8. Samoopravné kódy

Digitální přenos informací spočívá v přenosu dlouhých posloupností číslic 0 a 1 (bitů). Velmi schématicky lze jednotlivé fáze přenosu informací znázornit následovně:

$$\begin{array}{ccc}
\text{zdroj informace} & & \\
\downarrow & & \\
\text{vysílač (kódovací zařízení)} & & \\
\downarrow & & \\
\text{komunikační kanál} & \leftarrow & \textbf{ŠUM} \\
\downarrow & & \\
\text{přijímač (dekódovací zařízení)} & & \\
\downarrow & & \\
\text{přijatá informace} & &
\end{array}$$

Samoopravné kódy jsou matematický nástroj, jak se vypořádat se šumem v komunikačním kanálu, který může pokazit informaci během přenosu.

Představme si, že naše kódovací zařízení kóduje informaci do posloupnosti délky 11. Jednotlivé předávaná "slova" jsou tedy posloupnosti složené z 11 cifer 0 a 1 a každá taková posloupnost je opravdu "slovem" v "jazyku" našich zpráv. Přijmeme-li posloupnost z jedenácti 0 a 1, je to pro nás smysluplné slovo, které mohlo být opravdu vysláno.

Dojde-li nyní během přenosu k poškození informace vlivem šumu v komunikačním kanálu, nelze to na přijímací straně nijak poznat. Nemáme tak vůbec jistotu, že jsme přijali stejnou informaci, která byla vyslána.

Nejjednodušší způsob, jak zjistit, došlo-li během přenosu slova k chybě vlivem šumu, spočívá v přidání dvanácté číslice tak, aby v takto prodlouženém slově byl sudý počet číslic 1. Tak například slovo

01010011101

obsahuje sudý počet číslic 1, přidáme tedy 0 na dvanácté místo. Dostaneme tak slovo

01010011101**0**,

ve kterém je dvanáctá (kontrolní) číslice zvýrazněna. Tato číslice nenese žádnou informaci, neboť ji lze odvodit z předchozích jedenácti číslic. Říká se jí *kontrolní symbol.* Každé slovo délky 12 tak obsahuje 11 *informačních symbolů* a jeden kontrolní symbol. Ještě jeden příklad přidání kontrolního symbolu:

01010001000**1**.

Této metodě se říká *kontrola parity,* anglicky **parity check.** Umožňuje rozpoznat, došloli během přenosu informace k jedné chybě. Pokud se ale vyskytnou dvě chyby, přijímací strana nic nepozná. Je důležité uvědomit si také cenu, kterou za možnost detekovat jednu chybu platíme. Touto cenou je zpomalení přenosu informace. Po přidání jednoho kontrolního symbolu vysíláme slova délky 12, ve kterých ale pouze prvních 11 symbolů nese informaci, poslední symbol není informační, je kontrolní. Rychlost přenosu informace se přidáním jednoho kontrolního symbolu snížila na $\frac{11}{12}$ původní rychlosti. To je docela malá cena v porovnání se ziskem, který přidání jednoho symbolu přineslo. Můžeme se nyní daleko více spolehnout na správnost přijaté informace.

Ověříme si to na následujícím smyšleném příkladu. Náš informační kanál tedy přenáší slova délky 11, jeho kapacita je 10^7 symbolů (bitů) za sekundu. Musíme ještě zvolit spolehlivost přenosu informace naším hypotetickým kanálem. Zvolíme ji $p = 1 - 10^{-8}$. To je pravděpodobnost, že se vysílaný symbol přenese správně, neboli $1 - p = 10^{-8}$ je pravděpodobnost, se kterou se vysílaný symbol pokazí.

