Teorie Informace - text k přednášce

Štěpán Holub a Michal Kupsa

7. května 2025

1 Entropie jako cena informace

V běžném jazyce někdy říkáme, že jsme obdrželi spoustu informací, ale že byly všechny bezcenné. Taková formulace poukazuje na skryté rozlišení mezi zprávou a její informační hodnotou. Informační hodnotu zprávy lze přitom chápat jako příspěvek k naší schopnosti rozlišovat, identifikovat stav věcí. Bezcenná je zpráva, která nás informuje o něčem, co už víme, případně o něčem, co nijak nepřispívá k tomu, co vědět chceme (tak můžeme chápat zprávu nepřesnou či lživou, která nás sice informuje o stavu mysli našeho informátora, ale nikoli o stavu věcí, které nás zajímají). Malou informační hodnotu má zpráva očekávaná s velkou pravděpodobností, velkou naopak zpráva nečekaná a nepravděpodobná. Z toho je vidět, že pojem informace je spojen s náhodností a nejistotou a s mírou, s jakou se nejistota díky přijaté zprávě změní. Teorie informace je kvantitativní, matematické uchopení uvedených neformálních intuicí.

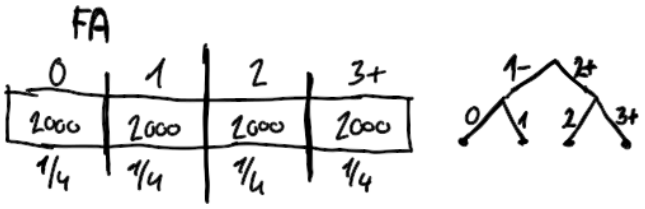
Začněme příkladem. Mějme nejmenovanou instituci, která má seznam 8000 lidí (říkejme jim souhrnně populace) a ráda by zjistila, kolik kreditních karet má každý z nich. Zadá tento důležitý úkol raději hned dvěma různým firmám.

Firma FA přinese seznam 8000 lidí, ve kterém jsou v kolonce počtu karet zapsány cifry 0,1,2 a 3+, kde 3+ znamená tři a více bez přesnějšího rozlišení. Počet lidí příslušející každé kategorii je 2000, viz Obrázek 1. Firma FB přinese stejný seznam, ve kterém je zapsáno 1-,2,3 a 4+, kde 1- znamená 1 nebo žádná, a 4+ znamená 4 a více. Počet lidí pro příslušné počty karet jsou po řadě 4000, 2000, 1000 a 1000. Viz Obrázek 2.

Otázka je, který seznam je z hlediska informační hodnoty kvalitnější. Toto samozřejmě souvisí s konkrétním využitím, ale zjednodušme si nyní situaci a měřme to pouze přes „náklady“ na získání informace v podobě počtu otázek. V zájmu rovných podmínek, uvažujme pouze otázky, na které se dá odpovědět ANO a NE. Kolik otázek musela položit první a druhá firma? Firma FA se mohla například ptát, zda má člověk 2 a více karet, a dle získané odpovědi následnou otázkou rozlišit 0 nebo 1, případně 2 a 3+. Tato strategie se dá popsat pomocí stromu otázek, případně pomocí posloupnosti „řezů“ populací, viz Obrázek 1b, kde pořadí řezů populací je indikován délkou svislé čáry (nejdelší je první).

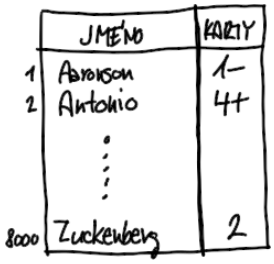
Tímto způsobem položí dohromady 16000 otázek, 2 otázky na člověka. Druhá firma může postupovat podobně, rozdělí si populaci první otázkou na dvě a dvě kategorie, a druhou otázkou se již dostane na jednu ze čtyř možností. Takto bude pokládat také 16000 otázek. Druhá firma má ale lepší možnost. Namísto strategie, která balancuje strom otázek z hlediska počtu možných odpovědí, může balancovat otázky z hlediska rozložení populace. Tento přístup vede k tomu, že první otázkou "1- vs 2+"rozdělí populaci napůl. V případě odpovědi 1- se už nemusí dále ptát, v druhém případě opět rozdělí příslušnou část populace napůl otázkou „2 vs 3+“. V případě odpovědi „3+“, je pak třeba položit ještě třetí otázku „3 vs 4+“, viz Obrázek 2b. Takto se sice stane, že budeme pokládat některým lidem tři otázky, ale celkem bude položeno jen $4000 + 2000 * 2 + 2000 * 3 = 14000$ otázek, v průměru $7/4$

(a) Seznam FA

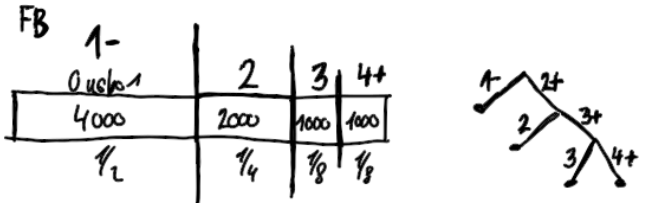


(b) Rozložení počtu kreditek, strom otázek

Obrázek 1: Průzkum dle firmy FA



(a) Seznam FB



(b) Rozložení počtu kreditek, strom otázek

Obrázek 2: Průzkum dle firmy FB

otázek na člověka. Firma FA tedy přinesla nákladnější informaci. Otázkou je, zda nákladnější znamená lepší.

Podívejme se nyní na věc pohledem klienta, neboli nejmenované instituce. Jak porovnat oba seznamy? Pokud by se jeden seznam dal zcela odvodit z druhého, jistě bychom prohlásili ten první za hodnotnější. To ovšem není tento případ, neboť seznamy se navzájem doplňují. Pokud je použijeme oba, jsme schopni u každého člověka říci, zda má 0,1,2,3 nebo 4+ kreditních karet. Takovéto spojení informací od obou firem do jednoho seznamu nabízí jiný způsob jak zjistit kvalitu informace. Konkrétně se můžeme ptát na dodatečné náklady, pokud by instituce chtěla doplnit seznam od firmy FA na sdruženou informaci. Firma FA by v takovém případě musela rozdělit kategorii 3+ na kategorie 3 a 4+. K tomu by stačilo položit $2000 * 1 = 2000$ otázek. Přepočteno na celkovou populaci to dává $1/4$ dodatečné otázky na člověka. Při doplňování seznamu od firmy FB na sdruženou informaci bychom potřebovali rozlišit kategorii $1-$ na kategorie 0 a 1. K tomu je třeba dodatečně položit $4000 * 1 = 4000$ otázek. V průměru 1/2 otázky na člověka (vztaženo k celé populaci). I z tohoto pohledu je tedy informace od firmy FA kvalitnější.

Dejme nyní tento příklad do souvislosti s matematickými pojmy. Populaci označíme Ω a seznamy od firem FA a FB budeme chápat jako funkce $X : \Omega \to A$ a $Y : \Omega \to B$, kde $A = \{0, 1, 2, 3+\}$ a $B = \{1-, 2, 3, 4+\}$. Na množině Ω dále uvažujme pravděpodobnost $\mathbb{P}$, která každému člověku přiřadí nějakou váhu, v našem případě všem stejnou. Tím se funkce X a Y stanou náhodnými veličinami v širším

slova smyslu. V teorii pravděpodobnosti se za náhodné veličiny standardně uvažují funkce z pravděpodobnostního prostoru do reálných čísel, což umožňuje správně definovat střední hodnotu, rozptyl atd. Pro teorii informace je přirozenější zabývat se zobrazeními do konečné, či spočetné množiny, kdy nepředpokládáme u možných hodnot žádnou další strukturu, či zavedené operace. Proto není problém, že je hodnotou například barva, typ vozu, či označení „3+“, jako v našem příkladě. Pro naše dvě náhodné veličiny pak definujeme entropii $\mathcal{H}(X)$ (resp. $\mathcal{H}(Y)$), sdruženou entropii $\mathcal{H}(X,Y)$ a podmíněnou entropii $\mathcal{H}(X|Y)$ (resp. $\mathcal{H}(Y|X)$) pomocí vzorců

$$\begin{aligned}
\mathcal{H}(X) &= -\sum_{a\in A} \mathbb{P}(X=a)\log \mathbb{P}(X=a),\\
\mathcal{H}(X,Y) &= -\sum_{a\in A,b\in B} \mathbb{P}(X=a,Y=b)\log \mathbb{P}(X=a,Y=b),\\
\mathcal{H}(X|Y) &= -\sum_{a\in A,b\in B} \mathbb{P}(X=a,Y=b)\log \mathbb{P}(X=a|Y=b),
\end{aligned}$$

kde nedefinované členy sum ($0\log 0, 0\log\frac{0}{0}$) ignorujeme, t.j. považujeme je za nulu. Základ logaritmu je fixní a obvykle ho volíme rovný 2. Jednoduchým výpočtem pak lze zjistit, že

$$\mathcal{H}(X)=2, \qquad \mathcal{H}(Y)=\frac{7}{4},$$
$$\mathcal{H}(X,Y)=\frac{9}{4}, \qquad \mathcal{H}(X|Y)=\frac{1}{2}, \qquad \mathcal{H}(Y|X)=\frac{1}{4}.$$

Veličiny $\mathcal{H}(X)$ a $\mathcal{H}(Y)$ číselně odpovídají průměrnému počtu otázek na člověk, které musí položit firma FA či FB. Platí dokonce, že ve vzorci průměrované hodnoty $-\log\mathbb{P}(X=a)$ odpovídají přesně počtu otázek položených člověku z kategorie „a“ v ideální strategii pro firmu FA, podobně $-\log\mathbb{P}(Y=b)$ odpovídá počtu otázek pro člověka z kategorie „b“ při ideální strategii pro firmu FB. Podobně, $-\log\mathbb{P}(X=a\mid Y=b)$ zde odpovídá počtu dodatečných otázek pro člověka z kategorie a, víme-li od firmy FB, že člověk patří do kategorie „b“, z čehož pak plyne, že $\mathcal{H}(X|Y)$ je rovno průměrnému počtu doplňujících otázek na člověka při znalosti seznamu od firmy FB.

Entropii se tedy dá rozumět jako průměrné kvantitě informace, kterou získáme, jsme-li schopni rozdělit jednolitou populaci do určitých skupin, například dle počtu kreditních karet. Jde tedy o kvantifikaci naší schopnosti rozlišovat. podobně se dá interpretovat sdružená entropie, kde rozlišujeme do jemnějších kategorií. Podmíněná entropie $\mathcal{H}(X|Y)$ pak je kvantifikace schopnosti rozlišit v populaci kategorie veličiny X, pokud už umíme rozlišovat do kategorií veličiny Y.

Na závěr ještě zmiňme, že rovnost průměrného počtu otázek v různých scénářích s příslušnou entropií bývá v obecném případě jen přibližná. Takto pěkně příklad vyšel díky tomu, že všechny uvažované pravděpodobnosti, včetně těch podmíněných, byly celočíselnými mocninami čísla, které jsme zvolili jako základ pro logaritmus ve vzorcích pro entropii.

2 Pravděpodobnostní prostor a náhodné veličiny

V těchto přednáškách budeme používat následující standardní definici pravděpodobnostního prostoru, pravděpodobnostní míry a náhodné veličiny.

Měřitelný prostor je dvojice $(\Omega, \mathcal{F})$, kde Ω je množina a $\mathcal{F}$ je **σ-algebra**. Pojem σ-algebry je vymezen požadavky, aby $\mathcal{F}$ obsahovala celé Ω a byla uzavřena na spočetná sjednocení a komplement, tj.

- $\Omega \in \mathcal{F}$
- $V \in \mathcal{F} \Rightarrow \Omega \setminus V \in \mathcal{F}$
- $V_i \in \mathcal{F}, i \in \mathbb{N} \Rightarrow \bigcup_{i\in\mathbb{N}} V_i \in \mathcal{F}$.

Podmnožiny Ω ležící v $\mathcal{F}$ se nazývají **měřitelné**. Z pravidel pro σ-algebru vyplývá také to, že $\mathcal{F}$ je uzavřená na spočetné průniky, konečná sjednocení a průniky a že obsahuje prázdnou množinu. Termín „měřitelný“ už sám naznačuje, že se na systému $\mathcal{F}$ chystáme zavést míru. V našem případě to bude navíc vždy míra pravděpodobnostní. **Pravděpodobnostní míra** $\mathbb{P}$ je zobrazení z $\mathcal{F}$ do intervalu $[0, 1]$, které splňuje:

- $\mathbb{P}(\Omega) = 1$;
- (σ**-aditivita**) je-li $V_i \in \mathcal{F}, i \in \mathbb{N}$, systém vzájemně disjunktních množin, pak

$$\mathbb{P}\left(\bigcup_{i\in M} V_i\right) = \sum_{i\in\mathbb{N}} \mathbb{P}\left(V_i\right).$$

Trojici $(\Omega, \mathcal{F}, \mathbb{P})$ nazýváme **pravděpodobnostní prostor**.

Platí, že i pravděpodobnost *konečného* sjednocení vzájemně disjunktních množin je součtem pravděpodobností; v definici σ-aditivity stačí zvolit všechny množiny až na konečný počet prázdné a všimnout si, že pravděpodobnost prázdné množiny je nula.

Je-li $(A, \mathcal{F}')$ měřitelný prostor a $(\Omega, \mathcal{F}, \mathbb{P})$ pravděpodobnostní prostor, pak **náhodná veličina** X je **měřitelné zobrazení** $X : \Omega \to A$, přičemž měřitelným nazýváme takové zobrazení, pro které je *vzor* každé měřitelné množiny měřitelný. Řekneme, že náhodná veličina je **konečná**, pokud je A konečná; **diskrétní**, pokud je A nejvýše spočetná; a **reálná**, pokud $A \subseteq \mathbb{R}$.

Definice náhodné veličiny vyžaduje, aby A byla množina s mírou, nebo aby na ní byl alespoň definován systém měřitelných množin. V případě konečných a diskrétních náhodných veličin můžeme (a bez výjimky budeme) předpokládat, že měřitelné jsou všechny podmnožiny A, tedy měřitelný systém $\mathcal{F}'$ je celá potenční množina množiny A, pro kterou zavedeme symbol 2^A.

Poznámka: Symbolem B^A značíme obvykle množinu $\{f \mid f : A \to B\}$ funkcí z A do B, což odpovídá představě, že taková funkce je „A-tic prvků z B“, jako je např. naše p výše šesticí reálných čísel. Podmnožinu $Y \subseteq A$ můžeme současně identifikovat s její charakteristickou funkcí $\chi_Y : A \to \{0, 1\}$ indikující, které prvky v Y leží, a které ne (pokud nulu a jedničku chápeme jako booleovské hodnoty „pravda“ „nepravda“, jedná se o predikát „leží v A“). Zbývá použít množinové kódování čísla 2 právě jako $\{0, 1\}$.

Touto notací si ušetříme symbol $\mathcal{P}$. Různých „p“ je v pravděpodobnosti už víc než dost.

U reálných náhodných veličin budeme za $\mathcal{F}'$ brát systém Borelovských množin $\mathcal{B}$, tedy nejmenší σ-algebru, která obsahuje intervaly. Borelovské množiny zejména obsahují všechny jednoprvkové množiny.

Pro danou náhodnou veličinu $X : \Omega \to A$ definujme zobrazení $P_X : \mathcal{F}' \to \mathbb{R}$ předpisem

$$P_X(V) := \mathbb{P}\left(X^{-1}(V)\right).$$

Důležitým pozorováním je, že P_X je pravděpodobnostní míra na $(A, \mathcal{F}')$ indukovaná náhodnou veličinou X. Máme tedy nový pravděpodobnostní prostor $\left(A, \mathcal{F}', P_X\right)$. Tomuto zobrazení říkáme **rozdělení náhodné veličiny** X.

Pro diskrétní veličiny navíc definujeme $P_X(a) := P_X(\{a\})$, čímž dostáváme zobrazení $P_X : A \to \mathbb{R}$, což je reálná náhodná veličina právě na pravděpodobnostním prostoru $\left(A, \mathcal{F}', P_X\right)$. Tomuto zobrazení budeme rovněž říkat **rozdělení náhodné veličiny** X.

Poznámka: Jedná se o typické „zneužití terminologie“: symbol P_X a pojem *rozdělení náhodné veličiny* mají nyní dva významy podle toho, jestli je argumentem prvek, nebo podmnožina A. Přestože např. tvrzení, že P_X je náhodná veličina na prostoru $\left(A, 2^A, P_X\right)$ může být na první pohled matoucí, málokdo by nakonec v této situaci volal po zavedení nového symbolu a nového termínu.

V programovacích jazycích se takový postup nazývá „function overloading“. I fakt, že to řada jazyků podporuje, ukazuje, že nedorozumění při dostatečné míře pozornosti nehrozí.

Oba významy pojmu rozdělení náhodné veličiny jsou ovšem úzce svázány, protože rozdělení definované na prvcích pomocí aditivity definuje rozdělení definované na podmnožinách.

Poznámka: Pro spočetné diskrétní veličiny to s sebou nese problematiku konvergence řad, která je ovšem v případě pravděpodobnostního prostoru triviální, protože sčítáme nezáporné hodnoty a součet je shora omezen jedničkou, takže všechny součty existují, jsou to limity rostoucí omezené posloupnosti částečných součtů.

Tento postup naopak nedává dobrý smysl ve spojitém případě $A = \mathbb{R}$, kdy obvykle platí $P_X(a) = 0$ pro každé $a \in \mathbb{R}$ (viz níže úvahu o nekonečných

posloupnostech nezávislých hodů mincí). V takovém případě se rozdělením náhodné veličiny rozumí pouze první význam P_X, tedy funkce $P_X : \mathcal{F}' \to \mathbb{R}$. Roli podobnou našemu rozdělení diskrétní náhodné veličiny $P_X : A \to \mathbb{R}$ plní *distribuční funkce*, definovaná pro $a \in \mathbb{R}$ jako $\mathbb{P}\left(X^{-1}(-\infty, a]\right)$. Tato funkce určuje míru vzorů pro intervaly (a, b), a tedy i pro systém Borelovských množin, který je ve spojitém případě naší volbou pro $\mathcal{F}'$.

Bude také často vhodné chápat nejvýše spočetnou množinu A jako pravděpodobnostní prostor $\left(A, 2^A, P\right)$, kde pravděpodobnostní míra P není nutně indukovaná nějakou náhodnou veličinou. Taková míra je opět ekvivalentní zobrazení $P : A \to [0, 1]$ splňujícímu

$$\sum_{a \in A} P(a) = 1,$$

které budeme nazývat **(diskrétní) rozdělení na** A. Množinu všech takových rozdělení budeme značit Δ_A.

O tvrzeních, která platí všude až na množinu míry nula, řekneme, že platí **skoro jistě** ($\mathbb{P}$-skoro jistě), což budeme zkracovat na s.j. a psát v případě potřeby nad znak rovnosti nebo za tvrzení. Pokud tedy např. řekneme, že dvě množiny $M, N \in \mathcal{F}$ jsou si rovny skoro jistě ($\mathbb{P}$-skoro jistě), zapsáno $M = N$ s.j. , znamená to, že jejich symetrická diference $M \triangle N$ má míru nula. Zároveň dovolíme, aby náhodné veličiny byly definované „skoro jistě“, tedy aby jejich definice byla korektní na množině s pravděpodobností 1. Všimněte si, že i takto definované veličiny mají např. dobře definovanou střední hodnotu v tom smyslu, že její hodnota nezáleží na případném dodefinovaní.

Abychom mohli mmožiny s nulovou mírou (nemožné jevy) výslovně ignorovat, zavedeme pro libovolné rozdělení P (na nejvýše spočetné množině A) pojem **nosiče rozdělení**:

$$s(P) = \{a \in A \mid P(a) > 0\}.$$

Definujeme také **nosič diskrétní náhodné veličiny** X jako

$$s(X) = X^{-1}\left(s\left(P_X\right)\right) = \left\{\omega \in \Omega \mid P_X(X(\omega)) > 0\right\} = \left\{\omega \in \Omega \mid X(\omega) \in s\left(P_X\right)\right\}.$$

V obou případech dostáváme ze σ-aditivity míry vztahy

$$P(s(P)) = 1 \quad \text{a} \quad P_X\left(s\left(P_X\right)\right) = \mathbb{P}(s(X)) = 1.$$

(Připomeňme si znovu, že takto jednoduchý přístup k nosiči není možný pro spojité veličiny, u kterých by podobně definovaný nosič mohl být i prázdný. Všimněte si také, že zatímco nosič rozdělení je nosičem v běžném smyslu množiny, na které je daná funkce nenulová, nosič náhodné veličiny je definován zprostředkovaně přes nosič rozdělení.)

Pro obor hodnot A náhodné veličiny X často uvažujeme nějaké další ohodnocení $f : A \to \mathbb{R}$ (případně $f : s\left(P_X\right) \to \mathbb{R}$), a to zejména (ale nejen) pokud veličina X není reálná. Je f měřitelné zobrazení (tedy zejména pokud je X diskrétní), je takto

definována nová, reálná náhodná veličina $f \circ X$. Střední hodnota této reálné veličiny pak splňuje vztah

$$\mathbb{E}\left(f(X)\right) = \sum_{a \in s(P_X)} P_X(a) \cdot f(a).$$

Připomeňme, že v klasické teorii pravděpodobnosti má konečná reálná náhodná veličina $Y : \Omega \to B \subset \mathbb{R}$, kde B je konečná množina, střední hodnotu definovanou vztahem

$$\mathbb{E}\left(Y\right) = \sum_{b \in B} P_Y(b) \cdot b.$$

Tento základní vzorec se pak v případě konečné náhodné veličiny $Y = f(X)$ liší od vztahu uvedeného výše pouze seskupením sčítanců dle jádra zobrazení f a ignorováním (některých) nulových členů sumy (zkuste si rozmyslet).

V dalším textu budeme využívat také následující vlastnosti střední hodnoty, které lze odvodit z výše uvedených vztahů. Pro konečné reálné náhodné veličiny X a Y a $\alpha \in \mathbb{R}$ platí

- $X \overset{\text{s.j.}}{=} Y$, pak $\mathbb{E}\left(X\right) = \mathbb{E}\left(Y\right)$.
- $X \leq Y$ s.j. , pak $\mathbb{E}\left(X\right) \leq \mathbb{E}\left(Y\right)$
- $X \leq Y$ s.j. a $\mathbb{E}\left(X\right) = \mathbb{E}\left(Y\right)$, pak $X \overset{\text{s.j.}}{=} Y$.
- $\mathbb{E}\left(\alpha X + Y\right) = \alpha \mathbb{E}\left(X\right) + \mathbb{E}\left(Y\right)$.
- $Y = X$ s.j. , pak $f(X) = f(Y)$ s.j. a $\mathbb{E}\left(f(X)\right) = \mathbb{E}\left(f(Y)\right)$ pro libovolnou měřitelnou funkci f.

> **Poznámka:** Uvedené vztahy platí pro libovolné reálné, nejen pro konečné, náhodné veličiny. Pro obecné reálné veličiny je však třeba použít teorii míry.

Všimněme si, že střední hodnota je vlastností rozdělení X a lze ji také chápat jako střední hodnotu náhodné veličiny f na pravděpodobnostním prostoru $\left(A, 2^A, P_X\right)$. Dává tedy smysl mluvit o střední hodnotě $\mathbb{E}_P\left(f\right)$ funkce $f : A \to \mathbb{R}$ i pro obecné rozdělení P na množině A. Hodnoty $P(a)$ jsou **váhy** prvků množiny A a střední hodnota

$$\mathbb{E}_P\left(f\right) = \sum_{a \in s(P)} P(a) \cdot f(a),$$

je **vážený průměr**, resp. **konvexní kombinace** jejích hodnot. Tyto dva pojmy jsou pro nás synonymy.

> **Poznámka:** Kromě váženého průměru (konvexní kombinace) reálných čísel, budeme v další kapitole uvažovat také vážený průměr souboru rozdělení, které bude opět rozdělením na dané konečné množině.

*

V těchto přednáškách budeme téměř výhradně pracovat s konečnými veličinami. Základní vlastnosti teorie informace jsou totiž nejlépe viditelné na takových veličinách a byly pro ně zavedeny. Možné rozšíření na spojité (nebo spočetné) veličiny proto pouze přináší komplikace, které výkladu nijak nepomáhají. Z toho také plyne, že nebudeme potřebovat téměř žádné netriviální poznatky z teorie míry. Pokusme se nyní vysvětlit, proč jsou přesto výše uvedené definice používající jazyk teorie míry vhodné, či dokonce nutné. Bude to tedy připomínka toho, jak matematika přistupuje k pojmu pravděpodobnosti a proč.

Uvažujme hod (ne nutně férovou) kostkou, jejíž stěny jsou popsány písmeny a, b, c, d, e, f. Chování kostky je jednoduše definováno šesticí pravděpodobností $(p_1, p_2, p_3, p_4, p_5, p_6)$ pro jednotlivé výsledky, což je prostě šestice nezáporných reálných čísel se součtem jedna. Pokud p chápeme jako zobrazení

$$p : \{1, 2, 3, 4, 5, 6\} \to \mathbb{R},$$

což je přirozený pohled pro jakékoli indexování, dostáváme pravděpodobnostní prostor

$$\Omega = \{1, 2, 3, 4, 5, 6\},$$

na němž je zobrazení p rozdělením, které definuje míru na celém 2^Ω předpisem

$$\mathbb{P}(\{n\}) = p(n).$$

Náhodnou veličinu popisující hod kostkou nyní definujeme jako zobrazení (bijekci)

$$X : \Omega \to \{a, b, c, d, e, f\}.$$

Pravděpodobnost, že na kostce padne např. c, pak je mírou vzoru jednoprvkové množiny $\{c\}$, formálně

$$\mathbb{P}(X = c) := \mathbb{P}\left(X^{-1}(\{c\})\right) = \mathbb{P}(\{3\}) = p(3),$$

kde $\mathbb{P}(X = c)$ je pouhá notační konvence odpovídající neformálnímu „pravděpodobnost, že X je rovno c“. Všimněme si, že pokud by pravděpodobnost byla definována na prvcích Ω, nebylo by možné definovat např. „pravděpodobnost, že hodnota X je samohláska“, což je $\mathbb{P}(\{1, 5\})$. Z aditivity pak dostáváme

$$\mathbb{P}(\{1, 5\}) = \mathbb{P}(\{1\}) + \mathbb{P}(\{5\}) = p_1 + p_5,$$

aniž bychom museli pravděpodobnost uvažované události „padla samohláska“ zvlášť definovat. To také vede k tomu, že za **jevy** označujeme (měřitelné) podmnožiny Ω. Poněkud neintuitivně je tedy v našem případě jevem „padla samohláska“ množina $\{1, 5\}$, tedy $X^{-1}(\{a, e\})$, nikoli samo $\{a, e\}$. To je obvyklým zdrojem zmatku,

zejména pro začátečníky, a také proto, že se tento formální přístup velmi často nedodržuje a zavádějí se různé konvence podobné námi výše zavedenému $\mathbb{P}(X = c)$.

Jednoprvkové podmnožiny Ω, budeme nazývat **elementárními jevy**, ale jen tehdy, *jsou-li měřitelné*. Poměrně častá konvence označující za elementární jevy *prvky* Ω je terminologicky nešťastná. O prvcích Ω budeme raději mluvit jako o **stavech**, a o Ω pak jako o **stavovém prostoru**.

*

Nejasným zůstává, proč jsme u naší kostky zavedli nové indexy $\{1, 2, 3, 4, 5, 6\}$ a nepsali raději p_a, p_b, p_c, p_d, p_e, p_f, což by opět byly zkratky pro $\mathbb{P}(\{a\})$ atd. Za stavový prostor bychom tak uvažovali přímo množinu $\{a, b, c, d, e, f\}$. Jevem např. „X je samohláska“ by pak v souladu s intuicí byla množina $\{a, e\}$. Takový naivní přístup je možný, a přesně odpovídá tomu, že místo pravděpodobnostního prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ pracujeme s prostorem $(A, 2^A, P)$, kde $A = \{a, b, c, d, e, f\}$ a P je míra daná hodnotami p_a, p_b, p_c, p_d, p_e, p_f, což je rozdělení na A.

Výhoda oddělení stavového prostoru od množiny hodnot se nicméně výrazně projeví, když začneme uvažovat více náhodných veličin. Uvažme např. vedle výše definovaného hodu kostkou ještě hod (opět ne nutně férovou) mincí s hodnotami $\{\texttt{panna}, \texttt{orel}\}$ s příslušnými pravděpodobnostmi $p_{\texttt{p}}$ a $p_{\texttt{o}}$. V běžném uvažování bez velké diskuse předpokládáme, že je možné obě veličiny kombinovat a mluvit např. o pravděpodobnosti jevu „na kostce padlo b a na minci `orel`“. To ale znamená, že uvažujeme novou náhodnou veličinu s hodnotami v množině $\{a, b, c, d, e, f\} \times \{\texttt{panna}, \texttt{orel}\}$. Pokud bychom neměli společný stavový prostor, musíme zavést novou pravděpodobnostní míru. Stavový prostor i míra se tak budou neustále měnit v závislosti na uvažovaných veličinách. V novém prostoru náhle nebude zcela zřejmé ani to, co míníme jevem, že na kostce padlo b, resp. jaký je vlastně mezi jednotlivými prostory vztah. Budeme muse ztotožnit staré $\{b\}$ s novým $\{(b, \texttt{panna}), (b, \texttt{orel}))\}$ apod. Výhody jednoho společného stavového prostoru, na němž jsou definovány všechny uvažované náhodné veličiny způsobem popsaným výše, se stávají zřejmými. Odpovídá to představě jednoho světa, jehož *stav* ω kompletně určuje hodnoty všech náhodných veličin. Je-li příslušná jednoprvková množina $\{\omega\}$ měřitelná, můžeme ji chápat jako jev, který plně charakterizuje stav světa.

Soubory více náhodných veličin definují nové náhodné veličiny, kterým říkáme **náhodné vektory** nebo **sdružené náhodné veličiny**. Jsou-li X_i, $i \in I$, náhodné veličiny s obory hodnot A_i, definují náhodný vektor $X_I := (X_i)_{i \in I}$, což je zobrazení z Ω do kartézského součinu množin A_i, dané předpisem

$$X_I(\omega) = \big(X_i(\omega)\big)_{i \in I} .$$

Není však jasné, zda můžeme toto zobrazení nějak přirozeně chápat jako náhodnou veličinu. K tomu je totiž potřeba definovat na kartézském součinu množin A_i nějakou σ-algebru měřitelných množin. To je vždy možné, ale v obecném případě to předpokládá pokročilejší úvahy z teorie míry. Nás bude nicméně zajímat jednoduchý

případ (konečných) vektorů

$$X_{[0..n)} := \left(X_0, X_1, \ldots, X_{n-1}\right),$$

kde všechna X_i jsou diskrétní náhodné veličiny. Pak bude i náhodný vektor diskrétní, a budeme se tedy moci držet našeho předpokladu, že jsou měřitelné všechny podmnožiny oboru hodnot. I přesto se však bez dostatečně bohatého pravděpodobnostního prostoru neobejdeme, pokud chceme uvažovat *neomezený* počet různých náhodných veličin, což bude případ studia náhodných procesů. Stačí uvážit neomezený počet nezávislých hodů mincí. Uvažme např. spočetně mnoho nezávislých hodů férovou mincí, tedy nezávislé binární uniformní náhodné veličiny, které označme X_i, $i \in \mathbb{N}$. Všimněme si, že každé $\omega \in \Omega$ definuje nekonečnou posloupnot z $\{0,1\}^{\mathbb{N}}$. Tím je definováno zobrazení $X_{\mathbb{N}} : \Omega \to \{0,1\}^{\mathbb{N}}$, které je limitou vektorů $X_{[0..n)}$. Protože měřitelné množiny tvoří σ-algebru, je pro každou nekonečnou posloupnost $s \in \{0,1\}^{\mathbb{N}}$ množina

$$M_s := \bigcap_{n\in\mathbb{N}} X^{-1}_{[0..n)}\left(s_{[0..n)}\right) = X^{-1}_{\mathbb{N}}(\{s\})$$

měřitelná, jsouc spočetným průnikem měřitelných množin. Z předpokladu nezávislosti X_i plyne, že

$$\mathbb{P}\left(X^{-1}_{[0..n)}\left(s_{[0..n)}\right)\right) = 2^{-n},$$

a tedy M_s má míru nula. Pro mnoho s může být M_s prázdné, nicméně soubor neprázdných M_s

$$\left\{M_s \mid s \in \{0,1\}^{\mathbb{N}}, M_s \neq \emptyset\right\}$$

je disjunktním rozkladem Ω na měřitelné množiny míry nula. Ze σ-aditivity míry plyne, že tento rozklad, a tedy ani Ω, nemohou být spočetné. Potřebujeme tedy nespočetně velký stavový prostor, v němž má každý elementární jev (tedy výše zmíněný úplný popis světa) nulovou pravděpodobnost.

Zobrazení $X_{\mathbb{N}}$ můžeme samo považovat za náhodnou veličinou, čímž se ovšem dostáváme na pole spojitých náhodných veličin, čemuž jsme se chtěli vyhnout. V těchto úvahách se tak ukazuje význam slova „téměř“ ve slibu, že se budeme zabývat pouze konečnými veličinami. Pokud totiž chceme takových konečných veličin uvažovat neomezeně mnoho, spojité veličiny se budou stále vznášet na pozadí jako limitní případy. Přesto platí, že v našich přednáškách budeme mluvit výhradně o konečných náhodných vektorech s konečně mnoha hodnotami. Otázky měřitelnosti tedy nebudou hrát v našich úvahách žádnou roli a můžeme (jak je v teorii pravděpodobnosti zvykem) předpokládat, že všechny uvažované veličiny a množiny jsou měřitelné.

*

Každá náhodná veličina $X : \Omega \to A$ indukuje disjunktní **rozklad** stavového prostoru Ω na množiny $X^{-1}(\{a\})$, $a \in A$, na měřitelné podmnožiny. Naopak, každý nejvýše spočetný rozklad $\mathcal{R}$ prostoru Ω na měřitelné podmnožiny definuje diskrétní

náhodnou veličinu R: stačí zvolit pro každou množinu v rozkladu nějaké jméno a stavu ω přiřadit jméno množiny, ve které leží. Tímto jménem může být i množina sama, v takovém případě mluvíme o **přirozené projekci**, kterou můžeme zapsat jako

$$\pi : \omega \mapsto [\omega]_R.$$

Z definic plyne,že $[\omega]_R = R^{-1}(R(\omega))$ a platí $\mathbb{P}\left([\omega]_R\right) = P_R(R(\omega))$. Díky předpokladu, že rozklad byl nejvýše spočetný, je zobrazení R měřitelné. Pro danou diskrétní náhodnou veličinu X nás často bude zajímat pouze rozklad, který indukuje, a obor hodnot pak můžeme podle libosti přejmenovávat.

> **Poznámka:** Disjunktní rozklad definičního oboru přirozeně indukuje každé zobrazení (nejen náhodná veličina). Ekvivalence $\omega \sim \omega'$ definovaná vztahem $X(\omega) = X(\omega')$ se v algebře nazývá *jádro zobrazení*. V případě homomorfismu se jedná o kongruenci a faktorizací podle této kongruence dostáváme strukturu $\Omega/_\sim$, která je podle věty o isomorfismu isomorfní oboru hodnot.
>
> Poznamenejme také, že v lineární algebře nebo v teorii grup se jako „jádro zobrazení“ obvykle označuje jedna význačná ekvivalenční třída, totiž ta odpovídající neutrálnímu prvku. Tato konvence je přirozená v tom smyslu, že tato význačná třída celou ekvivalenci, tedy *jádro zobrazení* v obecnějším smyslu, přímočaře určuje.

Jak už bylo řečeno, vektor diskrétních náhodných veličin $X_{[0..n)}$, kde $X_i : \Omega \to A_i$ je sám novou diskrétní náhodnou veličinou definovanou na kartézském součinu množin A_i. Její rozdělení

$$P_{X_{[0..n)}}(a_1, a_2, ..., a_n) = \mathbb{P}\left(\{\omega \mid X_i(\omega) = a_i, i = 0, 1, \dots, n-1\}\right)$$

se nazývá **sdružené rozdělení** a jednoznačně určuje rozdělení jednotlivých složek, kterým se říká **marginální rozdělení**. Z aditivity míry totiž dostáváme

$$P_{X_k}(b) = \mathbb{P}\left(X_k = b\right) = \sum_{a_i \in s(A_i),\, i \neq k} \mathbb{P}\left(X_k = b_k, X_i = a_i, i \neq k\right),$$

protože množina všech stavů ω, pro které $X_k = b$, se rozpadá na disjunktní podmnožiny určené hodnotami $X_i(\omega)$ pro $i \neq k$. Takovým součtům se někdy říká *věta o celkové pravděpodobnosti* a v jejich pozadí je fakt, že rozklad Ω definovaný náhodným vektorem je zjemněním rozkladů definovaných jeho složkami.

Budeme-li pracovat s vektory délky dva, budeme obvykle psát (X, Y), namísto (X_0, X_1). Všimněme si také, že pokud označíme $X = X_{[0..n-1)}$ a $Y = X_{n-1}$, můžeme náhodný vektor $X_{[0..n)}$ chápat jako dvojici (X, Y) a podobně pro jiná seskupení náhodného vektoru.

Náhodné veličiny $X_0, X_2, ..., X_{n-1}$ jsou **nezávislé**, právě když

$$P_{X_{[0..n)}}(a_1, a_2, ..., a_n) = P_{X_0}(a_0) P_{X_1}(a_1) \cdots P_{X_{n-1}}(a_{n-1})$$

pro všechna $a_0, a_1, \dots, a_{n-1}$.

Kontrolní otázky

1. Co je definičním oborem a oborem hodnot zobrazení $\mathbb{P}$, kterému říkáme pravděpodobnost?
2. V jakém vztahu jsou prvky množiny $\mathcal{F}$ k množině Ω, neboli o jakých (matematických) objektech má smysl říct, že mají pravděpodobnost?
3. Ujasněte si, že z definice σ-algebry skutečně vyplývá, že je $\mathcal{F}$ uzavřená na spočetné průniky, konečná sjednocení a průniky a že obsahuje prázdnou množinu.
4. Musí být množina, která má nulovou pravděpodobnost, prázdná?
5. Jak je na reálných náhodných veličinách a na rozděleních definováno sčítání a násobení reálným číslem?
6. Ověřte, že náhodné veličiny jsou uzavřeny na vážené průměry.
7. Ověřte, že rozdělení jsou uzavřena na vážené průměry.
8. Má množina, která neleží v $\mathcal{F}$, nulovou pravděpodobnost?
9. Zjistěte a formálně přesně dokažte, čemu se rovná π^{-1}, kde π je přirozená projekce.
10. Ověřte $[\omega]_{\mathcal{R}} = R^{-1}(R(\omega))$ a $\mathbb{P}\left([\omega]_R\right) = P_R(R(\omega))$.

3 Informace a entropie náhodných veličin

Nadále se budeme zabývat pouze konečnými náhodnými veličinami. Míra informace je definována **informačním obsahem** náhodného jevu:

Definice 3.1. ***Informační obsah*** *náhodného jevu $U \in \mathcal{F}$, $\mathbb{P}(U) > 0$, je*

$$\mathfrak{I}(U) = -\log \mathbb{P}(U).$$

Logaritmus v definici je dvojkový, což je v teorii informace nejobvyklejší volba, ačkoli by bylo možné uvažovat i jiné základy. Entropie je bezrozměrná veličina, ale (v případě dvojkového logaritmu) se obvykle používá jako jednotka **bit**, což je zkratka pro „binary digit“, kterou začal systematicky používat Claude Shannon. Jev, který má pravděpodobnost $1/2$ má tedy informační obsah jeden bit. To odpovídá informačnímu obsahu jedné binární cifry, ovšem *pouze tehdy*, jsou-li obě cifry *stejně pravděpodobné*. Z tohoto hlediska je také jasné, proč se v definici objevuje (dvojkový) logaritmu Daná posloupnost n cifer má *při rovnoměrném rozdělení* pravděpodobnost 2^{-n}, a informační obsah jevu, který taková posloupnost představuje, je tedy v souladu s očekáváním n bitů. Informační obsah charakterizuje, kolik binárních cifer je třeba ke specifikaci daného jevu. Pokud jev např. pokrývá jednu osminu stavového prostoru, je osm rovnocenných kandidátů na hodnotu jevu a je potřeba tří binárních cifer pro sdělení informace, že daný jev nastal. To je ovšem jen neformální intuice. Není např. jasné, jak interpretovat situaci, kdy dvojkový logaritmus není celé číslo, např. v případě jedné třetiny. Většina výsledků v teorii informace je nicméně potvrzením, že výchozí intuice funguje velmi dobře.

