20SK – Signály a kódy

Přednáška 8 – Prefixové kódy, komprese dat (19.11.2018)

Probíraná témata:

- Formální definice kódu.
- Kódy s pevnou délkou slova (blokové kódy).
- Kódy s proměnnou délkou slova.
- Jednoznačná dekódovatelnost.
- Kraftova nerovnost pro prefixové kódy.
- Optimální prefixový kód, Huffmanovo kódování.

Relevantní literatura je [1, kapitola 2], [2, kapitola 2], [3, odstavec 1.2] a [4, celá kapitola I].

Seznam literatury

[1] Gallager, R.: Course materials for 6.450 *Principles of Digital Communications I*, Fall 2006. MIT OpenCourseWare (http://ocw.mit.edu/), Massachusetts Institute of Technology.

[2] Adámek, J: *Foundations of Coding: Theory and Applications of Error-Correcting Codes with an Introduction to Cryptography and Information Theory*. Wiley Interscience, 1991, 352 pp.

[3] Seibt, P.: Algorithmic Information Theory – Mathematics of Digital Information Processing. Springer, 2006, 447 pp.

[4] Adámek, J: Kódování. Matematika pro vysoké školy technické, sešit XXXI. SNTL, 1989, 192 pp.

Kódování informace z diskrétních zdrojů

Typy zdrojů:

a) analogové zdroje: spojitý vlnový signál

b) analogové sekvence: diskrétní v čase, spojité v hodnotě

c) diskrétní zdroje: symboly z nějaké $\boxed{\text{abecedy } X}$ diskrétní v čase

libovolná konečná množina symbolů

$X = \{A, B, C, \dots, Z\}$

$X = \{\cdot, -, \sim\}$

$X = \{0, 1, 2, \dots, 255\}$

Základní pravidlo pro kódování

Kód musí být jednoznačně dekódovatelný!

Fungují pouze pro diskrétní zdroje!

Př: $A \rightarrow 010$, $C \not\rightarrow 010$; $010 < \begin{matrix} A? \\ C? \end{matrix}$

Kódování – definice

- zdrojová abeceda … $\mathcal{A}$ (vstup)
- kódová abeceda … $\mathcal{B}$ (výstup) … u nás vždy binární, $\mathcal{B} = \{0, 1\}$
- slovo v $\mathcal{A}$ či $\mathcal{B}$ … konečná posloupnost symbolů z $\mathcal{A}$ či $\mathcal{B}$, 'AHOJ' , '01001010'
- kódování: symbol z $\mathcal{A}$ mapuje na slovo v $\mathcal{B}$

  $A \rightarrow 010$

  $B \rightarrow 011$

  $\vdots$

  ale: lze i slovo v $\mathcal{A}$ → slovo v $\mathcal{B}$

  → 'AHOJ' → '01001010' → využití při kompresi

Př: Kód 2 z 5

$0 \to 00011$
$1 \to 11000$
$2 \to 10100$
$3 \to 01100$
$4 \to 10010$
$5 \to 01010$
$6 \to 00110$
$7 \to 10001$
$8 \to 01001$
$9 \to 00101$

$\mathcal{A}$ $\quad$ $\mathcal{B}$

$173 \to 11000\ 10001\ 01100$

↓

blokový kód – pevná délka slova
fixed-length

počet bitů blokového kódu:

M … počet symbolů
L … počet bitů

$$\boxed{M \leq 2^L}$$

$$L = \lceil \log_2 M \rceil$$

Př: $0 \to 0 \to 0000$
⋮
$9 \to 9 \to 1001$
$10 \to A \to 1010$
$11 \to B$
$15 \to F \to 1111$

$\boxed{\text{Základní pravidlo pro kódování}}$

Kód musí být jednoznačně dekódovatelný!
Fungují pouze pro diskrétní zdroj!