Nyní spočítáme pravděpodobnost, že se slovo délky 11 přenese s nějakou chybou. Tato chyba se může vyskytnout na libovolném z 11 míst a to s pravděpodobností $1 - p$. Zbývajících 10 symbolů se přenese správně s pravděpodobností p^{10}. Celková pravděpodobnost, že se slovo délky 11 přenese s jednou chybou, je tedy

$$11p^{10}(1-p) \doteq \frac{11}{10^8}.$$

Za jednu sekundu se tedy přenese nesprávně přibližně

$$\frac{11}{10^8}\frac{10^7}{11} = 0,1$$

slov, čili v průměru se během 10 sekund přenese jedno slovo nesprávně (6 slov za minutu, 8640 za jeden den), aniž by přijímací strana měla nějakou možnost si toho povšimnout.

Teď budeme k vysílaným slovům přidávat dvanáctý kontrolní symbol, který bude kontrolovat paritu. Při přenosu takových slov se vždy pozná jedna chyba, dvě chyby se ale nepoznají. Jaká je nyní pravděpodobnost, že během přenosu jednoho slova dojde ke dvěma chybám? Počet možných chyb ve dvou symbolech z dvanácti je $\binom{12}{2}$, každá s pravděpodobností $1 - p$. Zbylých 10 symbolů se přenese správně, každý s pravděpodobností p. Pravděpodobnost toho, že při přenosu jednoho slova délky 12 dojde současně ke dvěma chybám je tedy

$$\binom{12}{2}p^{10}(1-p)^2 \doteq \frac{66}{10^{16}}.$$

To se za jednu sekundu stane přibližně u

$$\frac{66}{10^{16}}\frac{10^7}{12} \doteq 5,5 \cdot 10^{-9}$$

slov. Jedno slovo se tak přenese se dvěma chybami (a přijímací strana si toho tudíž nevšimne) přibližně jednou za 2000 dní, tj. jednou za 6 let.

Kontrola parity tedy umožňuje rozpoznat, že během přenosu došlo k jedné chybě, nedává ale možnost rekonstruovat původně odeslané slovo. Pokud tedy byla informace během přenosu poškozena, nezbývá než si vyžádat znovu její odeslání. To ale může být velmi nepraktické. Jednoduchou metodou, jak získat možnost opravit jednu chybu je vyslat každé slovo délky 11 třikrát těsně za sebou. Přijmeme tedy tři slova a pokud dojde během přenosu tohoto prodlouženého slova délky 33 k jedné chybě, přijmeme dvě stejná slova délky 11 a jedno, které se od těch dvou stejných bude lišit na jednom místě. Správné slovo délky 11 tedy bude mít většinu a prosadí se jako přijaté slovo. Tento “opakovací kód” (**repetition code**) je ale velmi pomalý. Z každých 33 symbolů je pouze 11 informačních a plných 22 kontrolních. Rychlost přenosu informace při takovém kódování by tedy byla pouze $\frac{1}{3}$ původní rychlosti. Ve skutečnosti lze ale té samé spolehlivosti přenosu (tj. schopnosti odhalit a opravit jednu chybu) dosáhnout přidáním pouhých 4 kontrolních symbolů, z každých 15 vyslaných symbolů je tedy plných 11 informačních a rychlost přenosu informace je tedy $\frac{11}{15}$ původní rychlosti při stejné spolehlivosti.

První *samoopravný kód,* tj. kód schopný odhalit a opravit jednu chybu, vymyslel **Richard Wesley Hamming** těsně po druhé světové válce. Hamming zemřel 7.ledna 1998 v Kalifornii ve věku 82 let a lze jej bez nadsázky nazvat jedním z otců informačního věku.

Ukážeme si první a nejjednodušší z Hammingových kódů, který se označuje $(7, 4, 3)$. Znamená to, že každé kódové slovo má délku 7, obsahuje 4 informační symboly, a “vzdálenost” libovolných dvou kódových slov je aspoň 3. Co je to vzdálenost, si řekneme později.