Informační obsah může nabývat libovolných nezáporných hodnot (jistý jev má nulový informační obsah) a není definován pro množiny nulové míry. V literatuře je v takovém případě často dodefinován jako $+\infty$. My tento přístup následovat nebudeme.

Klíčovým pojmem teorie informace je **entropie náhodné veličiny**, což je její *průměrný informační obsah*. Tuto průměrnou hodnotu můžeme nahlížet dvěma ekvivalentními způsoby. První, názornější, definuje entropii pomocí jejího rozdělení:

Definice 3.2. ***Entropie rozdělení*** $P : A \to [0,1]$ *je*

$$\mathcal{H}(P) := \sum_{a \in s(P)} P(a) \cdot (-\log P(a)).$$

Entropie ***náhodné veličiny*** *je entropie jejího rozdělení*

$$\mathcal{H}(X) := \mathcal{H}\left(P_X\right) = \sum_{a \in s\left(P_X\right)} P_X(a) \cdot \left(-\log P_X(a)\right).$$

Pro konečnou náhodnou veličinu je entropie kladné reálné číslo. Pokud je $s(P_X)$ spočetné, je suma stále definovaná, ale entropie může být nekonečná. V literatuře, kde

se dodefinovává informační obsah nemožného jevu, se běžně setkáte se zjednodušeným zápisem

$$\mathcal{H}(X) = \sum_{a \in A} P_X(a) \cdot \left(-\log P_X(a)\right)$$

doplněným konvencí, že $0 \cdot (-\log \infty)$ je rovno nule. Hodnota na množině míry nula se tedy dodefinuje, aby se poté ignorovala. To je přístup, kterému se, jak jsme řekli, vyhneme.

Entropie je podle definice vážený průměr funkce $-\log \circ P_X$ na množině A. Tato funkce zjevně úzce souvisí s informačním obsahem, konkrétně s informačním obsahem jevů „$X = a$“. Chceme-li přenést definici entropie do základního stavového prostoru Ω, musíme definovat informační obsah daného stavu ω, jako informační obsah ekvivalenční třídy ω v rozkladu definovaném veličinou X:

Definice 3.3. ***Informační obsah náhodné veličiny*** $X : \Omega \to A$ *je náhodná veličina* $\mathfrak{I}_X : \Omega \to [0, \infty)$ *daná předpisem*

$$\mathfrak{I}_X(\omega) = -\log P_X(X(\omega)), \quad \omega \in s(X).$$

Informační obsah náhodné veličiny X je tedy definován skoro jistě a je složením zobrazení

$$\mathfrak{I}_X : \Omega \xrightarrow{X} A \xrightarrow{P_X} [0,1] \xrightarrow{-\log} [0, \infty).$$

Přímo z definic nyní za použití vztahu pro střední hodnotu diskrétní náhodné veličiny plyne

$$\mathcal{H}(X) = \mathbb{E}\left(\mathfrak{I}_X\right),$$

což ospravedlňuje naše tvrzení, že entropie je průměrný informační obsah veličiny X.

Je důležité si všimnout, že entropie diskrétní náhodné veličiny je pouze vlastností rozdělení, a to navíc bez ohledu na přejmenování prvků množiny A. Je to tedy vlastnost souboru (multimnožiny) pravděpodobností, jejichž součet je jedna.

> **Poznámka:** Definici 3.2 lze opět číst také jako střední hodnotu funkce $-\log \circ P_X$ chápané jako náhodná veličina na pravděpodobnostním prostoru $(A, 2^A, P_X)$. Hodnotu $-\log P_X(a)$ lze zároveň chápat jako informační obsah jednoprvkové množiny $\{a\}$ podle Definice 3.1, ve které místo prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ uvažujeme právě $(A, 2^A, P_X)$. Pokud bychom tedy označili $-\log \circ P_X$ např. jako $\mathfrak{I}_{P_X}$, dostáváme $\mathcal{H}_{P_X}(X) = \mathbb{E}\left(\mathfrak{I}_{P_X}\right)$, kde střední hodnotu chápeme opět vzhledem k $(A, 2^A, P_X)$. To je druhý možný význam tvrzení, že entropie X je její průměrný informační obsah, který odpovídá naivnímu přístupu k hodu kostkou popsanému v Kapitole 2. Připomeňme, že tento přístup je nevhodný právě proto, že pro každý náhodný jev zavádí jeho vlastní pravděpodobnostní prostor, což mimo jiné vyžaduje neustálé upřesňování vzhledem k jakému prostoru počítáme střední hodnotu.

Následující tvrzení dokazuje důležitý a intuitivní fakt, že manipulací s hodnotami náhodné veličiny nemůžeme získat víc informace než kolik je jí tam už obsaženo. Naopak tak můžeme informaci ztratit, a to právě tehdy, pokud „zapomeneme“ některé rozdíly mezi hodnotami.

Tvrzení 3.4. *Buď $X : \Omega \to A$ náhodná veličina a buď $f : A \to A'$ zobrazení. Potom $f \circ X$ je náhodná veličina s hodnotami v A' a platí $\mathfrak{I}_{f\circ X} \leq \mathfrak{I}_X$ s.j. a $\mathcal{H}(f(X)) \leq \mathcal{H}(X)$. Pokud je zobrazení f injektivní, pak platí rovnosti (ta první opět jen skoro jistě).*

Důkaz. Pro $\omega \in s(X)$, $\mathfrak{I}_{f\circ X}(\omega) \leq \mathfrak{I}_X(\omega)$ plyne z inkluze

$$X^{-1}(X(\omega)) \subseteq (f \circ X)^{-1}((f \circ X)(\omega)).$$

Pokud je zobrazení f injektivní, dostáváme rovnost množin a příslušné rovnosti. □

Typickým příkladem je projekce, tedy zobrazení „zapomínající“ jednu (nebo více) složek.

Tvrzení 3.5. *Pro dvě náhodné veličiny X, Y platí $\mathfrak{I}_X \leq \mathfrak{I}_{X,Y}$ s.j. a $\mathcal{H}(X) \leq \mathcal{H}(X, Y)$.*

Důkaz. Stačí aplikovat předchozí Tvrzení pro $f : A \times B \to A$ definované jako projekce na první souřadnici, tj. $f(a, b) = a$. □

Tvrzení 3.6. *$X_0, X_1, .., X_{n-1}$ jsou nezávislé náhodné veličiny, pak*

$$\mathfrak{I}_{X_{[0..n)}} = \sum_{i=0}^{n-1} \mathfrak{I}_{X_i} \text{ s.j.}, \qquad \mathcal{H}\left(X_{[0..n)}\right) = \sum_{i=0}^{n-1} \mathcal{H}\left(X_i\right).$$

Důkaz. Pokud jsou veličiny nezávislé, pak pro $\omega \in s\left(X_0, X_2, \ldots, X_{n-1}\right)$:

$$\begin{aligned}\mathfrak{I}_{X_{[0..n)}}(\omega) &= -\log P_{X_{[0..n)}}\left(X_{[0..n)}(\omega)\right) = -\log \prod_{i=0}^{n-1}\left(P_{X_i}\left(X_i(\omega)\right)\right) \\ &= \sum_{i=0}^{n-1} -\log\left(P_{X_i}(X_i(\omega))\right) = \sum_{i=0}^{n-1} \mathfrak{I}_{X_i}(\omega).\end{aligned}$$

□

3.1 Podmíněná informace a podmíněná entropie

Důležitým konceptem v teorii pravděpodobnosti je podmiňování. Základem je podmiňování určitým jevem, a z toho odvozené podmiňování diskrétní náhodnou veličinou. Pro náhodný jev $U \subseteq \Omega$, $\mathbb{P}(U) > 0$, definujeme **pravděpodobnost podmíněnou jevem U** následujícím předpisem:

$$\mathbb{P}(V \mid U) := \frac{\mathbb{P}(V \cap U)}{\mathbb{P}(U)}, \qquad V \in \mathcal{F}.$$

Na takto definované zobrazení je možné se neformálně dívat jako na zúžení Ω na nový stavový prostor, totiž na jev U. Všimněme si ale, že jsme ve skutečnosti na původním měřitelném prostoru $(\Omega, \mathcal{F})$ definovali novou pravděpodobnostní míru, označme ji na okamžik $\mathbb{P}_U$, která prvku $V \in \mathcal{F}$ přiřazuje $\mathbb{P}_U(V) := \mathbb{P}(U \mid V)$. Zápis $\mathbb{P}(U \mid V)$ je notační konvencí, která může být matoucí, protože se nejedná o hodnotu původní pravděpodobnostní míry na nějakém jevu, ale o hodnotu nové pravděpodobnostní míry.

Pro diskrétní náhodnou veličinu Y s hodnotami v B můžeme podle této nové, podmíněné míry definovat **podmíněné rozdělení** $P_{Y|U}$ následovně

$$P_{Y|U}(b) := \mathbb{P}\left(Y^{-1}(b) \mid U\right), \qquad b \in B.$$

Nejčastěji budeme podmiňovat jevem popsaným jinou náhodnou diskrétní náhodnou veličinou X. Máme-li tedy diskrétní náhodnou veličinu X s hodnotami v A, budeme symbolem $P_{Y|X}$ značit rodinu rozdělení $\left\{P_{Y|X=a}\right\}_{a \in s\left(P_X\right)}$, kde „$X = a$“ chápeme opět jako zápis jevu z $\mathcal{F}$. Pro dané $a \in s\left(P_X\right)$, je pak $P_{Y|X=a}$ korektně definované rozdělení na B, které je podmíněné jevem $X^{-1}(a)$. Pro jednoduchost pak budeme pro $P_{Y|X=a}(b)$ používat také zápis $P_{Y|X}(a, b)$. Všimněte si pořadí argumentů, které vyjadřuje, že nejprve vybíráme podmiňující jev; v literatuře se často používá sugestivní zápis argumentů $P_{X|Y}(b \mid a)$. Není těžké nahlédnout, že

$$P_{Y|X}(a, b) = \frac{P_{X,Y}(a, b)}{P_X(a)}, \qquad a \in s\left(P_X\right), b \in B.$$

Podmíněná pravděpodobnost je tedy v tomto klasickém přístupu vlastností rozdělení a je definována na $A \times B$. Je ale také možné ji zapsat jako náhodnou veličinu definovanou na Ω:

$$\mathbb{P}(Y = Y(\omega) \mid X = X(\omega)), \qquad \omega \in s(X).$$

Tento pohled budeme často používat a používají ho i následující definice. Kromě už dříve zmíněného varování, že se nejedná o hodnotu původní pravděpodobnostní míry na nějakém jevu, ale o hodnotu nové pravděpodobnostní míry $\mathbb{P}_{X=X(\omega)}(Y = Y(\omega))$, je nyní důležité si uvědomit, že použitá nové pravděpodobnostní míra není fixní, ale mění se v závislosti na ω. To vše za použití konvence, že např. $X = X(\omega)$ označuje jev $X^{-1}(X(\omega))$.

Definice 3.7. ***Informační obsah*** *náhodné veličiny Y za* ***podmínky*** *X je*

$$\begin{aligned}\mathfrak{I}_{Y|X}(\omega) &= -\log \mathbb{P}(Y = Y(\omega) \mid X = X(\omega)) \\ &= -\log P_{Y|X}(X(\omega), Y(\omega)), \qquad \omega \in s(X, Y).\end{aligned}$$

Náhodná veličina $\mathfrak{I}_{Y|X}$ vyjadřuje pro daný stav míru dodatečné informace, kterou přináší znalost Y při znalosti X. Je definována skoro jistě, konkrétně na nosiči vektoru (X, Y). Její střední hodnotu definujeme jako nový pojem.

Definice 3.8. ***Entropie*** *náhodné veličiny Y* ***za podmínky*** *X je*

$$\mathcal{H}(Y \mid X) = \mathbb{E}\left(\mathfrak{I}_{Y|X}\right).$$

Uvědomme si nejprve, že $\mathcal{H}(Y \mid X)$, podobně jako výše podmíněná pravděpodobnost, je nové značení. Není to zejména entropie $Y \mid X$, už proto že žádnou takovou náhodnou veličinu jsme nedefinovali a $P_{Y|X}$ není samo o sobě rozdělení, je to rodina podmíněných rozdělení. Tato rozdělení už mají bezprostřední vazbu na zmíněnou podmíněnou entropii.

Definice 3.9. *Pro náhodný jev $U \subseteq \Omega$ nenulové pravděpodobnosti nazveme* ***entropií*** *náhodné veličiny Y* ***za podmínky*** *U entropii podmíněného rozdělení $\mathcal{H}\left(P_{Y|U}\right)$. Tuto entropii budeme také značit $\mathcal{H}(Y \mid U)$.*

V případě $\mathcal{H}(Y \mid U)$ se sice jedná o notační konvenci, je to ale zápis standardně definované entropie rozdělení $P_{Y|U}$.

> **Poznámka:** Je užitečné si všimnout, že entropii za podmínky nelze rozumně interpretovat pomocí náhodné veličiny na původním prostoru. To souvisí s poznámkou výše, týkající se míry $\mathbb{P}_U$, která není přímočaře interpretovatelná jako základní míra $\mathbb{P}$ nějaké množiny.
>
> Můžeme být v pokušení se domnívat, že entropie za podmínky bude střední hodnotou náhodné veličiny
>
> $$\mathfrak{I}_{Y|U}(\omega) = -\log \mathbb{P}\left(Y = Y(\omega) \mid U\right) = -\log \mathbb{P}_U\left(Y = Y(\omega)\right) = -\log P_{Y|U}\left(Y(\omega)\right).$$
>
> Takový pokus má dvě úskalí. Logaritmus je jednak definován pouze na ω, pro které je $P_{Y|U}$ nenulové, což nemusí být množina míry jedna. Za druhé, i pokud to bude množina míry jedna, resp. pokud dodefinujeme funkci v nedefinovaných bodech nulou, nebude obecně platit $\mathbb{E}\left(\mathfrak{I}_{Y|U}\right) = \mathcal{H}(Y \mid U)$. Ukažme to na příkladu hodů kostkou, náhodné veličiny Y s hodnotami sudá/lichá a jevu "padlo malé číslo"(tedy jedna, dva nebo tři). Pak je
>
> $$\begin{aligned}\mathcal{H}(Y \mid U) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) = \\ = \mathbb{P}_U\left(Y = \text{sudá}\right) \cdot -\log\left(\mathbb{P}_U\left(Y = \text{sudá}\right)\right) + \\ \mathbb{P}_U\left(Y = \text{lichá}\right) \cdot -\log\left(\mathbb{P}_U\left(Y = \text{lichá}\right)\right),\end{aligned}$$
>
> což bychom mohli zapsat jako $\mathbb{E}_{\mathbb{P}_U}\left(\mathfrak{I}_{Y|U}\right)$, tedy jako střední hodnotu vůči jiné míře. To není totéž co
>
> $$\begin{aligned}\mathbb{E}\left(\mathfrak{I}_{Y|U}\right) = -\frac{1}{2}\log\left(\frac{1}{3}\right) - \frac{1}{2}\log\left(\frac{2}{3}\right) = \\ = \mathbb{P}\left(Y = \text{sudá}\right) \cdot -\log\left(\mathbb{P}_U\left(Y = \text{sudá}\right)\right) + \\ \mathbb{P}\left(Y = \text{lichá}\right) \cdot -\log\left(\mathbb{P}_U\left(Y = \text{lichá}\right)\right).\end{aligned}$$

Entropie $\mathcal{H}(Y \mid X = a)$, neboli $\mathcal{H}\left(P_{Y|X=a}\right)$, může být k entropii $\mathcal{H}(Y)$ v libovolném vztahu. Dozvíme-li se hodnotu náhodné proměnné X může se naše nejistota o Y jak zvýšit tak snížit.

Pro entropii veličiny za podmínky platí následující explicitní formule.

Tvrzení 3.10.

$$\mathcal{H}(Y \mid X) = \sum_{(a,b)\in s\left(P_{X,Y}\right)} P_{X,Y}(a,b) \cdot \log \frac{P_X(a)}{P_{X,Y}(a,b)} = \\ = \sum_{a\in s\left(P_X\right)} P_X(a) \cdot \mathcal{H}(Y \mid X = a) \ .$$

Důkaz. Veličina $\mathfrak{I}_{Y|X}(\omega)$ nabývá na definičním oboru $s(X,Y)$ hodnot

$$-\log P_{Y|X}(a,b) = \log \frac{P_Y(b)}{P_{X,Y}(a,b)} \qquad\qquad (a,b) \in s\left(P_{X,y}\right).$$

Konkrétně, pro dané (a,b) nabývá výše zmíněnou hodnotu na celé množině dané podmínkami „$X = a, Y = b$“. První rovnost je tedy přímočarým výpočtem střední hodnoty diskrétní náhodné veličiny podle definice.

Pro $a \in s\left(P_X\right)$ je rozdělení $P_{Y|X=a}$ definováno na B jako $\frac{P_{X,Y}(a,b)}{P_X(a)}$ a podle definic tedy

$$\mathcal{H}(Y \mid X = a) = \sum_{b\in s\left(P_{Y|X=a}\right)} P_{Y|X=a}(b) \cdot \left(-\log P_{Y|X=a}(b)\right) = \\ = \sum_{b\in s\left(P_{Y|X=a}\right)} \frac{P_{X,Y}(a,b)}{P_X(a)} \cdot \log \frac{P_X(a)}{P_{X,Y}(a,b)} \, .$$

Proto

$$\sum_{a\in s\left(P_X\right)} P_X(a) \cdot \mathcal{H}(Y \mid X = a) = \sum_{a\in s\left(P_X\right)} P_X(a) \cdot \sum_{b\in s\left(P_{Y|X=a}\right)} \frac{P_{X,Y}(a,b)}{P_X(a)} \cdot \log \frac{P_X(a)}{P_{X,Y}(a,b)} \\ = \sum_{a\in s\left(P_X\right)} \sum_{b\in s\left(P_{Y|X=a}\right)} P_{X,Y}(a,b) \cdot \log \frac{P_X(a)}{P_{X,Y}(a,b)} \, .$$

Pro druhou rovnost tedy zbývá ověřit, že sčítáme přes stejnou množinu

$$s\left(P_{X,Y}\right) = \left\{(a,b) \mid a \in s\left(P_X\right), b \in s\left(P_{Y|X=a}\right)\right\} \ .$$

□

Vidíme, že entropie Y za podmínky X je váženým průměrem z entropií podmíněných rozdělení. Je to tedy průměrné množství informace nutné k určení Y, pokud již známe X.

Nyní definujeme poslední informační veličinu tohoto oddílu.

Definice 3.11. ***Vzájemný informační obsah*** *náhodných veličin X a Y definujeme předpisem:*

$$\mathfrak{I}_{X:Y}(\omega) = \log \frac{\mathbb{P}(X = X(\omega), Y = Y(\omega))}{\mathbb{P}(X = X(\omega)) \cdot \mathbb{P}(Y = Y(\omega))} \, .$$

Funkce $\mathfrak{I}_{X:Y}$ je definovaná na $s(X,Y)$. Proto je definovaná skoro jistě. Neformální ospravedlnění této definice plyne z toho, že vzájemný informační obsah je nulový právě tehdy, kdy je pravděpodobnost jevu $(X,Y) = (a,b)$ rovna součinu pravděpodobností $X = a$ a $Y = b$, tedy v situaci, kdy jsou oba jevy nezávislé. Závislé jevy mohou mít vzájemnou informaci kladnou i zápornou, čímž se tato veličina liší od jiných informačních veličin definovaných v tomto textu. Ukážeme nicméně, že má nezápornou střední hodnotu, která je navíc definovaná i v případě spočetného $s\left(P_{X,Y}\right)$. Tato střední hodnota bývá nazývána vzájemnou informací. My se budeme držet pojmu „entropie“, abychom zdůraznili, že se jedná pouze o průměrnou hodnotu, ale alternativní název promítneme do notace.

Definice 3.12. ***Vzájemnou entropii*** *náhodných veličin X a Y definujeme předpisem:*

$$\mathcal{I}(X:Y) = \mathbb{E}\left(\mathfrak{I}_{X:Y}\right) = \sum_{(a,b)\in s\left(P_{X,Y}\right)} P_{X,Y}(a,b)\cdot\log\frac{P_{X,Y}(a,b)}{P_X(a)P_Y(b)}.$$

Příklad 3.13. Uvažujme pravděpodobnostní rozdělení

$$P_{X,Y}:\ \begin{array}{c|cc} X\backslash Y & b_0 & b_1 \\ \hline a_0 & 1/2 & 0 \\ a_1 & 1/4 & 1/4 \end{array},\quad P_{Y|X}:\ \begin{array}{c|cc} X\backslash Y & b_0 & b_1 \\ \hline a_0 & 1 & 0 \\ a_1 & 1/2 & 1/2 \end{array},\quad P_{X|Y}:\ \begin{array}{c|cc} X\backslash Y & b_0 & b_1 \\ \hline a_0 & 2/3 & 0 \\ a_1 & 1/3 & 1 \end{array}.$$

Pak marginální rozdělení jsou $P_X = \left(\frac{1}{2},\frac{1}{2}\right)$, a $P_Y = \left(\frac{3}{4},\frac{1}{4}\right)$. Dále $\mathcal{H}(X) = 1$, $\mathcal{H}(Y) \doteq 0.811$, $\mathcal{H}(X,Y) = \frac{3}{2}$, $\mathcal{I}(X:Y) = 0.311$. Pro podmíněné entropie dostáváme

$$0 = \mathcal{H}\left(Y \mid X = a_0\right) < \mathcal{H}(Y) < \mathcal{H}\left(Y \mid X = a_1\right) = 1.$$

Podmíněná entropie $\mathcal{H}(Y \mid X) = \frac{1}{2}$ je již menší než $\mathcal{H}(Y)$. Podobně $\mathcal{H}(X \mid Y) = 0.689 < \mathcal{H}(X)$.

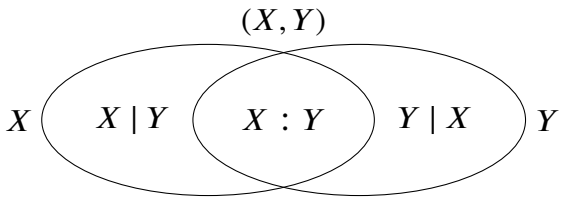
Přímo z definic a z linearity střední hodnoty dostáváme:

Tvrzení 3.14. *Nechť X, Y jsou náhodné veličiny. Pak na $s(X,Y)$ platí*

- $\mathfrak{I}_{X,Y} = \mathfrak{I}_{Y|X} + \mathfrak{I}_X$,
- $\mathfrak{I}_X = \mathfrak{I}_{X:Y} + \mathfrak{I}_{X|Y}$.

Dále platí

- $\mathcal{H}(X,Y) = \mathcal{H}(Y \mid X) + \mathcal{H}(X)$,
- $\mathcal{H}(X) = \mathcal{I}(X:Y) + \mathcal{H}(X \mid Y)$.



Obrázek 3: Informace a entropie vzájemná a za podmínky

Tyto vztahy dohromady tvoří užitečný diagram na Obrázku 3. Z něj lze názorně získávat další rovnosti, např. následující alternativní definici vzájemného informačního obsahu a vzájemné entropie:

$$\begin{aligned}\mathfrak{I}_{X:Y} &= \mathfrak{I}_X + \mathfrak{I}_Y - \mathfrak{I}_{X,Y},\\ \mathcal{I}(X:Y) &= \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X,Y)\ .\end{aligned}$$

Vztah lze přímočaře rozšířit na více veličin.

Tvrzení 3.15 (Řetězové pravidlo)**.** *Pro posloupnost náhodných veličin $X_0, \dots X_{n-1}$ platí rovnosti*

$$\begin{aligned}\mathfrak{I}_{X_{[0..n)}} &= \sum_{i=0}^{n-1} \mathfrak{I}_{X_i|X_{[0..i)}}\\ &= \mathfrak{I}_{X_0} + \mathfrak{I}_{X_1|X_0} + \mathfrak{I}_{X_2|X_0,X_1} + \cdots + \mathfrak{I}_{X_{n-1}|X_{[0,n-1)}}\end{aligned}$$

a

$$\begin{aligned}\mathcal{H}\left(X_{[0..n)}\right) &= \sum_{i=0}^{n-1} \mathcal{H}\left(X_i \mid X_{[0..i)}\right)\\ &= \mathcal{H}\left(X_0\right) + \mathcal{H}\left(X_1 \mid X_0\right) + \mathcal{H}\left(X_2 \mid X_0, X_1\right) + \cdots + \mathcal{H}\left(X_{n-1} \mid X_{[0,n-1)}\right),\end{aligned}$$

první z nich na s $\left(X_{[0..n)}\right)$.

Důkaz. Jedná se o opakovanou aplikaci Tvrzení 3.14. Neboli, postupujeme indukcí dle počtu veličin. Pro $n = 1$ platí triviálně. Pokud vztah platí pro jakoukoliv posloupnost délky n, pak pro posloupnost $X_0, \dots, X_n$ délky $n+1$ dostáváme

$$\begin{aligned}\mathfrak{I}_{X_{[0..n+1)}} &= \mathfrak{I}_{X_{[0..n)}} + \mathfrak{I}_{X_n|X_{[0..n)}}\\ &= \sum_{i=0}^{n-1} \mathfrak{I}_{X_i|X_{[0..i)}} + \mathfrak{I}_{X_n|X_{[0..n)}} = \sum_{i=0}^{n} \mathfrak{I}_{X_i|X_{[0..i)}}.\end{aligned}$$

Druhou rovnost odvodíme analogicky, nebo přímo z definice pomocí linearity střední hodnoty. □

Kontrolní otázky

1. Proč pro diskrétní náhodnou veličinu platí $\mathbb{P}(s(X)) = 1$? Jakou roli zde hraje předpoklad, že X je diskrétní?
2. Ověřte $\mathcal{H}(X) = \mathbb{E}(\mathfrak{I}(X))$.
3. Proč se v Tvrzení 3.4 používá „skoro jistě“?
4. Dokažte $s(X, Y) \subseteq s(X)$.
5. Ověřte $s\left(P_{X,Y}\right) = \left\{(a, b) \mid a \in s\left(P_X\right), b \in s\left(P_{Y|X=a}\right)\right\}$.
6. Vysvětlete notaci $\mathbb{P}_U(V)$, $\mathbb{P}(U \mid V)$, $\mathbb{P}(Y = a \mid U)$, $\mathbb{P}(Y = b \mid X = a)$, $\mathcal{H}(X)$, $\mathcal{H}\left(P_X\right)$, $\mathcal{H}(X, Y)$, $\mathcal{H}(Y \mid X = a)$ a $\mathcal{H}(Y \mid X)$. Kdy se jedná o dobře definované zobrazení aplikované na argument v závorce, a kdy o notační konvenci, které takový výklad neumožňuje? U dobře definovaných zobrazení určete definiční obor a obor hodnot. U notačních konvencí vysvětlete, o jaká zobrazení se vzhledem k použitým parametrům jedná.

3.2 Divergence entropie

Uvažujme nyní dvě pravděpodobnostní rozdělení P a Q na stejné konečné množině A.

Definice 3.16. ***Divergence entropie*** *rozdělení P vzhledem k rozdělení Q je definována vzorcem*

$$\mathcal{D}(P \parallel Q) = \sum_{a \in s(P)} P(a) \cdot \log \frac{P(a)}{Q(a)},$$

pokud $s(P) \subseteq s(Q)$. Jinak $\mathcal{D}(P \parallel Q) = +\infty$.

Význam této definice a důvod, proč se nazývá „divergencí“, lze ilustrovat následující úvahou. Všimněme si, že platí

$$\begin{aligned}\mathcal{D}(P \parallel Q) &= -\sum_{a \in s(P)} P(a) \cdot \log Q(a) + \sum_{a \in s(P)} P(a) \cdot \log P(a) \\ &= -\sum_{a \in s(P)} P(a) \cdot \log Q(a) - \mathcal{H}(P).\end{aligned}$$

Jedná se tedy o rozdíl mezi entropií P a jakousi „pomýlenou entropií“ při které přisuzujeme náhodným jevům informační obsah daný rozdělením Q. To odpovídá situaci, kdy kódování vhodné pro rozdělení Q aplikujeme na rozdělení P.

Vlastnosti divergence odvodíme z konvexity logaritmu. Připomeňme příslušné definice.

Definice 3.17. *Reálná funkce $f : I \to \mathbb{R}$ je **konvexní** na intervalu I, pokud její graf leží pod každou její sečnou, tj, pokud pro každé $x, y \in I$ a každé $t \in (0,1)$ platí*

$$f(tx + (1-t)y) \leq t \cdot f(x) + (1-t) \cdot f(y)\,.$$

*Funkce f je **striktně konvexní**, pokud pro každé $x \neq y \in I$ platí ostrá nerovnost*

$$f(tx + (1-t)y) < t \cdot f(x) + (1-t) \cdot f(y)\,.$$

*Funkce f je **(striktně) konkávní**, je-li funkce $-f$ (striktně) konvexní.*

Pro konvexní funkce platí tvrzení o Jensenově nerovnost, která citujeme bez důkazu.

Tvrzení 3.18 (Jensenova nerovnost I)**.** *Nechť $f : I \to \mathbb{R}$ je konvexní funkce, $x_1, \ldots, x_n \in I$ a nechť $t_1, \ldots, t_n$ jsou nezáporná čísla, jejichž součet je* 1*. Pak*

$$f\left(\sum_{i \leq n} t_i x_i\right) \leq \sum_{i=1}^{n} t_i \cdot f(x_i)$$

Pokud je f striktně konvexní a platí rovnost, potom je množina $\{x_i \mid t_i > 0\}$ jednoprvková.

Tvrzení 3.19 (Jensenova nerovnost II)**.** *Buď ξ reálná diskrétní veličina, s $\left(P_\xi\right) \subseteq I$, $f : I \to \mathbb{R}$ buď konvexní. Potom*

$$f(\mathbb{E}(\xi)) \leq \mathbb{E}(f(\xi))\,.$$

Pokud je f striktně konvexní a platí rovnost, potom je ξ triviální, t.j. s $\left(P_\xi\right)$ je jednoprvková.

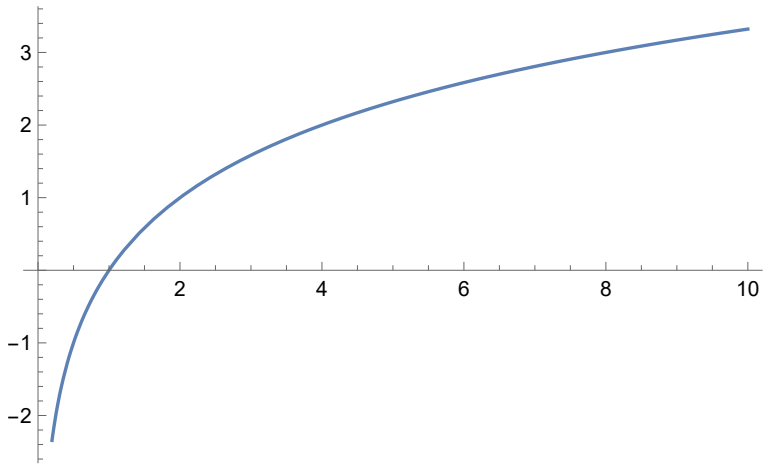
Je-li $f''(x) > 0$ na otevřeném intervalu I, pak f je striktně konvexní na I. Protože pro logaritmus platí $\log_a''(x) = -1/(x^2 \ln a)$ je logaritmus striktně konkávní na $(0, \infty)$ (viz obrázek 4).

Z konkávnosti logaritmu plyne klíčová vlastnost divergence, totiž její nezápornost. Ve smyslu předchozího neformálního popisu divergence to znamená, že „po -mýlená“ entropie je vždy větší než ta skutečná. To také předjímá budoucí poznatky o tom, že nejefektivněji se vždy kóduje kódy odpovídajícími skutečnému informačnímu obsahu.

Tvrzení 3.20. *$\mathcal{D}(P \parallel Q) \geq 0$ a rovnost nastává právě když $P = Q$.*

Důkaz. Stačí uvažovat případ $s(P) \subseteq s(Q)$. Jelikož je $-\log x$ striktně konvexní a striktně klesající, dostáváme

$$\begin{aligned}\mathcal{D}(P \parallel Q) &= \sum_{a \in s(P)} P(a)\left(-\log \frac{Q(a)}{P(a)}\right) \geq -\log\left(\sum_{a \in s(P)} P(a) \cdot \frac{Q(a)}{P(a)}\right) \\ &= -\log\left(\sum_{a \in s(P)} Q(a)\right) \geq -\log\left(\sum_{a \in s(Q)} Q(a)\right) = 0.\end{aligned}$$



Obrázek 4: Graf (konkávní) funkce $\log x$

Pokud bychom chtěli namísto nerovností rovnosti, musí být $s(P)$ rovno $s(Q)$ a podíl $\frac{Q(a)}{P(a)}$ musí být roven konstantě α pro všechna $a \in s(P)$. Z předchozího plyne,

$$\alpha = \sum_{a \in s(P)} P(a) \frac{Q(a)}{P(a)} = \sum_{a \in s(Q)} Q(a) = 1.$$

□

Z předchozího tvrzení plyne řada užitečných výsledků.

Příklad 3.21. Buďte $P = \left\{\frac{1}{5}, \frac{4}{5}\right\}$ a $Q = \left\{\frac{2}{3}, \frac{1}{3}\right\}$ binární rozdělení. Pak $Q/P = \left\{\frac{10}{3}, \frac{5}{12}\right\}$, střední hodnota Q/P vzhledem k rozdělení P je rovna jedné, zatímco střední hodnota funkce $-\log$ na Q/P vzhledem k rozdělení P je díky konvexnosti kladná (viz Obrázek 5).

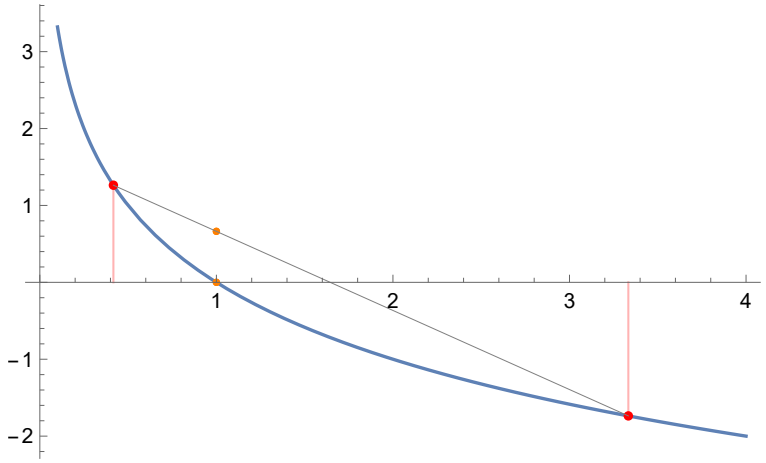
Tvrzení 3.22. *Nechť má veličina X hodnoty v n-prvkové množině A. Pak $\mathcal{H}(X) \leq \log(n)$. Rovnost nastane právě tehdy když je P_X rovnoměrně rozloženo na A.*

Důkaz. Na množině A uvažujme rovnoměrné rozložení U_n, dané předpisem $U_n(a) = \frac{1}{n}$ pro každé $a \in A$. Potom $s\left(P_X\right) \subseteq s\left(U_n\right)$ a

$$0 \leq \mathcal{D}\left(P_X \parallel U_n\right) = \sum_{a \in s\left(P_X\right)} P_X(a) \log\left(\frac{P_X(a)}{\frac{1}{n}}\right) = \log(n) - \mathcal{H}(X).$$

Kýžená rovnost nastane právě tehdy když P_X a U_n splývají. □

Následující věta dává divergenci do úzké souvislosti se vzájemnou informací. Jedná se tak o dalšího kandidáta na alternativní definici vzájemné entropie jako



Obrázek 5: Divergence rozdělení P a Q z příkladu 3.21

„divergence od nezávislosti“. V jeho znění vystupuje rozdělení $P_X \cdot P_Y$, což je rozdělení na $A \times B$, definované předpisem $(P_X \cdot P_Y)(a,b) = P_X(a)P_Y(b)$, jsou-li P_X a P_Y rozdělení definovaná na A a B.

Tvrzení 3.23. *$\mathcal{I}(X:Y) = \mathcal{D}\left(P_{X,Y} \parallel P_X \cdot P_Y\right)$. Tedy $\mathcal{I}(X:Y) \geq 0$ a rovnost nastává právě když X a Y jsou nezávislé.*

Důkaz. Lze lehko nahlédnout, že $s\left(P_{X,Y}\right) \subseteq s\left(P_X \cdot P_Y\right)$. Proto

$$\mathcal{I}(X:Y) = \sum_{(a,b)\in s\left(P_{X,Y}\right)} P_{X,Y}(a,b) \log \frac{P_{X,Y}(a,b)}{P_X(a)\cdot P_Y(b)} = \mathcal{D}(P_{X,Y} \parallel P_X \cdot P_Y).$$

Tedy $\mathcal{I}(X:Y) \geq 0$ a rovnost nastane právě tehdy když $P_{X,Y}$ je totožné s $P_X \cdot P_Y$, neboli když jsou veličiny X a Y nezávislé. □

Ze součtových vzorců nyní okamžitě plyne:

Tvrzení 3.24. *Nechť X, Y jsou náhodné veličiny. Pak*

(1) $\mathcal{H}(X,Y) \leq \mathcal{H}(X) + \mathcal{H}(Y)$

(2) $\mathcal{H}(X \mid Y) \leq \mathcal{H}(X)$.

Rovnosti nastanou právě tehdy, když jsou veličiny nezávislé.

Předchozí vlastnosti lze pak také odvodit jiným způsobem, který se ukáže přínosným v následující kapitole.

3.3 Entropie jako funkce na prostoru rozdělení

V této části kapitoly budeme uvažovat entropii jako funkci na prostoru rozdělení Δ_A. K tomuto účelu využijeme pojem konvexní kombinace rozdělení zavedený na str. 8. Uvažujme tedy nějaký konečný soubor rozdělení $\left(P_i\right)_{i\in I} \subseteq \Delta_A$ a soubor reálných nezáporných vah $\left(\alpha_i\right)_{i\in I}$, které se sečtou na jedničku. Konvexní kombinací daných rozdělení s danými vahami nyní rozumíme rozdělení $P \in \Delta_A$ dané předpisem

$$P(a) = \sum_{i\in I} \alpha_i P_i(a), \qquad a \in A.$$

Tuto kombinaci také zapisujeme jako $P = \sum_{i\in I} \alpha_i P_i$. Není těžké ověřit, že se opět jedná o rozdělení, čili, že $(P(a))_{a\in A}$ je soubor nezáporných reálných čísel, jejichž součet je jedna.

Rozšiřme nyní pojem konvexní a konkávní funkce i na prostor Δ_A takto:

Definice 3.25. *Reálná funkce $f : \Delta_A \to \mathbb{R}$ je **konvexní**, pokud pro každé $P, Q \in \Delta_A$ a každé $t \in (0,1)$ platí*

$$f(tP + (1-t)Q) \leq t \cdot f(P) + (1-t) \cdot f(Q)\,.$$

*Funkce f je **striktně konvexní**, pokud pro každé $P \neq Q \in I$ platí ostrá nerovnost*

$$f(tP + (1-t)Q) < t \cdot f(P) + (1-t) \cdot f(Q)\,.$$

*Funkce f je **(striktně) konkávní**, je-li funkce $-f$ (striktně) konvexní.*

Všimněte si, že $tP + (1-t)Q$ je konvexní kombinace rozdělení P a Q. Při této definici konvexity pak platí také následující varianta Jensenovy nerovnosti.

Tvrzení 3.26 (Jensenova nerovnost III)**.** *Nechť $P_1, \ldots, P_n \in \Delta_A$ a nechť $t_1, \ldots, t_n$ jsou nezáporná čísla, jejichž součet je 1, $f : \Delta_A \to \mathbb{R}$. Pokud je f konvexní, pak*

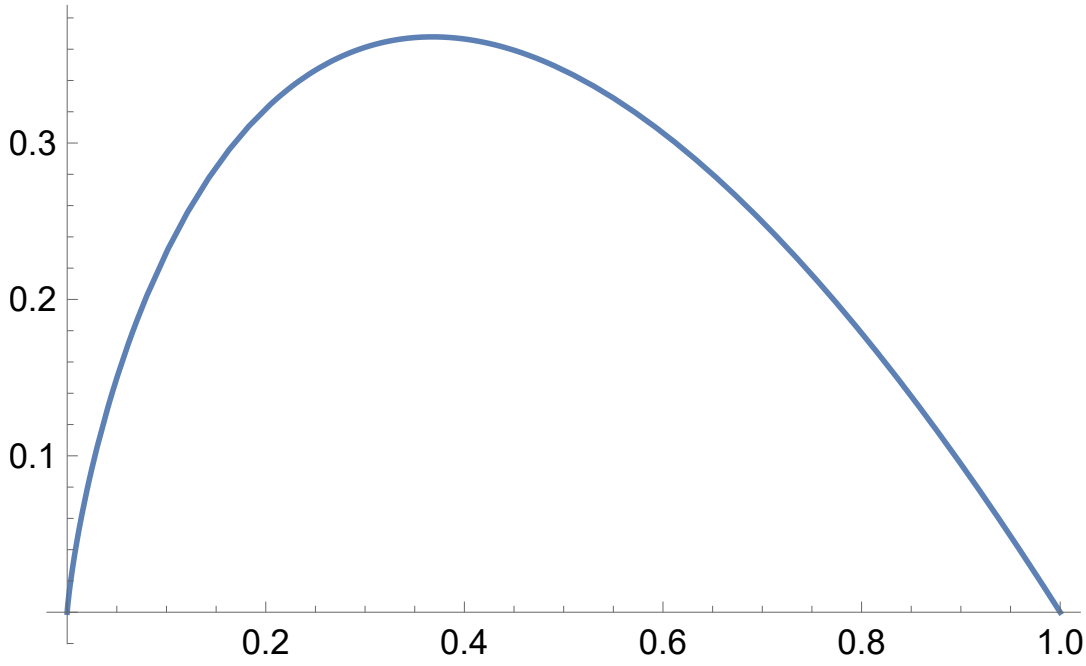
$$f\left(\sum_{i\leq n} t_i P_i\right) \leq \sum_{i=1}^{n} t_i \cdot f(P_i)$$

Pokud je f striktně konvexní a platí rovnost, potom je množina $\{P_i \mid t_i > 0\}$ jednoprvková.