$\boxed{\text{Kódování – definice}}$

- Zdrojová abeceda … $\mathcal{A}$ (vstup)
- kódová abeceda … $\mathcal{B}$ (výstup) … u nás vždy binární, $\mathcal{B} = \{0,1\}$
- slovo v $\mathcal{A}$ či $\mathcal{B}$ … konečná posloupnost symbolů z $\mathcal{A}$ či $\mathcal{B}$ 'AHOJ', '01001010'
- kódování: symbol z $\mathcal{A}$ mapuje na slovo v $\mathcal{B}$
  ale: lze i slovo v $\mathcal{A}$ → slovo v $\mathcal{B}$
  'AHOJ' → '01001010' → využívání při kompresi

$A \to 010$
$B \to 011$
⋮

Př: $A \to 010$
$C \not\to 010$

$010 <$ A? C?

Př: Neprefixový kód

a → 1
b → 10
c → 100

abbca → 110101001

→ Prefixové kódy (prefix-free ; instantaneous codes)

- žádné kódové slovo není prefixem jiného kódového slova
- v okamžiku, kdy přijmu poslední bit kódového slova, mohu jej dekódovat
- pokud mám blokový kód s M různými kódovými slovy, je prefixový
- pokud znám pravděpodobnostní rozdělení symbolů zdroje, mohu zkonstruovat prefixový kód optimální délky → Huffmanovo kódování

Kódy s proměnnou délkou slova

Idea: ne všechny symboly mají stejnou pravděpodobnost výskytu

⇒ entropie zdroje je vyšší, než nezbytně nutná

Řešení: kódovat časté symboly jako krátká slova a méně časté symboly jako dlouhá slova

Př: Morseovka

E → .
T → –
A → .–
⋮

Př: Jednoduchý bin. kód pro tři symboly

a → 0
b → 10
c → 11

abbca → 01010110

a → 0
b → 1
c → 10

abbca → 011100 → abbbaa

Není jednoznačně dekódovatelné!

Př: Neprefixový kód

a → 1
b → 10
c → 100

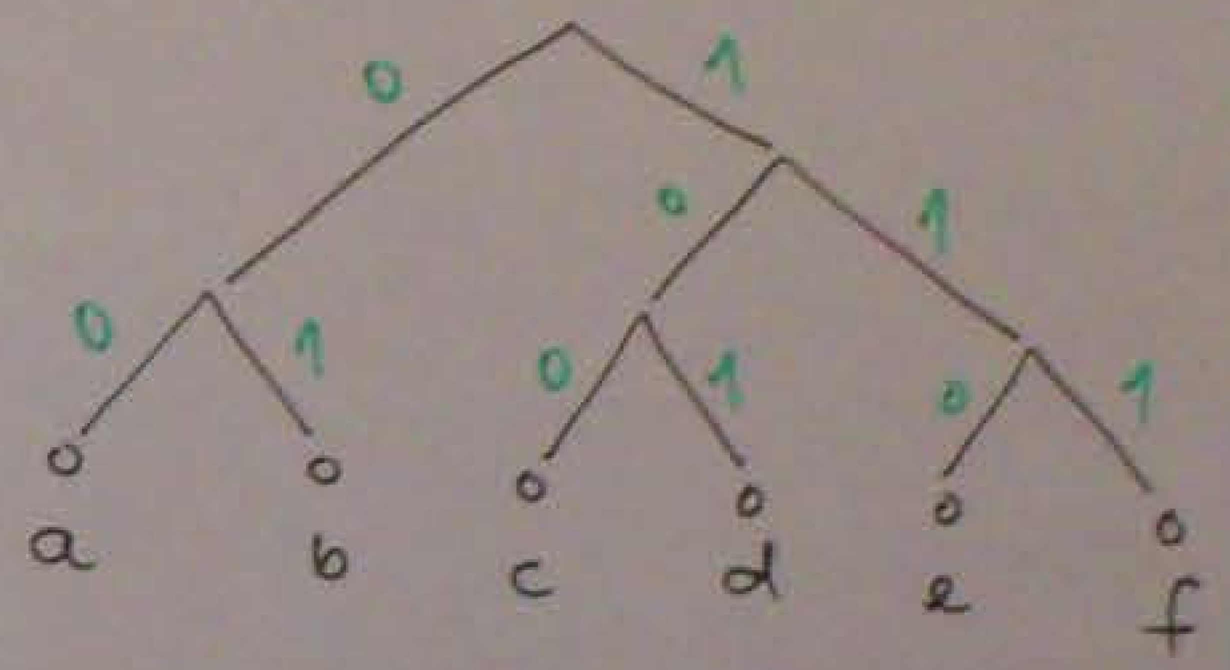
abbca → 110101001

→ Prefixové kódy (prefix-free ; instantaneous codes)