Čtyři informační symboly znamenají, že informaci v každém slově nesou čtyři číslice 0 nebo 1. Zbývající tři kontrolní symboly doplňují vysílaná slova na celkovou délku 7. Vysílaná slova jsou tedy tvaru

$$x_1x_2x_3x_4x_5x_6x_7,$$

kde každé $x_i = 0, 1$. Informační symboly budou symboly x_3, x_5, x_6, x_7 a kontrolní symboly budou x_1, x_2, x_4. Kontrolní symboly dopočítáme důmyslným použitím kontroly parity. Paritu budeme kontrolovat na určitých podmnožinách míst v posloupnosti $x_1x_2x_3x_4x_5x_6x_7$. Jsou-li dány informační symboly x_3, x_5, x_6, x_7, dopočítáme kontrolní symboly x_1, x_2, x_4 z rovností

$$x_1 + x_3 + x_5 + x_7 = 0,$$
$$x_2 + x_3 + x_6 + x_7 = 0,$$
$$x_4 + x_5 + x_6 + x_7 = 0.$$

Tyto rovnosti si snadno zapamatujeme. První rovnice sčítá symboly na místech, která mají v binárním vyjádření číslici 1 na místě jednotek. Druhá rovnice sčítá symboly na místech, která mají v binárním vyjádření číslici jedna na místě desítek. A konečně třetí rovnice sčítá symboly na místech, která mají v binárním vyjádření číslici 1 na místě stovek.

Informační symboly $x_3 = 0, x_5 = 1, x_6 = 0, x_7 = 1$ tedy zakódujeme do slova

010**0**101.

Podobně informační symboly $x_3 = 1, x_5 = 0, x_6 = 0, x_7 = 1$ zakódujeme takto:

001**1**001.

Jak nyní poznáme, na kterém místě došlo k chybě, pokud předpokládáme, že během přenosu zakódovaného slova komunikačním kanálem byl poškozen jeden symbol? To znamená, že aspoň jedna ze tří rovnic, které jsme použili při kódování, nebude pro přijaté slovo platit. Předpokládejme tedy, že při přenosu posledního slova byl poškozen třetí symbol, místo slova **0011**001 jsme tedy přijali slovo

$$\mathbf{0001}001 = x_1x_2x_3x_4x_5x_6x_7.$$

Spočítáme tři kontrolní rovnosti

$$\begin{aligned} x_1 + x_3 + x_5 + x_7 &= 1 = a_1 \\ x_2 + x_3 + x_6 + x_7 &= 1 = a_2 \\ x_4 + x_5 + x_6 + x_7 &= 0 = a_3. \end{aligned}$$

Výsledky si napíšeme jako jedno číslo ve dvojkové soustavě $a_3a_2a_1 = 011$. Výsledek je tedy binární zápis polohy poškozeného symbolu - změnili jsme symbol x_3 na třetím místě.

To není žádná náhoda. Zakódování slova proběhlo tak, aby se všechny tři součty rovnaly 0. Pokud tedy pro přijaté slovo vyjde první součet 1, znamená to, že byl poškozen symbol na některém z míst, které mají v binárním vyjádření 1 na místě jednotek. Pokud vyjde první součet 0, znamená to, že všechny symboly na místech, které mají 1 na místě jednotek, se přenesly správně. Podobně pro druhý součet. Pokud vyjde 1, byl poškozen symbol na některém z míst, které mají v binárním vyjádření 1 na místě desítek. Pokud vyjde 0, byly všechny symboly na místech, které mají v binárním vyjádření jednotku na místě desítek, přeneseny správně. A podobně pro třetí rovnost. Hammingův kód je tedy schopen lokalizovat chybu, pokud byl během přenosu kódového slova poškozen jeden symbol.

Hammingův kód má jasný geometrický význam. Máme-li kód délky n, jsou jeho prvky posloupnosti n symbolů 0 a 1. Jsou-li $\mathbf{x} = x_1x_2\cdots x_n$ a $\mathbf{y} = y_1y_2\cdots y_n$ dvě takové posloupnosti, pak říkáme, že jejich *Hammingova vzdálenost* se rovná počtu indexů i, pro které platí $x_i \neq y_i$. Formálně je Hammingova vzdálenost slov $\mathbf{x}$ a $\mathbf{y}$ rovna

$$d(\mathbf{x}, \mathbf{y}) = |\{i : x_i \neq y_i\}|,$$

kde pro nějakou množinu X symbol $|X|$ označuje počet prvků množiny X. Takto definovaná vzdálenost binárních posloupností délky n má všechny obvyklé vlastnosti metriky, zejména platí trojúhelníková nerovnost

$$d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$$

pro libovolné tři binární posloupnosti délky n.