Pokud je f konkávní, platí opačná nerovnost. V případě striktně konkávní funkce f platí rovnost právě tehdy když je množina $\{P_i \mid t_i > 0\}$ jednoprvková.

Uvažujeme nyní funkci $\phi : [0,\infty) \to \mathbb{R}$ definovanou předpisem

$$\phi(x) = \begin{cases} 0 & x = 0, \\ -x \log x, & x > 0. \end{cases}$$



Obrázek 6: Graf (konkávní) funkce $\phi(x) = -x \log x$
(s maximem v bodě $1/e$)

Tato funkce je spojitá, mimo jiné je tedy spojitá zprava v nule. Navíc má na intervalu $(0, \infty)$ zápornou druhou derivaci $\phi''(x) = -\log_2(e)/x$. Funkce je tedy striktně konkávní na $(0, \infty)$. Díky spojitosti v nule je pak striktně konkávní na celém svém definičním oboru $[0, \infty)$. Na intervalu $[0, 1]$, který nás zajímá, je graf funkce znázorněn na Obrázku 6.

Entropii rozdělení $P \in \Delta_A$ lze nyní zapsat jako

$$\mathcal{H}(P) = \sum_{a \in A} \phi(P(a)).$$

Díky vhodnému dodefinování v nule tak v tomto vztahu nepotřebujeme omezení na nosič rozdělení. To je užitečné v důkazu následujícího lemmatu.

Lemma 3.27. *Entropie je striktně konkávní na prostoru rozdělení* Δ_A.

Důkaz. Buď $P, Q \in \Delta_A$, $t \in (0, 1)$. Pak platí

$$\begin{aligned}
\mathcal{H}(tP + (1-t)Q) &= \sum_{a \in A} \phi(tP(a) + (1-t)Q(a)) \\
&\geq \sum_{a \in A} (t\phi(P(a)) + (1-t)\phi(Q(a))) \\
&= t \sum_{a \in A} \phi(P(a)) + (1-t) \sum_{a \in A} \phi(Q(a)) = t\mathcal{H}(P) + (1-t)\mathcal{H}(Q),
\end{aligned}$$

kde nerovnost vyplývá z konkavity funkce ϕ. Pokud $P \neq Q$, pak existuje $a_0 \in A$ takové, že se $P(a_0)$ liší od $Q(a_0)$. Ze striktní konkavity ϕ plyne, že se příslušné členy v sumách budou lišit, a proto vyjde celkově ostrá nerovnost. Funkce entropie je tedy striktně konkávní na prostoru rozdělení Δ_A. □

Tato vlastnost entropie může posloužit k alternativním důkazům některých tvrzení v této kapitole. Důležité je následující pozorování, podle kterého je (nepodmíněné) rozdělení konvexní kombinací rozdělení podmíněných:

$$P_Y = \sum_{a \in s(P_X)} P_X(a) \cdot P_{Y|X=a}\,,$$

kde příslušné váhy jsou pravděpodobnosti jednotlivých podmínek. Důsledkem striktní konkávnosti entropie pak je, že

$$\mathcal{H}(Y) = \mathcal{H}\left(P_Y\right) \geq \sum_{a \in s(P_X)} P_X(a) \cdot \mathcal{H}\left(P_{Y|X=a}\right) = \mathcal{H}(Y \mid X)\,,$$

kde rovnost nastává pouze v případech, kdy $P_{Y|X=a} = P_Y$ pro všechna $a \in s\left(P_X\right)$. Taková podmínka je ovšem ekvivalentní s nezávislostí (je třeba si trochu rozmyslet). Dokázali jsme tedy takto, že je podmíněná entropie vždy menší nebo rovna nepodmíněné a rovná se pouze v případě nezávislosti. Z tohoto faktu pak lze odvodit nezápornost vzájemné entropie a další podobná tvrzení, včetně charakterizace nezávislých veličin pomocí těchto pojmů.

Přímo z konkávnosti funkce ϕ také snadno plyne, že entropie je na množině A velikosti n maximální pro uniformní rozdělení U_n.

$$H(U_n) = \sum_{a \in A} \phi\left(\frac{1}{n}\right) = \sum_{a \in A} \phi\left(\sum_{a' \in A} \frac{1}{n} P(a')\right) \geq \sum_{a \in A} \sum_{a' \in A} \frac{1}{n} \phi(P(a')) = H(P).$$

Není-li P uniformní, je nerovnost díky striktní konkávnosti ostrá. Všimněte si, že zde nehraje roli, že ϕ nenabývá maxima v bodě $\frac{1}{2}$.

Jiná úvaha vedoucí ke stejnému výsledku využívá konkávnost entropie a nezávislost entropie na permutaci hodnot (tedy její symetrie). Pro rozdělení $P \in \Delta_A$ uvažujme cyklickou permutaci $\sigma : A \to A$. Dále definujme P_i, $i = 1 \ldots n$, jako rozdělení dané předpisem

$$P_i(a) = P(\sigma^i(a)), \qquad a \in A.$$

Ze symetrie entropie plyne $\mathcal{H}\left(P_i\right) = \mathcal{H}(P)$, $i = 1 \ldots n$. Vzhledem k cykličnosti permutace navíc pro každé $a \in A$ projdou obrazy iterovaných permutací $(\sigma^i)(a)$, $i = 1 \ldots n$, postupně celou množinu A (každý prvek jednou). Proto je aritmetickým průměrem permutovaných rozdělení rozdělení uniformní na A. Opravdu, pro $a \in A$,

$$\frac{1}{n} \sum_{i=1}^{n} P_i(a) = \frac{1}{n} \sum_{a \in A} P(a) = \frac{1}{n} = U_n(a).$$

Nakonec

$$\mathcal{H}(P) = \sum_{i=1}^{n} \frac{1}{n} \mathcal{H}\left(P_i\right) \leq \mathcal{H}\left(\sum_{i=1}^{n} \frac{1}{n} P_i\right) = \mathcal{H}\left(U_n\right).$$

Kontrolní otázky

1. Dokažte, že $P_{Y|X=a} = P_Y$ pro všechna $a \in s\left(P_X\right)$, právě když jsou veličiny X a Y nezávislé.

3.4 Vztah tří náhodných veličin. Podmíněná vzájemná entropie a podmíněná nezávislost.

Příklad 3.28. Nechť $X_1, X_2 : \Omega \to B$, kde B je nějaká dvouprvková množina, jsou nezávislé náhodné veličiny s rozdělením $\left(\frac{1}{2}, \frac{1}{2}\right)$ a $X_3 = (X_1 + X_2) \mod 2$.

Pak X_i, X_j jsou nezávislé, kdykoliv $i \neq j$, ale trojice X_1, X_2, X_3 nezávislá není. Pro entropii platí (i,j různá)

$$\mathcal{H}\left(X_i\right) = 1, \qquad \mathcal{H}\left(X_i, X_j\right) = 2, \qquad \mathcal{H}\left(X_1, X_2, X_3\right) = 2.$$

Tvrzení 3.29. *Nechť X, Y, Z jsou náhodné veličiny. Pak*

(1) $\mathfrak{I}_{X,Y,Z} = \mathfrak{I}_X + \mathfrak{I}_{Y|X} + \mathfrak{I}_{Z|X,Y}$, s.j.

(2) $\mathfrak{I}_{Y,Z|X} = \mathfrak{I}_{Y|X} + \mathfrak{I}_{Z|X,Y}$, s.j.

(3) $\mathcal{H}(X, Y, Z) = \mathcal{H}(X) + \mathcal{H}(Y \mid X) + \mathcal{H}(Z \mid X, Y)$

(4) $\mathcal{H}(Y, Z \mid X) = \mathcal{H}(Y \mid X) + \mathcal{H}(Z \mid X, Y)$

Důkaz. Body (1) a (3) jsou řetězové pravidlo (Tvrzení 3.15) pro tři veličiny. Body (2) a (4) plynou z (1) a (3) po odečtení $\mathfrak{I}_X$, resp. $\mathcal{H}(X)$ ze součtového vzorce (Tvrzeni 3.14) pro veličiny X a (Y, Z). □

Poznamenejme ještě, že zápisem $\mathcal{H}(X, Y \mid Z)$ rozumíme entropii sdružené veličiny (X, Y) za podmínky Z. Alternativní uzávorkování $\mathcal{H}(X, (Y \mid Z))$ by ostatně nedávalo dobrý smysl, protože $Y \mid Z$ není, jak jsme viděli, náhodná veličina.

Uvažujeme-li tři náhodné veličiny, bude pro nás důležité kvantifikovat, jak výsledek jedné z nich ovlivní vzájemnou entropii zbylých dvou. Příklad 3.28 ukazuje, že tento účinek může být dramatický. Pro tento účel zavedeme podmíněnou vzájemnou entropii.

Definice 3.30. *Pro náhodný jev $U \subseteq \Omega$ s nenulovou pravděpodobností definujeme vzájemnou entropii veličin X a Y* ***za podmínky*** *U vztahem*

$$\mathcal{I}(X : Y \mid U) := \mathcal{H}(X \mid U) + \mathcal{H}(Y \mid U) - \mathcal{H}(X, Y \mid U).$$

Vzájemnou entropii veličin X a Y ***za podmínky*** *Z, kde Z je náhodná veličina, definujeme vztahem*

$$\mathcal{I}(X : Y \mid Z) := \mathcal{H}(X \mid Z) + \mathcal{H}(Y \mid Z) - \mathcal{H}(X, Y \mid Z).$$

Vzájemná entropie podmíněná náhodnou veličinou Z je opět průměrem vzájemné entropie podmíněné příslušnými jevy, které Z popisuje:

Lemma 3.31.

$$\mathcal{I}(X : Y \mid Z) = \sum_{c \in s(P_Z)} P_Z(c) \cdot \mathcal{I}(X : Y \mid Z = c).$$

Důkaz. Plyne z podobného vztahu pro entropii za podmínky v Tvrzení 3.10. □

V dalším chceme především ukázat, že podmíněná vzájemná entropie je nezáporná, a charakterizovat, kdy je nulová.

Lemma 3.32. *Pro náhodný jev $U \subseteq \Omega$, $\mathbb{P}(U) > 0$, platí*

$$\mathcal{I}(X : Y \mid U) = \mathcal{D}\left(P_{X,Y|U} \parallel P_{X|U} \cdot P_{Y|U}\right).$$

Důkaz. Použijeme známý vzorec pro rozklad podmíněné pravděpodobnosti, podle kterého pro dané $a \in A$ platí

$$P_{X|U}(a) = \mathbb{P}(X = a \mid U) = \sum_{b \in B} \mathbb{P}(X = a, Y = b \mid U) = \sum_{b \in B} P_{X,Y|U}(a, b).$$

Tedy

$$\begin{aligned}
\mathcal{H}\left(P_{X|U}\right) &= -\sum_{a \in s\left(P_{X|U}\right)} P_{X|U}(a) \log\left(P_{X|U}(a)\right) \\
&= -\sum_{a \in s\left(P_{X|U}\right)} \sum_{b \in B} P_{X,Y|U}(a, b) \log\left(P_{X|U}(a)\right) \\
&= -\sum_{(a,b) \in s\left(P_{X,Y|U}\right)} P_{X,Y|U}(a, b) \log\left(P_{X|U}(a)\right).
\end{aligned}$$

Poslední rovnost lze nahlédnout takto: pokud je $P_{X,Y|U}(a, b)$ nenulové, pak je také nenulové $P_{X|U}(a)$, a proto se v první sumě sčítá přes bohatší množinu. Ovšem ve členech, které jsou v první sumě navíc, je zastoupena nulová pravděpodobnost. Tudíž jsou nulové.

Pouhou výměnou náhodných veličin dostáváme, že

$$\mathcal{H}\left(P_{Y|U}\right) = -\sum_{(a,b)\in s\left(P_{X,Y|U}\right)} P_{X,Y|U}(a,b)\log\left(P_{Y|U}(a)\right)\,.$$

Nakonec,

$$\begin{aligned}
\mathcal{I}\left(X:Y\mid U\right) &= \mathcal{H}\left(P_{X|U}\right)+\mathcal{H}\left(P_{Y|U}\right)-\mathcal{H}\left(P_{X,Y|U}\right)\\
&= \sum_{a,b\in s P_{X,Y|U}} P_{X,Y|U}(a,b)\log\left(\frac{1}{P_{X|U}(a)}\right)\\
&\quad+\sum_{a,b\in s P_{X,Y|U}} P_{X,Y|U}(a,b)\log\left(\frac{1}{P_{Y|U}(b)}\right)\\
&\quad+\sum_{a,b\in s P_{X,Y|U}} P_{X,Y|U}(a,b)\log\left(P_{X,Y|U}(a,b)\right)\\
&= \sum_{a,b\in s P_{X,Y|U}} P_{X,Y|U}(a,b)\log\left(\frac{P_{X,Y|U}(a,b)}{P_{X|U}(a)P_{X|U}(b)}\right)\\
&= \mathcal{D}\left(P_{X,Y|U}\parallel P_{X|U}\cdot P_{Y|U}\right),
\end{aligned}$$

kde u poslední rovnosti je třeba dodat, že $s\left(P_{X,Y|U}\right)\subseteq s\left(P_{X|U}\cdot P_{Y|U}\right)$. □

Následující definice přenášejí klíčový pojem nezávislosti do kontextu podmiňování. Řekneme, že jevy $U, V \subseteq \Omega$ jsou **podmíněně nezávislé** vzhledem k jevu $W \subseteq \Omega$, pokud $\mathbb{P}(W) > 0$ a platí

$$\mathbb{P}(U\cap V\mid W) = \mathbb{P}(U\mid W)\cdot\mathbb{P}(V\mid W)\,.$$

Pojem podmíněné nezávislosti rozšíříme na náhodné veličiny přirozeným způsobem.

Definice 3.33. *Veličiny X a Y jsou **podmíněně nezávislé** vzhledem k veličině Z, pokud pro všechna $c \in s\left(P_Z\right)$, $(a,b)\in A\times B$, platí*

$$\mathbb{P}(X=a, Y=b\mid Z=c) = \mathbb{P}(X=a\mid Z=c)\cdot\mathbb{P}(Y=b\mid Z=c)\,. \qquad (\perp).$$

Tento vztah značíme zápisem $X \perp Y \mid Z$.

Dvě veličiny jsou tedy podmíněně nezávislé, pokud elementární jevy, které popisují, jsou podmíněně nezávislé vzhledem ke každému možnému jevu popsanému veličinou, kterou podmiňujeme.

Podmíněnou nezávislost lze, podobně jako klasickou nezávislost, zapsat pomocí součinů nepodmíněných pravděpodobností. Takové vyjádření je méně srozumitelné z hlediska významu, ale technicky příjemné, neboť se není třeba omezovat na nosiče veličin.

Lemma 3.34. *Veličiny X a Y jsou podmíněně nezávislé vzhledem k veličině Z, právě když pro každé $(a, b, c) \in A \times B \times C$ platí*

$$\mathbb{P}(X = a, Y = b, Z = c) \cdot \mathbb{P}(Z = c) = \mathbb{P}(X = a, Z = c) \cdot \mathbb{P}(Y = b, Z = c).$$

Důkaz. Pokud platí rovnost v lemmatu, dostaneme pro každé $c \in s(P_Z)$ vztah pro podmíněnou nezávislost vydělením této rovnosti druhou mocninou nenulové hodnoty $\mathbb{P}(Z = c)$.

Pokud naopak předpokládáme podmíněnou nezávislost, musíme uvážit dva případy. Pro $c \in s(P_Z)$ dostaneme rovnost z lemmatu analogicky k předchozí části důkazu, totiž vynásobením vztahu pro podmíněnou nezávislost druhou mocninou hodnoty $\mathbb{P}(Z = c)$. V případě $c \notin s(P_Z)$ platí rovnost v lemmatu také, protože jsou všechny pravděpodobnosti v ní nulové. □

Nyní můžeme zodpovědět naši otázku ohledně nezápornosti podmíněné vzájemné entropie.

Tvrzení 3.35. *Pro náhodné veličiny platí $\mathcal{I}(X : Y \mid Z) \geq 0$. Rovnost nastane právě tehdy když $X \perp Y \mid Z$, tedy právě když rozdělení $P_{X,Y|Z=c}$ a $P_{X|Z=c} \cdot P_{X|Z=c}$ splývají pro každé $c \in s\left(P_Z\right)$.*

Důkaz. Podle Lemmatu 3.31 je $\mathcal{I}(X : Y \mid Z)$ konvexní kombinací vzájemných entropií podmíněných jednotlivými jevy $Z = c$, $c \in s(Z)$. Tyto vzájemné entropie jsou podle Lemmatu 3.32 a Tvrzení 3.20 nezáporné. Nule se zmíněná konvexní kombinace bude podle Tvrzení 3.20 rovnat právě tehdy, když budou pro všechna $c \in s\left(P_Z\right)$ rozdělení $P_{X,Y|Z=c}$ a $P_{X|Z=c} \cdot P_{Y|Z=c}$ splývat. To je definice podmíněné nezávislosti. □

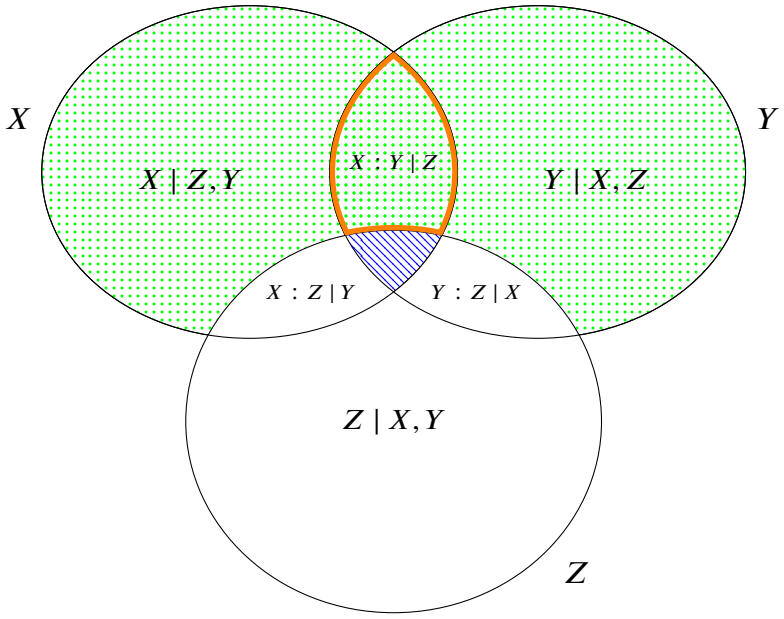
Podmíněná verze součtových vzorců je ilustrována na Obrázku 7, který umožňuje rychlé odvozování různých vztahů. Pro tečkovanou oblast např. dostáváme dvojí vyjádření

$$\mathcal{H}(X, Y \mid Z) = \mathcal{H}(X \mid Y, Z) + \mathcal{H}(Y \mid X, Z) + \mathcal{I}(X : Y \mid Z).$$

Obrázek ovšem vyžaduje důrazné varování. V Příkladu 3.28 je vzájemná entropie $\mathcal{H}\left(X_i : X_j\right)$ nulová, ale podmíněná vzájemná entropie $\mathcal{H}\left(X_i : X_j \mid X_k\right)$ je jeden bit, pro $\{i, j, k\} = \{1, 2, 3\}$. Šrafovaná oblast v Obrázku 7, jakási „vzájemná informace tří veličin“, tedy může mít „zápornou plochu“. Příslušná hodnota, kterou bychom mohli být v pokušení značit $\mathcal{I}(X : Y : Z)$, nemá příliš jasný intuitivní význam. Můžeme ji ale např. chápat jako vliv třetí veličiny na vzájemnou informaci druhých dvou. Vidíme, že tento vliv je symetrický, tedy platí

$$\begin{aligned}\mathcal{I}(X : Y) - \mathcal{I}((X : Y \mid Z) &= \mathcal{I}(Y : Z) - \mathcal{I}(Y : Z \mid X) \\ &= \mathcal{I}(Z : X) - \mathcal{I}(Z : X \mid Y).\end{aligned}$$

Dokažme tedy následující lemma, které poskytuje alternativní vyjádření oblasti vyznačené v Obrazku 7 tučně.



Obrázek 7: Informace a entropie vzájemná a za podmínky pro tři veličiny

Lemma 3.36.

$$\begin{aligned}\mathcal{I}(X : Y \mid Z) &= \mathcal{H}(X,Z) + \mathcal{H}(Y,Z) - \mathcal{H}(Z) - \mathcal{H}(X,Y,Z)\\ &= \mathcal{H}(X \mid Z) - \mathcal{H}(X \mid Y,Z)\\ &= \mathcal{I}(X : (Y,Z)) - \mathcal{I}(X : Z)\,.\end{aligned}$$

Důkaz. Z definice vzájemné entropie za podmínky a ze součtových vztahů pro entropie za podmínky dostáváme následující řadu rovností, obsahující i rovnosti z tvrzení lemmatu:

$$\begin{aligned}\mathcal{I}(X : Y \mid Z) &= \mathcal{H}(X,Z) - \mathcal{H}(Z) + \mathcal{H}(Y,Z) - \mathcal{H}(Z) - \big(\mathcal{H}(X,Y,Z) - \mathcal{H}(Z)\big)\\ &= \mathcal{H}(X,Z) + \mathcal{H}(Y,Z) - \mathcal{H}(Z) - \mathcal{H}(X,Y,Z)\\ &= \mathcal{H}(X,Z) - \mathcal{H}(Z) - \big(\mathcal{H}(X,Y,Z) - \mathcal{H}(Y,Z)\big)\\ &= \mathcal{H}(X \mid Z) - \mathcal{H}(X \mid Y,Z)\\ &= \mathcal{H}(X) - \mathcal{H}(X \mid Y,Z) - \big(\mathcal{H}(X) - \mathcal{H}(X \mid Z)\big)\\ &= \mathcal{I}(X : (Y,Z)) - \mathcal{I}(X : Z)\,.\end{aligned}$$

□

Okamžitým důsledkem definice vzájemné entropie za podmínky, Tvrzení 3.35 a druhé a třetí rovnosti předchozího tvrzení jsou tvrzení následující.

Tvrzení 3.37. *Pro tři náhodné veličiny X, Y, Z platí*

(1) $\mathcal{H}(X,Z) + \mathcal{H}(Y,Z) - \mathcal{H}(X,Y,Z) - \mathcal{H}(Z) \geq 0$ (submodularita entropie)*;*

(2) $\mathcal{H}(X, Y \mid Z) \leq \mathcal{H}(X \mid Z) + \mathcal{H}(Y \mid Z)$ (subaditivita podmíněné entropie)*;*

(3) $\mathcal{H}(X \mid Y, Z) \leq \mathcal{H}(X \mid Z)$ (monotonie entropie vzhledem k podmínce)*;*

(4) $\mathcal{I}(X : Z) \leq \mathcal{I}(X : (Y, Z))$ *(*monotonie vzájemné informace).

Rovnosti nastanou současně a to právě v případě, že $X \perp Y \mid Z$.

Na konec této kapitoly uvedeme pojem, který si blíže všímá třetí podmínky v předchozím tvrzení, tedy monotonie entropie vzhledem k podmínce. Ta říká, že entropie klesá se vzrůstající apriorní informací. Případ, kdy větší apriorní informace nepřináší snížení entropie popisuje tzv. markovská vlastnost. Ta pro nás bude důležitá i v kontextu náhodných procesů. Nejprve uveďme definici.

Definice 3.38. *Uspořádaná trojice veličin* X, Z, Y *tvoří* ***markovský řetězec****, značíme* $X \to Z \to Y$*, pokud platí*

$$\mathbb{P}(Y = b \mid Z = c, X = a) = \mathbb{P}(Y = b \mid Z = c),$$

pro všechna $b \in B$*,* $(a, c) \in s\left(P_{X,Z}\right)$.

Značení pro markovský řetězec nás zve, abychom se dívali na náhodné veličiny X, Y a Z a jako na náhodný proces, tedy jako na posloupnost náhodných událostí v čase, jejichž podmíněná pravděpodobnost je dána nějakými kauzálními souvislostmi. Markovská vlastnost pak znamená, že X ovlivňuje Y výhradně svým účinkem na Z. Známe-li Z, víme vše, čím X přispívá ke znalosti Y. Následující lemma potvrzuje, co naznačuje uvedený neformální popis, totiž že markovská vlastnost je ekvivalentní podmíněné nezávislosti.

Tvrzení 3.39. *Následující podmínky jsou ekvivalentní:*

(1) $X \to Z \to Y$ *je markovský řetězec*

(2) $Y \to Z \to X$ *je markovský řetězec*

(3) $X \perp Y \mid Z$.

(4) $Y \perp X \mid Z$.

Důkaz. Z definice podmíněné nezávislosti okamžitě plyne, že jsou poslední dvě podmínky ekvivalentní. Vzhledem k symetrii podmínek stačí pro platnost celého lematu dokázat ekvivalenci podmínek (1) a (3).

Dokažme nejprve implikaci (1) $\Rightarrow$ (3). Předpokládejme tedy, že $c \in P_Z$, $a \in A$, $b \in B$. Nejprve prozkoumejme triviální možnost, $(a, c) \notin s\left(P_{X,Z}\right)$. V takovém případě jsou pravděpodobnosti $\mathbb{P}(X = a \mid Z = c)$ a $\mathbb{P}(X = a, Y = b \mid Z = c)$ nulové a

definiční vztah podmíněné nezávislosti je splněn, neboť na obou stranách rovnice dostáváme nulu. Pro netriviální případ $(a, c) \in s\left(P_{X,Z}\right)$ platí definiční vztah markovského řetězce, který vynásobíme dobře definovaným výrazem $\mathbb{P}(X = a \mid Z = c)$ a dostaneme:

$$\begin{aligned}\frac{\mathbb{P}(X = a, Z = c)}{\mathbb{P}(Z = c)} \cdot \frac{\mathbb{P}(Y = b, Z = c, X = a)}{\mathbb{P}(Z = c, X = a)} = \\ = \mathbb{P}(X = a \mid Z = c) \cdot \mathbb{P}(Y = b \mid Z = c),\end{aligned}$$

a krácením nenulové hodnoty $\mathbb{P}(X = a, Z = c)$ dostáváme požadované

$$\mathbb{P}(X = a, Y = b \mid Z = c) = \mathbb{P}(X = a \mid Z = c) \cdot \mathbb{P}(Y = b \mid Z = c) .$$

Nyní dokažme implikaci (3) $\Rightarrow$ (1). Nechť $(a, c) \in s(P_{Y,Z})$, $b \in B$. Pak také $c \in s\left(P_Z\right)$ a definiční vztah podmíněné nezávislosti můžeme vydělit nenulovou pravděpodobností $\mathbb{P}(X = a \mid Z = c)$ a tím dostaneme definiční vztah pro markovský řetězec, viz předchozí část důkazu. □

Předchozí lemma ukazuje důležitou symetrii vztahu $X \to Z \to Y$, který bychom mohli díky tomu zapisovat jako $X \leftrightarrow Z \leftrightarrow Y$. Tím je zpochybněno výše uvedené výlučně kauzální chápání procesu. Přesto je kauzální interpretace vztahu členů řetězce základní možností aplikace, což vyjadřuje i fakt, že následující nerovnosti, které odpovídají naší neformální diskusi a z markovské vlastnosti přímo plynou, se anglicky označují jako **data processing inequality** (nejčastěji v podobě první z nich, viz Theorem 2.8.1 v CT). Zpracováním dat nemůžeme získat žádnou informaci o jejich zdroji, která již v datech není před zpracováním.

Tvrzení 3.40. *Pro markovský proces $X \to Z \to Y$ platí*

(1) $\mathcal{I}(X : Y) \leq \mathcal{I}(X : Z)$

(2) $\mathcal{I}(Y : X) \leq \mathcal{I}(Y : Z)$

(3) $\mathcal{H}(X \mid Y) \geq \mathcal{H}(X \mid Z)$

(4) $\mathcal{H}(Y \mid X) \geq \mathcal{H}(Y \mid Z)$

Důkaz. Platí

$$\begin{aligned}\mathcal{I}(X : Z) - \mathcal{I}(X : Y) &= H(X) - H(X \mid Z) - (H(X) - H(X \mid Y)) = \\ &= \mathcal{H}(X \mid Y) - \mathcal{H}(X \mid Z) \geq H(X \mid Y, Z) - H(X \mid Z) = \\ &= -I(X : Y \mid Z) = 0,\end{aligned}$$

kde použitá nerovnost byla dokázána v Tvrzení 3.37. Celkově dostáváme platnost nerovností (1) a (3). Zbylé dvě plynou ze symetrie X a Y. □

Kauzální povahu markovského řetězce lze upřesnit jako posloupnost náhodných veličin, kdy následující vznikne vždy „zašuměním“ veličiny předchozí. Neboli $Z = f_1(X, \xi_1)$, kde ξ_1 reprezentuje šum, což je v matematickém pojetí náhodná veličina, a f_1 je deterministická funkce popisující vliv šumu na veličinu X. Podobně $Y = f_2(Z, \xi_2)$, kde ξ_2 reprezentuje další šum a f_2 je opět deterministická funkce. Markovská vlastnost je vyjádřena požadavkem, aby tento šum ξ_2 „nevěděl nic o minulosti“, tedy aby byl nezávislý na páru (X, ξ_1). Uvědomme si, že nestačí, aby byl šum ξ_2 nezávislý na X. Mohlo by se totiž stát, že se nejedná o další šum, ale naopak o odstranění šumu ξ_1.

Kapitolu jsme zahájili Příkladem 3.28, který popisuje situaci, kdy jsou veličiny nezávislé, ale nejsou podmíněně nezávislé. Na závěr poznamenejme, že je možný i opačný případ. Například podmíněná entropie $I(X : Y \mid X)$ je nulová, neboli $X \to X \to Y$ je markovský řetězec, bez ohledu na to, zda jsou X a Y nezávislé, či nikoli.

Kontrolní otázky

1. Dokažte součtové vztahy nad Lemmatem 3.36.

2. Vyjadřují ostatní plochy v Obrázku 7, kromě šrafované, kladné hodnoty?

4 Náhodné procesy a jejich entropie

Matematické pojetí informace vyložené v předchozích kapitolách lze shrnout takto: informace je informační obsah zprávy, což je mínus logaritmus pravděpodobnosti této zprávy. Zprávu tedy chápeme jako hodnotu náhodné veličiny, která vybírá mezi (konečně mnoha) možnými zprávami. Střední hodnota informačního obsahu dané veličiny je její entropie.

Tento popis chceme nyní rozšířit na v praxi všudypřítomnou situaci (potenciálně) nekonečné posloupnosti zpráv. Jedná se tedy o posloupnost $\left(X_n\right)_{n\in\mathbb{N}}$ náhodných veličin. Pro jednoduchost můžeme předpokládat, že všechny náhodné veličiny nabývají hodnot ze stejné abecedy A. Taková posloupnost se nazývá **náhodný proces**. Klasickým příkladem náhodného procesu jsou písmena v souboru nějakých textů, např. knih v českém jazyce. Z tohoto příkladu je vidět, že mezi jednotlivými veličinami procesu mohou být velmi komplexní závislosti (jak je pravděpodobnostní rozdělení jednadvacátého písmene knihy v závislosti na předchozích dvaceti?), a to i v případě, kdy mají jednotlivé veličiny stejné rozdělení.

Náhodný proces

Náhodný proces bychom mohli zkusit chápat jako jednu náhodnou veličinu nabývající hodnot z množiny A^ω nekonečných posloupností nad abecedou A. Jak už jsme však upozorňovali v Kapitole 2, opustili bychom tím bezpečné vody diskrétní pravděpodobnosti: možných hodnot je (v netriviálních případech) nespočetně mnoho a každá jednotlivá hodnota má nulovou pravděpodobnost. To vyžaduje mnohem robustnější teorii, která se musí opřít o celou teorii pravděpodobnosti včetně celého nezbytného aparátu σ-algebry $\mathcal{F}$ měřitelných množin (v tomto případě nemohou mít všechny množiny konzistentně definovanou pravděpodobnost). Proto budeme v souladu s Kapitolou 2 uvažovat náhodný proces $X = \left(X_n\right)_{n\in\mathbb{N}}$ jako posloupnost prodlužujících se náhodných vektorů $X_{[0..n)}$ a případně vyšetřovat asymptotické chování takové posloupnosti.

Entropie procesu

V rámci této přednášky nás zajímá především informační obsah náhodného procesu. Chceme pro něj tedy rozumně definovat pojem entropie. Vzhledem k tomu, že je proces limitou náhodných vektorů $X_{[0..n)}$, pro které je entropie dobře definovaná, je jedinou rozumnou možností definice

$$\mathcal{H}(X) := \lim_{n\to\infty} \frac{1}{n}\mathcal{H}\left(X_{[0..n)}\right),$$

která samozřejmě dává smysl jen tehdy, pokud limita existuje (což je ekvivalentní rovnosti $\liminf$ a $\limsup$, které jsou definovány vždy). Jinak říkáme, že náhodný proces entropii nemá.

Entropie procesu je tedy limita entropie připadající průměrně na jeden symbol. Stejnou myšlenku lze zapsat následovně, kdy je entropie počátečního vektoru zapsána přímo jako součet příspěvků jednotlivých symbolů.

Lemma 4.1.

$$\mathcal{H}(X) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathcal{H}\left(X_i \mid X_{[0..i)}\right).$$

Důkaz. Přímo z řetězového pravidla (Tvrzení 3.15). □

Pro limitu aritmetického průměru platí následující jednoduché lemma.

Lemma 4.2. *Má-li posloupnost reálných čísel* $\left(a_n\right)_{n\in\mathbb{N}}$ *limitu, pak má tutéž limitu i posloupnost částečných průměrů* $b_n := \frac{1}{n}\sum_{i=0}^{n-1} a_i$.

Důkaz. Nechť je a limita $\left(a_n\right)_{n\in\mathbb{N}}$ a pro $\varepsilon > 0$ nechť je $N(\varepsilon) \in \mathbb{N}$ číslo takové, že $\left|a_i - a\right| < \varepsilon$ pro každé $n > N(\varepsilon)$. Pak pro každé $\varepsilon > 0$ a každé $n > N(\varepsilon/2)$ platí

$$\begin{aligned}
\left|b_n - a\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}(a_i - a)\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left|a_i - a\right| \\
&= \frac{1}{n}\sum_{i=1}^{N(\varepsilon/2)}\left|a_i - a\right| + \frac{1}{n}\sum_{i=N(\varepsilon/2)+1}^{n}\left|a_i - a\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{N(\varepsilon)}\left|a_i - a\right| + \varepsilon/2,
\end{aligned}$$

kde první sčítanec jde k nule. Pro dostatečně velká n je tedy $\left|b_n - a\right| < \varepsilon$, což jsme chtěli ukázat. □

Odtud plyne následující kritérium pro existenci entropie procesu:

Důsledek 4.3. *Pokud* $\mathcal{H}\left(X_n \mid X_{[0..n)}\right)$ *konverguje, má proces* X *entropii.*

Druhy konvergence

V případě entropie jsme vystačili s běžným pojmem limity. U procesu *reálných* náhodných veličin $\left(Y_i\right)_{i\in\mathbb{N}}$ budeme ale také někdy mluvit o tom, že posloupnost takových veličin konverguje k nějaké limitní náhodné veličině Y. Rozeznáváme přitom dva různé typy konvergence. Vzhledem k tomu, že všechny náhodné veličiny reálného procesu $\left(Y_i\right)_{i\in\mathbb{N}}$ jsou zobrazení $\Omega \to \mathbb{R}$, nabízí se jednak konvergence definovaná bodově. Pro každé $\omega \in \Omega$ je tvrzení $\lim_{n\to\infty} Y_n(\omega) = Y(\omega)$ opět standardní tvrzení o limitě reálné posloupnosti. V souladu s naší zavedenou terminologií tedy řekneme, že $\left(Y_i\right)_{i\in N}$ k Y **konverguje skoro jistě**, pokud

$$\mathbb{P}\left(\left\{\omega \in \Omega \mid \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\right\}\right) = 1.$$

To lze také zapsat vztahem

$$\forall \varepsilon > 0, \qquad \lim_{n\to\infty} \mathbb{P}\left(\sup_{k\geq n} \left|Y_k - Y\right| > \varepsilon\right) = 0\,,$$

neboli míra množiny, na které se posloupnost $\left(Y_i\right)_{i\in\mathbb{N}}$ *ještě někdy v budoucnu* odchýlí od Y o více než ε, se pro rostoucí n postupně snižuje až k nule.

Konvergence skoro jistě je výhodná v situaci, kdy máme kontrolu nad bodovým chováním náhodných veličin procesu. Často ale umíme dokázat, že se náhodné veličiny Y_n k veličině Y blíží na vzdálenost ε s velkou pravděpodobností, tedy na množinách míry jdoucí k jedné, nemáme ale dobrou kontrolu nad tím, jaké množiny to jsou. Může se dokonce stát, že body ω se ve zbytkové „ε-neukázněné“ množině pro rostoucí n střídají tak, že posloupnost nekonverguje bodově nikde: pro každé ω se $Y_n(\omega)$ výrazně odchýlí od $Y(\omega)$ nekonečněkrát. Pro tyto situace definujeme slabší pojem **konvergence v pravděpodobnosti**, která nastává, pokud pro každé $\varepsilon > 0$ platí

$$\lim_{n\to\infty} \mathbb{P}\left(\left|Y_n - Y\right| > \varepsilon\right) = 0\,.$$

Pro uvedené konvergence budeme používat zkrácený zápis:

$$Y_n \to Y \text{ v.p.} \qquad Y_n \to Y \text{ s.j.}$$

resp.

$$\lim_{n\to\infty} Y_n = Y \text{ v.p.} \qquad \lim_{n\to\infty} Y_n = Y \text{ s.j.}$$

AEP

Připomeňme, že entropie je střední hodnota informačního obsahu. Pokud má proces entropii, střední hodnota informačního obsahu přepočítaná na symbol se pro prodlužující se výstupy procesu ustaluje. Zajímavá otázka je, co se při tom děje s jednotlivými informačními obsahy. Ukazuje se, že pro řadu významných procesů máme o chování informačních obsahů jednotlivých výstupů velmi dobrý přehled, který bude navíc hrát klíčovou roli jak při kompresi, tak při přenosu informace kanálem. Konkrétně dochází k tomu, že informační obsahy na jeden symbol konvergují (v jednom z uvedených významů) ke své střední hodnotě, tedy k entropii. To motivuje následující definici.

Definice 4.4. *Řekneme, že náhodný proces $X = (X_i)_{i\in\mathbb{N}}$ s entropií má **asymptoticky rovnoměrné rozdělení**, pokud* v.p. *platí*

$$\lim_{n\to\infty} \frac{1}{n} \mathfrak{I}_{X_{[0..n)}} = \mathcal{H}(X)\ .$$

Pokud konvergence platí s.j. *, řekneme že X má **silně asymptoticky rovnoměrné rozdělení**.*

Tuto vlastnost budeme v krátkosti označovat jako AEP (z anglického „asymptotic equipartition property“), resp. sAEP. Jak už bylo řečeno, AEP znamená, že výstupy procesu mají s prodlužující se délkou s čím dál větší pravděpodobností informační obsah odpovídající průměru, tedy entropii. Tuto množinu výstupů zachycuje následující definice.

