \oplus Z_3 \oplus Z_4$ i $W_2^{(4)} = Z_2 \oplus Z_4$ możemy użyć tej samej obserwacji: $W_1^{(4)} = (Z_1 \oplus Z_3) \oplus (Z_2 \oplus Z_4)$ i zmienne $(Z_1 \oplus Z_3), (Z_2 \oplus Z_4)$ spełniają $\mathbb{P}(Z_1 \oplus Z_3 = 1) = \mathbb{P}(Z_2 \oplus Z_4 = 1) \approx 0{,}2$, co daje $\mathbb{P}(W_1^{(4)} = 1) \approx 2 \cdot 0{,}2 \cdot 0{,}8 = 0{,}32$ i entropię $H(W_1^{(4)}) \approx 0{,}899$. Z równości $H(W_1^{(4)}) + H(W_2^{(4)}|W_1^{(4)}) = 2 \cdot H(0{,}2)$ wyliczamy $H(W_2^{(4)}|W_1^{(4)}) \approx 0{,}528$. Rachunki dla $H(W_3^{(4)}|W_{<3}^{(4)})$ są dużo bardziej skomplikowane, a ich wynik przedstawiony jest na diagramie.

Zauważmy, że P' *polaryzuje* entropię warunkową: pewne m współrzędnych ma entropię większą niż 0,5, a pozostałe $n - m$ mniejszą. W ogólności, dla dużych n pewne $n - m$ współrzędnych ma entropię bliską 0, a pozostałe m współrzędnych entropię prawie 1. Te właśnie współrzędne wybierzemy do funkcji P.

Czytelnik Uważny spostrzegł zapewne, że w przykładzie wybraliśmy *m pierwszych* współrzędnych. Był to jednak jedynie szczęśliwy zbieg okoliczności. W ogólności właściwe współrzędne rozłożone są dość chaotycznie i jest to istotna wada z punktu widzenia efektywnej implementacji tej konstrukcji. Dla ilustracji zauważmy, że zamiast $W_1^{(4)}, W_2^{(4)}$ moglibyśmy wybrać $W_1^{(4)}, W_3^{(4)}$ i obliczone entropie byłyby takie same (co łatwo wytłumaczyć symetrią między zmiennymi Z_2 i Z_3).

Istnienie funkcji dekompresującej Dec można pokazać, korzystając z polaryzacji. Dla uproszczenia notacji przyjmijmy, że do P wybraliśmy m pierwszych współrzędnych P'. Zauważmy najpierw, że potrafimy z dużym prawdopodobieństwem z $\vec{W}_{\leqslant m}^{(n)}$ obliczyć całe $\vec{W}^{(n)}$. Pierwsze m wartości po prostu znamy, a dla $j = m+1, \ldots, n$, skoro $H(W_j^{(n)}|\vec{W}_{i<j}^{(n)}) \approx 0$, to wartość zmiennej $W_j^{(n)}$ jest prawie pewna dla znanych $\vec{W}_{i<j}^{(n)}$: łatwe rachunki (patrz margines) pokazują, że bardziej prawdopodobna z wartości $W_j^{(n)}$ ma prawdopodobieństwo przynajmniej $1 - H(W_j^{(n)}|\vec{W}_{i<j}^{(n)}) \approx 1$ i to ją wybieramy na $W_j^{(n)}$. Można pokazać, że sumaryczne prawdopodobieństwo błędu jest małe; dowód jest prosty, ale wymaga dokładnej formalizacji własności polaryzacji. Gdy już mamy całe $\vec{W}^{(n)} = P'(\vec{Z})$, pozostaje zastosować funkcję odwrotną do P' i otrzymamy $\vec{Z}$.

Jeśli $A \in \mathbb{Z}_2$ oraz $p_0 = \mathbb{P}(A = 0)$ i $p_1 = \mathbb{P}(A = 1) = 1 - p_0$, to $\max(p_0, p_1) \geqslant 1 - H(A)$: lewa strona jest liniowo malejąca dla $p_0 \in [0, \frac{1}{2}]$ i liniowo rosnąca dla $p_0 \in [\frac{1}{2}, 1]$. Jednocześnie prawa strona jest wypukła i dla $p_0 = 0$, $p_0 = \frac{1}{2}$ oraz $p_0 = 1$ mamy równość.

Te bardziej prawdopodobne wartości musimy wcześniej obliczyć, dokonując obliczeń dla wszystkich możliwych wektorów wejściowych dla P'.

Dowód własności polaryzacji, a nawet dokładne sformułowanie tej własności, są trudne i – niestety – wykraczają zdecydowanie poza ramy tego artykułu.

Konstrukcja *kodów polaryzujących*, autorstwa Erdala Arıkana, była przełomem w teorii informacji: wszystkie poprzednie konstrukcje tego typu miały gorsze własności teoretyczne i zwykle były losowe, co utrudniało wydajną implementację. Kody polaryzujące są pierwszymi kodami, które nie tylko mają optymalne gwarancje teoretyczne, ale też nadają się do praktycznej implementacji, choć to drugie jest wyzwaniem. Zostały wprowadzone do standardu 5G NR (New Radio), a prace nad ich praktycznymi modyfikacjami i implementacjami trwają. Co ciekawe, kody korekcyjne i kodowanie entropijne były dwoma oddzielnymi, klasycznymi działami teorii informacji. Dopiero po kilkudziesięciu latach pokazano głęboki związek między nimi.

Erdal Arıkan, *Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels*. IEEE Transactions of Information Theory 55(7): 3051-3073 (2009).

Prezentację oparłem głównie na notatkach z wykładu Madhura Tulsianiego. Pełen dowód znajduje się w dostępnej w sieci książce: Venkatesan Guruswami, Atri Rudra, Madhu Sudan, *Essential Coding Theory*.