Hammingův kód, který jsme uvedli v předchozích odstavcích, má tu vlastnost, že Hammingova vzdálenost libovolných dvou správně zakódovaných slov je aspoň 3. Proto je Hammingův kód označován $(7, 4, 3)$. Pokud přijímací strana přijme slovo $\mathbf{y}$ s poruchou, existuje mezi správně zakódovanými slovy nejvýše jedno slovo $\mathbf{x}$, které má od $\mathbf{y}$ vzdálenost 1. Kdyby totiž existovalo ještě jedno takové kódové slovo $\mathbf{z}$, platilo by podle trojúhelníkové nerovnosti pro Hammingovu vzdálenost

$$d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) = 1 + 1 = 2$$

a slova $\mathbf{x}, \mathbf{z}$ by byla dvě kódová slova se vzdáleností nejvýše 2, což pro Hammingův kód není možné. Každá dvě slova Hammingova kódu se liší aspoň ve třech symbolech, mají vzdálenost aspoň 3.

Podobnou úvahou dospějeme k závěru, že kód $(n, k, 2t+1)$ je schopný opravit jakékoliv slovo, ve kterém bylo pokaženo nejvýše t symbolů. V takovém kódu je tedy Hammingova vzdálenost libovolných dvou kódových slov aspoň $2t+1$.

V teorii kódování bývá zvykem uvádět při vysílání slov nejdříve informační symboly a teprve potom symboly kontrolní. Také Hamming takto upravil pořadí symbolů ve svém kódu. To lze udělat řadou různých způsobů, nejčastěji se používá následující pořadí:

$$x_7x_3x_5x_6x_1x_2x_4 = y_1y_2y_3y_4y_5y_6y_7,$$

kde $x_1x_2x_3x_4x_5x_6x_7$ je slovo zakódované způsobem, který jsme si uvedli. V přerovnaném slově jsou informační symboly na prvních čtyřech místech a po nich následují tři kontrolní symboly. Kontrolní rovnosti pro výpočet kontrolních symbolů jsou potom

$$\begin{aligned} y_5 + y_1 + y_2 + y_3 &= 0, \\ y_6 + y_1 + y_2 + y_4 &= 0, \\ y_7 + y_1 + y_3 + y_4 &= 0. \end{aligned}$$

Hammingův kód má následující vlastnost, která vyplývá z toho, jak počítáme kontrolní symboly (pomocí lineárních rovnic). *Jsou-li* $\mathbf{y}, \mathbf{z}$ *dvě slova Hammingova kódu, pak také jejich součet* $\mathbf{y}+\mathbf{z}$ *je slovem Hammingova kódu. Hammingův kód je tedy čtyřdimenzionální podprostor aritmetického prostoru* $\mathcal{Z}_2^7$. Tato vlastnost je společná řadě důležitých kódů.

Definice 8.19. Binární lineární kód dimenze k a délky n je k-dimenzionální podprostor $\mathcal{C}$ aritmetického prostoru $\mathcal{Z}_2^n$.

Obecnější lineární kódy jsou podprostory aritmetických prostorů $\mathcal{Z}_q^n$ nad konečným tělesem, které má q prvků. Zejména jsou důležité kódy nad tělesy, které mají 2^m prvků. Například NASA používá kód $(255, 223, 33)$ nad konečným tělesem, které má $2^8 = 256$ prvků. Každé slovo v tomto kódu je posloupnost 255 prvků tělesa velikosti 2^8. Prvky tohoto tělesa jsou posloupnosti složené z osmi symbolů 0,1. Libovolné slovo je tedy tvořené posloupností z 2040 symbolů 0,1, z nichž 256 je kontrolních.