Definice 4.5. *Nechť je X proces s entropií. Množinu*

$$\mathcal{M}_\varepsilon^n(X) = \left\{ u \in A^n \,:\, \left| \frac{1}{n} \mathfrak{I}\left(X_{[0..n)} = u\right) - \mathcal{H}(X) \right| < \varepsilon \right\},$$

nazveme ***ε-typickou množinou výstupů délky*** *n procesu X.*

Předchozí definice používá informační obsah jevu (viz Definice 3.1 a diskuse před Definicí 3.3). Prvky typické množiny budeme také neformálně označovat jako „typické výstupy“. Typická množina procesu s AEP má následující vlastnosti.

Tvrzení 4.6. *Buď X proces s AEP a $\varepsilon, \delta > 0$.*

(1) Pro $u \in \mathcal{M}_\varepsilon^n$ platí

$$2^{-n(\mathcal{H}(X)+\varepsilon)} \leq \mathbb{P}\left(X_{[0..n)} = u\right) \leq 2^{-n(\mathcal{H}(X)-\varepsilon)}.$$

(2) Pro dostatečně velká n platí $\mathbb{P}\left(X_{[0..n)} \in \mathcal{M}_\varepsilon^n\right) > 1 - \delta$.

(3) $\left|\mathcal{M}_\varepsilon^n\right| \leq 2^{n(\mathcal{H}(X)+\varepsilon)}$

(4) $\left|\mathcal{M}_\varepsilon^n\right| > (1-\delta) \cdot 2^{n(\mathcal{H}(X)-\varepsilon)}$ pro dostatečně velká n.

Důkaz. (1) plyne z definice typické množiny a (2) z definice AEP.

Podle (1) pro každé $u \in \mathcal{M}_\varepsilon^n$ platí $\mathbb{P}\left(X_{[0..n)} = u\right) \geq 2^{-n(\mathcal{H}(X)+\varepsilon)}$. Protože jevy $X_{[0..n)} = u$ jsou pro různá u disjunktní, nemůže jich být víc než tvrdí (3).

Podobně pro n, pro které podle bodu (2) platí $\mathbb{P}\left(X_{[0..n)} \in \mathcal{M}_\varepsilon^n\right) > 1 - \delta$, musí vzhledem k $\mathbb{P}\left(X_{[0..n)} = u\right) \leq 2^{-n(\mathcal{H}(X)-\varepsilon)}$ mohutnost $\mathcal{M}_\varepsilon^n$ splňovat (4). □

Podle předchozího tvrzení mají typické výstupy podobnou pravděpodobnost a počáteční vektory procesu se tedy čím dál víc podobají uniformnímu rozdělení na typické množině. Onu zdánlivou rovnoměrnost je ale třeba vnímat v logaritmické škále: blízké jsou informační obsahy spíše než pravděpodobnost.

Všimněme si paradoxního faktu, že typické výstupy jsou téměř jisté, ale početně naopak tvoří stále menší část všech výstupů, jak je vidět z jejich počtu. Podíl „typických výstupů“ délky n na všech výstupech je totiž zhruba

$$\frac{2^{n\mathcal{H}(X)}}{|A|^n} = 2^{n\left(\mathcal{H}(X) - \log|A|\right)}$$

a tedy exponenciálně rychle klesá k nule, s výjimkou procesu s entropií $\log|A|$. Ten odpovídá procesu nezávislých uniformních rozdělení, u kterého jsou typické *všechny* výstupy.

Stacionární proces

Standardním případem, kdy příspěvek jednotlivých symbolů konverguje, je **stacionární proces**, tedy proces, pro který platí

$$\mathbb{P}\left(X_{[0..n)} = u\right) = \mathbb{P}\left(X_{[k,n+k)} = u\right) \qquad k, n \in \mathbb{N},\ u = u_0 u_1 \cdots u_{n-1} \in A^n.$$

Pro $k = 1$ dostáváme, že všechny veličiny stacionárního procesu mají stejné rozdělení. Všimněme si ovšem, že to pro stacionaritu procesu nestačí, uvažme např. proces $\left(Y_n\right)_{n\in\mathbb{N}}$, kde $Y_{2i} = Y_{2i+1} = X_i$, kde $\left(X_n\right)_{n\in\mathbb{N}}$ je i.i.d. proces. Pak jsou všechna Y_n stejně rozdělená, ale existují dvě odlišná rozdělení pro vektory (Y_i, Y_{i+1}).

Tvrzení 4.7. *Nechť X je stacionární proces.*

(1) Pro všechna $k, m, \ell \in \mathbb{N}$, platí

$$s\left(P_{X_{[k..m)}}\right) = s\left(P_{X_{[k+\ell..m+\ell)}}\right), \qquad \mathcal{H}\left(X_{[k..m)}\right) = \mathcal{H}\left(X_{[k+\ell..m+\ell)}\right),$$

(2) posloupnost $\mathcal{H}\left(X_n \mid X_{[0..n)}\right)$, $n \in \mathbb{N}$, je nerostoucí,

(3) entropie procesu $\mathcal{H}(X)$ je dobře definovaná a je rovna limitě

$$\mathcal{H}(X) = \lim_{n\to\infty} \mathcal{H}\left(X_n \mid X_{[0..n)}\right).$$

Důkaz. Buď X stacionární proces $k, m, \ell \in \mathbb{N}$, $k < m$, $u \in A^{m-k}$. Z definice stacionarity dostáváme, že se pravděpodobnosti $\mathbb{P}\left(X_{[k,m)} = u\right)$ a $\mathbb{P}\left(X_{[k+\ell,m+\ell)} = u\right)$ rovnají. Z toho plyne

$$s\left(P_{X_{[k..m)}}\right) = s\left(P_{X_{[k+\ell..m+\ell)}}\right)$$

a

$$\begin{aligned}\mathcal{H}\left(X_{[k..m)}\right) &= \sum_{u\in s\left(P_{X_{[k..m)}}\right)} -\mathbb{P}\left(X_{[k,m)} = u\right) \log \mathbb{P}\left(X_{[k,m)} = u\right) \\ &= \sum_{u\in s\left(P_{X_{[k+\ell..m+\ell)}}\right)} -\mathbb{P}\left(X_{[k+\ell..m+\ell)} = u\right) \log \mathbb{P}\left(X_{[k+\ell..m+\ell)} = u\right) = \mathcal{H}\left(X_{[k+\ell..m+\ell)}\right).\end{aligned}$$

Tím je dokázáno (1). Z toho a z Tvrzení 3.37 dále dostáváme

$$\begin{aligned}\mathcal{H}\left(X_n \mid X_{[0..n)}\right) &\le \mathcal{H}\left(X_n \mid X_{[1..n)}\right) \\ &= \mathcal{H}\left(X_{[1..n+1)}\right) - \mathcal{H}\left(X_{[1..n)}\right) = \mathcal{H}\left(X_{[0..n)}\right) - \mathcal{H}\left(X_{[0..n-1)}\right) \\ &= \mathcal{H}\left(X_{n-1} \mid X_{[0..n-1)}\right).\end{aligned}$$

Posloupnost $\mathcal{H}\left(X_n \mid X_{[0..n)}\right)$ je tedy nerostoucí a nezáporná, má tedy limitu a (3) plyne z Důsledku 4.3. □

I.i.d. procesy

Nejjednodušším případem, kterému se budeme věnovat, jsou **i.i.d. procesy** (independent and identically distributed), tedy posloupnosti, ve kterých mají všechny veličiny $X_i : \Omega \to A$ stejné rozdělení P, které budeme nazývat **základní rozdělění**, a jsou nezávislé, neboli

$$\mathbb{P}\left(X_i = a\right) = \mathbb{P}\left(X_j = a\right) = P(a), \qquad i, j \in \mathbb{N}, a \in A$$
$$\mathbb{P}\left(X_{[0..n)} = u\right) = \prod_{i=0}^{n-1} \mathbb{P}\left(X_i = u_i\right) = \prod_{i=0}^{n-1} P(u_i) \quad n \in \mathbb{N}, u = u_0 u_1 \cdots u_{n-1} \in A^n.$$

Připomeňme, že takto definovaná nezávislost je silnější, než nezávislost po dvojicích. Viz Příklad 3.28, ve kterém jsou veličiny nezávislé po dvou, ale ne po třech.

Z nezávislosti (viz Tvrzení 3.6) okamžitě dostáváme

Lemma 4.8. *I.i.d. proces $X = \left(X_n\right)_{n\in\mathbb{N}}$ je stacionární a pro všechna i a n platí*

$$\frac{1}{n}\mathcal{H}\left(X_{[0..n)}\right) = \mathcal{H}\left(X_i\right), \qquad \mathcal{H}(X) = \mathcal{H}\left(X_i\right).$$

Základní nástroj pro studium i.i.d. procesů je silný zákon velkých čísel, který připomínáme bez důkazu.

Věta 4.9 (Silný zákon velkých čísel)**.** *Buď $\left(X_n\right)_{n\in\mathbb{N}}$ i.i.d. proces se základním rozdělením P a hodnotami v A, a buď $f : A \to \mathbb{R}$ libovolná funkce. Pak*

$$\frac{1}{n}\sum_{i=0}^{n} f(X_i) \to \mathbb{E}_P(f) \text{ s.j.}$$

Uvedený zákon se nazývá „silný“, protože v něm figuruje konvergence skoro jistě. Připomeňme, že v Kapitole 2 jsme definovali $\mathbb{E}_P(f)$ jako $\sum_{a\in s(P)} P(a)f(a)$. Vzhledem k předpokladu identického rozdělení navíc pro každé i platí

$$\mathbb{E}_P(f) = \mathbb{E}\left(f(X_i)\right).$$

Ze zákona velkých čísel jednoduše plyne, že i.i.d. procesy mají AEP.

Tvrzení 4.10. *Každý i.i.d. proces má silně asymptoticky rovnoměrné rozdělení.*

Důkaz. Nechť je $X = \left(X_n\right)_{n\in\mathbb{N}}$ i.i.d. proces se základním rozdělením $P = P_{X_0}$. Zvolme v silném zákonu velkých čísel za f funkci definovanou na $a \in s(P)$ jako $-\log P(a)$. Pak $\mathbb{E}_P(f) = \mathcal{H}(X)$ a pro všechna i na $s\left(X_i\right)$ platí $f(X_i) = \mathfrak{I}_{X_i}$. Z nezávislosti pak na nosiči každého vektoru $X_{[0..n)}$ plyne

$$\frac{1}{n}\mathfrak{I}_{X_{[0..n)}} = \frac{1}{n}\sum_{i=0}^{n-1}\mathfrak{I}_{X_i} = \frac{1}{n}\sum_{i=0}^{n-1} f(X_i).$$

Zákon velkých čísel je tedy pro naše f jen jinou formulací konvergence z definice sAEP. □

Podívejme se blíže na to, jak vypadají prvky typické množiny i.i.d. procesu. Označme $P = P_{X_0}$ a předpokládejme $A = s(P)$ (uvažujme jen možné hodnoty). Pak máme

$$\mathcal{H}(P) = -\sum_{a \in A} P(a) \log P(a)$$

a pro slovo $u = u_0 \cdots u_{n-1} \in A^n$ platí díky nezávislosti X_i

$$\mathfrak{I}\left(X_{[0..n)} = u\right) = -\sum_{i=1}^{n} \log P(u_i) \; = -\sum_{a \in A} |u|_a \log P(u_a)\,,$$

kde $|u|$ značí délku slova u a $|u|_a$ značí počet výskytů znaku (neboli *písmene*) a ve slově u. Z toho je vidět, že slovo patří do typické množiny, pokud relativní četnost $\frac{|u|_a}{|u|}$ jednotlivých písmen v u odpovídá (zhruba) jejich pravděpodobnosti $P(a)$ (taková slova nazýváme **frekvenčně typická** pro rozdělení P). To dává pojmu „typická množina“ další intuitivní smysl, který je navíc úzce propojen se zákonem velkých čísel: slovo je prvkem typické množiny, pokud v něm jsou frekvence výskytu písmen blízké očekávání; taková slova se navíc objevují nejčastěji.

Několik varování:

- V typické množině se nemusejí vyskytovat jen frekvenčně typická slova. Očekávaný informační obsah lze dosáhnout i jinak.
- Slova z typické množiny se objevují nejčastěji, ale početně (obvykle) tvoří velmi malou část všech výstupů (viz výše).
- Typické slovo nemá samo o sobě maximální pravděpodobnost. Uvědomme si, že nejpravděpodobnějším slovem ze všech je posloupnost opakujícího se nejpravděpodobnějšího písmene. Takové slovo je ovšem jen jedno a pravděpodobnost, že se objeví, je tedy mizivá. Každé jednotlivé slovo z typické množiny je sice také málo pravděpodobné, vysoce pravděpodobné ovšem je, že se objeví *nějaké* slovo z typické množiny.

Příklad 4.11. Uvažujme rozdělení na čtyřpísmenné abecedě $A = \{a, b, c, d\}$, kde

$$P_Y(a) = P_Y(b) = \frac{1}{8}, \quad P_Y(c) = \frac{1}{4}, \quad P_Y(d) = \frac{1}{2}\,.$$

Frekvenčně typické slovo délky 8 obsahuje 1 áčko, 1 béčko, 2 céčka a 4 déčka. Celkový informační obsah takového slova je $1 \cdot 3 + 1 \cdot 3 + 2 \cdot 2 + 4 \cdot 1 = 14$ bitů. To je očekávaná hodnota osmi kopií P_Y, protože

$$\mathcal{H}\left(P_Y\right) = \frac{1}{8}\log 8 + \frac{1}{8}\log 8 + \frac{1}{4}\log 4 + \frac{1}{2}\log 2 = \frac{7}{4}\,.$$

Následující tabulka ukazuje ve druhém řádku počty slov s informačním obsahem uvedeným v prvním řádku. Ve třetím řádku je souhrnná pravděpodobnost slov s daným informačním obsahem. Pravděpodobnost jednotlivého slova u je z definice informačního obsahu rovna $2^{-\mathfrak{I}(u)}$.

8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	8	44	168	518	1288	2716	4824	7393	9648	10 864	10 304	8288	5376	2816	1024	256
0.39	1.6	4.3	8.2	13.	16.	17.	15.	11.	7.4	4.1	2.0	0.79	0.26	0.067	0.012	0.0015

Další tabulka ukazuje složení množiny slov s očekávaným informačním obsahem 14. Je vidět, že se nejedná pouze o slova s typickou frekvencí $(1, 1, 2, 4)$, ačkoli ty tvoří největší skupinu.

{0, 0, 6, 2}	{0, 1, 4, 3}	{0, 2, 2, 4}	{0, 3, 0, 5}	{1, 0, 4, 3}	{1, 1, 2, 4}	{1, 2, 0, 5}	{2, 0, 2, 4}	{2, 1, 0, 5}	{3, 0, 0, 5}
28	280	420	56	280	840	168	420	168	56

Pokud za ε v definici typické množiny zvolíme $\frac{1}{8}$, dostanou se do ní všechna slova s informačním obsahem 13 až 15. Těch je zhruba 13.5%, ale jejich souhrnná pravděpodobnost je zhruba 48%.

Markovské procesy

Mírně složitější jsou **procesy markovské** (řádu jedna), které zobecňují pojem krátkého markovského řetězce ze závěru Kapitoly 3.4: závislost každé veličiny na dosavadním průběhu procesu lze v markovském procesu redukovat na závislost na bezprostředně předcházející veličině. Jsou to tedy procesy, které splňují podmínku

$$\mathbb{P}\left(X_n = a \mid X_{[0..n)} = u\right) = \mathbb{P}\left(X_n = a \mid X_{n-1} = u_{n-1}\right)$$

pro všechna $a \in A$, $n \geq 1$, $u = u_0 u_1 \cdots u_{n-1} \in A^n$ taková, že jevy v podmínce mají nenulovou pravděpodobnost. Jinými slovy,

$$X_{[0..n-1)} \to X_{n-1} \to X_n, \qquad n \geq 2.$$

Je-li navíc závislost X_{n+1} na X_n stejná pro všechna n, t.j. $\mathbb{P}\left(X_{n+1} = b \mid X_n = a\right)$ je konstantní vzhledem k n pro všechna n, pro která je definovaná, mluvíme o **homogenním markovském procesu**. Takový proces je plně určen hodnotami $m_{a,b}$, $a, b \in A$, které určují pravděpodobnost toho, že X_{n+1} je rovno b, pokud je $X_n = a$. Tyto hodnoty tvoří čtvercovou matici M, která se nazývá **přechodovou maticí** markovského procesu. Je to tzv. **stochastická matice**, tedy matice jejíž každý řádek je **stochastický vektor**, tedy vektor nezáporných reálných čísel se součtem jedna. Jinak řečeno, stochastický vektor je pravděpodobnostní rozdělení (v našem případě na A) a stochastická matice je soubor takových rozdělení (v našem případě indexovaný opět množinou A). Snadno ověříme, že pro přechodovou matici M platí klíčový vztah

$$P_{X_{n+1}} = P_{X_n} \cdot M\,.$$

Mocniny přechodové matice M popisují podmíněné pravděpodobnosti dalších veličin na počátečním rozdělení a její vlastnosti tak určují vlastnosti procesu.

Všimněme si drobné technické komplikace: pokud je $\mathbb{P}\left(X_n = a\right)$ nulová, nemůžeme psát $m_{a,b} = \mathbb{P}\left(X_{n+1} = b \mid X_n = a\right)$. Proto jsme výše museli podmínku nenulové pravděpodobnosti stále opakovat. V krajním případě, kdy je $\mathbb{P}\left(X_n = a\right) = 0$ pro všechna n, je dokonce $m_{a,b}$ pro proces irelevantní. Zatímco tedy každá stochastická matice definuje spolu s X_0 jednoznačně markovský proces, jehož je přechodovou maticí, pro daný markovský proces může existovat přechodových maticí, které ho určují, více. Nejednoznačnost je právě v členech $m_{a,b}$, pokud je hodnota a v celém procesu nemožným jevem. Komplikaci s nedefinovanými podmíněnými pravděpodobnostmi lze opět obejít tím, že markovskou podmínku zapíšeme jako

$$\begin{aligned}&\mathbb{P}\left(X_n = a, X_{[0..n)} = u\right) \cdot \mathbb{P}\left(X_{n-1} = u_{n-1}\right)\\ =\ &\mathbb{P}\left(X_n = a, X_{n-1} = u_{n-1}\right) \cdot \mathbb{P}\left(X_{[0..n)} = u\right)\end{aligned}$$

a podmínku pro přechodovou matici jako

$$\mathbb{P}\left(X_n = b, X_{n-1} = a\right) = m_{a,b} \cdot \mathbb{P}\left(X_{n-1} = a\right).$$

Opakovaným použitím tohoto vztahu dostáváme

Lemma 4.12.

$$\mathbb{P}\left(X_{[n..n+k)} = u_0 u_1 \cdots u_{k-1}\right) = P_{X_n}(u_0) \cdot \prod_{i=1}^{k-1} m_{u_{i-1},u_i}.$$

Homogenní markovský proces je výhodné chápat jako náhodnou procházku po grafu, jehož vrcholy reprezentují prvky z A a kde hrany jsou ohodnoceny podmíněnými pravděpodobnostmi $m_{a,b}$. Druhým klíčovým parametrem je **počáteční rozdělení** P_{X_0}, které určuje pravděpodobnosti, s jakými začínáme v konkrétním vrcholu grafu. Máme tedy randomizovaný start. Právě na této randomizaci záleží, zda bude celý proces stacionární. Pokud například přiřadíme jednomu vrcholu a_0 pravděpodobnost startu 1, dostaneme deterministický start, a ten (až na triviální výjimky) nepovede ke stacionárnímu procesu. Je snadné ověřit, že homogenní markovský proces je stacionární právě tehdy, když počáteční rozdělení $P = P_{X_0} \in \Delta_A$, splňuje podmínku

$$\sum_{a \in A} P(a) \cdot m_{a,b} = P(b)$$

pro všechna b, tedy $P \cdot M = P$, a stochastický vektor P je tedy vlastním vektorem stochastické matice $M = \left(m_{a,b}\right)_{a,b \in A}$. Takovému vektoru říkáme **stacionární (počáteční) rozdělení**. Pro stacionární homogenní markovský proces lze jednodušše určit entropii procesu.

Lemma 4.13. *Pro stacionární homogenní markovský proces $X = \left(X_n\right)_{n \in \mathbb{N}}$ a každé $i \in \mathbb{N}$ platí*

$$\mathcal{H}(X) = \mathcal{H}\left(X_{i+1} \mid X_i\right) = \mathcal{H}\left(X_1 \mid X_0\right), \quad \mathcal{H}\left(X_{[0..n)}\right) = \mathcal{H}\left(X_0\right) + (n-1) \cdot \mathcal{H}(X).$$

Důkaz. Z řetězového pravidla dostáváme, že

$$\mathcal{H}\left(X_{[0..n)}\right) = \mathcal{H}\left(X_0\right) + \sum_{i=1}^{n-1} \mathcal{H}\left(X_i \mid X_{[0..i)}\right),$$

kde markovská vlastnost $X_{[0..i-1)} \to X_{i-1} \to X_i$ a stacionarita dává pro členy sumy vše potřebné:

$$\begin{aligned}\mathcal{H}\left(X_i \mid X_{[0..i)}\right) &= \mathcal{H}\left(X_i \mid X_{i-1}\right) = \mathcal{H}\left(P_{X_i,X_{i-1}}\right) - \mathcal{H}\left(P_{X_{i-1}}\right) \\ &= \mathcal{H}\left(P_{X_1,X_0}\right) - \mathcal{H}\left(P_{X_0}\right) = \mathcal{H}\left(X_1|X_0\right).\end{aligned}$$

Protože je proces stacionární, první tvrzení plyne z Tvrzení 4.7(3). □

Na závěr se sluší poznamenat, že entropie existuje pro všechny homogenní markovské procesy, i ty nestacionární. Entropie takového procesu je pak rovna entropii stacionárního markovského procesu se stejnou přechodovou maticí. Takových ale může být více, a my se zde touto obecnější teorií nebudeme zabývat.

Markovské procesy - AEP

Také pro homogenní markovské procesy lze zformulovat obdobu zákona velkých čísel a AEP. My se zde zaměříme na jednoduchý případ, pro který již máme definovanou entropii, tedy na stacionární případ. Označme tedy základní rozdělení P_{X_i} jako P.

K tomu budeme navíc požadovat, aby byl tento markovský proces souvislý, neboli aby v přechodovém grafu existovala „nenulová“ cesta mezi každými dvěma pravděpodobnostně možnými stavy. Řekneme tedy, že markovský proces je **souvislý**, pokud je stacionární a homogenní a pokud pro každé dva stavy $a, b \in A$, pro které jsou $P(a)$ a $P(b)$ nenulové, existuje slovo $u = aub \in A^+$ takové, že $\mathbb{P}\left(X_{[0..n)} = u\right) > 0$ (kde n je délka aub).

Pro souvislé markovské procesy platí zákon velkých čísel v podobě analogické Větě 4.9. My ale pro jednoduchost vyslovíme zákon velkých čísel pro souvislé markovské procesy přímo ve formě, která sleduje výskyt digramů, tedy dvojic po sobě jdoucích symbolů. Ty odpovídají hranám v přechodovém grafu. Ze stacionarity a homogenity plyne, že příslušné rozdělení $P_{X_i,X_{i+1}}$ je stejné pro všechna i, označme ho P_2.

Tvrzení 4.14. *Je-li $X = \left(X_n\right)_{n\in\mathbb{N}}$ souvislý markovský proces, pak pro každou funkci $f : A \times A \to \mathbb{R}$ platí*

$$\frac{1}{n}\sum_{i=0}^{n-1} f\left(X_i, X_{i+1}\right) \to \mathbb{E}_{P_2}(f) \text{ s.j.}$$

Nyní můžeme pro souvislé markovské procesy dokázat silné AEP.

Tvrzení 4.15. *Je-li $X = \left(X_n\right)_{n\in\mathbb{N}}$ souvislý markovský proces, pak má silně asymptoticky rovnoměrné rozdělení.*

Důkaz. Nechť je $X = \left(X_n\right)_{n\in\mathbb{N}}$ souvislý markovský proces. Definujme reálnou funkci f na $s\left(P_2\right)$ předpisem

$$f(a,b) = -\log m_{a,b},$$

kde $m_{a,b}$ je přechodová pravděpodobnost daná homogenitou procesu, tedy $m_{a,b} = \mathbb{P}\left(X_{i+1} = b \mid X_i = a\right)$, pro každé $i \in \mathbb{N}$. Na nosiči (X_i, X_{i+1}) tedy platí $f(X_i, X_{i+1}) = \mathfrak{I}_{X_{i+1}|X_i}$, a tedy $\mathbb{E}_{P_2}(f) = \mathcal{H}\left(X_{i+1} \mid X_i\right)$, což je rovno $\mathcal{H}(X)$.

Pro $\omega \in s\left(X_{[0..n)}\right)$ označme $u = X_{[0..n)}(\omega) \in A^n$. Pak z Lemmatu 4.12 plyne

$$\begin{aligned}\mathfrak{I}_{X_{[0..n)}}(\omega) &= -\log \mathbb{P}\left(X_{[0..n)} = u\right) = -\log P_{X_0}(u_0) + \sum_{i=1}^{n-1} -\log m_{u_{i-1},u_i} \\ &= \mathfrak{I}_{X_0}(\omega) + \sum_{i=0}^{n-2} f\left(X_i(\omega), X_{i+1}(\omega)\right).\end{aligned}$$

Podělením přes n,

$$\frac{1}{n}\mathfrak{I}_{X_{[0..n)}}(\omega) = \frac{1}{n}\mathfrak{I}_{X_0}(\omega) + \frac{n-1}{n}\left(\frac{1}{n-1}\sum_{i=0}^{n-2} f\left(X_i(\omega), X_{i+1}(\omega)\right)\right).$$

Dle zákona velkých čísel pro markovský proces jde pravá strana skoro jistě k $\mathbb{E}_{P_2}(f) = \mathcal{H}(X)$, což jsme chtěli ukázat.

□

Přechodový graf a nestacionární procesy (nepovinné)

Nejčastější způsob, jak poznat souvislost markovského procesu je zvážení vlastností **přechodového grafu**, který je určen přechodovou maticí M. Upřesněme, že se jedná o ohodnocený graf, kde vrcholy jsou stavy markovského procesu, tedy množina A a ohodnocení hrany (a, b) je rovno $M_{a,b}$, přičemž hrany s nulovou vahou neuvažujeme. Potom platí jednoduché pozorování.

Tvrzení 4.16. *Nechť $X = \left(X_n\right)_{n\in\mathbb{N}}$ je stacionární homogenní markovský proces s přechodovou maticí M. Pokud je přechodový graf odvozený z matice M silně souvislý, pak je i proces souvislý.*

Získali jsme tedy grafové kritérium pro souvislost procesu. Toto kritérium je zároveň možno použít pro formulaci AEP i pro nestacionární procesy.

Věta 4.17. *Nechť $X = \left(X_n\right)_{n\in\mathbb{N}}$ je homogenní markovský proces s přechodovou maticí M. Pokud je přechodový graf odvozený z matice M silně souvislý, pak proces má entropii i silně asymptoticky rovnoměrné rozdělení.*

Důkaz již vyžaduje hlubší pochopení markovských procesů, a proto ho zde neuvádíme. Podobně bez důkazu uvádíme způsob, jak entropii daného procesu získat.

Tvrzení 4.18. *Nechť $X = \left(X_n\right)_{n\in\mathbb{N}}$ je homogenní markovský proces s přechodovou maticí $\boldsymbol{M}$. Nechť je přechodový graf odvozený z matice $\boldsymbol{M}$ silně souvislý. Pak pro matici přechodu $\boldsymbol{M}$ existuje jediné stacionární počáteční rozdělení $p = (p_a)_{a\in A}$. Entropie procesu X je pak rovna entropii stacionárního homogenního markovského procesu Y s počátečním rozdělením $P_{Y_0} = p$ a maticí přechodu $\boldsymbol{M}$. Platí tedy*

$$H(X) = H(Y) = H(Y_1|Y_0).$$

5 Komprese dat

Kódy

Buď M nějaká množina zpráv, které chceme přenášet nějakým kanálem, případně množina dat, která chceme někde ukládat. K tomu musíme M nejprve zakódovat do symbolů, které jsou pro přenos nebo zápis vhodné. Je-li M množina zpráv, zobrazení $f : M \to B^*$, kde B^* je množina posloupností prvků B, nazýváme **kód** a obrazy f nazýváme **kódová slova**. Přirozeným požadavkem je, aby byl kód **prostý**, kódová slova totiž chceme umět dekódovat zpět na původní zprávy.

V této kapitole budeme zkoumat případ, kdy je množina zpráv konečná. Základní definice kódu se ale vztahuje i na spočetně nekonečné množiny zpráv. Důležitý a přirozený případ spočetné množiny zpráv jsou libovolně dlouhé posloupnosti symbolů z nějaké množiny A (např. všechny konečné binární posloupnosti). Takové posloupnosti obvykle nazýváme **slova** a množinu A **abeceda**. Množina slov délky n nad abecedou A je A^n, pro množinu všech slov jsme již výše zavedli značení A^*. Pokud vyloučíme prázdné slovo, které značíme ε, značíme výslednou množinu A^+. Důležité pro nás bude neostré částečné uspořádání slov pomocí **relace prefix**. Pokud je tedy u prefixem v, budeme psát $u \leq v$.

Nejpřirozenější konstrukce kódu na nekonečné množině zpráv $M = A^*$ je **blokové rozšíření** nějakého kódu $f : A \to B^*$ na abecedě A na zobrazení $f^* : A^* \to B^*$ definované předpisem

$$f^*(a_1 a_2 \dots a_n) := f(a_1) f(a_2) \dots f(a_n).$$

Kód na A^*, který je takto blokovým rozšířením svých hodnot na abecedě A (či přesněji na slovech délky jedna, která ovšem obvykle s písmeny ztotožňujeme), nazýváme **blokovým kódem**. Obrazem prázdného slova je prázdné slovo, což souhlasí s přirozeným chápáním prázdné konkatenace. Kód f na A nazýváme **jednoznačně dekódovatelným**, pokud je jeho blokové rozšíření prosté na A^*. Samozřejmě, ne každý kód na A^* je blokový, ve většině případů dokonce blokové rozšíření nemůže dosahovat optimálních výsledků a je nutné použít jiné postupy, které budeme zkoumat v Kapitole 6.

Kód f se nazývá **prefixový**, pokud obraz žádné zprávy není prefixem obrazu jiné zprávy. Důležitou výhodou prefixového kódu je to, že i bez nějakého speciálního zakončovacího symbolu poznáme konec zprávy. Speciálně je prefixový každý kód **uniformní**, tedy takový, který má všechna kódová slova stejné dlouhá. Pokud kód prefixový není, existuje situace, kdy jsme dostali kódové slovo, ale nevíme, jestli přenos nebude pokračovat na jiné kódové slovo.

Lemma 5.1. *Jednoznačně dekódovatelný kód nepoužívá prázdné slovo. Je to tedy zobrazení $A \to B^+$.*

Každý prefixový kód $f : A \to B^*$ je prostý. Je také jednoznačně dekódovatelný, s výjimkou jednoprvkové množiny $A = \{a\}$ a $f(a) = \varepsilon$.

Důkaz. Pokud $f(a) = \varepsilon$, pak $f^*(a) = f^*(aa) = \varepsilon$ a f není jednoznačně dekódovatelný.

Pokud je kód f prefixový, pak pro $a \neq b$ zejména $f(a) \neq f(b)$, protože každé slovo je svým vlastním prefixem. Kód f je tedy prostý.

Nechť je kód f prefixový a $f^*(u) = f^*(v)$ pro $u \neq v$. Můžeme předpokládat, že u a v jsou nejkratší možná, tedy zejména nezačínají stejným písmenem. Jsou-li u i v neprázdná, pak jsou obrazy jejich prvních písmen prefixově srovnatelné, spor s prefixovostí. Je-li jedno ze slov prázdné a druhé neprázdné, pak pro všechna písmena a obsažená v tom neprázdném platí $f(a) = \varepsilon$. Protože ε je prefixem každého slova, musí díky prefixovosti f platit $A = \{a\}$. □

Příklad 5.2. Kód $f_1 : \{a, b, c\} \to \{0, 1\}^*$, definovaný jako

$$a \mapsto 0 \qquad b \mapsto 01 \qquad c \mapsto 10,$$

je prostý, ale není jednoznačně dekódovatelný, protože $f_1^*(ac) = f_1^*(ba) = 010$. Kódové slovo 010, což je kódové slovo pro kód f_1^*, nelze jednoznačně dekódovat. Kód f_1 tedy také není prefixový. Jak je vidět i z rovnosti $f_1^*(ac) = f_1^*(ba)$, obrazy $f_1(a)$ a $f_1(b)$ jsou prefixově srovnatelné.

Kód $f_2 : \{a, b, c\} \to \{0, 1\}^*$, definovaný jako

$$a \mapsto 0 \qquad b \mapsto 01 \qquad c \mapsto 11,$$

také není prefixový, ale jak f_2, tak f_2^* jsou prosté, jak lze snadno ověřit pokusem o konstrukci kódového slova, které by nemělo jednoznačný vzor. Dekódování f_2^* je přesto velmi nepraktické, protože všechna kódová slova ac^n a bc^k jsou prefixově srovnatelná. Musíme tedy čekat na konec celého přenosu, abychom zjistili první písmeno zprávy.

> **Poznámka:** Na slovech je přirozeně definována (asociativní) operace zřetězení, která slovům u a v přiřazuje slovo uv. S touto operací tvoří A^+ volnou pologrupu (resp. A^* volný monoid) s bází tvořenou slovy délky jedna. Blokové rozšíření f^* lze pak stručně definovat jako homomorfismus monoidů $A^* \to B^*$.
>
> V literatuře se často „kódem“ rozumí přímo množina kódových slov, navíc obvykle pouze pokud je f^* prosté, tedy pokud je f jednoznačně dekódovatelné. Bez zmínky o zobrazení, pouze v termínech abecedy B, to lze vyjádřit slovy, že kódová slova nesplňují žádnou netriviální rovnost.

Existence prefixových kódů

Základním požadavkem na kódy je co nejkratší výstup. Pokud chceme kód blokově rozšiřovat, je navíc nutné, aby byl jednoznačně dekódovatelný. Mezi jednoznačně dekódovatelnými kódy navíc preferujeme z výše uvedených důvodů prefixové kódy. Prozkoumejme tedy nejprve kombinatorické podmínky pro existenci prefixových

kódů na konečné množině zpráv, kterou budeme v tomto oddílu značit A, protože nás bude zajímat blokové rozšíření na $M = A^*$.

Prefixové kódy je možné snadno znázornit pravidelným orientovaným stromem stupně $D := |B|$ s hranami popsanými písmeny, jehož kořenem je prázdné slovo, každý vrchol reprezentuje slovo popisující cestu z kořene a kódová slova jsou reprezentována listy. Je-li $f : A \to B^*$ prefixový kód s nejdelším kódovým slovem délky L, má příslušný strom hloubku L. Vrchol odpovídající kódovému slovu délky ℓ vzhledem k prefixovosti „blokuje“ všechny své následníky, zejména všechna svá prodloužení délky ℓ, kterých je $D^{L-\ell}$. Z toho prostým počítáním listů a po následném vydělení D^L plyne, že pro prefixový kód f musí platit

$$\sum_{a \in M} D^{-|f(a)|} \leq 1\,,$$

což je tzv. **Kraftova nerovnost**, která ovšem podle následující věty platí i pro jednoznačně dekódovatelné kódy.

Tvrzení 5.3 (McMillanova nerovnost)**.** *Nechť je $f : A \to B^*$ jednoznačně dekódovatelný kód. Pak platí Kraftova nerovnost.*

Důkaz. Pro každé $u \in B^*$ označme $\rho(u) = D^{-|u|}$ a rozšiřme toto značení aditivně na množiny:

$$\rho(S) = \sum_{u \in S} \rho(u).$$

Zobrazení ρ, které vystupuje v Kraftově nerovnosti, vyjadřuje jakousi „hustotu“ množiny S. Zejména pro slova stejné délky k platí

$$\rho\left(S \cap B^k\right) \leq 1,$$

což je podíl množiny slov z S na množině všech slov z B^k. Je-li délka slov z S nejvýše L, platí tedy také

$$\rho\left(S\right) \leq L. \qquad (\diamondsuit)$$

Pro ρ navíc platí $\rho(uv) = \rho(u)\rho(v)$, tedy zejména pro každé $a_1 \cdots a_k \in A^k$ také

$$\rho\left(f^*(a_1 \cdots a_k)\right) = \prod_{i=1}^{k} \rho\left(f(a_i)\right). \qquad (\spadesuit)$$

Věta, kterou dokazujeme, tvrdí, že $\rho\left(f(A)\right) \leq 1$. Postupujme sporem a předpokládejme $\rho(f(A)) > 1$. Označme ℓ délku nejdelšího slova z $f(A)$. Buď n dostatečně velké číslo, aby platilo $(\rho(f(A)))^n > n\ell$, a uvažujme množinu $f^*(A^n)$ všech kódových slov, která vzniknou jako obrazy zpráv délky n. Délka těchto kódových slov je nejvýše $n\ell$. Podle ($\diamondsuit$) platí

$$n\ell \geq \rho\left(f^*(A^n)\right) = \sum_{u \in A^n} \rho\left(f^*(u)\right) = (\rho(f(A)))^n > n\ell,$$

kde první rovnost je důsledkem předpokladu jednoznačné dekódovatelnosti: na levé straně rovnosti totiž sčítáme hustoty přes množinu obrazů, na pravé straně přes množinu vzorů; pokud by f^* nebylo na A^n prosté, platila by ostrá nerovnost $<$. Druhá rovnost plyne z (♠) po dosazení definice $\rho(f(A))$ a roznásobení sum.

Dostali jsme spor. □

Příklad 5.4. Uvažujme kódy f_1 a f_2 z Příkladu 5.2. U obou dostáváme Kraftovu sumu rovnu jedné. Je-li tedy Kraftova nerovnost splněna, o tom, zda je kód jednoznačně dekódovatelný nám to nic neříká.

Uvažme nyní kód f_4, definovaný na $A = \{a, b, c, d\}$ jako

$$a \mapsto 0 \qquad b \mapsto 01 \qquad c \mapsto 10 \qquad d \mapsto 11.$$

Kraftova suma je nyní $\rho\left(f_4(A)\right) = \frac{5}{4}$, a kód tedy podle McMillanovy-Kraftovy nerovnosti nemůže být jednoznačně dekódovatelný. Ilustrujme na tomto příkladu myšlenku důkazu Tvrzení 5.3. Pro $n = 16$ platí

$$\left(\frac{5}{4}\right)^{16} = \sum_{u \in A^{16}} \rho\left(f^*(u)\right) > 32.$$

Protože slova z $f^*\left(A^{16}\right)$ mají délku nejvýše 32, existuje alespoň jedna délka $1 \leq d \leq 32$, pro kterou je

$$\sum_{|f(w)|=d} \rho\left(f^*(w)\right) = \sum_{|f(w)|=d} \frac{1}{2^d} > 1\,.$$

Existuje tedy více než 2^d slov w která f_4^* zobrazuje na binární slova délky d. Zobrazení f_4^* tedy není prosté.

Jako u většiny početních argumentů, je i zde odhad velmi hrubý a velkorysý. Všimněme si například, že výpočet nebere v úvahu fakt, že neexistují žádné obrazy délky méně než 16. Bližším výpočtem lze ověřit, že obrazů délky 26 je více než sedmkrát víc, než je binárních slov této délky. Lze též snadno ověřit, že existuje 270 obrazů délky osm, takže již na slovech této délky lze uplatnit jemnější početní argument, podle nějž jakýkoli binární kód s délkami $(1, 2, 2, 2)$ má nejednoznačné dekódovatelné slovo délky osm. Ve skutečnosti však z kombinatorické analýzy snadno plyne, že kolize nastává nejpozději na slově délky tři (obrazu slova délky dva).

Přestože je argument takto velkorysý, samotná mez je přesná, jak uvidíme v následujícím tvrzení.

Pro dané slovo u a přirozené číslo $n \geq |u|$, označíme $[u]_n^B$ množinu všech slov délky n nad abecedou B, které jsou prodloužením slova u, t.j. slovo u je jejich prefixem (index pro abecedu budeme vynechávat, pokud bude jasná z kontextu). Tato slova odpovídají listům ve stromě hloubky n, které u „blokuje“. Mohutnost $[u]_n^B$ je $D^{n-|u|}$, kde $D = |B|$.

Tvrzení 5.5 (Existence prefixového kódu)**.** *Pokud soubor přirozených čísel ℓ_a, $a \in A$ splňuje Kraftovu nerovnost, tedy*

$$\sum_{a \in A} D^{-\ell_a} \leq 1,$$

pro $D = |B|$, pak existuje prefixový kód $f : A \to B^*$ splňující $|f(a)| = \ell_a$, $a \in A$.

Důkaz. Tvrzení dokážeme konstrukcí požadovaného kódu, tedy volbou kódových slov. Nejprve seřadíme prvky A do posloupnosti a_1, a_2 až a_n tak, aby délky $\ell_i := \ell_{a_i}$ byly seřazeny vzestupně. Popíšeme nyní induktivní proces volby kódových slov. Nejprve pro prvek a_1 vyberme libovolné kódové slovo délky ℓ_1. Předpokládejme, že máme prefixový kód f dobře definovaný na prvcích $a_1, a_2, .., a_k$, $k < n$, a chceme ho rozšířit na prvek a_{k+1} tak, aby byl stále prefixový. K tomu stačí zvolit slovo délky $\ell := \ell_{k+1}$, které není prodloužením žádného z dosud zvolených obrazů f, ani není žádnému z nich rovno. Pro „zakázané“ množiny slov $f(a_i)$, $i \leq k$ délky ℓ platí

$$\left| \bigcup_{i=1}^{k} \left[f(a_i) \right]_{\ell}^{B} \right| = \sum_{i=1}^{k} \left| \left[f(a_i) \right]_{\ell}^{B} \right| = \sum_{i=1}^{k} D^{\ell - \ell_i} < \sum_{a \in A} D^{\ell - \ell_a} = D^{\ell} \sum_{a \in A} D^{-\ell_a} \leq D^{\ell}.$$

Z ostré nerovnosti plyne, že hledané slovo $f(a_{k+1}) \in B^{\ell}$, které nenáleží žádné z množin $\left[f(a_i) \right]_{\ell}^{B}$, $i \leq k$, existuje. Toto slovo zároveň není prefixem žádného předchozího kódového slova, neboť z nich má maximální délku. (Všimněte si, že konstrukce je korektní i pro $k = 0$, kdy je množina na levé straně rovnosti prázdná. V tomto případě tedy volíme slovo a_1 délky ℓ_1 libovolně.) □

Zdůrazněme, že důkaz existence kódu v předchozím tvrzení je konstruktivní. Deterministická konstrukce může volit mezi kandidáty nejmenší slovo v nějakém předem daném uspořádání slov (například lexikografickém).

Tvrzení 5.3 a Tvrzení 5.5 ukazují, že silnějším požadavkem na prefixovost oproti požadavku na pouhou jednoznačnou dekódovatelnost nic neztrácíme:

Důsledek 5.6. *Pro každý jednoznačně dekódovatelný kód existuje prefixový kód se stejnými délkami kódových slov.*

Střední délka kódu

V souvislostí s kompresí nás zajímají kódy s co nejkratšími obrazy, přičemž rozhodujícím měřítkem bude *průměrná* délka zprávy, což poukazuje na to, že zprávy jsou hodnotami nějaké náhodné veličiny $X : \Omega \to M$, resp. že na množině M je definované nějaké pravděpodobnostní rozdělení. V případě konečné množiny M nás tedy bude zajímat **střední délka kódu**

$$\mathbb{E}\left(|f(X)|\right) = \sum_{a \in M} P_X(a) \cdot |f(a)|,$$

kterou se budeme snažit minimalizovat a kterou označíme $[\![f]\!]_X$.

Definice 5.7. *Řekneme, že kód $f : M \to B^*$ je pro danou náhodnou veličinu $X : \Omega \to M$ **optimální** pokud*

- *je prefixový;*
- *má mezi prefixovými kódy nejmenší možnou střední délku. Tedy pro každý prefixový kód $f' : M \to B^*$ je $[\![f]\!]_X \leq [\![f']\!]_X$.*

Pro každé X s hodnotami v konečné množině existuje (alespoň jeden) optimální kód.

Lemma 5.8. *Mějme náhodnou veličinu X s hodnotami v konečné množině M. Pak existuje optimální kód $f : M \to B^*$ pro X.*

Důkaz. Jistě pro X existuje alespoň jeden prefixový kód g, stačí např. zvolit kód uniformní s délkou kódových slov $n = \lceil \log_{|B|} M \rceil$. Tím získáváme horní odhad pro střední délku optimálního kódu f. Musí platit

$$[\![f]\!]_X = \sum_{a \in M} |f(a)| P_X(a) \leq n = [\![g]\!]_X.$$

Z uvedeného vzorce pro střední délku kódu dále plyne horní odhad délky $|f(a)|$ pro každé $a \in s\left(P_X\right)$, konkrétně $f(a) \leq n/P_X(a)$. Pro zprávy z nosiče tedy vybíráme z konečné množiny možných délek kódových slov, přičemž tyto délky vždy již určují průměrnou délku kódu. Můžeme tedy vybrat ty délky, které poskytují nejmenší střední délku s podmínkou, že splňují Kraftovu nerovnost, která zaručuje existenci prefixového kódu daných délek.

Pokud existují zprávy mimo nosič, musíme i jim přiřadit nějaká kódová slova. Ačkoli jejich délka na střední délku kódu nemá vliv, zprávy mimo nosič tím vynucují, že v předchozím kroku musíme z množiny kandidátů vyloučit ty, kteří splňují Kraftovu *rovnost*.

Pokud délky slov na nosiči splňují ostrou Kraftovu nerovnost, je naopak vždy možné je doplnit délkami pro slova mimo nosič tak, aby Kraftova nerovnost stále platila. □

Příklad 5.9. V předchozím důkazu se ukázalo, že i zprávy mimo nosič mohou ovlivnit střední délku kódu. Pokud máme například zakódovat do binární abecedy zprávy $\{a, b, c\}$, přičemž pravděpodobnost zprávy c je nulová, nemůžeme zvolit výhodné $a \mapsto 0$, $b \mapsto 1$, protože pro c by nezbylo žádné kódové slovo, které by neporušovalo prefixovost. Mohli bychom navrhnout, aby se zpráva c nekódovala vůbec. Tím ale měníme zadání pro konstrukci kódu. Je tedy vhodnější rozlišit situaci, kdy se kód konstruuje pro všechny zprávy, a situaci, kdy výslovně požadujeme kód jen pro nosič.

Příklad 5.10. Všimněme si také patologické situace, kdy množina zpráv je jednoprvková. V takové situaci je veličina X konstantní a její entropie je nulová. Posílat

takovou zprávu tedy nedává dobrý smysl. Nicméně předchozí lemma platí i pro tento případ, optimální kód přiřazuje jediné zprávě prázdné slovo. Střední délka takového kódu je nula, což jistě zaručuje druhou podmínku optimality. Současně v případě jediného kódového slova platí, že kód je triviálně prefixový. Všimněme si ale, že není jednoznačně dekódovatelný. (To je jediný případ, kdy prefixový kód není jednoznačně dekódovatelný.)

Následující lemma vyslovuje zřejmé pravidlo, že optimální kód musí přiřazovat pravděpodobnějším slovům kratší kódová slova.

Lemma 5.11. *Je-li f optimální kód pro X, pak pro žádné $a, b \in M$ neplatí současně $P_X(a) < P_X(b)$ a $|f(a)| < |f(b)|$.*

Důkaz. Pokud by pro nějaký pár a, b nastala zakázaná situace, stačilo by vyměnit kódová slova, čímž by se snížila střední délka kódu o

$$\big(P_X(b) - P_X(a)\big)\,(|f(b)| - |f(a)|)\,.$$

□

Jako dosud budeme písmenem D označovat mohutnost výstupní abecedy. Tato abeceda bude nadále konečná a netriviální, tedy **minimálně dvouprvková** ($D \geq 2$). Entropii a divergenci pak bude vhodné měřit „D"-itech, tedy zavedeme

$$\mathcal{H}_D(X) = \sum_{a \in s(P_X)} P_X(a)\left(-\log_D P_X(a)\right), \qquad \mathcal{D}_D(P \parallel Q) = \sum_{a \in s(P)} P(a) \log_D \frac{P(a)}{Q(a)}.$$

Tedy $\mathcal{H}(X) = \mathcal{H}_2(X)$ a $\mathcal{D}(P \parallel Q) = \mathcal{D}_2(P \parallel Q)$. Jistě platí

$$\mathcal{H}_D(X) = \frac{\mathcal{H}(X)}{\log D}, \qquad \mathcal{D}_D(P \parallel Q) = \frac{\mathcal{D}(P \parallel Q)}{\log D}.$$

Následující tvrzení je první, které uvádí kompresi do již několikrát ohlašované souvislosti s entropií.

Tvrzení 5.12. *Nechť X je náhodná veličina s hodnotami v konečné množině M a f je prefixový kód $f : M \to B^*$. Potom je střední délka kódu zdola odhadnuta entropií:*

$$[\![f]\!]_X \geq \mathcal{H}_D(X)\,.$$

Rovnost nastává právě tehdy když $P_X(a) = D^{-|f(a)|}$, $a \in M$.

Důkaz. Označme $\ell_a = |f(a)|$, $c = \sum_{a \in s(P_X)} D^{-\ell_a}$ a definujme rozdělení $Q \in \Delta_A$ předpisem

$$Q(a) = \begin{cases} \frac{D^{-\ell_a}}{c}, & a \in s\left(P_X\right) \\ 0, & a \notin s\left(P_X\right). \end{cases}$$

Jelikož mají P_X a Q stejné nosiče, můžeme rozdíl obou stran nerovnosti zapsat následovně:

$$\begin{aligned}\llbracket f \rrbracket_X - \mathcal{H}_D(X) &= \sum_{a \in s(P_X)} P_X(a) \left(|f(a)| + \log_D P_X(a) \right) \\ &= \sum_{a \in s(P_X)} P_X(a) \left(-\log_D(cQ(a)) + \log_D P_X(a) \right) \\ &= -\log_D c + \mathcal{D}_D(P_X \parallel Q) \geq 0.\end{aligned}$$

Z Kraftovy nerovnosti dostáváme, že $c \leq 1$, tedy $-\log c$ je nezáporný. Dokázali jsme požadovanou nerovnost.

Pokud platí $P_X(a) = D^{-|f(a)|}$, pro všechna $a \in M$, pak lze pouhým dosazením ověřit rovnost entropie a střední délky kódu. Naopak, pokud je střední délka kódu rovna entropii, musí být $c = 1$ a $\mathcal{D}(P_X \parallel Q) = 0$. Z toho plyne $P_X(a) = Q(a) = D^{-|f(a)|}$ pro každé $a \in s\left(P_X\right)$. Zároveň dostáváme, že $s\left(P_X\right) = M$. Jinak by totiž součet $\sum_{a \in M} D^{-\ell_a}$ byl ostře větší než $c = 1$ a to by byl spor s Kraftovou nerovností. □

Všimněte si, že odhad z předchozího tvrzení platí i pro kódy jednoznačně dekódovatelné, protože ke každému takovému kódu existuje prefixový se stejnými délkami kódových slov, viz Důsledek 5.6.

Shannonův kód

Následující tvrzení poskytuje první horní odhad na střední délku kódu. Všimněte si, že zde nejprve ignorujeme zprávy s nulovou pravděpodobností.

Tvrzení 5.13. *Nechť X je náhodná veličina s hodnotami v konečné množině M. Potom existuje prefixový kód $f : s\left(P_X\right) \to B^*$, pro který platí $|f(a)| = \lceil -\log_D P_X(a) \rceil$, $a \in s\left(P_X\right)$, a*

$$\llbracket f \rrbracket_X < \mathcal{H}_D(X) + 1.$$

Důkaz. Položme $\ell_a = \lceil -\log_D P_X(a) \rceil$, pro $a \in s\left(P_X\right)$. Potom

$$\sum_{a \in s(P_X)} D^{-\ell_a} \leq \sum_{a \in s(P_X)} P_X(a) = 1.$$

Existuje tedy prefixový kód s danými délkami kódových slov. Pro takový kód platí

$$\llbracket f \rrbracket_X = \sum_{a \in s(P_X)} P_X(a) \cdot \ell_a < \sum_{a \in s(P_X)} P_X(a) \left(-\log_D P_X(a) + 1 \right) = \mathcal{H}_D(X) + 1.$$

□

Každý (prefixový) kód splňující podmínky z předchozího tvrzení se nazývá **Shannonův**. Přesněji o takovém kódu řekneme, že je to Shannonův kód pro náhodnou veličinu X (pokud není výstupní abeceda B daná z kontextu, musíme specifikovat i ji).

Tvrzení 5.14. *Je-li $f : s\left(P_X\right) \to B^*$ Shannonův kód pro náhodnou veličinu X, jsou následující podmínky ekvivalentní:*

(1) $[\![f]\!]_X = \mathcal{H}_D(X)$

(2) $|f(a)| = -\log_D(P_X(a)) \in \mathbb{N}$, $a \in s\left(P_X\right)$,

(3) $-\log_D(P_X(a)) \in \mathbb{N}$, $a \in s\left(P_X\right)$,

(4) $\sum_{a \in s\left(P_X\right)} D^{-|f(a)|} = 1$.

Důkaz. Dle Tvrzení 5.12, jsou podmínky (1) a (2) ekvivalentní pro každý prefixový kód. Ekvivalence podmínky (3) s podmínkou (2) vyplývá přímo z definice Shannonova kódu.

Platí

$$\sum_{a \in s\left(P_X\right)} D^{-\lceil -\log_D P_X(a) \rceil} \leq \sum_{a \in s\left(P_X\right)} D^{\log_D P_X(a)} = \sum_{a \in s\left(P_X\right)} P_X(a) = 1.$$

Rovnost platí, právě když $\lceil -\log_D P_X(a) \rceil = -\log_D P_X(a)$ pro všechna $a \in s\left(P_X\right)$. Tím je dokázána ekvivalence (2) a (4). □

Tvrzení 5.15. *Nechť je X náhodná veličina s hodnotami v konečné množině M a nechť B je abeceda mohutnosti D. Potom existuje prefixový kód $f : s\left(P_X\right) \to B^*$ takový, že $[\![f]\!]_X = \mathcal{H}_D(X)$, právě když $-\log_D(P_X(a)) \in \mathbb{N}$, $a \in s\left(P_X\right)$. Takový kód pak bude mít nejmenší střední délku kódu, bude Shannonův a bude pro něj platit $|f(a)| = -\log_D(P_X(a))$, $a \in s\left(P_X\right)$.*

Důkaz. Věta je důsledkem předchozích vět. □

Shannonův kód je vytvořen pro konkrétní veličinu. Následující věta ukazuje, že pro jinou veličinu může být, a povětšinou je, krajně neoptimální.

Tvrzení 5.16. *Nechť X, Y je náhodná veličina s hodnotami v konečné množině M a $f : s\left(P_Y\right) \to B^*$ je Shannonův kód pro veličinu Y. Pokud je $s\left(P_X\right) \subseteq s\left(P_Y\right)$, platí*

$$\mathcal{H}_D(X) + \mathcal{D}_D(P_X \parallel P_Y) \leq [\![f]\!]_X.$$

Důkaz. Rozdíl dvou členů z opačných stran nerovnosti zapíšeme pomocí divergence:

$$\begin{aligned}\llbracket f \rrbracket_X - \mathcal{H}_D(X) &= \sum_{a \in s(P_X)} P_X(a) \left(|f(a)| + \log_D P_X(a) \right) \\ &\geq \sum_{a \in s(P_X)} P_X(a) \left(-\log_D P_Y(a) + \log_D P_X(a) \right) \\ &= \mathcal{D}_D(P_X \parallel P_Y).\end{aligned}$$

□

Nyní ukážeme, že Shannonův kód lze rozšířit i mimo nosič náhodné veličiny, aniž bychom ztratili omezení na střední délku kódu dokázané pro nosič.

Tvrzení 5.17. *Nechť X je náhodná veličina s hodnotami v konečné množině M. Potom existuje prefixový kód $f : M \to B^*$, pro který platí*

$$\llbracket f \rrbracket_X \leq \mathcal{H}_D(X) + 1.$$

Důkaz. Vezměme Shannonův kód $f : s\left(P_X\right) \to B^*$ a označme $\ell_a = |f(a)|$, $a \in s\left(P_X\right)$. Pro ten je nerovnost splněna. Pokud je $\sum_{a \in s(P_X)} D^{-\ell_a} < 1$, můžeme najít dostatečná velká ℓ_a, $a \in M \setminus s\left(P_X\right)$, taková, že $\sum_{a \in M} D^{-\ell_a}$ je stále menší než jedna. Dle Tvrzení 5.5 existuje prefixový kód $f' : M \to B^*$ takový, že $\ell_a = |f(a)|$ pro všechna $a \in M$. Jelikož se délky kódových slov pro f a f' shodují na $s\left(P_X\right)$, mají stejnou střední délku kódu a nerovnost platí i pro f'.

Pokud je $\sum_{a \in s(P_X)} D^{-\ell_a} = 1$, dostáváme z Tvrzení 5.14, že $\llbracket f \rrbracket_X = \mathcal{H}_D(X)$. Vezměme prvek $c \in s\left(P_X\right)$ (pokud chceme co nejmenší střední hodnotu, volme tak, aby $|f(c)|$ bylo maximální). Platí $P_X(c) = D^{-|f(c)|} \leq D^0 \leq 1$. Definujme $\ell'_c = \ell_c + 1$ a $\ell'_a = \ell_a$, pro $a \in s\left(P_X\right)$ různá od c. Potom platí $\sum_{a \in s(P_X)} D^{-\ell'_a} < 1$ a opět lze najít čísla ℓ'_a, $a \in M \setminus s\left(P_X\right)$ taková, že existuje prefixový kód f' na M s délkami slov ℓ'. Pro takový kód platí

$$\llbracket f' \rrbracket_X = \sum_{a \in s(P_X)} P_X(a) \cdot \ell'_a = \sum_{a \in s(P_X)} P_X(a) \cdot \ell_a + P_X(c) \cdot 1 \leq \mathcal{H}_D(X) + 1.$$

□

Kontrolní otázka

- Ověřte, že Q v důkazu Tvrzení 5.12 je opravdu rozdělení.

Huffmanův kód - nejkratší binární prefixový kód

Podle Tvrzení 5.15 je Shannonův kód optimální ve speciální situaci, kdy logaritmy všech pravděpodobností jsou celočíselné. V obecné situaci horní odhad $\mathcal{H}_D(X) + 1$ pro jeho délku optimalitu nezaručuje, jak ukazuje následující příklad.

Příklad 5.18. Rozdělení $P = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}\right)$ má entropii

$$\mathcal{H}(P) = \log(3) + \frac{1}{3} \approx 1.918.$$

Záporné logaritmy pravděpodobností jsou $-\log P = (1.58, 1.58, 2.58, 2.58)$, takže délky pro Shannonův kód jsou $d = (2, 2, 3, 3)$. Tomu odpovídá například prefixový kód $g = (00, 01, 100, 101)$ s délkou

$$[\![g]\!]_P = \frac{7}{3} \approx 2.333.$$

Minimální kód pro P je ale například $f = (00, 01, 10, 11)$ s délkou

$$[\![f]\!]_P = 2 < [\![g]\!]_P.$$

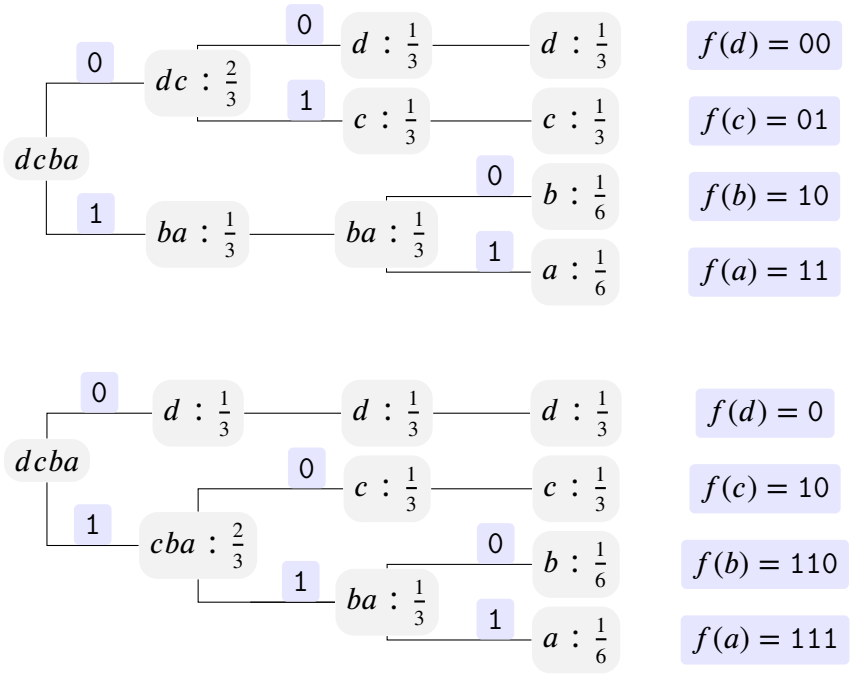
Nyní popíšeme **Huffmanův algoritmus**, který konstruuje optimální prefixové kódy pro *binární* výstupní abecedu, tedy pro případ kdy $D = 2$. Zvolme $B = \{0, 1\}$. Algoritmus pracuje rekurzivně následujícím způsobem. Sloučí dvě písmena s nejmenší pravděpodobností do jediného nového písmene a získá tak rozdělení na menší abecedě. Potom sestrojí minimální kód pro tuto menší abecedu a z kódu složeného písmene získá kód původních dvou písmen tak, že k němu přidá jeden rozlišovací bit. Viz pseudokód algoritmu níže.

Algorithm 1: `Huffman`

Data: P, M
Result: $f : M \to \{0, 1\}^*$
if $M = \{c\}$ **then**
  $g(c) = \varepsilon$;
else
  $a, b \leftarrow \arg\min_{x \neq y \in M} P(x) + P(y)$; `(* možná nejednoznačnost *)`
  $C \leftarrow M \setminus \{a, b\} \cup \{c\}$; `(*` $c \notin M$ `*)`
  $Q(x) \leftarrow \begin{cases} P(x), & x \in C,\ x \neq c \\ P(a) + P(b), & x = c \end{cases}$;
  $g \leftarrow$ `Huffman`(Q, C) ;
  $f(x) \leftarrow \begin{cases} g(x), & x \in C \setminus \{c\}, \\ g(c)0, & x = a, \\ g(c)1, & x = b \end{cases}$
end

Algoritmus není deterministický a zkonstruovaný **Huffmanův kód** tak není vždy určen jednoznačně, a to ani co do délek kódových slov. V případě, že někdy v průběhu Huffmanova algoritmu není dvojice písmen s nejmenšími pravděpodobnostmi

jen jedna, vedou všechny alternativy k optimálnímu kódu. K této situaci dochází i v Příkladu 5.18 a dva možné kódy s minimální střední délkou jsou sestrojeny na Obrázku 8.



Obrázek 8: Dva různé Huffmanovy kódy pro rozdělení $P(a) = P(b) = \frac{1}{6}$, $P(c) = P(d) = \frac{1}{3}$.

Nyní dokážeme správnost Huffmanova algoritmu.

Tvrzení 5.19. *Pro náhodnou veličinu X s hodnotami v konečné množině M vrací Huffmanův algoritmus optimální kód $f : M \to \{0,1\}^*$ pro X a v Kraftově nerovnosti platí rovnost, tedy*

$$\sum_{x \in M} 2^{-|f(x)|} = 1.$$

Důkaz. Postupujme indukcí podle velikosti M. Je-li $M = \{c\}$ jednoprvková, pak Huffmanův algoritmus vrací kód $c \mapsto \varepsilon$ a tvrzení platí, viz Příklad 5.10.

Nechť $|M| \geq 2$ a nechť a, b jsou písmena zvolená Huffmanovým algoritmem. Pro libovolné $x, y \in M$, $x \neq y$, tedy platí $P(a)+P(b) \leq P(x)+P(y)$. Buďte C, Q, c a g definovány jako v algoritmu. Tedy i f je definováno pomocí g jako v algoritmu. Pak $Q = P_{\alpha \circ X}$, kde $\alpha : M \to C$ přejmenovává a a b na nové písmeno c a jinde je identitou.

Z indukčního předpokladu plyne, že g je optimální a prefixový kód pro náhodnou veličinu $\alpha \circ X$ a platí pro něj Kraftova rovnost. Pak

$$\sum_{x \in M} 2^{-|f(x)|} = 2^{-|g(c0)|} + 2^{-|g(c1)|} + \sum_{x \in C \setminus \{c\}} 2^{-|g(x)|} = \sum_{x \in C} 2^{-|g(x)|} = 1.$$

Že je kód f prefixový, je zřejmé z prefixového stromu g a jeho rozšíření na prefixový strom f přidáním dvou listů. Uveďme ještě formálnější důkaz. Z definic vidíme, že $g(\alpha(z)) \leq f(z)$ pro všechna $z \in M$. Postupujme sporem a předpokládejme, že jsou $f(x)$ a $f(y)$, $x \neq y$, prefixově srovnatelné. Pak jsou i jejich prefixy $g(\alpha(x))$ a $g(\alpha(y))$ prefixově srovnatelné, což vzhledem k prefixovosti g znamená, že $\alpha(x) = \alpha(y)$, a tedy $\{x, y\} = \{a, b\}$. Protože $f(a) = g(c)0$ a $f(b) = g(c)1$ prefixově srovnatelné nejsou, dostáváme spor s předpokladem, že f není prefixový.

Zbývá ukázat, že f má mezi prefixovými kódy nejmenší možnou střední délku. Postupujme sporem. Nechť je h optimální kód pro X a platí $[\![h]\!] < [\![f]\!]$. Předpokládejme dále, že h má mezi optimálními kódy pro X nejmenší možný součet délek všech kódových slov. Mezi písmeny s minimální pravděpodobností zvolme $a' \in M$, pro které je $|h(a')|$ maximální. Z Lemmatu 5.11 plyne, že $|h(a')|$ je maximální přes všechna kódová slova (nejen ta jejichž vzory mají minimální pravděpodobnost). Předpokládáme, že je M alespoň dvouprvková, což pro prefixový kód znamená, že maximální délka kódového slova je minimálně jedna. Slovo $h(a')$ je tedy tvaru uc, kde $u \in B^*$ a $c \in B$. Z prefixovosti kódu plyne, že mezi kódovými slovy není žádné, které by bylo prefixem slova u. Z předpokládané minimality součtu délek kódových slov plyne, že pokud zkrátíme kódové slovo $h(a')$ na u, porušíme prefixovost kódu. To znamená, že u je prefixem nějakého dalšího kódového slova $h(b')$. Z maximality délky uc plyne, že $h(b') = uc'$ pro $c \neq c' \in B$. Definujme nyní h' záměnou písmen $a \leftrightarrow a'$ a $b \leftrightarrow b'$ (jsou-li různá), neboli $h'(a) = h(a')$, $h'(a') = h(a)$, $h'(b) = h(b')$, $h'(b') = h(b)$ a $h'(x) = h(x)$ jinak. Protože h' má stejnou množinu kódových slov jako h, je to prefixový kód. Předpokládejme, bez újmy na obecnosti, že $P_X(a) \leq P_X(b)$. Z minimality $P_X(a) + P_X(b)$ pak plyne $P_X(a) = P_X(a')$ a $P_X(b) \leq P_X(b')$. Rovněž platí $|h'(b')| \leq |h'(b)|$, protože $|h'(b)| = |h(b')|$ je maximální délka mezi všemi kódovými slovy. Střední hodnota se tedy výměnou písmen nezvětší a h' je stejně jako h optimální.

Definujme nyní $g' : C \to B^*$ předpisem $g'(c) = u$ a $g'(x) = h'(x)$, $x \neq c$. Z prefixovosti h' snadno odvodíme, že i g' je prefixový; používáme při tom fakt, že B je binární, a pokud je tedy u vlastním prefixem nějakého slova, je jeho prefixem i $u0$ nebo $u1$. Dále platí

$$\begin{aligned}
[\![f]\!]_X &= P_X(a) \cdot |g(c)0| + P_X(b) \cdot |g(c)1| + \sum_{x \in M \setminus \{a,b\}} P_X(x) \cdot |f(x)| \\
&= P_X(a) + P_X(b) + P_{\alpha \circ X}(c) \cdot |g(c)| + \sum_{x \in C \setminus \{c\}} P_{\alpha \cdot X}(x) \cdot |g(x)| \\
&= P(a) + P(b) + [\![g]\!]_{\alpha \circ X}
\end{aligned}$$

a podobně

$$\begin{aligned}
[\![h']\!]_X &= P_X(a) \cdot |u0| + P_X(b) \cdot |u1| + \sum_{x \in M \setminus \{a,b\}} P_X(x) \cdot \left|h'(x)\right| \\
&= P_X(a) + P_X(b) + P_{\alpha \circ X}(c) \cdot |g'(c)| + \sum_{x \in C \setminus \{c\}} P_{\alpha \cdot X}(x) \cdot \left|g'(x)\right| \\
&= P(a) + P(b) + [\![g']\!]_{\alpha \circ X}.
\end{aligned}$$

Z optimality g plyne $[\![g]\!]_{\alpha \circ X} \leq [\![g']\!]_{\alpha \circ X}$, tedy $[\![f]\!]_X \leq [\![h']\!]_X$ a f tedy splňuje druhou podmínku optimality. □

Všimněme si, že pokud pro rozdělení z Příkladu 5.18, kde $P(a) = P(b) = \frac{1}{6}$ a $P(c) = P(d) = \frac{1}{3}$, zvolíme optimální prefixový kód

$$a \mapsto \texttt{10}, \qquad b \mapsto \texttt{01}, \qquad c \mapsto \texttt{11}, \qquad d \mapsto \texttt{00},$$

pak neexistují žádná písmena a a b s vlastnostmi z předchozího lemmatu. Záměna $b \leftrightarrow b'$ v důkazu je nutná. To zejména znamená, že takový optimální kód nemůže být výstupem Huffmanova algoritmu.

Poznamenejme, že pro jednoprvkovou množinu zpráv, připouštějící pouze triviální rozdělení s nulovou entropií, vrací algoritmus jako jediné kódové slovo slovo prázdné, tedy kód s nulovou střední délkou. To je současně již výše zmíněný (patologický) prefixový kód.

Poznamenejme konečně, že Kraftova rovnost (která opět platí i pro jednoprvkovou množinu zpráv) zaručuje, že žádné kódové slovo nelze zkrátit, aniž by byla znemožněna jednoznačná dekódovatelnost. Huffmanův kód má tedy nejen optimální střední délku, ale také minimální vektor délek vzhledem k součinovému uspořádání, a to i mimo nosič. Jinak řečeno, pokud pro nějaký jiný prefixový (či jednoznačně dekódovatelný) kód f' platí $|f'(a)| \leq |f(a)|$, pro všechna $a \in M$, pak $|f'(a)| = |f(a)|$, pro všechna $a \in M$.

6 Komprese - asymptotické chování

V případě, kdy $M = A^*$, budeme jednotlivá písmena chápat jako náhodný proces ve smyslu Kapitoly 4. V předchozí kapitole jsme zavedli jeden jednoduchý postup, jak na nekonečné množině $M = A^*$ definovat kód, totiž blokovým rozšířením nějakého kódu ma A. V této kapitole prozkoumáme vlastnosti kódování procesů obecněji.

Asymptotickou průměrnou délku kódového slova na jeden vstupní symbol nazveme v případě procesů „kompresní poměr“ a definujeme ho takto:

Definice 6.1. *Pro náhodný proces X s hodnotami v abecedě A a pro kód $f : A^* \to B^*$ definujeme **dolní** a **horní kompresní poměr** f jako*

$$\underline{[\![f]\!]}_X = \liminf_{n\to\infty} \frac{[\![f]\!]_{X_{[0..n)}}}{n}.$$

$$\overline{[\![f]\!]}_X = \limsup_{n\to\infty} \frac{[\![f]\!]_{X_{[0..n)}}}{n}.$$

***Kompresní poměr** je limita (pokud existuje):*

$$[\![f]\!]_X = \lim_{n\to\infty} \frac{[\![f]\!]_{X_{[0..n)}}}{n}.$$

Nejjednodušším případem je opět i.i.d. proces, kde pro délku obrazu blokového rozšíření platí následující vztah.

Lemma 6.2. *Pro i.i.d. proces X s hodnotami v abecedě A a pro kód $f : A \to B^*$ platí*

$$[\![f^*]\!]_{X_{[0..n)}} = n \cdot [\![f]\!]_{X_0},$$
$$[\![f^*]\!]_X = [\![f]\!]_{X_0}.$$

Důkaz. Vzhledem k tomu, že v f^* řetězíme výstupy pro jednotlivá písmenka, platí

$$\left|f^*\left(X_{[0..n)}\right)\right| = \sum_{i=0}^{n-1} \left|f\left(X_i\right)\right|.$$

Jelikož jsou X_i a X_0 stejně rozdělené veličiny, jsou stejně rozdělené i veličiny $\left|f(X_i)\right|$ a $\left|f(X_0)\right|$ a mají tudíž stejnou střední hodnotu. Platí tedy $[\![f]\!]_{X_i} = [\![f]\!]_{X_0}$. Z rovnosti jednotlivých středních hodnot a z linearity dostaneme požadovanou rovnost pro střední hodnotu délky blokového kódu:

$$[\![f^*]\!]_{X_{[0..n)}} = \sum_{i=0}^{n-1} \mathbb{E}\left(\left|f\left(X_i\right)\right|\right) = n \cdot \left|f\left(X_0\right)\right| = n[\![f]\!]_{X_0}.$$

Přechod k limitě, tedy důkaz druhé rovnosti z lemmatu, je už triviální. □

Z předchozího lemmatu je vidět, proč blokové rozšíření nemusí být optimální volbou. Pokud totiž na A není možné dosáhnout entropie $\mathcal{H}_D\left(X_0\right)$, bude se nedokonalost kódu f násobit a pro dostatečně dlouhé vektory nebude kódování optimální (bude např. překročena mez pro Shannonovy kódy). Přirozenou myšlenkou, jak konstruovat optimální kódy pro A^*, je zkontruovat pro každý vektor $X_{[0..n)}$ vždy nový kód f_n a tyto parciální kódy sloučit. Problém ovšem je, že výsledný kód nebude prefixový a nemusí být ani prostý. Lze očekávat, že pro různé délky budou volena stejná kódová slova. Pro překonání této obtíže je důležitá následující myšlenka.

Lemma 6.3. *Předpokládejme, že $f_n : A^n \to B^*$, $n \geq 0$, je soubor prefixových kódů, a nechť $\alpha : \mathbb{N} \to B^*$ je prefixový kód. Potom je kód $g : A^* \to B^*$, definovaný předpisem*

$$g(u) = \alpha\left(|u|\right) f_{|u|}(u),$$

prefixový.

Důkaz. Buďte slova $g(u)$ a $g(v)$ prefixově srovnatelná. Pak i $\alpha(|u|)$ a $\alpha(|v|)$ jsou prefixově srovnatelná, a protože α je prefixový kód, mají u a v stejnou délku, označme ji k. Po zkrácení $g(u)$ a $g(v)$ o společný prefix $\alpha(k)$ vidíme, že i $f_k(u)$ a $f_k(v)$ jsou prefixově srovnatelné. Z toho plyne $u = v$, protože f_k je prefixový. Tedy g je prefixový. □

Eliasovy kódy

Nyní je třeba najít vhodný kód α. Naznačíme zde postup, jak generovat posloupnost kódů jejichž asymptotické vlastnosti se zlepšují. Jedná se o variaci na tzv. Eliasovy γ-kódy. Uveďme nejprve pomocné lemma.

Lemma 6.4. *Předpokládejme, že $f : M \to B^*$ je prostý kód a $\alpha : \mathbb{N} \to B^*$ je prefixový kód. Potom je kód $g : M \to B^*$, definovaný předpisem*

$$g(a) = \alpha\left(|f(a)|\right) f(a),$$

prefixový.

Důkaz. Podobně jako Lemma 6.3. □

Začneme s kódy γ_0 a β. Fixujme dva speciální znaky z abecedy B a označme je pro jednoduchost 0 a 1. Pak prefixový kód $\gamma_0 : \mathbb{N} \to B^*$ je definován předpisem

$$\gamma_0(n) = 1^n 0.$$

Zvolme nějaké uspořádání na množině B. Zobrazení $\beta : \mathbb{N} \to B^*$, definujeme tak, že nulu zobrazíme na prázdné slovo, prvních D kladných přirozených čísel zobrazíme bijektivně na množinu B při zvoleném uspořádání, dalších D^2 čísel se zobrazí na B^2 v lexikografickém uspořádání, a tak dále.

Zobrazení β je nyní bijekce mezi $\mathbb{N}$ a B^*. Pokud písmena B identifikujeme s jejich pořadovými čísly, tedy pokud uvažujeme $B = \{1, 2, \ldots, D\}$, dostaneme přirozený, byť málo používaný D-adický numerační systém, u kterého stejně jako u standardního D-árního zápisu platí, že $a_{k-1}a_{k-2}\cdots a_1a_0$ reprezentuje číslo

$$\sum_{i=0}^{k-1} a_i D^i\,,$$

jak lze snadno ověřit indukcí. Všimněme si, že zde používáme jiné označení písmen ciframi, než byla výše uvedená volba 0 a 1. To můžeme chápat jako připomínku rozdílné povahy kódů γ a kódu β.

Posloupnost slov kladné délky k seřazená lexikograficky začíná k jedničkami a končí k ciframi D. Tato slova tedy odpovídají číslům n splňujícím

$$D^{k-1} \le \sum_{i=0}^{k-1} D^i = \sum_{i=1}^{k-1} D^i + 1 \le n \le \sum_{i=1}^{k} D^i < D^{k+1} = D + (D-1)\sum_{i=1}^{k} D^i.$$

D-adický zápis je samozřejmě nejvýše tak dlouhý jako D-ární. Vidíme např., že číslo D^{k+1}, což je v D-árním zápisu první číslo s délkou zápisu $k+2$, má v D-adickém zápisu délku $k+1$, konkrétně k cifer $D-1$, následovaných jedním D (např. $8 = 112$).

Podle výše uvedené řady nerovností tedy pro kladné n platí $k-1 \le \log_D n < k+1$, neboli

$$\log_D n - 1 < |\beta(n)| \le \log_D n + 1,$$

kde druhá nerovnost je neostrá pouze pro $n = 1$.

Lemma 6.5. *Kódy $\gamma_1, \gamma_2 : \mathbb{N} \to B^*$, definované předpisem*

$$\gamma_1(n) := \gamma_0(|\beta(n)|)\beta(n), \qquad \gamma_2(n) := \gamma_1(|\beta(n)|)\beta(n),$$

jsou prefixové,

$$\gamma_0(0) = \gamma_1(0) = \gamma_2(0) = 0$$

a pro každé kladné n platí

$$\big|\gamma_1(n)\big| \le 2\log_D n + 3, \qquad \big|\gamma_2(n)\big| \le \log_D n + 2\log_D(\log_D n + 1) + 4.$$

Funkce $|\gamma_1|, |\gamma_2| : \mathbb{N} \to \mathbb{N}$ jsou neklesající.

Důkaz. Jelikož je γ_0 prefixový a β je prostý, je dle Lemmatu 6.4 γ_1 prefixový kód. Stejné argumenty dokazují, že je γ_2 prefixový.

Hodnoty na nule plynou z konvence $\beta(0) = \varepsilon$. Je také vidět, že hodnota v nule neporušuje prefixovost: nula je jediné kódové slovo začínající nulou.

Omezení délek kódů vyplývají z konstrukce a z omezení pro délku kódu β výše. □

Věta 6.6. *Buď X náhodný proces s výstupní abecedou A, který má entropií. Nechť $f : A^* \to B^*$ je prefixový kód (nemusí být blokový). Potom*

$$\mathcal{H}_D(X) \leq \underline{[\![f]\!]}_X .$$

Důkaz. Kód f zúžený na $M = A^n$ je opět prefixový. Z Tvrzení 5.12 plyne, že

$$[\![f]\!]_{X_{[0..n)}} \geq \mathcal{H}_D\left(X_{[0..n)}\right).$$

Vydělením obou stran rovnice číslem n a přechodem k liminf dostáváme tvrzení. □

Věta 6.7. *Buď X náhodný proces s výstupní abecedou A, který má entropii. Pak existuje prefixový kód $f : A^* \to B^*$, takový, že*

$$[\![f]\!]_X = \mathcal{H}_D(X) .$$

Důkaz. Aplikací Tvrzení 5.17 dostáváme, že pro každé n existuje prefixový kód f_n z A^n do B^* se střední délkou kódu menší než $\mathcal{H}\left(X_{[0..n)}\right) + 1$. Definujme

$$f(u) = \gamma_1(|u|) f_{|u|}(u), \qquad u \in A^*,$$

což je podle Lemmatu 6.3 prefixový kód, pro který platí

$$[\![f]\!]_{X_{[0..n)}} = |\gamma_1(n)| + [\![f_n]\!]_{X_{[0..n)}} \leq 2\log_D n + 4 + \mathcal{H}_D\left(X_{[0..n)}\right).$$

Poděleníím n a přechodem k limsup, dokončíme důkaz.

□

Věta 6.8. *Buď X náhodný proces s výstupní abecedou A. Nechť $f : A^* \to B^*$ je prostý kód. Potom existuje prefixový kód, který má dolní i horní kompresní poměr stejný jako f.*

Speciálně, pokud má f kompresní poměr, existuje prefixový kód se stejným kompresním poměrem jako f.