Hammingův kód $(7, 4, 3)$ je tedy čtyřdimenzionální podprostor $\mathcal{Z}_2^7$. K popisu tohoto podprostoru stačí uvést nějakou jeho bázi, čili nějaké čtyři lineárně nezávislé prvky tohoto podprostoru. Dostaneme je například tak, že postupně zakódujeme čtveřice informačních symbolů $1000, 0100, 0010, 0001$. Dostaneme tak čtyři lineárně nezávislá slova Hammingova kódu:

$$1000\mathbf{111}, \qquad 0100\mathbf{110}, \qquad 0010\mathbf{101}, \qquad 0001\mathbf{011}.$$

Zapíšeme-li tato čtyři slova jako řádky do matice, dostaneme matici

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

Tato matice se nazývá *generující matice* Hammingova kódu $(7, 4, 3)$.

Generující matici H můžeme použít k zakódování libovolného slova $\mathbf{x} = a_1a_2a_3a_4$ tvořeného čtyřmi informačními symboly, které chceme odeslat. Vezmeme vektor $\mathbf{x} = (a_1, a_2, a_3, a_4)$ a vynásobíme jej maticí H zprava, tj. spočítáme 7-dimenzionální vektor

$$\mathbf{x}H.$$

A jak je to s dekódováním přijatého slova $\mathbf{y} = y_1y_2y_3y_4y_5y_6y_7$? K tomu je nutné použít matici kontroly parity pro Hammingův kód $(7, 4, 3)$

$$H^{\perp} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Všimněte si, že v sedmi řádcích matice $H^{\perp}$ je všech sedm nenulových binárních slov délky 3. A tři sloupce této matice obsahují koeficienty kontrolních rovností, pomocí kterých jsme ověřovali správnost přijatého slova (už po přeuspořádání informačních a kontrolních symbolů). Přijmeme-li slovo $y_1y_2y_3y_4y_5y_6y_7$, považujeme jej opět za binární vektor $\mathbf{y} = (y_1, y_2, y_3, y_4, y_5, y_6, y_7)$. “Přejedeme” jej maticí $H^{\perp}$ a dostaneme vektor

$$\mathbf{y}H^{\perp}$$

dimenze 3. Co z něho usoudíme? Rovná-li se $(0, 0, 0)$, přijaté slovo $\mathbf{y}$ splňuje všechny testy parity, přišlo tedy nepoškozené a přineslo informaci $y_1y_2y_3y_4$. Pokud je vektor $\mathbf{y}H^{\perp}$ nenulový, vyskytuje se v kontrolní matici $H^{\perp}$ na i-tém řádku. V přijatém slově $y_1y_2y_3y_4y_5y_6y_7$ byl potom poškozen symbol y_i. Ten opravíme a z opraveného slova si ponecháme první čtyři informační symboly.

Hammingův kód $(7, 4, 3)$ byl v praxi používán do konce padesátých let. Bellova společnost jej používala v telefonních ústřednách, firma IBM v počítačích. Firma IBM mohla Hammingův kód začít používat teprve poté, kdy Bellova společnost přišla o výhradní práva na používání Hammingova kódu po porážce v antimonopolním procesu v první polovině padesátých let. V současnosti jsou používány důmyslnější kódy, které nejenom opravují více chyb, ale používají rafinovanější metody kódování a dekódování založené na počítání s polynomy, které mají koeficienty z konečného tělesa velikosti 2^m.

Kontrolní matice $H^{\perp}$ je vyobrazena na Hammingově medaili, kterou uděluje americký Institut elektrických a elektronických inženýrů (The Institute of Electrical and Electronics Engineers, Inc., zkráceně IEEE) za “mimořádné výsledky v informačních vědách a systémech.” Hamming sám se stal v roce 1988 jejím prvním držitelem.