Důkaz. Definujme $g(u) = \gamma_1(|f(u)|) f(u)$. Dle Lemmatu 6.4 je takový kód prefixový.

Zvolme libovolné $\delta > 0$ a k němu k takové, že $\frac{|\gamma_1(m)|}{m} < \delta$, pro všechna $m \geq k$. Množina všech slov délky nejvýše k nad abecedou B je konečná a kód f je prostý, proto existuje $\ell > 0$ takové, že pro všechna $n \geq \ell$, $f(A^n)$ obsahuje jen slova délky minimálně $k + 1$.

Z toho plyne,

$$|f(u)| \leq |g(u)| = |\gamma_1(|f(u)|)| + |f(u)| \leq \delta|f(u)| + |f(u)|.$$

Neboli

$$[\![f]\!]_{X_{[0..n)}} \leq [\![g]\!]_{X_{[0..n)}} \leq (1+\delta) \cdot [\![f]\!]_{X_{[0..n)}}.$$

Dostáváme, že $\overline{[\![g]\!]}_X$ je mezi $\overline{[\![f]\!]}_X$ a $(1+\delta)\overline{[\![f]\!]}_X$. Stejné nerovnosti platí i pro dolní kompresní poměr. Jelikož bylo $\delta > 0$ libovolné, jsou horní kompresní poměry f a g stejné. Totéž platí pro dolní kompresní poměry, a tedy i pro kompresní poměr, pokud existuje. □

Stejně jako u konečné abecedy tedy požadavek prefixovosti neznamená žádnou nevýhodu oproti požadavku prostoty. Negativně to vyjadřuje následující tvrzení.

Důsledek 6.9. *Buď X náhodný proces s výstupní abecedou A, který má entropii. Nechť $f : A^* \to B^*$ je prostý kód. Potom $\mathcal{H}_D(X) \leq \underline{[\![f]\!]}_X$.*

Důkaz. Přímo z Věty 6.6 a Věty 6.8. □

6.1 Bodový kompresní poměr

V předchozí části jsme zkoumali limitu posloupnosti středních hodnot, tedy jak se limitně chová průměrný poměr délky výstupního slova vůči délce vstupu. V této kapitole zesílíme naši pozornost a budeme se zabývat tím, jak se chová daný poměr bodově.

Kokrétně definujeme bodový dolní a horní kompresní poměr takového kódu pro kód $f : A^* \to B^*$ a náhodný proces X v bodě $\omega \in \Omega$ předpisem

$$\underline{(\!|f|\!)}_X(\omega) = \liminf_{n\to\infty} \frac{\left|f\left(X_{[0..n)}(\omega)\right)\right|}{n}.$$

$$\overline{(\!|f|\!)}_X(\omega) = \limsup_{n\to\infty} \frac{\left|f\left(X_{[0..n)}(\omega)\right)\right|}{n}.$$

Bodový kompresní poměr je limita (pokud existuje):

$$(\!|f|\!)_X(\omega) = \lim_{n\to\infty} \frac{\left|f\left(X_{[0..n)}(\omega)\right)\right|}{n}.$$

Pokud zkoumáme střední hodnotu bodového kompresního poměru, mohli bychom očekávat, že se rovná kompresnímu poměru definovanému dříve. Tak tomu bohužel vždy není, neboť nelze vždy prohodit pořadí střední hodnoty a limitního přechodu, aniž by to mělo vliv na výslednou hodnotu. Vzhledem k nezápornosti zkoumaných funkcí lze alespoň říci, že střední hodnota dolního bodového poměru je menší než dolní kompresní poměr - jedná se o důsledek Fatouova lemmatu. My zde toto lemma neuvádíme a nebudeme ani nikde dále využívat zmíněnou nerovnost.

Nejprve ukážeme, že bodový dolní kompresní poměr pro prostý kód je skoro všude větší nebo roven informačnímu obsahu a tedy i entropii, pokud existuje. Dolní a horní informační obsah procesu X v bodě $\omega \in \Omega$ definujeme předpisem

$$\underline{\mathfrak{I}}_X(\omega) = \liminf_{n\to\infty} \frac{\mathfrak{I}_{X_{[0..n)}}(\omega)}{n}.$$

$$\overline{\mathfrak{I}}_X(\omega) = \limsup_{n\to\infty} \frac{\mathfrak{I}_{X_{[0..n)}}(\omega)}{n}.$$

Informační obsah procesu je limita (pokud existuje):

$$\mathfrak{I}_X(\omega) = \lim_{n\to\infty} \frac{\mathfrak{I}_{X_{[0..n)}}(\omega)}{n}.$$

Začneme důležitým lemmatem z teorie pravděpodobnosti.

Lemma 6.10 (Borel-Cantelliho lemma)**.** *Buď V_n, $n \in \mathbb{N}$, posloupnost měřitelných podmnožin Ω takových, že $\sum_{n=0}^{\infty} \mathbb{P}\left(V_n\right) < +\infty$. Pak pro skoro každý bod $\omega \in \Omega$ platí, že náleží jen do konečného počtu těchto množin, t.j.*

$$\mathbb{P}\left(\bigcap_{n\in\mathbb{N}} \bigcup_{k\geq n} V_k\right) = 0.$$

Důkaz. Průnik zmenšujících se sjednocení popisuje právě množinu bodů, které se vyskytují v nekonečně mnoha množinách V_n. Formulace slovní a pomocí formule si tedy odpovídá. Rovnost dokážeme následovně: jelikož je suma pravděpodobností konečná, existuje pro každé $\varepsilon > 0$ přirozené číslo m takové, že $\sum_{k=m}^{\infty} \mathbb{P}\left(V_k\right) < \varepsilon$. Dále platí

$$\mathbb{P}\left(\bigcap_{n\in\mathbb{N}} \bigcup_{k\geq n} V_k\right) \leq \mathbb{P}\left(\bigcup_{k\geq m} V_k\right) \leq \sum_{k=m}^{\infty} \mathbb{P}\left(V_k\right) < \varepsilon.$$

Ovšem $\varepsilon > 0$ bylo libovolné. □

Tvrzení 6.11. *Buď X náhodný proces s výstupní abecedou A a buď $f : A^* \to B^*$ prostý kód. Potom platí*

$$\liminf_{n\to\infty} \frac{1}{n}\left(\left|f\left(X_{[0..n)}\right)\right| - \mathfrak{I}_{X_{[0..n)}}\right) \geq 0 \text{ s.j.}$$

Mimo jiné,

$$\underline{(\!|f|\!)}_X \geq \underline{\mathfrak{I}}_X, \quad \overline{(\!|f|\!)}_X \geq \overline{\mathfrak{I}}_X, \qquad \text{s.j.}$$

Důkaz. Uvažujme množiny

$$V_n = \left\{\omega \in s\left(X_{[0..n)}\right) \mid \left|f\left(X_{[0..n)}(\omega)\right)\right| < \mathfrak{I}_{X_{[0,n)}}(\omega) - 2\log n\right\}, \qquad n \in \mathbb{N},$$

tedy množiny, na jejichž obraz kóduje f překvapivě dobře, přičemž za překvapivou odchylku od očekávané hodnoty (informačního obsahu) volíme $2\log n$, z důvodu, který se brzy ukáže. Označme tyto obrazy, tedy příslušné množiny slov z A^n, jako W_n. Z definice informačního obsahu dostáváme, že $u \in W_n$, právě když

$$P_{X_{[0..n)}}(u) < 2^{-|f(u)|-2\log n}.$$

Díky předpokladu, že kód f je prostý, pro něj, a tedy tím spíš i pro jeho zúžení na W_n, platí Kraftova nerovnost. Dostáváme tedy

$$\mathbb{P}\left(V_n\right) = \sum_{u\in W_n} P_{X_{[0..n)}}(u) < \sum_{u\in W_n} 2^{-|f(u)|-2\log n} \leq \frac{1}{n^2} \sum_{u\in W_n} 2^{-|f(u)|} \leq \frac{1}{n^2}.$$

Suma $\sum_{n=1}^{\infty} \frac{1}{n^2}$ je konečná, což vyplývá z integrálního kritéria, případně ze srovnání s teleskopickou řadou $\sum_{n=2}^{\infty} \left(\frac{1}{n-1} - \frac{1}{n} \right)$. Z Borelova-Cantelliho lemmatu nyní plyne, že množina

$$V = \bigcap_{n \in \mathbb{N}} \bigcup_{k \geq n} V_k$$

nekonečných slov, na kterých bude nekonečně často docházet k překvapivě krátkému kódování, má nulovou pravděpodobnost. Skoro pro všechna slova se tedy kódování dříve či později začne chovat „normálně“. Neboli, pro $\omega \in \Omega \setminus V$ platí, že patří jen do konečně mnoha množin V_n, a proto

$$\liminf_{n \to \infty} \frac{1}{n} \left(\left| f(X_{[0..n)}) \right| - \mathfrak{F}_{X_{[0,n)}}(\omega) \right) \geq \liminf_{n \to \infty} \frac{2 \log n}{n} = 0.$$

Druhá část tvrzení plyne z první skrze jednoduché pozorování, že rozdíl dvou dolních limit (taktéž horních limit a limit) je větší nebo roven dolní limitě rozdílu. □

Pokud informační obsah konverguje skoro jistě k entropii, neboli pokud má proces sAEP, dostáváme okamžitě následující důležitý fakt.

Věta 6.12. *Buď X náhodný proces s výstupní abecedou A a entropií, který má silně asymptoticky rovnoměrné rozložení (např. i.i.d. proces). Nechť $f : A^* \to B^*$ je prostý kód. Potom*

$$\underline{(\!|f|\!)}_X \geq \mathcal{H}_D(X) \text{ s.j. .}$$

Kapitolu zakončíme pozitivním tvrzením, že pro rozumně se chovající procesy komprimují kódy zkonstruované v úvodu optimálně skoro všude i bodově.

Tvrzení 6.13. *Buď X náhodný proces s výstupní abecedou A. Potom existuje prefixový kód $f : A^* \to B^*$ takový, že*

$$\lim_{n \to \infty} \frac{1}{n} \left(\left| f\left(X_{[0..n)}\right) \right| - \mathfrak{F}_{X_{[0..n)}} \right) = 0 \text{ s.j.}$$

Mimo jiné,

$$\underline{(\!|f|\!)}_X \leq \underline{\mathfrak{F}}_X, \quad \overline{(\!|f|\!)}_X \leq \overline{\mathfrak{F}}_X, \qquad \text{s.j.}$$

Důkaz. Pro každé přirozené n buď f_n Shannonův kód pro $X_{[0..n)}$. Definujme

$$f(u) = \gamma_1(|u|) f_{|u|}(u), \qquad u \in A^*,$$

což je podle Lemmatu 6.3 prefixový kód. Pro $\omega \in s\left(X_{[0..n)}\right)$ platí

$$\frac{\left| f\left(X_{[0..n)}(\omega)\right) \right|}{n} \leq \frac{2 \log_D n + 4 + \mathfrak{F}_{X_{[0..n)}}(\omega) + 1}{n}.$$

Rozdíl stran nerovnosti je tedy shora omezen výrazem $(2\log_D n + 5)/n$, který konverguje k nule. Tedy

$$\limsup_{n\to\infty} \frac{1}{n}\left(\left|(f(X_{[0..n)})\right| - \mathfrak{I}_{X_{[0..n)}}\right) \leq 0 \text{ s.j.}$$

Z Tvrzení 6.11 máme opačné omezení pro dolní limitu, a proto existuje limita a je rovna nule. Tím je dokázána první část tvrzení. Druhá část je okamžitým důsledkem. □

Opět můžeme formulovat silnější verzi v případě silně asymptoticky rovnoměrného rozdělení. Důkaz je přímým důsledkem výše uvedeného tvrzení a samotné definice sAEP.

Věta 6.14. *Buď X náhodný proces s výstupní abecedou A a entropií, který má silně asymptoticky rovnoměrné rozložení (např. i.i.d. proces). Potom existuje prefixový kód $f : A^* \to B^*$ takový, že*

$$\langle\!|f|\!\rangle_X = \mathcal{H}_D(X) \text{ s.j.}$$

7 Komprese - Univerzální kódy

V minulé části jsme ukázali, že prosté kódy nejsou schopny komprimovat v rychlejším poměru, než je entropie procesu, a zároveň jsme zkonstruovali prefixové kódy, které optimální kompresní poměr dosahovaly. Naše výsledky (týkající se průměrné délky kódu) platily pro všechny procesy s entropií. Na druhou stranu, optimální kód byl vždy zkonstruován pro jeden náhodný proces a pro jiné procesy optimální nebyl. V této kapitole ukážeme, že lze zkonstruovat kód, či kódy, který bude mít kompresní poměr roven entropii zároveň pro všechny i.i.d. procesy. Poznamenejme bez precizní formulace a důkazu, že takový kód bude optimální asymptoticky, ale pro omezenou délku zpráv a konkrétní proces optimální nebude. Důvodem je zvyšující se divergence mezi jednotlivými procesy.

7.1 Frekvenční kód

Nejjednodušší kód, který je asymptoticky optimální pro všechny i.i.d. procesy s hodnotami v abecedě B velikosti D, je frekvenční kód, který popisuje dané slovo tak, že udá počty výskytů jednotlivých písmen, spolu s pořadím daného slova v množině všech slov se stejnými počty výskytů.

Předpokládejme, že se kódující i dekódující strana shodne na konkrétním uspořádání všech slov nad abecedou A. Můžeme například zvolit uspořádání na A a lexikograficky ho rozšířit na A^+. Pokud potom pošleme informaci o počtu výskytů písmen v dané zprávě a relativní pořadí této zprávy mezi slovy se stejným počtem výskytů, není problém pro dekódujícího dohledat si v seznamu těchto slov danou zprávu.

Nechť tedy $(a_1, a_2, \dots, a_m)$ posloupnost je posloupnost písmen z A ve zvoleném uspořádání. Pro slovo $u \in A^+$ a $a \in A$, značí $|u|_a$ počet výskytů písmene a ve slově u. Zobrazení $\boldsymbol{p} : A^+ \to \mathbb{N}^m$,

$$\boldsymbol{p} : u \mapsto \left(|u|_{a_1}, |u|_{a_1}, \dots, |u|_{a_m}\right),$$

přiřazuje každému slovu jeho **Parikhův vektor**. Označme

$$T(\boldsymbol{v}) = \{u \in A^+ \mid \boldsymbol{p}(u) = \boldsymbol{v}\},$$

množinu slov, s daným Parikhovým vektorem $\boldsymbol{v}$. Tato slova jsou tedy permutací jedno druhého. Prvky v $T(\boldsymbol{v})$ seřadíme v danému uspořádání na A^+ a zápisem $T(\boldsymbol{v})_i$, $0 \leq i < |T(\boldsymbol{v})|$, budeme značit i-tý prvek množiny $T(\boldsymbol{v})$ v daném uspořádání. Symbolem $\ell(u)$ budeme značit pořadové číslo slova u mezi slovy se stejným Parikhovým vektorem, tedy

$$T(\boldsymbol{p}(u))_{\ell(u)} = u.$$

Zároveň označme jako $\mathcal{P}_u$ empirické pravděpodobností rozdělení na A dané zprávou u, konkrétně $\mathcal{P}_u(a) = \frac{|u|_a}{|u|}$, $a \in A$.

Frekvenční kód $\mathfrak{f} : A^+ \to B^+$ je nyní definován předpisem

$$\mathfrak{f}(u) = \gamma_1(|u|_1)\gamma_1(|u|_2)\dots\gamma_1(|u|_m)\gamma_2(\ell(u)),$$

kde pro jednoduchost zápisu předpokládáme, že $A = \{1, 2, \dots m\}$.

Lemma 7.1. *Frekvenční kód* $\mathfrak{f} : A^+ \to B^+$ *je prefixový.*

Důkaz. Nechť jsou $\mathfrak{f}(u)$ a $\mathfrak{f}(v)$ prefixově srovnatelné. Pak jsou i $\gamma_1(|u|_1)$ a $\gamma_1(|v|_1)$ prefixově srovnatelné a z prefixovosti γ_1 plyne $|u|_1 = |v|_1$. Tento společný prefix odstraníme a opakováním téhož postupu dostaneme, že slova u a v mají stejný Parikhův vektor a že $\ell(u) = \ell(v)$. Tedy $u = v$, což jsme chtěli ukázat. □

V dalším lemmatu a také v další části, která se týká Lempelova-Zivova kódu, budeme používat pravděpodobnostní argument, který se opírá o takzvanou (tenzorovou) mocninu rozdělení, které definujeme následovně: pro dané rozdělení P na konečné množině A a kladné přirozené n, je n**-tá mocnina** P rozdělení na A^n definované předpisem

$$P^n(v) = \prod_{i=0}^{|u|-1} P(v_i), \qquad v \in A^n.$$

Není těžké ověřit, že se opravdu jedná o rozdělení, tedy že $P^n(u)$ je vždy nezáporné a součet přes všechna možná slova z A^n je jedna. Jak jsme si již mohli všimnout, takovéto rozdělení se objevuje v případě i.i.d. procesů. Pokud definujeme i.i.d. proces Y s hodnotami v A pomocí podmínky $P_{Y_0} = P$, pak je pro každé kladné n, rozdělení $P_{Y_{[0..n)}}$ n-tou mocninou rozdělení P. Pro náš pravděpodobnostní argument ovšem proces Y nepotřebujeme.

Lemma 7.2. *Pro* $u \in A^+$ *platí*

$$|T(\boldsymbol{p}(u))| \leq D^{|u|\cdot\mathcal{H}_D(\mathcal{P}_u)}.$$

Důkaz. Mějme $v \in T(\boldsymbol{p}(u))$. Vzhledem k tomu, že má v stejný parikhův vektor, musí se každý symbol v_i, $i < |v|$, vyskytovat také ve slově u. Z tohoto triviálního pozorování plyne, že $\mathcal{P}_u(v_i) > 0$ pro všechna $i < |u|$. Můžeme tedy přejít k logaritmu,

$$\log_D(\mathcal{P}_u)^{|u|}(v) = \sum_{i=0}^{|u|-1} \log_D \mathcal{P}_u(v_i) = \sum_{a\in s(\mathcal{P}_u)} |v|_a \log_D((\mathcal{P}_u(a)))$$

$$= |u| \sum_{a\in s(\mathcal{P}_u)} \frac{|u|_a}{|u|} \log_D((\mathcal{P}_u(a))) = -|u| \cdot \mathcal{H}_D\left(\mathcal{P}_u\right),$$

kde druhá rovnost vznikla seskupením stejných sčítanců a využitím faktu, že $v_i \in s\left(\mathcal{P}_u\right)$ pro všechna $i < |v|$. Všechna slova z $T(\boldsymbol{p}(u))$ tedy mají stejnou pravděpodobnost

$$(\mathcal{P}_u)^{|u|}(v) = D^{-|u|\cdot\mathcal{H}_D(\mathcal{P}_u)},$$

a proto jich nemůže být víc než tvrdí lemma. □

Lemma 7.3. *Pro $x \in A^{\mathbb{N}}$ platí*

$$\limsup_{n\in\mathbb{N}} \frac{\left|\mathfrak{f}\left(x_{[0..n)}\right)\right|}{n} \leq \limsup_{n\in\mathbb{N}} \mathcal{H}_D\left(\mathcal{P}_{x_{[0..n)}}\right).$$

Důkaz. Označme $u = x_{[0..n)}$. Platí

$$\begin{aligned}\left|\mathfrak{f}\left(x_{[0..n)}\right)\right| &= |\gamma_2(\ell(u))| + \sum_{a\in A} |\gamma_1(|u|_a)| \leq |\gamma_2(|T(u)|)| + \sum_{a\in A} |\gamma_1(n)| \\ &\leq |A|(2\log_D n + 3) + \log_D\left(D^{n\mathcal{H}_D(\mathcal{P}_u)}\right) + 2\log_D\left(\log_D D^{n\mathcal{H}_D(\mathcal{P}_u)} + 1\right) + 4. \\ &\leq |A|(2\log_D n + 3) + n\mathcal{H}_D\left(\mathcal{P}_{x_{[0..n)}}\right) + 2\log_D\left(n\log_D(|A|) + 1\right) + 4.\end{aligned}$$

Poděleníčíslem n a přechodem k limsup dostáváme tvrzení. □

Nyní můžeme ukázat, že frekvenční kód je (dokonce bodově) optimální pro všechny procesy, u kterých empirická frekvence odpovídá entropii.

Věta 7.4. *Nechť proces X s hodnotami v abecedě A splňuje*

$$\lim_{n\to\infty} \mathcal{H}_D\left(\mathcal{P}_{X_{[0..n)}(\omega)}\right) = \mathcal{H}_D(X) \qquad \text{s.j.}.$$

Pak

$$(\!|\mathfrak{f}|\!)_X = \mathcal{H}_D(X) \text{ s.j.} \quad a \quad [\![\mathfrak{f}]\!]_X = \mathcal{H}_D(X).$$

Důkaz. Z předchozího lemmatu díky předpokladu dostáváme

$$\underline{(\!|\mathfrak{f}|\!)}_X \leq \overline{(\!|\mathfrak{f}|\!)}_X \leq \mathcal{H}_D(X) \text{ s.j.}$$

Z Věty 6.12 naopak plyne, že $\underline{(\!|\mathfrak{f}|\!)}_X \geq \mathcal{H}_D(X)$ skoro jistě. Tedy i $(\!|\mathfrak{f}|\!)_X = \mathcal{H}_D(X)$ skoro jistě. □

Které procesy jsou v tomto smyslu rozumné, zde nebudeme podrobněji zkoumat. Nebudeme ani zkoumat, zda je frekvenční kód optimální i za nějaké slabší podmínky (např. sAEP). Omezíme se na nejjednodušší případ i.i.d.

Lemma 7.5. *Pro i.i.d. proces X platí, že*

$$\lim_{n\to\infty} \mathcal{H}_D\left(\mathcal{P}_{X_{[0..n)}}\right) = \mathcal{H}_D(X) \qquad \text{s.j.}.$$

Důkaz. Podle zákona velkých čísel platí pro každé $a \in s\left(P_{X_0}\right)$, že

$$\lim_{n\to\infty} \mathcal{P}_{X_{[0..n)}} = P_{X_0} \text{ s.j.}$$

Ze spojitosti součtu a součinu a spojitosti logaritmu pro kladné hodnoty tedy dostáváme

$$\begin{aligned}&\lim_{n\to\infty} \sum_{a\in s\left(P_{X_{[0..n)}}\right)} \mathcal{P}_{X_{[0..n)}}(a)\left(-\log_D \mathcal{P}_{X_{[0..n)}}(a)\right)\\ &= \sum_{a\in s\left(P_{X_0}\right)} P_{X_0}(a)\left(-\log_D P_{X_0}(a)\right) = \mathcal{H}_D\left(P_{X_0}\right) = \mathcal{H}_D(X).\end{aligned}$$

□

Frekvenční kód je univerzální v tom smyslu, že dosahuje optimální kompresi dat pro i.i.d. proces. Na rozdíl od blokových kódů má však jednu velkou nevýhodu. Kódovací i dekódovací algoritmus má exponenciální výpočtovou složitost.

Navíc má velké dekódovací zpoždění: nemůžeme dekódovat postupně, zprávu můžeme začít dekódovat až poté, co je celá přenesena.

Zároveň tento algoritmus, tak jak jsme ho zde představili, nefunguje obecně pro stacionární procesy, či souvislé Markovské procesy. Dal by se rozšířit tím, že bychom zvažovali frekvence výstupu slov, nikoliv jen písmen, ale toto rozšíření by vyžadovalo jemnou práci s mnoha parametry. Raději představíme jiný typ kódu.

7.2 Lempelův-Zivův rekurenční kód

Další typ kódů je založen na rekurenci. Podslovo kódovaného textu je určeno adresou svého předcházejícího výskytu. Tyto kódy jsou stejně jako frekvenční kód univerzální: dosahují optimální kompresi dat pro každý i.i.d. proces.

Časová složitost kódovacího algoritmu je $\mathcal{O}(n \log n)$, kde n je délka komprimovaného slova. Časová složitost dekódovacího algoritmu je dokonce lineární.

Rekurenční kódy lze konstruovat pro libovolnou abecedu. Dané slovo $u \in B^*$ rozložíme na bloky proměnné délky tak, že začneme prázdným slovem a pokračujeme vždy nejkratším úsekem, který se mezi předcházejícími bloky nevyskytuje. Pokud po ukončení takového algoritmu zbude neprázdný koncový blok z u přidáme ho do souboru bloků, viz pseudokód níže.

Výslednému posloupnosti $\{y_0, y_1, \ldots, y_C\}$, kde C závisí na u, říkáme **rozbor**. Platí $y = y_1 y_2 \cdots y_C$ a pro $1 \leq i \leq C$ je y_i tvaru $y_{a_i} c_i$, kde $c_i \in A$. Zvolme pevně prefixový kód $f : A \to B^+$. **Lempelův-Zivův kód** slova u pak definujeme jako:

$$\mathfrak{r}(u) = \gamma_2(a_1) f(c_1) \gamma_2(a_2) f(c_2) \ldots \gamma_2(a_C) f(c_C).$$

Příklad 7.6. Zakódujme slovo

$$u = 000011001110011001101100$$

nad binární abecedou $A = \{0, 1\}$ opět do binární abecedy $B = \{0, 1\}$, kde za f volíme identitu. Výstup algoritmu `LZ-Rozbor` je shrnut v následující tabulce:

i	0	1	2	3	4	5	6	7	8	9	10
$y^{(i)}$	ε	0	00	01	1	001	11	0011	00110	111	00
a_i		0	1	1	0	2	4	5	7	6	1
c_i		0	0	1	1	1	1	1	0	1	0
k_i	0	1	3	5	6	9	11	15	20	23	25

Algorithm 2: `LZ Rozbor`

Data: $u \in A^+$
Result: $(y_0, y_1, \ldots, y_C); (k_0, k_1, \ldots, k_C); (a_1, \ldots, a_C); (c_1, \ldots, c_C)$
$y_0 \leftarrow \varepsilon$;
$z \leftarrow u$;
$j \leftarrow 1$;
$k_0 \leftarrow 0$;
while $z \neq \varepsilon$ **do**
  $y_j \leftarrow \begin{cases} \text{nejkratší prefix } z \text{ neležící v } \{y_0, \ldots, y_{j-1}\}, \text{ pokud existuje,} \\ z \text{ jinak} \end{cases}$;
  $(a_j, c_j) \leftarrow\ c_j \in A, a_j < j$ takové, že $y_j = y_{a_j} c_j$;
  $k_j \leftarrow \left| y_1 \cdots y_j \right|$;
  $z \leftarrow y_C^{-1} z$;
  $C \leftarrow C + 1$;
end

Připomeňme, že naše kódování γ_2 používá pro kódování čísel nejprve unární zápis čísla zakončený oddělovacím symbolem (to odpovídá γ_0) a poté dva po sobě jdoucí dyadické zápisy čísel pomocí β. Symboly abecedy B tedy plní tři různé role. Pro γ_0 (tam jsme se rozhodli je značit 1 a 0), poté pro β (kde jsme používali jména 1 a 2) a konečně slouží pro zakódování abecedy A. Pro větší přehlednost zachováváme různá značení pro prvky B i v následujícím zápisu.

$$(y)_{i=0}^{10} = (\varepsilon, 0, 00, 01, 1, 001, 11, 0011, 00110, 111, 00)$$
$$(a_i, c_i)_{i=1}^{10} = ((0,0), (1,0), (1,1), (0,1), (2,1), (4,1), (5,1), (7,0), (6,1), (1,0))$$
$$\mathfrak{r}(u) = \overbrace{0}^{0}\ \overbrace{\mathbf{0}\,10\,11}^{1}\ \overbrace{\mathbf{0}\,10\,11}^{1}\ \mathbf{1}\ \overbrace{0}^{0}\ \overbrace{\mathbf{1}\,10\,12}^{2}\ \overbrace{\mathbf{1}\,10\,212}^{4}\ \overbrace{\mathbf{1}\,10\,221}^{5}\ \mathbf{1}$$
$$\overbrace{110\,11\,11\,11}^{7}\ \mathbf{0}\ \overbrace{10\,22\,2}^{6}\ \mathbf{1}\ \overbrace{10\,11}^{1}\ \mathbf{0}$$

Pokud tedy $B = \{0, 1\}$, pro symbol 2 dyadického zápisu použijeme 0 a pro **0**, **1** postupně 0, 1, bude u ve výsledku zakódováno jako

$$\mathfrak{r}(u) = 00101101011101101011001011000111101111101000011010\,.$$

Lemma 7.7. *Lempelův-Zivův kód je prostý.*

Důkaz. Nechť $\mathfrak{r}(u) = \mathfrak{r}(v)$. Protože jsou obě slova zřetězením posloupností kódových slov prefixových kódů, jsou tyto posloupnosti pro obě slova stejné. Posloupnost slov definujících Lempel-Zivův kód jednoznačně definuje rozbor slov u a v. Rozbory obou slov se tedy rovnají, a proto i $u = v$. □

Optimalita Lempelova-Zivova kódu (nepovinné)

Lemma 7.8. *Délka Lempelova-Zivova kódu splňuje omezení*

$$|\mathfrak{r}(u)| \leq C(u) \left(\log_D C(u) + 2\log_D(\log_D C(u) + 1) + 4 + K\right),$$

kde K je konstanta, která dominuje délky kódových slov kódu $|f|$.

Důkaz. Z konstrukce kódu vyplývá, že

$$|\mathfrak{r}(u)| = \sum_{i=1}^{C(u)} \left(|\gamma_2(a_i)| + |f(c_i)|\right),$$

pro čísla $C(u)$, a_i a c_i odvozené z rozkladu slova u. Z definice plyne, že $a_i < C(u)$. Z monotónnosti logaritmu a z odhadu pro γ_2 z lematu 6.5 pak získáváme pro kladná a_i:

$$|\gamma_2(a_i)| \leq \log_D C(u) + 2\log_D(\log_D C(u) + 1) + 4.$$

Ovšem platnost stejného odhadu pro případ $a_i = 0$ lze ověřit pouhým dosazením.

Máme tedy uniformní odhad, který již zaručuje platnost tvrzení. □

Další lemma říká, že $C(u)$ roste spolu s délkou u do nekonečna.

Lemma 7.9. *Pro $u \in A^+$ platí $C(u) \geq \sqrt{|u|}$.*

Důkaz. Buď $\left(y^{(i)}\right)_{0 \leq i \leq C(u)}$ rozbor slova u. Z konstrukce rozboru plyne, že $\left|y^{(i)}\right| \leq i$, pro všechna $i \leq C(u)$. Tedy

$$|u| = \sum_{i=1}^{C(u)} \left|y^{(i)}\right| \leq \sum_{i=1}^{C(u)} i = \frac{C(u)(C(u)+1)}{2} \leq C^2(u).$$

Z toho již dokazované tvrzení plyne. □

V následujících lemmatech a tvrzeních budeme opět využívat pravděpodobnostního argumentu spojeného n-tou mocninou daného rozdělení, která byla definována v předchozí části při dokazování efektivity komprese frekvenčním kódem.

Lemma 7.10. *Buď P pravděpodobnost na A, $q = \max_{a\in A} P(a)$, $u \in A^+$. Potom*

$$C(u)\left(\log_D C(u) - \log_D\left(\log_D^2 C(u) + C(u)q^{\log_D^2 C(u)} + 1\right)\right) \le -\log_D P^{|u|}(u).$$

Důkaz. Vzhledem k vlastnostem rozboru, platí

$$P^{|u|}(u) = \prod_{i=1}^{C(u)} P^{|y^{(i)}|}(y^{(i)}).$$

Označme $m = C(u)$, $n = |u|$. Z konkávnosti logaritmu dostáváme

$$\begin{aligned}\frac{1}{m}\log_D P^{|u|}(u) &= \frac{1}{m}\log_D \prod_{i=1}^{m} P^{|y^{(i)}|}(y^{(i)}) = \frac{1}{m}\sum_{i=1}^{m}\log_D P^{|y^{(i)}|}(y^{(i)}) \\ &\le \log_D \frac{1}{m}\sum_{i=1}^{m} P^{|y^{(i)}|}(y^{(i)}) = -\log_D m + \log_D \sum_{i=1}^{m} P^{|y^{(i)}|}(y^{(i)}).\end{aligned}$$

Z toho plyne,

$$m\left(\log_D m - \log_D \sum_{i=1}^{m} P^{|y^{(i)}|}(y^{(i)})\right) \le -\log_D P^{|u|}(u).$$

Pro odhad sumy v předchozím výrazu je důležité, že jsou slova z rozboru různá, až na to poslední. V následujícím odhadu jsou slova seskupena dle délky a společná pravděpodobnost slov stejné délky je odhadnuta shora pravděpodobností 1. Označme $R'(u) = \{y^{(i)}, 1 \le i \le m-1\}$. Pak platí

$$\begin{aligned}\sum_{i=1}^{m} P^{|y^{(i)}|}\left(y^{(i)}\right) &= P^{|y^{(m)}|}\left(y^{(m)}\right) + \sum_{k=1}^{\ell}\sum_{y\in R'(u),|y|=k} P^k(y) + \sum_{y\in R'(u),|y|>\ell} P^{|y|}(y) \\ &\le 1 + \sum_{k=1}^{\ell} 1 + mq^{\ell+1} \le 1 + \ell + mq^{\ell+1}.\end{aligned}$$

V aplikacích lemmatu budeme potřebovat, že logaritmus pravé strany bude zanedbatelný vzhledem k logaritmu m. Vhodnou volbou je zde $\ell = \left\lfloor \log_D^2 m \right\rfloor$. Při této volbě vychází z předchozích nerovností požadovaný odhad. □

Předchozí dvě lemmata by se dala, při zanedbání některých členů, číst jako nerovnosti $|\mathfrak{r}(u)| \le C(u)\log_D C(u) \le -\log_D P^{|u|}(u)$. V dalším lemmatech ukážeme, že asymptoticky je toto zanedbání členů v pořádku.

Lemma 7.11. *Platí*

$$\lim_{m\to\infty} \frac{m\left(\log_D m + 2\log_D\left(\log_D m + 1\right) + 4 + K\right)}{m\log_D m} = 1.$$

Pokud $0 < q < 1$*, pak*

$$\lim_{m\to\infty} \frac{m\left(\log_D m - \log_D\left(\log_D^2 m + mq^{\log_D^2 m} + 1\right)\right)}{m \log_D m} = 1.$$

Důkaz. První limita je důsledkem faktu, že $\log_D(\log_D m)$ roste pomaleji než $\log_D m$ a

$$\lim_{m\to\infty} \frac{2\log_D\left(\log_D m + 1\right) + 4 + K}{\log_D m} = 0.$$

U druhé limity, nejprve analyzujme chování posloupnosti $\left(mq^{\log_D^2 m}\right)_{m=1}^{\infty}$,

$$mq^{\log_D^2 m} = D^{\log_D\left(mq^{\log_D^2 m}\right)} = D^{\log_D m + \log_D^2 m \log_D q}.$$

Jelikož jde $\log_D m$ k nekonečnu a $\log_D(q) < 0$, jde exponent u výrazu vpravo k $-\infty$. Proto jde celý výraz k nule. Vyšetřovaná posloupnost je tedy omezená pomocí konstanty, která závisí jen na q. Označme tuto konstantu K'. Dostáváme,

$$\lim_{m\to\infty} \frac{\log_D\left(\log_D^2 m + mq^{\log_D^2 m} + 1\right)}{\log_D m} = \lim_{m\to\infty} \frac{\log_D\left(\log_D^2 m + K' + 1\right)}{\log_D m} = 0.$$

Z toho již plyne platnost hodnoty druhé limity z lemmatu. □

Uvědomme si, že v první limitě je v čitateli horní odhad pro délku kódu. V druhé limitě v čitateli zase vystupuje dolní odhad pro $-\log_D P^{|u|}(u)$, kde P je vlastně libovolná. Pokud přidáme fakt, že počet slov v rozboru jde do nekonečna, pokud do nekonečna jde délka slova, které kódujeme, dostaneme následující důsledek.

Důsledek 7.12. *Pro libovolnou netriviální pravděpodobnost* P *na abecedě* A *a pro libovolnou posloupnost* $x \in A^{\mathbb{N}}$ *platí:*

$$\limsup_{n\to\infty} \frac{\left|\mathfrak{r}\left(x_{[0..n)}\right)\right|}{n} \leq \limsup_{n\to\infty} \frac{-\log_D P^n\left(x_{[0..n)}\right)}{n}.$$

Důkaz. Buď $q = \max_{a \in A} P(a) < 1$, $c_n = C(x_{[0..n)})$. Použijeme-li odhady

$$\begin{aligned}
\limsup_{n\to\infty} \frac{|\mathfrak{r}(x_{[0..n)})|}{n} &\leq \limsup_{n\to\infty} \frac{c_n \left(\log_D c_n + 2\log_D(\log_D c_n + 1) + 4 + K\right)}{n} \\
&= \lim_{n\to\infty} \frac{c_n \left(\log_D c_n + 2\log_D(\log_D c_n + 1) + 4 + K\right)}{c_n \log_D c_n} \\
&\quad \cdot \lim_{n\to\infty} \frac{c_n \log_D c_n}{c_n \left(\log_D c_n - \log_D \left(\log_D^2 c_n + c_n q^{\log_D^2 c_n} + 1\right)\right)} \\
&\quad \cdot \limsup_{n\to\infty} \frac{c_n \left(\log_D c_n - \log_D \left(\log_D^2 c_n + c_n q^{\log_D^2 c_n} + 1\right)\right)}{n} \\
&\leq 1 \cdot 1 \cdot \limsup_{n\to\infty} \frac{-\log_D P^n(x_{[0..n)})}{n}.
\end{aligned}$$

□

Věta 7.13. *Pro každý i.i.d. proces X s hodnotami v abecedě A platí*

$$\lim_{n\to\infty} \frac{\left|\mathfrak{r}\left(X_{[0..n)}\right)\right|}{n} = \mathcal{H}_D(X) \text{ s.j.}$$

Tedy

$$L(X, \mathfrak{r}) = \mathcal{H}_D(X).$$

Důkaz. Z Tvrzení 4.10 dostáváme, že

$$\lim_{n\to\infty} \frac{-\log_D P^n\left(X_{[0..n)}\right)}{n} = \mathcal{H}_D(X) \text{ s.j. .}$$

Aplikací předchozího lemmatu dostáváme, že

$$\limsup_{n\to\infty} \frac{|\mathfrak{r}(X_{[0..n)})|}{n} \leq \mathcal{H}_D(X) \text{ s.j. .}$$

Označme $g(\omega) = \liminf_{n\to\infty} \frac{|\mathfrak{r}(X_{[0..n)}(\omega))|}{n}$, podobně $g'(\omega)$ označuje limsup. Z předchozích lemmat dostáváme, že $g \leq g' \leq \mathcal{H}_D(X)$ skoro jistě. Z Věty 6.12 naopak plyne, že $g(\omega) \geq \mathcal{H}_D(X)$ skoro jistě. Tedy $g = g' = \mathcal{H}_D(X)$ skoro jistě. □

Podařilo se nám tedy dokázat, že podobně, jako v případě frekvenční kódu, je i Lempelův-Zivův kód univerzální v tom smyslu, že dosahuje optimální kompresi dat pro i.i.d. proces.

Univerzalita tohoto kódu jde ale mnohem dále. Tento kód dosahuje, v průměru i bodově, optimálního kompresního poměru také pro všechny stacionární procesy a pro všechny, i nestacionární, Markovské procesy. A ani to není definitivní omezení.

I to je důvod pro jeho reálné využití v kompresním formátu gzip, ale také v algoritmech použitých při ukládání souborů TIFF a PDF (zde je na uživateli, zda kompresi chce využít).

Uveďme tato fakta přesněji, v podobě následujícího tvrzení a věty. Důkaz tvrzení spadá do teorie markovských procesů, důkaz Věty pak do ergodické teorie. Oboje vybočuje z rámce našeho předmětu, proto důkazy neuvádíme. Nejprve tvrzení, které říká, že i nestacionární markovské procesy mají dobře definovanou entropii.

Tvrzení 7.14. *Každý homogenní markovský proces X má entropii.*

Věta 7.15. *Pro každý proces X s hodnotami v abecedě A, který je stacionární nebo homogenní markovský, je bodový kompresní poměr roven skoro jistě bodovému informačnímu obsahu na symbol, i.e.*

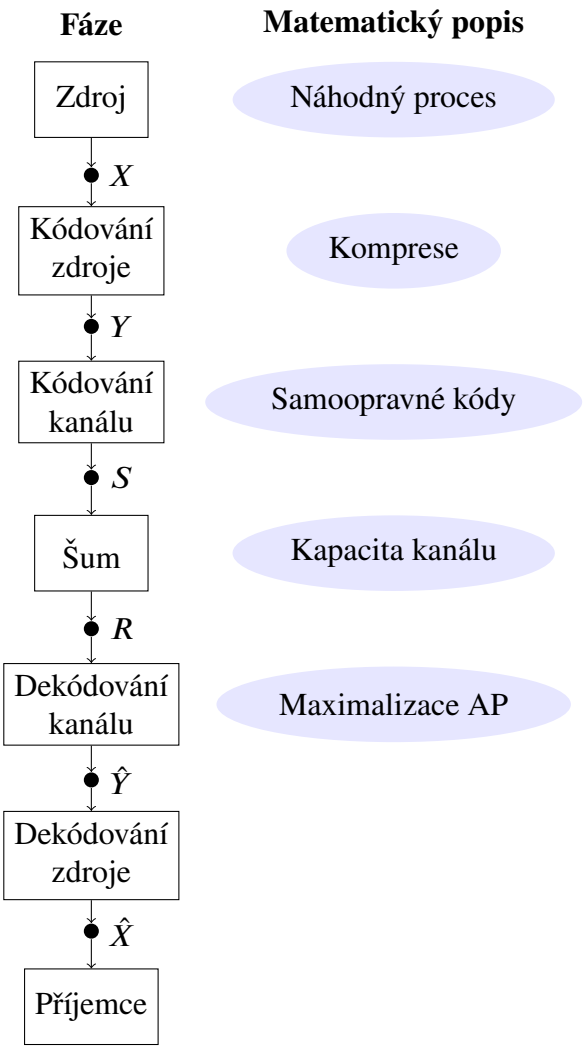
$$\lim_{n\to\infty} \frac{\left|\mathfrak{r}\left(X_{[0..n)}\right)\right|}{n} = \lim_{n\to\infty} \frac{\mathfrak{I}_{X_{[0..n)}}}{n} \text{ s.j.}$$

(obě limity skoro jistě existují a jsou si rovny). Z toho plyne, že

$$L(X, \mathfrak{r}) = \mathcal{H}_D(X).$$

8 Komunikační kanál

Základní aplikací teorie informace je komunikace, která přitom obsahuje několik fází, ke kterým se vážou různé požadavky a s tím související různé matematické otázky.



- Za zdroj informace v naší přednášce považujeme diskrétní náhodný proces $(X_i)_{i\in\mathbb{N}}$. To je významné zjednodušení. Mnohé informační zdroje (např. hudební produkce) jsou spojité. U spojitých procesů je jejich převod do diskrétní podoby již součástí kódování zdroje pro přenos.
- V případě diskrétního procesu už by žádné kódování zdroje být nemuselo. Důvodem pro něj může být prostá změna abecedy (např. převod písmen anglické abecedy do binárních symbolů ASCII). Součástí a hlavním smyslem kódování zdroje je ovšem obvykle komprese, tedy optimalizace využití kanálu.

- Pokud máme na mysli kanál bez šumu, je kompresí „přenos“ informace dokončen. Uvozovky naznačují, že v takovém případě často ani o přenosu nemluvíme. Je ovšem přirozené např. digitální medium, na které zprávu zapisujeme (a předáváme příjemci), za kanál považovat. Komprese je pak snahou kapacitu takového kanálu/media plně využít.

- Výsledkem kódování zdroje je náhodný proces $(Y_i)_{i\in\mathbb{N}}$. Komprese vede k procesu (co nejbližšímu) uniformně rozložených nezávislých veličin. Takový proces má totiž maximální entropii.

- Pokud uvažujeme kanál se šumem (např. i pokud bereme v úvahu možnou chybovost digitálního média), musíme naopak přidat „kontrolní bity“, neboli zprávu prodloužit tak, aby bylo možné chyby opravit. To je téma samoopravných kódů.

- Cílem kódování kanálu je opět maximálně využít možnosti kanálu. Je klíčovým příspěvkem Shannonova článku z roku 1948, že definoval kapacitu kanálu a ukázal, že je možné se jí kódováním libovolně přiblížit.

- Dekódování poškozené zprávy je často popisováno jako hledání nejbližšího kódového slova. Obecněji ovšem platí, že dekódování je hledáním odeslané zprávy, která je při znalosti zprávy přijaté nejpravděpodobnější (tzv. maximalizace aposteriorní pravděpodobnosti).

- Dekomprese a případný převod do formátu zdroje už je typicky matematicky přímočará.

8.1 Kanál a jeho kapacita

Diskrétní komunikační kanál má reprezentovat situaci, kdy zpráva v podobě konečné posloupnosti znaků ze vstupní abecedy A je převedena do výstupní abecedy B s tím, že může být poškozena šumem. Do deterministického procesu předání informace tak vstupuje šum v podobě jisté míry náhody a nejistoty na výstupu. Namísto deterministické funkce z A^+ do B^+, která by danému vstupu $u \in A^+$ přiřadila jeden výstup $v \in B^+$, budeme pro dané $u \in A^+$ reprezentovat výstup po průchodu kanálem náhodnou veličinou $\Gamma(u) : \Omega \to B^+$. Pro jednoduchost budeme předpokládat, že tato veličina nabývá jen konečně mnoho hodnot (má konečný nosič). Ekvivalentně se dá říci, že výstup kanálu je popsán funkcí $\Gamma : A^+ \times \Omega \to B^+$, která zohledňuje fakt, že výstup není ovlivněn jen vstupním slovem $u \in A^+$, ale také šumem, určitou náhodou, která je zde zachycena tím, že Γ je také funkcí $\omega \in \Omega$. Pár $(u, \omega) \in A^+ \times \Omega$ tak reprezentuje dvě ingredience, které determinují výstup kanálu, jednak vstupní slovo u, které jsme se použili na vstupu, a pak ω, které charakterizuje vnější podmínky, momentální teplotu, mikrostavy součástek v zařízení atd. Pokud jsme svědky šumu, tedy určitým způsobem nepředvídatelného chování, nemůžeme očekávat, že

bychom znali předpis funkce Γ. Opět je třeba uvažovat pravděpodobnostním způsobem. Pro jednotlivá vstupy $u \in A^+$, můžeme odvodit z Γ jednotlivé náhodné veličiny $\Gamma(u) : \Omega \to B^+$ předpisem $\Gamma(u)(\omega) = \Gamma(u, \omega)$. Tyto náhodné veličiny pak máme většinou zadány přes jejich rozdělení, ať už jednotlivá, či sdružená, například podmínkami na jejich nezávislost. To ale předbíháme. Obecný kanál tedy charakterizujeme jeho rozdělením Γ, které můžeme chápat jako funkci z $A^+ \times \Omega \to B^+$, nebo jako soubor náhodných veličin $(\Gamma(u))_{u \in A^+}$. Přechod mezi jednotlivými pojetími je okamžitý.

Základním úkolem v teorii kanálu je navrhnout takové schéma použití, které by dokázalo zpětně odstranit šum a rekonstruovat zaslanou zprávu, a to buď zcela bezchybně, nebo s minimální chybou. Schématem se pak myslí výběr vhodné množiny zpráv, které budeme kanálem posílat, a k tomu příslušné dekódovací funkce, která bude výstupní slovo interpretovat zpět na jednu z těchto možných zpráv.

Jeden z používaných způsobů (nikoliv optimální, jak ukážeme později), jak se vypořádat s chybovostí přenosu jsou kontrolní bity (předpokládáme binární vstupní i výstupní abecedu, $A = B = \{0, 1\}$). Vstupní zprávu délky n, kterou chceme poslat, doplníme m kontrolními bity. Tyto kontrolní bity mohou být definovány například pomocí parity vybraných pozic v původní zprávě. Řekněme tedy, že posíláme zprávy délky 4 a pro každou zprávu $u = u_0u_1u_2u_3 \in A^4$ definujeme tři kontrolní bity předpisem $u_4 = u_0 + u_1$, $u_5 = u_2 + u_3$ a $u_6 = u_1 + u_3$, kde sčítání se provádí modulo 2. Kanálem tedy neposíláme původní zprávu, ale odvozené slovo $f(u) = u_0u_1 \dots u_6$, které je obohaceno o dopočtené bity. Vzhledem k šumu, můžou být písmena ve výstupním slovu $v = v_0v_1 \dots v_6$ odlišná od vstupních. Pokud ovšem i ve výstupním slovu fungují kontrolní bity, t.j. pokud platí $v_4 = v_0 + v_1$, $v_5 = v_2 + v_3$ a $v_6 = v_1 + v_3$, pak budeme věřit tomu, že $v_0v_1v_2v_3$ je původní zpráva u. V případě neshody v kontrolních bitech povolíme samoopravnou proceduru, tedy to, že u výstupního slova povolíme opravu variací malého počtu bitů tak, abychom dostali slovo v', pro které už kontroly budou fungovat. Pak budeme za vyslanou zprávu považovat prefix takto opraveného slova, t.j. $v'_0v'_1v'_2v'_3$. V tuto chvíli ponechme stranou pravděpodobnost chyby přenosu, jeho efektivitu, a zaměřme se pouze na strukturu, kterou zde vytváříme. Zásadním pozorováním je, že kanálem neposíláme původní zprávu u, ale upravený vstup $f(u)$, kterému budeme říkat kódové slovo. Na výstupu kanálu pak konstruujeme funkci, která konkrétnímu výstupu přiřadí zprávu, o které se domnívá, že byla vyslána. U některých výstupních slov pak rozhodne, že je ani dekódovat nebude vzhledem k riziku velké chyby.

Uvedený postup zachycuje pojem kódu, který v sobě zahrnuje tři fáze z obrázku z počátku kapitoly, $Y \to S \to R \to \hat{Y}$, kde je Y náhodná veličina, která definuje rozdělení na množině zpráv. Je ovšem výhodné nejprve definovat kód pouze jako množinu zpráv. V pozadí je představa, že zprávy se vybírají uniformně náhodně, a záleží tak pouze na jejich počtu. V literatuře je tento předpoklad dokonce často pouze implicitní a o žádné náhodné veličině se nemluví.

Z hlediska působení kanálu je zároveň nepodstatná naše původní zpráva, ale kódové slovo. To prochází kanálem a jeho struktura, tedy například kontrolní bity, roz-

hodují o naší dekódovací schopnosti. Přesněji řečeno, podstatná je vybraná množina kódových slov. Předpokládáme, že ta je známa dekódující straně. Dekódování je tedy dobré chápat jako sofistikovaný výběr z kódových slov s ohledem na slovo obdržené, tedy s ohledem na výstup kanálu. Přiřazení kódových slov zprávám je oběma stranám komunikace známo, a reprodukovat tedy z kódového slova danou zprávu je již triviální záležitost, od které odhlížíme. Předpokládáme přitom jen přirozenou vlastnost takového přiřazení, totiž, že je prosté.

Dále nás tedy bude zajímat množina kódových slov a jejich dekódování. V této souvislosti tedy definujeme ***n*-kód** pro kanál se vstupní abecedou A a výstupní abecedou B jako dvojici (M, g), kde M je konečná podmnožina A^n a g je zobrazení z podmnožiny $B' \subseteq B^*$ do M (připouštíme, že některá výstupní slova dekódovat nebudeme).

Samotné zobrazení $g : B' \to A^n$ je tedy možno chápat jako kód ve smyslu definice z kapitoly o kompresi a kódovými slovy jsou potom přirozeně všechny prvky z oboru hodnot zobrazení g. My tu ovšem nazýváme kódovými slovy prvky z nadmnožiny M. Tento nesoulad je možno vyřešit požadavkem, aby M bylo rovno oboru hodnot g. My ale necháme otevřenou možnost, že na některé prvky množiny M nebude nikdy žádný výstup dekódován. Činíme tak z praktických důvodů: klasickou strategií v teorii kanálu totiž je nejprve vymezit množinu kódových slov, a teprve dodatečně specifikovat způsob dekódování.

V případě bezpaměťového kanálu využijeme vědomosti, že výstupem může být zase jen zpráva délky n, proto definujeme g jako zobrazení z $B' \subseteq B^n$ do M. Pro daný kód definujeme jeho **velikost** a **délku**, velikost je $|M|$ a délka je n.

Jednou z klíčových charakteristik kódu je jeho **přenosová rychlost**, kterou definujeme jako podíl

$$\frac{1}{n} \log |M| .$$

Význam této definice je jasný, jedná se o průměrný počet bitů přenášený jedním symbolem kódového slova. To je jeden z momentů, kde se tiše předpokládá, že zprávy jsou rozděleny rovnoměrně. Pak teprve totiž opravdu platí, že přenosová rychlost n-kódu je $\mathcal{H}(S)/n$. V motivačnímu příkladu s kontrolními bity jsme tedy vytvořili prostý 7-kód (M, g), kde $|M| = |A^4|$. Pro snížení chyby při přenosu 4 bitů informace, jsme raději přenesli bitů 7. Tomu dobře odpovídá přenosová rychlost, která je z definice $4/7$, tedy 4 bity informace za sedm realizovaných přenosů.

Pro daný kód (M, g) definujeme **maximální a průměrnou pravděpodobnost chyby přenosu** předpisem

$$\lambda = \max_{m \in M} \lambda(m), \qquad \hat{\lambda} = \frac{1}{|M|} \sum_{m \in M} \lambda(m),$$

kde

$$\lambda(m) = \mathbb{P}\left(g\left(\Gamma(m)\right) \neq m\right)$$

je chyba dekódování při přenosu zprávy $m \in M$.

Naším cílem je přirozeně najít kód, pomocí něhož dokážeme posílat co největší množství zpráv, neboli kódových slov, (tedy maximalizovat přenosovou rychlost) s co nejmenší pravděpodobností chyby. S tím souvisí definice přenosové rychlosti a kapacity samotného kanálu.

Definice 8.1. *Řekneme, že* ***kanál dosahuje rychlosti přenosu*** $r \geq 0$ *(anglicky „rate"), pokud existuje posloupnost n-kódů* (M_n, g_n)*,* $n \in \mathbb{N}$*, pro něž maximální pravděpodobnost chyby přenosu tímto kanálem konverguje k nule a* $|M_n| \geq 2^{rn}$*,* $n \in \mathbb{N}$ *(t.j. přenosová rychlost kódů je alespoň r).*

Kapacita kanálu $C(\Gamma)$ *je supremum dosažitelných rychlostí přenosu daným kanálem.*

Jelikož pro n-kód (M, g) platí $|M| \leq |A^n|$, jsou dosažitelné rychlosti a tím i kapacita přirozeně omezeny shora $\log |A|$.

Zároveň stačí k dosažení určité přenosové rychlosti $r \leq \log |A|$ najít vhodné kódy (M_n, g_n) pro n dostatečně velká n. Pro malá n lze dodefinovat (M_n, g_n) libovolně v souladu s podmínkou $|M_n| \geq 2^{rn}$. Toto dodefinování nemá vliv na konvergenci maximální pravděpodobnosti chyby.

Kapacita kanálu $C(\Gamma)$ je tedy definovaná pomocí následující vlastnosti:

- Pro každé nezáporné $r < C(\Gamma)$ existuje posloupnost n-kódů (M_n, g_n), $n \in \mathbb{N}$, jejichž maximální pravděpodobnost chyby přenosu konverguje k nule a $|M_n| \geq 2^{rn}$, $n \in \mathbb{N}$.
- Pro každé nezáporné $r > C(\Gamma)$ a každou posloupnost n-kódů (M_n, g_n), $n \in \mathbb{N}$, pro něž platí $|M_n| \geq 2^{rn}$, $n \in \mathbb{N}$, maximální pravděpodobnost chyby přenosu nekonverguje k nule.

Jinými slovy tedy kapacitou měříme asymptoticky maximální možnou rychlost přenosu pomocí kódů rostoucí délky. Nemělo by být překvapivé, že podobně jako při bezztrátovém kódování procesů mohou delší kódy dosahovat lepších přenosových rychlostí než krátké, přinejmenším u bezpaměťového kanálu, na který se brzy zaměříme.

8.2 Fanova nerovnost a horní mez pro kapacitu kanálu

V této kapitole už bude třeba chápat zprávu jako hodnotu náhodné veličiny. Tomu je třeba přizpůsobit naše chápání definice kanálu, ve které jsme dosud předpokládali, že posíláme konkrétní slovo u a pro něj byl výstup, ovlivněný náhodným šumem, reprezentován náhodnou veličinou $\Gamma(u) : \Omega \to B^*$. Jak ale popsat výstup kanálu pokud i odesílané slovo je hodnotou nějaké náhodné veličiny $S : \Omega \to A^*$? Nejprve uveďme formální definici a pak si ji rozeberme.

Definice 8.2. *Výstup kanálu s rozdělením* Γ *pro náhodný vstup* $S : \Omega \to A^*$ *je náhodná veličina* $\Gamma(S) : \Omega \to B^*$*, která je definovaná vztahem*

$$\Gamma(S)(\omega) = \Gamma(S(\omega), \omega).$$

Uvědomme si, že jevy $S = u$, $u \in A^*$, tvoří rozklad pravděpodobnostního prostoru Ω. Jakoukoliv náhodnou veličinu tedy můžeme popsat takzvaně po částech, tedy zvlášť pro každý ze zmíněných jevů, z nichž se celé Ω skládá. Takto dostáváme ekvivalentní definici $\Gamma(S)$ pomocí podmínek

$$\Gamma(S)(\omega) := \Gamma(u)(\omega), \qquad \text{pokud } S(\omega) = u\,.$$

Náhodná veličina $\Gamma(S)$ je tedy identická s $\Gamma(u)$, kdykoliv $S = u$. To vypadá jako samozřejmost, ale nesmíme zapomenout, že výraz $S = u$ zde neznamená rovnost dvou objektů, které tudíž můžeme ve výrazech libovolně zaměňovat. V našem případě výraz $S = u$ popisuje jev, tedy určitou podmnožinu Ω, totiž množinu, na které má veličina S hodnotu u. A právě na tomto jevu splyne $\Gamma(S)$ s $\Gamma(u)$.

Všimněme si zejména, že definice po částech, přestože se v ní objevují podmínky, se liší od definice podmíněných pravděpodobností (už proto, že žádnou „podmíněnou náhodnou veličinu“ jsme nedefinovali).

Pro daný n-kód (M, g) popsaný v předchozí kapitole nyní uvažujeme náhodnou veličinu S s hodnotami v M a příslušný výstup $R = \Gamma(S)$. Důležité pojmy, jako např. chybu přenosu, lze nyní vyjádřit jako konkrétní jevy. Zde je ale zásadní vyslovit a zdůraznit základní předpoklad celé Shannonovy teorie kanálu, který v případě kódů chápaných jako množiny může snadno zůstat mimo centrum pozornosti. Tímto základním předpokladem je *nezávislost volby zprávy a šumu*. Dokud jsme chápali kanál jako soubor náhodných veličin, docházelo k volbě kódového slova jaksi předem, mimo teorii pravděpodobnosti. My jsme nyní učinili volbu kódového slova (v duchu základního principu vysvětleného v Kapitole 2) součástí stejného pravděpodobnostního prostoru jako šum. Tento fundamentální pojem nezávislosti vyslovíme pomocí definice.

Definice 8.3. *Náhodná veličina X je **nezávislá na kanálu** Γ, pokud X a $\Gamma(u)$ jsou nezávislé, pro všechna $u \in A^+$. Tuto vlastnost budeme zapisovat $X \perp \Gamma$.*

Pro sdružené a podmíněné pravděpodobnosti z definice po částech dostáváme následující rozumné vlastnosti.

Lemma 8.4. *Buď S náhodná veličina s hodnotami v A^+ nezávislá na kanálu Γ a $R = \Gamma(S)$ buď odpovídající výstup kanálu s rozdělením Γ. Pak*

$$\mathbb{P}\left(R = v, S = u\right) = \mathbb{P}\left(\Gamma(u) = v\right)\mathbb{P}\left(S = u\right), u \in A^+, v \in B^+.$$

a

$$\mathbb{P}\left(R = v \mid S = u\right) = \mathbb{P}\left(\Gamma(u) = v \mid S = u\right) = \mathbb{P}\left(\Gamma(u) = v\right)$$

pro každé $u \in s\left(P_S\right)$ a $v \in B^+$.

Důkaz. Důkaz je velmi přímočarý, pokud vše rozepíšeme v řeči příslušných podmnožin Ω. Konkrétně je třeba nahlédnout následující rovnost množin (tedy jevů)

$$\{\omega \in \Omega \mid R(\omega) = v, S(\omega) = u\} = \{\omega \in \Omega \mid \Gamma(u)(\omega) = v, S(\omega) = u\}\,.$$

Každá z těchto množin je dána konjunkcí podmínek, z nichž jedna je zastoupena v obou. Jde tedy o to, zda podmínka $S(\omega) = u$ implikuje ekvivalenci podmínek $R(\omega) = v$ a $\Gamma(u)(\omega) = v$. Ovšem za podmínky $S(\omega) = u$ je

$$R(\omega) = \Gamma(S)(\omega) = \Gamma(S(\omega), \omega) = \Gamma(u, \omega) = \Gamma(u)(\omega).$$

Proto, pokud $S(\omega) = u$, pak $R(\omega) = v$ a $\Gamma(u)(\omega) = v$ jsou ekvivalentní. Uvedené jevy jsou tedy totožné a mají tím spíš stejnou pravděpodobnost. Z nezávislosti S a Γ_u dostáváme

$$\mathbb{P}(R = v, S = u) = \mathbb{P}(\Gamma(u) = v, S = u) = \mathbb{P}(\Gamma(u) = v)\,\mathbb{P}(S = u).$$

Pokud $u \in s\left(P_S\right)$, pak podělením obou stran právě dokázané rovnosti výrazem $\mathbb{P}(S = u)$ dostaneme rovnost podmíněných pravděpodobností. Poslední rovnost plyne z nezávislosti. □

Jak jsme avizovali, můžeme nyní vyjádřit průměrnou pravděpodobnost chyby pomocí konkrétního jevu.

Lemma 8.5. *Buď (M, g) n-kód s průměrnou pravděpodobností chyby $\hat{\lambda}$, S buď náhodná veličina rovnoměrně rozložená na M, nezávislá na Γ.*

Označme $R = \Gamma(S)$. Potom je $\hat{\lambda}$ rovno pravděpodobnosti špatného dekódování náhodně zvoleného kódového slova S, t.j.

$$\hat{\lambda} = \mathbb{P}(g(\Gamma(S)) \neq S) = \mathbb{P}(g(R) \neq S).$$

Důkaz. Označme $R = \Gamma(S)$.

$$\begin{aligned}
\mathbb{P}(g(\Gamma(S)) \neq S) &= \sum_{m \in M} \mathbb{P}(g(\Gamma(S)) \neq S, S = m) \\
&= \sum_{m \in M} \mathbb{P}(g(\Gamma(m)) \neq m, S = m) \\
&= \sum_{m \in M} \mathbb{P}(g(\Gamma(m)) \neq m)\,\mathbb{P}(S = m) \\
&= \frac{1}{|M|} \sum_{m \in M} \mathbb{P}(g(\Gamma(m)) \neq m) = \hat{\lambda}
\end{aligned}$$

Předpoklad nezávislosti jsme použili u třetí rovnosti. □

8.2.1 Poznámka o nezávislosti transformací

Důležitou součástí teorie kanálu, kterou využijeme zejména pro kanál bez paměti v následující kapitole, jsou zobrazení aplikovaná na výstupy náhodných veličin, pomocí nichž lze definovat nové náhodné veličiny. Veličina $X : \Omega \to A$ je tak zobrazením $f : A \to A'$ *transformována* na veličinu $f \circ X : \Omega \to A'$. Tuto novou veličinu nazýváme **transformací** náhodné veličiny X a zapisujeme ji obvykle jako $f(X)$. S některými vlastnostmi transformací už jsme se setkali např. v Tvrzení 3.4. Vzhledem k zásadní roli pojmu nezávislosti veličin v teorii kanálu bude pro nás důležité následující lemma.

Lemma 8.6. *Nechť $X : \Omega \to A$ a $Y : \Omega \to B$ jsou náhodné veličiny a nechť $f : A \to A'$ a $g : B \to B'$ jsou zobrazení. Pokud jsou X a Y nezávislé, jsou nezávislé i transformace $f(X)$ a $g(Y)$.*

Důkaz. Předpokládejme, že jsou X a Y nezávislé. Důkaz nezávislosti $f(X)$ a $g(Y)$ je prostým výpočtem založeným na tom, že jev $f(X) = u$ je disjunktním sjednocením jevů $X = a$ přes všechna $a \in f^{-1}(u)$, tedy přes všechna $a \in A$, pro která je $f(a) = u$.

Pro každé $u \in A'$ a $v \in B'$ platí

$$\begin{aligned}
&\mathbb{P}\left(f(X) = u, f(Y) = v\right) \\
&\quad = \sum_{\substack{a \in f^{-1}(u) \\ b \in g^{-1}(v)}} \mathbb{P}\left(X = a, Y = b\right) = \sum_{\substack{a \in f^{-1}(u) \\ b \in g^{-1}(v)}} \mathbb{P}\left(X = a\right) \cdot \mathbb{P}\left(Y = b\right) \\
&\quad = \sum_{a \in f^{-1}(u)} \mathbb{P}\left(X = a\right) \cdot \left(\sum_{b \in g^{-1}(v)} \mathbb{P}\left(Y = b\right) \right) \\
&\quad = \sum_{a \in f^{-1}(u)} \mathbb{P}\left(X = a\right) \cdot \mathbb{P}\left(g(Y) = v\right) = \mathbb{P}\left(f(X) = u\right) \cdot \mathbb{P}\left(g(Y) = v\right).
\end{aligned}$$

Transformace jsou tedy nezávislé. □

Indukcí můžeme rozšířit předchozí tvrzení na nezávislé soubory.

Lemma 8.7. *Nechť $X_i : \Omega \to A_i$, $i < n$, jsou vzájemně nezávislé veličiny a $f : A_i \to A_i'$ jsou zobrazení. Pak jsou i transformace $f_i(X_i)$ vzájemně nezávislé.*

Důkaz. Chceme ukázat

$$\mathbb{P}\left(X_{[o..n+1)}\right) = (a_0, a_1, \ldots, a_n) = \prod_{i<n+1} \mathbb{P}\left(X_i = a_i\right).$$

Podle indukčního předpokladu platí

$$\mathbb{P}\left(X_{[o..n)}\right) = (a_0, a_1, \ldots, a_{n-1}) = \prod_{i<n} \mathbb{P}\left(X_i = a_i\right)$$

a nyní můžeme použít předchozí lemma pro $X = X_n$ a $Y = X_{[0..n)}$ s transformacemi f_n a transformací

$$g : (a_0, a_1, \ldots, a_{n-1}) \mapsto \left(f_0(a_0), f_1(a_1), \ldots, f_{n-1}(a_{n-1})\right)$$

□

Předchozí tvrzení říká, že mezi nezávislá data nelze vnést dodatečným zpracování žádnou novou závislost. Následující tvrzení říká, že zpracováním nelze do X vnést žádnou informaci o Y, která tam již není.

Lemma 8.8. *Nechť X a Y jsou náhodné veličiny, $f(X)$ je transformace X. Pak $Y \to X \to f(X)$.*

Důkaz. Pro $(x, y) \in s\left(P_{X,Y}\right)$ zjevně platí,

$$\mathbb{P}\left(f(X) = z \mid X = x, Y = y\right) = \mathbb{P}\left(f(X) = z \mid X = x\right) = \begin{cases} 1, \text{ pokud } z = f(x) \\ 0, \text{ pokud } z \neq f(x). \end{cases}$$

Dostáváme tedy markovskou vlastnost $Y \to X \to f(X)$. □

Lemma 8.9. *Nechť $X : \Omega \to A$ a $Y : \Omega \to A$ jsou náhodné veličiny a nechť $f : A \to A'$. Pokud jsou X a Y stejně rozdělené (tj. $P_X = P_Y$), pak jsou i transformace $f(X)$ a $f(Y)$ stejně rozdělené (tj. $P_{f(X)} = P_{f(Y)}$).*

Důkaz. Předpokládejme, že jsou X a Y stejně rozdělené. Pro každé $b \in B$ platí

$$P_{f(X)}(b) = \sum_{a \in f^{-1}(b)} P_X(a) = \sum_{a \in f^{-1}(b)} P_Y(a) = P_{f(Y)}(b)$$

Vskutku, $P_{f(X)} = P_{f(Y)}$. □

8.2.2 Fanova nerovnost

Následující Fanovu nerovnost použijeme pro horní odhad rychlosti přenosu kanálem. Jedná se o jednoduchý a fundamentální nástroj, který lze neformálně vyjádřit takto. Budeme chtít vyjádřit entropii zprávy S za podmínky R, tedy zbytkovou nejistotu o vstupu po přijetí výstupu. Ta pochází z přítomnosti šumu a bude zdrojem možné chyby dekódování. Fanova nerovnost tuto entropii odhaduje následující úvahou: základem nejistoty je to, že nevíme, zda došlo k chybě. Pokud k ní nedošlo, žádná nejistota nezbývá. Pokud k ní došlo, je naše nejistota nejvýše rovna nutnosti slepě uhodnout, které z *nepřijatých* kódových slov bylo vysláno.

Nejprve Fanovu nerovnost vyslovíme v obecné podobě, potom ji aplikujeme na právě popsanou situaci přenosu zprávy.

Tvrzení 8.10 (Fanova nerovnost)**.** *Nechť $S, S' : \Omega \to A$ jsou náhodné veličiny. Pak*

$$\mathcal{H}\left(S \mid S'\right) \leq h\left(\mathbb{P}\left(S \neq S'\right)\right) + \mathbb{P}\left(S \neq S'\right) \log\left(|A| - 1\right).$$

Důkaz. Definujme

$$E(\omega) = \begin{cases} 0 & \text{pokud} \quad S'(\omega) = S(\omega) \\ 1 & \text{pokud} \quad S'(\omega) \neq S(\omega). \end{cases}$$

Protože je E jednoznačně určena veličinami S a S', platí $\mathcal{H}\left(S, S', E\right) = \mathcal{H}\left(S, S'\right)$. Ze součtových vzorců nyní dostáváme

$$\begin{aligned} \mathcal{H}\left(S \mid S'\right) &= \mathcal{H}\left(S, S'\right) - \mathcal{H}\left(S'\right) = \mathcal{H}\left(E, S, S'\right) - \mathcal{H}\left(S'\right) \\ &= \mathcal{H}\left(S \mid S', E\right) + \mathcal{H}\left(S', E\right) - \mathcal{H}\left(S'\right) = \\ &= \mathcal{H}\left(S \mid S', E\right) + \mathcal{H}\left(E \mid S'\right) \leq \mathcal{H}\left(S \mid S', E\right) + \mathcal{H}\left(E\right). \end{aligned}$$

Pro E jakožto binární veličinu platí

$$\mathcal{H}(E) = h\left(\mathbb{P}\left(S \neq S'\right)\right).$$

Zbývá tedy odhadnout $\mathcal{H}(S \mid S', E)$, a to jako vážený průměr entropií podmíněných hodnotami podmiňujících veličin (Tvrzení 3.10). Pokud k chybě nedošlo, entropie je nulová. Pokud k ní došlo, zbývá pro S na výběr z $|A| - 1$ hodnot, což lze shora odhadnout entropií uniformního rozdělení. Celkově:

$$\begin{aligned}
\mathcal{H}\left(S \mid S', E\right) &= \sum_{(a,i)\in s\left(P_{S',E}\right)} \mathbb{P}\left(S' = a, E = i\right) \cdot \mathcal{H}\left(S \mid (S', E\right) = (a, i)) \\
&= \sum_{(a,0)\in s\left(P_{S',E}\right)} \mathbb{P}\left(S' = a, E = 0\right) \cdot \mathcal{H}\left(S \mid (S', E\right) = (a, 0)) \\
&\quad + \sum_{(a,1)\in s\left(P_{S',E}\right)} \mathbb{P}\left(S' = a, E = 1\right) \cdot \mathcal{H}\left(S \mid (S', E\right) = (a, 1)) \\
&\leq 0 + \sum_{(a,1)\in s\left(P_{S',E}\right)} \mathbb{P}\left(S' = a, E = 1\right) \cdot \log(|A| - 1) \\
&= \mathbb{P}\left(E = 1\right) \cdot \log(|A| - 1) = \mathbb{P}\left(S \neq S'\right) \cdot \log(|A| - 1).
\end{aligned}$$

Tím je nerovnost dokázána. □

Aplikace Fanovy nerovnosti na dekódování přenosu je založena na tom, že vyslané kódové slovo je při dekódování přijatého slova v odhadnuto jako $g(v)$.

Tvrzení 8.11 (Fanova nerovnost pro kód)**.** *Nechť (M, g) je n-kód a S náhodná veličina s hodnotami v M. Označme $R = \Gamma(S)$. Potom*

$$\mathcal{H}(S \mid R) \leq 1 + \hat{\lambda}_n \log(|M| - 1).$$

Důkaz. Z Fanovy nerovnosti plyne

$$\mathcal{H}(S \mid g(R)) \leq 1 + \mathbb{P}(g(R) \neq S) \log(|M| - 1).$$

Zároveň z Lemmatu 8.8 dostáváme markovský řetězec $S \to R \to g(R)$ a podle Tvrzení 3.40(4) platí $\mathcal{H}(S \mid R) \leq \mathcal{H}(S \mid g(R))$. □

8.2.3 Shannonova mez

Definujme **Shannonovu mez** pro kanál Γ následovně:

$$C_{\mathrm{Sh}}(\Gamma) = \liminf_{n\to\infty} C_{\mathrm{Sh}}^{(n)}(\Gamma),$$

kde

$$C_{\mathrm{Sh}}^{(n)}(\Gamma) = \frac{1}{n} \sup_{S:\Omega\to A^n, S\perp\Gamma} \mathcal{I}\left(S : \Gamma(S)\right).$$

Pro dané n tedy uvažujeme maximální vzájemnou informaci mezi vstupem a výstupem kanálu pro volby vstupu nezávislé na kanálu.

Prvním hlavním výsledkem naší teorie kanálu je, že Shannonova mez je opravdu horním odhadem kapacity kanálu. Intuitivní význam tohoto výsledku je jasný. Kanálem se nepřenáší více informace, než kolik jí mezi vstupem a výstupem je.

Věta 8.12 (Shannonova věta o nepropustnosti kanálu)**.** *Pro kanál* Γ *platí* $C(\Gamma) \leq C_{\mathtt{Sh}}(\Gamma)$.

Důkaz. Buď (M_n, g_n), $n \in \mathbb{N}$, posloupnost n-kódů, jejichž maximální pravděpodobnost chyby přenosu λ_n jde k nule (k nule jde tedy i průměrná pravděpodobnost chyby přenosu).

Chceme shora omezit rychlost přenosu, tedy $\log \left|M_n\right|$, což je rovno $\mathcal{H}\left(S_n\right)$, kde za S_n volíme náhodnou veličinu rovnoměrně rozdělenou na M_n a nezávislou na kanálu. Máme tedy

$$\log |M_n| = \mathcal{H}\left(S_n\right) = \mathcal{I}(S_n : R_n) + \mathcal{H}\left(S_n \mid R_n\right) .$$

Ze základního součtového vzorce tu vidíme, že entropie zprávy má dvě složky: kromě vzájemné informace s výstupem ještě entropii vstupu za podmínky výstupu. Neformálně řečeno, máme zde triviální pozorování, že zpráva obsahuje to, co se z ní přeneslo do výstupu, a to co se do výstupu nepřeneslo. Dokazované tvrzení (jak plyne z definice kapacity a definice Shannonovy meze) říká, že nepřenesená část zprávy jde asymptoticky k nule. Ale to je přesně to, co tvrdí Fanova nerovnost pro kanál spolu s předpokladem asymptoticky nulové průměrné chyby.

Přesněji a formálně, máme

$$\begin{aligned}
\log |M_n| = \mathcal{H}\left(S_n\right) &= \mathcal{I}(S_n : R_n) + \mathcal{H}\left(S_n \mid R_n\right) \\
&\leq \mathcal{I}(S_n : R_n) + 1 + \hat{\lambda}_n \log(|M_n| - 1) \\
&\leq \mathcal{I}(S_n : R_n) + 1 + n\hat{\lambda}_n \log |A| \\
&\leq \sup_{S : \Omega \to A^n, S \perp \Gamma} \mathcal{I}(S : \Gamma(S)) + 1 + n\hat{\lambda}_n \log |A| .
\end{aligned}$$

Nerovnost na druhém řádku je Fanova. Nerovnost na třetím řádku plyne z $\left|M_n\right| \leq |A^n|$, což je důsledkem prostoty f. Třetí nerovnost předpokládá nezávislost S_n na kanálu. Vydělením n a přechodem k liminf na obou stranách nerovnosti dostáváme

$$\liminf_{n\to\infty} \frac{1}{n} \log |M_n| \leq C_{\mathtt{Sh}}(\Gamma).$$

To znamená, že pro jakékoliv $r > C_{\mathtt{Sh}}(\Gamma)$ platí $|M_n| < 2^{nr}$ pro dostatečně velká n. Tedy kanál nedosahuje rychlosti přenosu r. Proto $C(\Gamma) \leq C_{\mathtt{Sh}}(\Gamma)$. □

Právě dokázaná věta je v literatuře obvykle označována jako „slabá inverzní forma věty o kódování kanálu“ (weak converse of the Channel Coding Theorem).

„Inverzní“ proto, že mluví o nemožnosti překročit kapacitu. Zde tuto inverznost vyjadřujeme záporem ve slově „nepropustnost“. „Slabá“ proto, že nezkoumá otázku, co se stane, pokud nějakou malou chybu připustíme. Zbývá totiž otázka, jaká je kapacita kanálu, pokud asymptoticky připustíme nějakou malou chybu. tedy pokud požadujeme aby chyba nepřekročila pro velká n nějaké $\varepsilon > 0$. Lze za této podmínky rychlost přenosu asymptoticky zvýšit? To by znamenalo najít pro nějaké malé $\varepsilon > 0$ posloupnost kódů, pro něž

$$\log |M_n| > n \left(C_{\mathrm{Sh}}(\Gamma) + \delta \right),$$

pro nějaké kladné δ, přičemž chyba přenosu je pro velká n nejvýše ε.

Odpověď na tuto otázku je komplikovanější z závisí na tom, jak si definujeme chybu přenosu. Pokud ji definujeme pomocí maxima, či aritmetického průměru chyb dekódování jednotlivých zpráv, pak nelze dosáhnout zlepšení. Tento výsledek se nazývá v literatuře silnou konverzí.

Na druhou stranu, pokud by chybou byla pravděpodobnost jevu $g(R) \neq S$, při náhodné S, pak by kapacita překročit šla. Chyba přenosu by zde také byla průměrem chyb při poslání jednotlivých zpráv, ale byl by to průměr vážený, který by zprávám, které se špatně dekódují přiřazoval sice kladnou, ale velmi malou pravděpodobnost ve srovnání s těmi, které se dekódují dobře.

Zde je třeba se tedy chovat opatrně z hlediska formulace takového tvrzení. My se podrobnostmi silné konverzní věty v této přednášce nezabýváme.

8.3 Kanál bez paměti

Kanál definovaný obecně v předchozí kapitole je těžké zkoumat, navíc představuje málo realistický případ, kdy šum bere v úvahu celou zprávu. Nyní se tedy omezíme na speciální případ, takzvaný **časově invariantní diskrétní kanál bez paměti**. Takový kanál působí odděleně a nezávisle na jednotlivá písmena zprávy (je „bez paměti“) a navíc stále stejným způsobem (je „časově invariantní“).

Dále budeme „časově invariantní diskrétní kanál bez paměti“ označovat stručně jako „kanál bez paměti“. Pro tento kanál umíme dokázat, že se jeho kapacita rovná Shannonově mezi a také můžeme lépe popsat, jak Shannonovu mez najít.

V souladu s neformálním popisem je kanál bez paměti definován pomocí i.i.d. procesu $(\Gamma_i)_{i\in\mathbb{N}}$, který za své hodnoty přijímá zobrazení z A do B. Kanál tedy náhodně vylosuje, který prvek se kam zobrazí. Formálně tedy máme náhodné veličiny

$$\Gamma_i : \Omega \to B^A, i \in \mathbb{N}.$$

Každá taková náhodná veličina tedy předepisuje, jak bude reagovat na konkrétní vstup z A. Můžeme tedy odvodit příslušné veličiny $\Gamma_i(a) : \Omega \to B$, $a \in A$, předpisem

$$\Gamma_i(a)(\omega) = (\Gamma_i(\omega))(a), \qquad \omega \in \Omega.$$

Tyto veličiny popisují výstup náhodného zobrazení $\Gamma_i(\omega) : A \to B$, při vstupu $a \in A$.

Připomeňme, že obecný kanál byl definován jako zobrazení $\Gamma : A^+ \times \Omega \to B^+$, a kanál bez paměti jsme definovali proces $\left(\Gamma_i\right)_{i\in\mathbb{N}}$, přičemž obě definice jsou v případě kanálu bez paměti spojeny předpisem

$$\Gamma(u) = \Gamma_0(u_0)\Gamma_1(u_1)\dots\Gamma_{n-1}(u_{n-1}) \qquad n \in \mathbb{N}, u \in A^n,$$

kde u_i značí jako obvykle písmena, ze kterých sestává u. Neboli

$$\Gamma(S) = \Gamma_0(\pi_0(S))\Gamma_1(\pi_1(S))\dots\Gamma_{n-1}(\pi_{n-1}(S)) \qquad n \in \mathbb{N}, u \in A^n,$$

kde $\pi_i : A^n \to A$ jsou projekce. Takto definovaný kanál je kanálem ve smyslu Definice 8.2.

Přechod od Γ_i k $\Gamma_i(a)$ proběhl pomocí dosazení prvku a do náhodného zobrazení Γ_i, což je vhodné chápat jako transformaci původní náhodné veličiny. Konkrétně, pro fixní $a \in A$, definujeme zobrazení $\texttt{eval}_a : B^A \to B$, předpisem $\texttt{eval}_a(z) = z(a)$. Z definice pak plyne

$$\Gamma_i(a) = \texttt{eval}_a(\Gamma_i).$$

Díky tomuto pozorování dostáváme nezávislost jednotlivých výstupů kanálu, která je okamžitým důsledkem Lemmatu 8.7.

Lemma 8.13. *Pro kanál bez paměti Γ, $n \geq 1$, $u \in A^n$, platí, že $(\Gamma_i(u_i))$, $i < n$, jsou vzájemně nezávislé veličiny, tj. pro každé $v \in B^n$,*

$$P_{\Gamma(u)}(v) = \prod_{i=0}^{n-1} P_{\Gamma_i(u_i)}(v_i),$$

či

$$\mathbb{P}\left(\Gamma(u) = (v)\right) = \prod_{i=0}^{n-1} \mathbb{P}\left(\Gamma_i(u_i) = (v_i)\right).$$

Zdůrazněme ovšem na tomto místě, že neplatí obrácená úvaha, tedy že by z nezávislosti $(\Gamma_i(u_i))$, $i < n$, pro všechna $n \geq 1$ a všechna $u \in A^n$ už plynula nezávislost veličin Γ_i. Jednoduše řečeno, z nezávislosti transformací se obecně nedá usuzovat nezávislost původních veličin. My ovšem tento silnější typ nezávislosti, tedy nezávislost veličin Γ_i, $i \in \mathbb{N}$, předpokládáme.

Připomeňme ještě druhý důležitý předpoklad, totiž že jsou Γ_i stejně rozdělené. Stejné rozdělení říká, že libovolné pevné zobrazení $z : A \to B$ je jako hodnota Γ_i stejně pravděpodobné pro libovolné i.

Pokud znovu uvážíme, že pro fixní a, je $\Gamma_i(a) = \texttt{eval}_a(\Gamma_i)$, dostáváme aplikací Lemmatu 8.9 následující lemma.

Lemma 8.14. *Pro kanál bez paměti Γ, $a \in A$, jsou veličiny $(\Gamma_i(a))$, $i \in \mathbb{N}$, stejně rozdělené.*

V souladu s naším zacílením na jednotlivé instance přenosu, tedy přenos jednotlivých písmen, definujeme následující pojem. Řekneme, že rozdělení P na množině $A \times B$ je **přenosové rozdělení** kanálu bez paměti Γ, pokud pro všechna $a \in A$, $b \in B$ platí

$$P(a,b) = P_A(a) \cdot P_{\Gamma_0(a)}(b),$$

kde P_A je marginální rozdělení P, tedy rozdělení vstupních písmen, k nimž kanál definuje rozdělení písmen výstupních. Pokud má dvojice náhodných veličin (X,Y), kde X má hodnoty v A a Y má hodnoty v B, sdružené rozdělení rovné přenosovému rozdělení kanálu, řekneme jednoduše, že tato dvojice má přenosové rozdělení.

Přenosové rozdělení je definováno opravdu tak, aby popisovalo sdružené rozdělení jednopísmenkového vstupu a výstupu pro daný kanál. Přímo z definice totiž plyne, že pokud je S náhodná veličina s hodnotami v A, která je nezávislá na kanálu, pak má dvojice $(S, \Gamma_0(S))$, neboli $(S_0, \Gamma(S_0))$ přenosové rozdělení. (Pozor obrácená implikace neplatí. Pokud má dvojice (X,Y) přenosové rozdělení, nemusí být $Y = \Gamma_0(X)$).

Ve zbytku kapitoly bude S značit náhodnou veličinu s hodnotami v A^n nezávislou na kanálu Γ bez paměti, a symbolem S_i, $i = 0, \ldots, n-1$, budeme označovat jednotlivá písmena S. Veličinu S tedy můžeme chápat jako náhodný vektor $S_{[0..n)}$. Podobně pro $R : \Omega \to B^n$, přičemž R bude značit $\Gamma(S)$, kde S bude vždy zřejmé z kontextu. Pro kanál bez paměti tedy dostáváme $R_i = \Gamma_i(S_i)$.

Lemma 8.15. *Pro každé $i < n$ je S_i nezávislá na kanálu. Zároveň je S_i nezávislá na $\Gamma_j(a)$, pro všechna $j \in \mathbb{N}$, $a \in A$.*

Důkaz. Nezávislost veličiny $S : \Omega \to A^n$ znamená, že je S a $\Gamma(u)$ nezávislá pro všechna $u \in \mathbb{N}$. Ovšem $S_i = \pi_i(S)$ je transformací veličiny S. Z Lemmatu 8.6 plyne, že také S_i a $\Gamma(u)$ jsou nezávislé veličiny, pro všechna $u \in A^n$. Tedy S_i je také nezávislá na kanálu. Pro pevné $j \in \mathbb{N}$ a $a \in A$, uvažujme slovo ua pro nějaké $u \in A^j$. Potom S_i a $\Gamma(ua)$ jsou nezávislé. Ovšem $\Gamma_j(a) = \pi_j(\Gamma(ua))$ je transformací veličiny $\Gamma(ua)$. Musí být tedy také nezávislá na S_i. □

Lemma 8.16. *Buďte S a S' náhodné veličiny s hodnotami A, které mají stejné rozdělení a jsou nezávislé na kanálu. Potom mají dvojice $(S, \Gamma_i(S))$ a $(S', \Gamma_j(S'))$, $i,j \in \mathbb{N}$, stejná rozdělení a platí*

$$\mathcal{I}\left(S : \Gamma_i(S)\right) = \mathcal{I}\left(S' : \Gamma_j(S')\right).$$

Důkaz. Předpokládejme, že S, S' a i, j, splňují předpoklady lemmatu. Pak

$$P_{S,\Gamma_i(S)}(a,b) = P_S(a) \cdot P_{\Gamma_i(a)}(b) = P_{S'}(a) \cdot P_{\Gamma_j(a)}(b) = P_{S',\Gamma_j(S')}(a,b),$$

kde první a poslední rovnost plyne z nezávislosti veličin na kanálu, druhá pak ze stejného rozdělení S a S' a $\Gamma_i(a)$ a $\Gamma_j(a)$.

Rovnost sdružených rozdělení zaručuje také rovnost marginálních rozdělení, mimo jiné, $P_{\Gamma_i(S)}$ je totožné s $P_{\Gamma_j(S)}$. Toho využijeme v dokončení důkazu, kde rovnost vzájemných informací bude přímým důsledkem rovnosti příslušných rozdělení,

$$\begin{aligned}\mathcal{I}(S : \Gamma(S)) &= \mathcal{H}\left(P_S\right) + \mathcal{H}\left(P_{\Gamma_i(S)}\right) - \mathcal{H}\left(P_{S,\Gamma_i(S)}\right) \\ &= \mathcal{H}\left(P_{S'}\right) + \mathcal{H}\left(P_{\Gamma_j(S')}\right) - \mathcal{H}\left(P_{S',\Gamma_j(S')}\right) = \mathcal{I}\left(S' : \Gamma_j(S')\right) .\end{aligned}$$

□

Tvrzení 8.17. *Pokud jsou $\left(S_i\right)_{i\leq n-1}$ vzájemně nezávislé, pak jsou i $\left(S_i, R_i\right)_{i\leq n-1}$ vzájemně nezávislé a $\left(R_i\right)_{i\leq n-1}$ jsou vzájemně nezávislé.*

Důkaz. Dokažme nezávislost souboru $\left(S_i, R_i\right)_{i\leq n-1}$. Pro $u \in A^n$ a $v \in B^n$ platí

$$\begin{aligned}P_{S,R}(u,v) = P_S(u)P_{\Gamma(u)}(v) &= \prod_{i=0}^{n-1} P_{S_i}(u_i) \prod_{i=0}^{n-1} P_{\Gamma_i(u_i)}(v_i) = \prod_{i=0}^{n-1} P_{S_i}(u_i)P_{\Gamma(u_i)}(v_i) \\ &= \prod_{i=0}^{n-1} P_{S_i,R_i}(u_i, v_i),\end{aligned}$$

kde první rovnost je důsledkem Lemmatu 8.4, druhá rovnost plyne z nezávislosti souboru $(S_i)_{i<n}$ a souboru $(\Gamma_i(u_i))_{i<n}$ viz Lemma 8.13. Veličiny $(S_i, R_i)_{i<n}$ jsou tedy vzájemně nezávislé. Přechodem k druhým projekcím pak dostaneme, že je nezávislý i soubor veličin $\left(R_i\right)_{i\leq n-1}$, viz Lemma 8.8. □

Důsledek 8.18. *Buď S buď náhodná veličina s hodnotami v A^n nezávislá na bezpaměťovém kanálu Γ. Pak platí:*

$$\mathcal{H}(R \mid S) = \sum_{i=0}^{n-1} \mathcal{H}\left(R_i \mid S_i\right) .$$

Dále

$$\mathcal{I}(S : R) \leq \sum_{i=0}^{n-1} \mathcal{I}\left(S_i : R_i\right) .$$

a rovnost nastane právě tehdy když jsou veličiny R_i, $i < n$, vzájemně nezávislé.

Důkaz. Buď $\omega \in s(S, R)$ a označme $u = S(\omega)$ a $v = R(\omega)$. Tedy $S_i(\omega) = u_i$ a $R_i(\omega) = v_i$. Pak z Lemmatu 8.17 dostáváme

$$\begin{aligned}\mathfrak{I}_{R|S}(\omega) &= -\log \mathbb{P}(R = v \mid S = u) = -\log \prod_{i=0}^{n-1} \mathbb{P}\left(R_i = v_i \mid S_i = u_i\right) \\ &= \sum_{i=0}^{n-1} -\log \mathbb{P}\left(R_i = v_i \mid S_i = u_i\right) = \sum_{i=0}^{n-1} \mathfrak{I}_{R_i|S_i}(\omega).\end{aligned}$$

Přechodem ke střední hodnotě:

$$\mathcal{H}(R \mid S) = \mathbb{E}\left(\mathfrak{I}_{R|S}\right) = \mathbb{E}\left(\sum_{i=0}^{n-1} \mathfrak{I}_{R_i|S_i}\right) = \sum_{i=0}^{n-1} \mathbb{E}\left(\mathfrak{I}_{R_i|S_i}\right) = \sum_{i=0}^{n-1} \mathcal{H}\left(R_i \mid S_i\right).$$

Pro vzájemnou informaci dostáváme,

$$\begin{aligned}\mathcal{I}(S : R) = \mathcal{H}(R) - \mathcal{H}(R \mid S) &\leq \sum_{i=0}^{n} \mathcal{H}\left(R_i\right) - \sum_{i=0}^{n} \mathcal{H}\left(R_i \mid S_i\right) \\ &= \sum_{i=0}^{n} \mathcal{I}\left(S_i : R_i\right).\end{aligned}$$

V předchozím výpočtu jsme sdruženou entropii $\mathcal{H}(R)$ odhadli pomocí součtu entropií jednotlivých veličin R_i, $i < n$. Rovnost nastane právě tehdy když jsou R_i vzájemně nezávislé. □

Tvrzení 8.19. *Pro bezpaměťový kanál platí*

$$C_{\mathrm{Sh}}(\Gamma) = C_{\mathrm{Sh}}^{(1)}(\Gamma) = C_{\mathrm{Sh}}^{(n)}(\Gamma),$$

$n \in \mathbb{N}$.

Důkaz. Stačí dokázat $C_{\mathrm{Sh}}^{(1)}(\Gamma) = C_{\mathrm{Sh}}^{(n)}(\Gamma)$, $n \in \mathbb{N}$, druhá rovnost pak vyplyne z definice. Nejprve si uvědomme, že $\Gamma(a) = \Gamma_0(a)$, pro všechna $a \in A$. Tedy

$$C_{\mathrm{Sh}}^{(1)}(\Gamma) = \sup_{S:\Omega\to a, S\perp\Gamma} \mathcal{I}\left(S : \Gamma_0(S)\right).$$

Nyní uvažujme $S : \Omega \to A^n$, která je nezávislá na kanálu. Z Lemmatu 8.15 plyne, že je veličina S_i nezávislá na kanálu a Z Lemmatu 8.16 dostáváme, že $\mathcal{I}\left(S_i : R_i\right)$ a $\mathcal{I}\left(S_i : \Gamma_0(S_i)\right)$ jsou totožné. Ze zmíněné nezávislosti také plyne, že S_i je veličinou, která je brána v potaz v definici $C_{\mathrm{Sh}}^{(1)}(\Gamma)$. Souhrnem,

$$\mathcal{I}\left(S_i : R_i\right) = \mathcal{I}\left(S_i : \Gamma_0(S_i)\right) \leq C_{\mathrm{Sh}}^{(1)}(\Gamma).$$

Z Důsledku 8.18 dostáváme

$$\mathcal{I}(S : R) \leq \sum_{i=0}^{n-1} \mathcal{I}\left(S_i : R_i\right) = \sum_{i=0}^{n-1} \mathcal{I}\left(S_i : \Gamma_0(S_i)\right) \leq n \cdot C_{\mathrm{Sh}}^{(1)}(\Gamma).$$

Vydělením přes n a přechodem k supremu přes všechna S nezávislá na kanálu dostaneme, že $C_{\mathrm{Sh}}^{(n)}(\Gamma) \leq C_{\mathrm{Sh}}^{(1)}(\Gamma)$.

Nyní vezměme libovolné $S_0 : \Omega \to A$, které je nezávislé na kanálu, a jeho vzájemně nezávislé kopie S_i, $i = 0, \ldots, n-1$, takové, že jejich konkatenace $S = S_{[0..n)}$ je také nezávislá na kanálu.

Veličiny $(S_i)_{i\leq n-1}$ jsou stejně rozdělené. Z Lemmatu 8.16 plyne, že všechny vzájemné informace $\mathcal{I}\left(S_i : R_i\right)$ jsou stejné. Z nezávislosti $(S_i)_{i\leq n-1}$ dostáváme nezávislost $(\Gamma_i(S_i))$. Z důsledku 8.18 pak plyne druhá rovnost v následujícím výpočtu,

$$\mathcal{I}\left(S_0 : \Gamma_0(S_0)\right) = \frac{1}{n}\sum_{i=0}^{n-1}\mathcal{I}\left(S_i : R_i\right) = \frac{1}{n}\mathcal{I}\left(S : \Gamma(S)\right) \leq C_{\mathtt{Sh}}^{(n)}(\Gamma).$$

Protože S_0 bylo libovolné, dostáváme $C_{\mathtt{Sh}}^{(1)}(\Gamma) \leq C_{\mathtt{Sh}}^{(n)}(\Gamma)$. □

8.4 Dosažení kapacity kanálu

Základní myšlenka dosažení kapacity kanálu je postavena na dekódování vzhledem ke společně typické množině. Než vysvětlíme tuto myšlenku, připomeneme nejprve samoopravné kódy pro binární symetrický (bezpaměťový) kanál. Ty jsou již konkrétní aplikací kódu založeného na typické množině, ačkoliv se tato terminologie v řešení přímo neobjevuje.

Binární symetrický kanál s pravděpodobností chyby $e \in [0,1]$ funguje tak, že každé odeslané písmeno ze vstupní abecedy $\{0,1\}$ přenese bezchybně s pravděpodobností $1-e$. To jinými slovy znamená, že s pravděpodobností e změní odesílaný symbol na symbol opačný, tedy 0 na 1 a 1 na 0. Pravděpodobnosti výstupu pro dané vstupy se obvykle vyjadřují maticí

$$\Gamma \sim \begin{pmatrix} 1-e & e \\ e & 1-e \end{pmatrix}.$$

Další úvahy platí pro případ $e < \frac{1}{2}$, tedy pro variantu, kdy bezchybný přenos převažuje nad chybovým. V tomto případě ochrana před chybami spočívá ve volbě kódových slov, která jsou od sebe dostatečně daleko v takzvané Hammingově vzdálenosti. Vzhledem k této vzdálenosti, je *koule* $B_d(u)$ se středem u a poloměrem d, množina slov, které se od u liší na méně než d místech. Střední hodnota počtu chyb při přenosu slova délky n je en. Ze zákona velkých čísel plyne, že pravděpodobnost více než $n(e+\varepsilon)$ chyb konverguje k nule pro $n \to \infty$. Z toho plyne, že pokud pro velké n posíláme slovo $u \in A^n$, bude s velkou pravděpodobností výstup $\Gamma(u)$ náležet kouli $B_d(u)$ pro $d_n = n(e+\varepsilon)$. Malé pravděpodobnosti chyby při přenosu m zpráv lze tedy dosáhnout pokud zvolíme vzájemně disjunktní koule s poloměrem d_n. Kódovými slovy budou středy těchto koulí a dekódovací zobrazení zobrazí celou kouli na svůj střed. Takovéto dekódování se nazývá samoopravným kódem.

Z hlediska rychlosti přenosu a kapacity je základní otázkou samoopravných kódů nalezení maximální množiny koulí, které se dají disjunktně vměstnat do prostoru A^n.

Objem koule s poloměrem d_n, neboli její počet prvků, lze aproximovat jako

$$1 + n + \cdots + \binom{n}{d_n} \approx \binom{n}{d_n} \approx \exp n \cdot h\left(d_n/n\right) = \exp n \cdot h\left(e+\varepsilon\right),$$

kde pro předposlední aproximaci potřebujeme nerovnost $e + \varepsilon \leq 1/2$.

Pokud tedy odhlédneme od geometrie prostoru a uděláme zjednodušenou úvahu, tedy celkový objem prostoru vydělíme objemem jednotlivých části, dostaneme, že takových disjunktních koulí by mohlo být až

$$\exp n / \exp n \cdot h\left(e+\varepsilon\right) = \exp n \cdot \left(1 - h\left(e+\varepsilon\right)\right) \approx \exp n \cdot C_{\mathrm{Sh}}\left(\Gamma\right).$$

Poslední aproximace zanedbává ε. Připomeňme, že $C_{\mathrm{Sh}}(\Gamma)$ je rovna $1 - h(e)$.

Tento heuristický argument dává ale ve skutečnosti jen horní mez pro rychlost přenosu pro tento typ kódů. Pokud bychom chtěli ukázat, že lze této meze dosáhnout, je třeba lehce pozměnit naše požadavky. Při dekódování musíme dodržet malou

chybu, tu lze dodržet nejen v momentu, kdy budou koule v našem výběru navzájem disjunktní, ale také tehdy, když se vzájemně budou překrývat, ale tyto průniky budou malé relativně k velikosti koule. Při takovém požadavku již lze do prostoru vměstnat dostatek koulí.

Zvolme $\varepsilon < 1/4$, $n \in \mathbb{N}$, $d = n(e+\varepsilon)$ a buď K počet prvků v kouli o poloměru d. Tato kardinalita jistě nezávisí na zvoleném středu koule.

Nyní vybírejme koule s malým vzájemným průnikem. Konkrétně můžeme postupně přidávat $B_d(u^{(i)})$ tak, že velikost průniku $B_d(u^{(i+1)})$ se sjednocením všech předchozích koulí $B_d(u^{(j)})$, $j \leq i$ je menší než $\frac{\varepsilon}{2}K$.

Postupným přidáváním dostaneme M koulí. Uvažujme případ, kdy již nejde přidat žádná koule tak, abychom neporušili podmínku o velikosti průniku. Označme sjednocení všech koulí symbolem V. Kódovými slovy našeho kódu budou středy koulí a dekódovat budeme všechny prvky z příslušné koule na její střed. Pokud bude slovo náležet do více koulí, má přednost ta s nižším pořadovým číslem. Pokud vyšleme zprávu, může dojít ke dvěma chybám, buď nastane příliš mnoho chyb a výstup bude mimo danou kouli, nebo bude i v nějaké předchozí kouli a bude tedy mylně dekódován jako jiný střed. Pravděpodobnost první chyby jde k nule s rostoucím n. Pravděpodobnost druhé chyby je omezena podmínkou na výběr koulí. Nemůže přesáhnout $\varepsilon/2$. Zvolme n dost velké, aby pravděpodobnost prvního druhu chyby byla menší než $\varepsilon/2$. Celková chyba tedy bude menší než ε.

Co se týče velikosti M, ta souvisí s velikostí V. Jelikož předpokládáme, že již nelze přidat žádnou další kouli, musí pro všechna nevybraná slova $u \in A^n$ platit, že průnik $B_d(u)$ a V má kardinalitu alespoň $\frac{\varepsilon}{2}K$. Pro ta vybraná to platí triviálně také (překryv je v tomto případě přesně K). Pro velikost překryvů tedy dostáváme

$$|A^n|\frac{\varepsilon}{2}K \leq \sum_{u\in A^n} \left|B_d(u) \cap V\right| \leq K|V|,$$

kde druhá nerovnost je velkorysým odhadem založeným na pozorování, že každý bod V může být pokryt nejvýše K různými koulemi (definovanými svými středy uvnitř koule se středem v pokrývaném bodě). Dostáváme $\frac{\varepsilon}{2}|A^n| \leq |V|$. Zároveň ale $|V| \leq M \cdot K$, a celkově tedy

$$M \geq \frac{\varepsilon}{2}|A^n|/K \approx \frac{\varepsilon}{2}2^n \exp -n(h(e+\varepsilon)) = \frac{\varepsilon}{2}\exp n(1-h(e+\varepsilon)).$$

Tento výraz roste exponenciálně rychlostí $\exp n(1-h(e+\varepsilon))$. Pokud půjdeme s požadovanou chybou k nule, dostaneme požadovanou kapacitu přenosu $C_{\text{Sh}}(\Gamma) = 1 - h(e)$. (Zde ignorujeme multiplikativní faktor $\varepsilon/2$, a vynecháváme tak detaily diskuse o vztahu mezi ε v exponentu a ε v tomto multiplikativním členu.)

V následující kapitole vysvětlíme tvorbu kódu pro obecný bezpaměťový kanál opřený o myšlenku typického přenosu. Zpětně se dá hledět na výběr středů jako na výběr z podmnožiny typických slov pro rovnoměrné rozložení zdroje. V takovém případě u BSC nastává následující

- výstup je rozložen rovnoměrně,

- všechna vstupní slova jsou typická,
- všechna výstupní slova jsou typická,
- typické výstupy pro vstup u náleží do sférické vrstvy $B_{n(e+\varepsilon)}(u) \setminus B_{n(e-\varepsilon)}(u)$.

Třetí bod naznačuje, že jsme měli v předchozím kódu používat nikoliv koule, ale sférické vrstvy. To jsme opravdu mohli, při zachování stejné chyby přenosu. To vyplývá z faktu, že pro dlouhá slova u je výstup $\Gamma(u)$ s velkou pravděpodobností nejen uvnitř výše zvolené koule, ale dokonce právě ve zmíněné sférické vrstvě. To odpovídá tomu, že je nejen nepravdpodobný příliš velký počet chyb, ale také příliš malá chybovost.

Zároveň musíme také upozornit na fakt, že typické množiny se sférickými vrstvami, které odpovídají frekvenčně typickým množinám, zcela neshodují. Zásadní je ale fakt, že do každé sférické vrstvy jsme schopni vměstnat typickou množinu a naopak.

8.4.1 Typické vstupy a výstupy

Zafixujme konkrétní i.i.d. proces $(S_i)_{i=1}^{\infty}$, a uvažujme odpovídající výstupní proces $R_i := \Gamma(S_i)$, $i \in \mathbb{N}$. Předpokládáme, že S je nezávislé na kanálu, z čehož plyne vlastnost i.i.d. pro procesy $(R_i)_{i=1}^{\infty}$ i sdružený proces $((S_i, R_i))_{i=1}^{\infty}$.

Budeme pracovat s typickými množinami dle Definice 4.5 pro vstup, výstup a sdruženě pro dvojici vstup-výstup. Tím myslíme množiny

$$\begin{aligned}
\mathcal{M}_{\varepsilon}^{n}(S) &= \left\{u \in A^n \mid \exp -n(\mathcal{H}(S)+\varepsilon) \le \mathbb{P}\left(S_{[0..n)} = u\right) \le \exp -n(\mathcal{H}(S)-\varepsilon)\right\},\\
\mathcal{M}_{\varepsilon}^{n}(R) &= \left\{v \in B^n \mid \exp -n(\mathcal{H}(R)+\varepsilon) \le \mathbb{P}\left(R_{[0..n)} = v\right) \le \exp -n(\mathcal{H}(R)-\varepsilon)\right\},\\
\mathcal{M}_{\varepsilon}^{n}(S,R) &= \{(u,v) \in A^n \times B^n \mid\\
&\qquad \exp -n(\mathcal{H}(R,S)+\varepsilon) \le \mathbb{P}\left(S_{[0..n)} = u, R_{[0..n)} = v\right) \le \exp -n(\mathcal{H}(R,S)-\varepsilon)\right\}.
\end{aligned}$$

Pro sdružený proces pak definujeme takzvanou společně typickou množinu:

$$\overline{\mathcal{M}_{\varepsilon}^{n}}(R,S) = \left(\mathcal{M}_{\varepsilon}^{n}(S) \times \mathcal{M}_{\varepsilon}^{n}(R)\right) \cap \mathcal{M}_{\varepsilon}^{n}(S,R).$$

Tato společně typická množina je podmnožinou $\mathcal{M}_{\varepsilon}^{n}(S,R)$, která po svých prvcích, coby dvojicích vstup a výstup, požaduje nejen typičnost páru pro sdružený proces, ale také typičnost samotného vstupu a výstupu vzhledem k příslušným procesům S a R.

8.4.2 Kódování a dekódování na základě typického přenosu

Nyní můžeme vysvětlit, jak pro dané ε a n budeme konstruovat kód, tedy jak budeme vybírat kódová slova a jak je budeme dekódovat.

Zaměříme se na typická slova, u kterých je navíc velká pravděpodobnost, že budou se svým obrazem tvořit společně typický pár. Definujme tedy

$$\hat{M}_\varepsilon^n := \left\{ u \in A^n \mid u \in \mathcal{M}_\varepsilon^n \wedge \mathbb{P}\left((u, \Gamma(u)) \in \overline{\mathcal{M}_\varepsilon^n}(S, R) \right) > 1 - \frac{\varepsilon}{2} \right\}.$$

Konstrukce kódu je následující. Pro $k > 0$ vybíráme vstupní slovo $u^{(k)}$ z $\hat{M}_\varepsilon^n$ tak, aby platilo

$$\mathbb{P}\left(\Gamma\left(u^{(k)}\right) \in \bigcup_{j<k} V_j \right) \leq \frac{\varepsilon}{2}, \tag{1}$$

kde V_i jsou množiny typických výstupů při typickém přenosu slova $u^{(i)}$, tedy

$$V_i := \left\{ v \in B^n \mid (u^{(i)}, v) \in \overline{\mathcal{M}_\varepsilon^n}(S, R) \right\},$$

a to tak dlouho, dokud takové slovo existuje. Dekódování je definováno tak, že slovo v dekódujeme na $u^{(i)}$, kde i je nejmenší takové, že $v \in V_i$ (pokud takové i neexistuje, dekódování selže).

Základní myšlenka dekódování je jasná: dekódujeme na vstupy, pro které je přijatý výstup typický. Na průnicích množin V_i přitom uplatníme pravidlo „Kdo dřív přijde, ten dřív mele“. Prioritu bude mít v takovém případě kódové slovo s nejmenším indexem. Jinak řečeno, na $u^{(i)}$ jsou dekódována právě výstupní slova z množiny

$$V_i \setminus \bigcup_{j=1}^{i-1} V_j.$$

Konstrukce snadno poskytuje omezení pravděpodobnosti chyby dekódování parametrem ε. Chyba totiž může být způsobena dvěma důvody: prvním z nich je netypický přenos kódového slova kanálem, kv;li kterému nebude výstup ležet v příslušném V_i, druhým je konflikt s jiným (dříve vybraným) kódovým slovem. Chyba prvního typu je prostřednictvím definice $\hat{M}_\varepsilon^n$ výslovně shora omezena hodnotou $\varepsilon/2$. Chyba druhého typu je rovněž výslovně omezena hodnotou $\varepsilon/2$, tentokrát požadavkem (1) malého překryvu. Klíčovou otázkou pro dosažení kapacity tedy je, zda tyto požadavky umožňují zkonstruovat kódy požadované velikosti, kterou prozkoumáme v následující části.

8.4.3 Velikost kódu

Nechť je dáno kladné $\varepsilon < 1/2$. Zajímá nás, jak velké je N, tedy jaká je velikost kódu zkonstruovaného v předchozí části. Odhad vyplyne z úvah o množině $V = \bigcup_{j \leq N} V_j$. Jedná se o množinu typických hodnot náhodné veličiny $R_{[0..n)} = \Gamma\left(S_{[0..n)}\right)$, kterou budeme měřit její pravděpodobností, tedy hodnotou $\mathbb{P}\left(R_{[0..n)} \in V\right)$.

Proveďme nejprve dolní odhad $\mathbb{P}\left(R_{[0..n)} \in V\right)$. Ten primárně vyplývá z faktu, že konstrukce kódu se zastaví až ve chvíli, kdy kódovou množinu nelze rozšířit. To je situace, kdy pro všechna $u \in \hat{M}_\varepsilon^n$ platí

$$\mathbb{P}\left(\Gamma(u) \in V\right) > \frac{\varepsilon}{2}.$$

Pro slova ze zkonstruovaného kódu to platí díky definici $\hat{M}_\varepsilon^n$, podle které padne výstup s pravděpodobností alespoň $1 - \varepsilon/2 > \varepsilon/2$ do příslušné množiny V_i. Pro slova mimo kód to platí z podmínky konstrukce a faktu, že kód již nelze rozšířit. Dostáváme

$$\begin{aligned}
\mathbb{P}\left(R_{[0..n)} \in V\right) &\geq \sum_{u \in \hat{M}_\varepsilon^n} \mathbb{P}\left(R_{[0..n)} \in V, S_{[0..n)} = u\right) \\
&= \sum_{u \in \hat{M}_\varepsilon^n} \mathbb{P}\left(R_{[0..n)} \in V \mid S_{[0..n)} = u\right) \mathbb{P}\left(S_{[0..n)} = u\right) \\
&= \sum_{u \in \hat{M}_\varepsilon^n} \mathbb{P}\left(\Gamma(u) \in V\right) \mathbb{P}\left(S_{[0..n)} = u\right) \\
&> \sum_{u \in \hat{M}_\varepsilon^n} \frac{\varepsilon}{2} \mathbb{P}\left(S_{[0..n)} = u\right) = \frac{\varepsilon}{2} \mathbb{P}\left(S_{[0..n)} \in \hat{M}_\varepsilon^n\right).
\end{aligned}$$

Musíme tedy ještě odhadnout pravděpodobnost množiny $\hat{M}_\varepsilon^n$. To není překvapivé, protože kódová slova vybíráme pouze z ní, a pokud by např. byla prázdná, byl by prázdný i konstruovaný kód. Také množinu $\hat{M}_\varepsilon^n$ odhadujeme podle pravděpodobnosti, že do ní padne hodnota $S_{[0..n)}$. Tuto pravděpodobnost si vyjádříme jako

$$\mathbb{P}\left(S_{[0..n)} \in \hat{M}_\varepsilon^n\right) = \mathbb{P}\left(S_{[0..n)} \in \hat{M}_\varepsilon^n \mid S_{[0..n)} \in \mathcal{M}_\varepsilon^n(S)\right) \cdot \mathbb{P}\left(S_{[0..n)} \in \mathcal{M}_\varepsilon^n(S)\right).$$

Jde tedy o pravděpodobnost typické množiny veličiny S a pravděpodobnost, že pro $u \in \mathcal{M}_\varepsilon^n(S)$ bude platit $\mathbb{P}\left((u, \Gamma(u)) \notin \overline{\mathcal{M}_\varepsilon^n}(S, R)\right) < \varepsilon/2$, viz definici $\hat{M}_\varepsilon^n$. Tvrzení o typické množině zaručuje, že pro n dostatečně velká budou pravděpodobnosti

$$\mathbb{P}\left(S_{[0..n)} \notin \mathcal{M}_\varepsilon^n(S)\right), \mathbb{P}\left(R_{[0..n)} \notin \mathcal{M}_\varepsilon^n(R)\right), \mathbb{P}\left(\left(S_{[0..n)}, R_{[0..n)}\right) \notin \mathcal{M}_\varepsilon^n(S, R)\right)$$

menší než $\frac{\varepsilon}{24}$. Pro taková n bude platit

$$\mathbb{P}\left(\left(S_{[0..n)}, R_{[0..n)}\right) \notin \overline{\mathcal{M}_\varepsilon^n}(S, R)\right) < \frac{\varepsilon}{8}$$

a

$$\mathbb{P}\left(\left(S_{[0..n)}, R_{[0..n)}\right) \notin \overline{\mathcal{M}_\varepsilon^n}(S, R) \mid S_{[0..n)} \in \mathcal{M}_\varepsilon^n(S)\right) < \frac{\varepsilon/8}{1 - \varepsilon/24} < \frac{\varepsilon}{4}.$$

Tvrdíme, že pro taková n zkoumaná pravděpodobnost splňuje

$$\mathbb{P}\left(S_{[0..n)} \in \hat{M}_\varepsilon^n \mid S_{[0..n)} \in \mathcal{M}_\varepsilon^n(S)\right) \geq \frac{1}{2}.$$

Důkaz má podobu argumentu, který vyslovíme jako obecné tvrzení.

Lemma 8.20. *Mějme náhodné jevy $U, W \subset \Omega$, kde U je disjunktním sjednocením možných jevů U_i, $i \in I$, pro nějakou konečnou množinu indexů I. Buď $r > 1$ a I' podmnožina I, která obsahuje ty indexy, pro které platí*

$$\mathbb{P}\left(W \mid U_i\right) \geq r \cdot \mathbb{P}\left(W \mid U\right).$$

Pokud $\mathbb{P}\left(W \mid U\right) > 0$, pak $\mathbb{P}\left(\bigcup_{i\in I'} U_i \mid U\right) \leq \frac{1}{r}$.

Důkaz. Důkaz vychází z faktu, že pravděpodobnost $\mathbb{P}\left(W \mid U\right)$ je váženým průměrem hodnot $\mathbb{P}\left(W \mid U_i\right)$ a hodnotu průměru, při nezáporných příspěvcích, nelze překročit r-krát více než v $1/r$ případech. Konkrétně

$$\begin{aligned}\mathbb{P}\left(W \mid U\right) &= \sum_{i\in I} \mathbb{P}\left(W \mid U_i\right)\mathbb{P}\left(U_i \mid U\right) \geq \sum_{i\in I'} \mathbb{P}\left(W \mid U_i\right)\mathbb{P}\left(U_i \mid U\right)\\ &\geq \sum_{i\in I'} r \cdot \mathbb{P}\left(W \mid U\right)\mathbb{P}\left(U_i \mid U\right)\\ &= r \cdot \mathbb{P}\left(W \mid U\right)\mathbb{P}\left(\bigcup_{i\in I'} U_i \mid U\right).\end{aligned}$$

Poděleným výrazem $r \cdot \mathbb{P}\left(W \mid U\right)$ dostáváme tvrzení lemmatu. □

Nyní stačí zvolit $I = \mathcal{M}_\varepsilon^n(S)$, $I' = \mathcal{M}_\varepsilon^n(S) \setminus \hat{M}_\varepsilon^n$,

$$\begin{aligned}W &= \left(\left(S_{[0..n)}, R_{[0..n)}\right) \notin \overline{\mathcal{M}_\varepsilon^n}\left(S, R\right)\right),\\ U &= \left(S_{[0..n)} \in \mathcal{M}_\varepsilon^n\left(S\right)\right)\end{aligned}$$

a $U_u = \left(S_{[0..n)} = u\right)$. Stručně lze provedenou úvahu shrnout takto: pravděpodobnost, že prvek $\mathcal{M}_\varepsilon^n$ dává společně netypický výstup s pravděpodobností alespoň $\varepsilon/2$, nemůže být jedna polovina či víc, protože pak by pravděpodobnost takových netypických výstupů na $\mathcal{M}_\varepsilon^n$ byla alespoň $\varepsilon/4$.

Celkem dostáváme hledaný spodní odhad na míru V, totiž

$$\begin{aligned}\mathbb{P}\left(R_{[0..n)} \in V\right) &> \frac{\varepsilon}{2}\,\mathbb{P}\left(S_{[0..n)} \in \hat{M}_\varepsilon^n \mid S_{[0..n)} \in \mathcal{M}_\varepsilon^n(S)\right) \cdot \mathbb{P}\left(S_{[0..n)} \in \mathcal{M}_\varepsilon^n(S)\right) \geq\\ &> \frac{\varepsilon}{2} \cdot \frac{1}{2} \cdot (1 - \frac{\varepsilon}{24}) > \frac{\varepsilon}{5}.\end{aligned}$$

Dolní odhad pro N vyplyne z nutnosti pokrýt tuto pravděpodobnost pravděpodobnostmi jednotlivými koulí V_i. Uvažujme tedy nějaké pevné $u^{(i)} \in \hat{M}_\varepsilon^n$. Z definic V_i a $\hat{M}_\varepsilon^n$ plyne $u^{(i)} \in \mathcal{M}_\varepsilon^n(S)$ a pro každé $v \in V_i$ také $v \in \mathcal{M}_\varepsilon^n(R)$ a $(u^{(i)}, v) \in \mathcal{M}_\varepsilon^n(R, S)$. Tudíž

$$\begin{aligned}\mathbb{P}\left(R_{[0..n)} = v \mid S_{[0..n)} = u^{(i)}\right) &= \frac{\mathbb{P}\left(R_{[0..n)} = v, S_{[0..n)} = u^{(i)}\right)}{\mathbb{P}\left(S_{[0..n)} = u^{(i)}\right)} \geq \frac{\exp -n(\mathcal{H}\left(R, S\right) + \varepsilon)}{\exp -n(\mathcal{H}\left(S\right) - \varepsilon)}\\ &= \exp -n\left(\mathcal{H}\left(R \mid S\right) + 2\varepsilon\right).\end{aligned}$$

Z nerovnosti

$$1 > \mathbb{P}\left(\Gamma(u^{(i)}) \in V_i\right) = \mathbb{P}\left(R_{[0..n)} \in V_i \mid S_{[0..n)} = u^{(i)}\right)$$
$$= \sum_{v \in V_i} \mathbb{P}\left(R_{[0..n)} = v \mid S_{[0..n)} = u^{(i)}\right) \geq \left|V_i\right| \cdot \exp -n\left(\mathcal{H}\left(R \mid S\right) + 2\varepsilon\right),$$

nyní plyne

$$|V_i| < \exp n\left(\mathcal{H}\left(R \mid S\right) + 2\varepsilon\right),$$

a tedy

$$\mathbb{P}\left(R_{[0..n)} \in V_i\right) = \sum_{v \in V_i} \leq \left|V_i\right| \exp -n(\mathcal{H}\left(R\right) - \varepsilon) \leq \exp -n(\mathcal{H}\left(R : S\right) - 3\varepsilon).$$

Celkem tedy dostáváme

$$\frac{\varepsilon}{5} < \mathbb{P}\left(R_{[0..n)} \in \bigcup_{j \leq N} V_j\right) < N \cdot \exp -n(\mathcal{H}\left(R : S\right) - 3\varepsilon).$$

a konečně

$$N > \frac{\varepsilon}{5} \exp n\left(\mathcal{H}\left(R : S\right) - 3\varepsilon\right).$$

Můžeme shrnout do následující tvrzení.

Tvrzení 8.21. *Pro každé $S : \Omega \to A$ a pro $\delta > 0$, dosahuje bezpaměťový kanál Γ přenosové rychlosti $\mathcal{I}(S : \Gamma(S)) - \delta$.*

Důkaz. Zvolme $\varepsilon > 0$ ostře menší než $\delta/3$. Výše uvedená konstrukce poskytuje kód, který lze kanálem posílat a dekódovat s chybou nejvýše ε a jeho velikost je pro dostatečně velká n více než

$$\exp n\left(\mathcal{H}\left(R : S\right) - \delta\right) = \exp n\left(\mathcal{H}\left(R_0 : S_0\right) - \delta\right).$$

□

Přechodem k supremu přes všechna možná S a přechodem δ k nule dostaneme Shannonovu větu.

Věta 8.22 (Shannonova Věta o propustnosti kanálu)**.** *Kapacita bezpaměťového kanálu Γ je větší nebo rovna než $C_{\mathrm{Sh}}(\Gamma)$.*