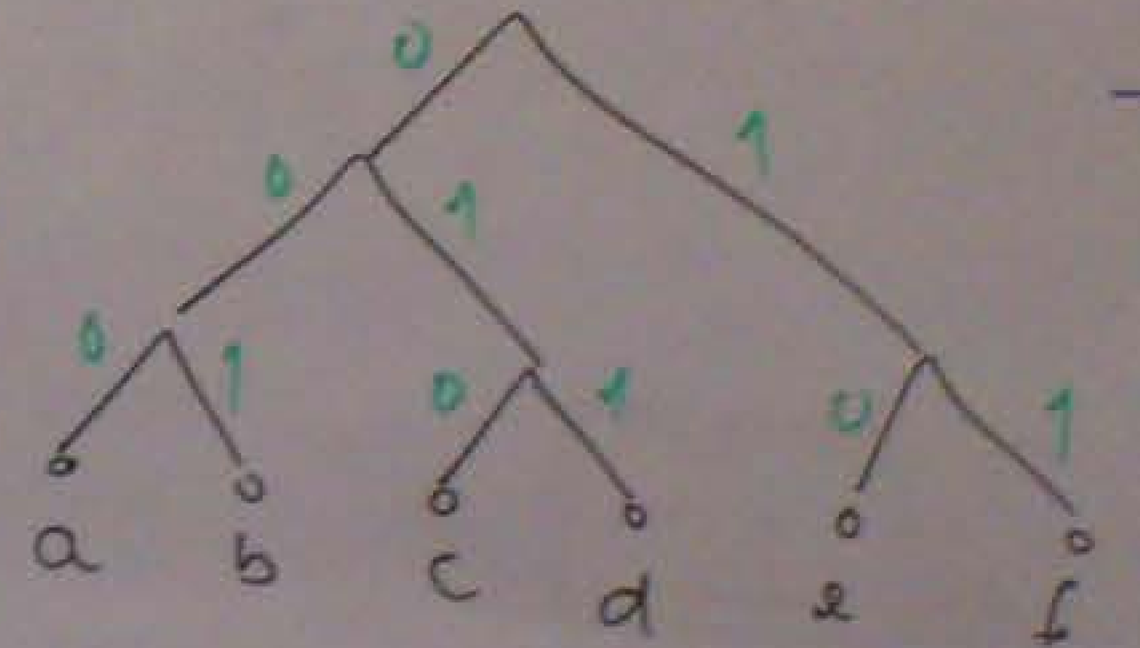
- žádné kódové slovo není prefixem jiného kódového slova
- v okamžiku, kdy přijmu poslední bit kódového slova, mohu jej dekódovat
- pokud mám blokový kód s M různými kódovými slovy, je prefixový
- pokud znám pravděpodobnostní rozdělení symbolů zdroje, mohu zkonstruovat prefixový kód optimální délky → Huffmannovo kódování

Př: $A = \{a, b, c, d, e, f\}$

→ prefix. kód ⇒ binární strom se symboly z A v listech



a → 00
b → 01
c → 100
d → 101
e → 110
f → 111



→ optimální kód: konstrukce stromu podle pravděpodobnostního zastoupení symbolů

(?) Lze zkonstruovat binární pref. kód s danou množinou délek kódových slov?

↓

Kraftova nerovnost

$\mathcal{L} = \{ l_1, l_2, \dots, l_n \}$ ... délky kódových slov

Pokud $\boxed{\sum_{i=1}^{n} 2^{-l_i} \leq 1}$, kód sestrojit lze!

→ $\sum \dots = 1$ ... jde o úplný pref. kód

- nelze do něj přidat žádný další symbol
- pokud je n liché, kód nebude úplný nikdy!

Huffmanovo kódování .... aproximovat délku kódového slova jako $-\log_2 p_i$

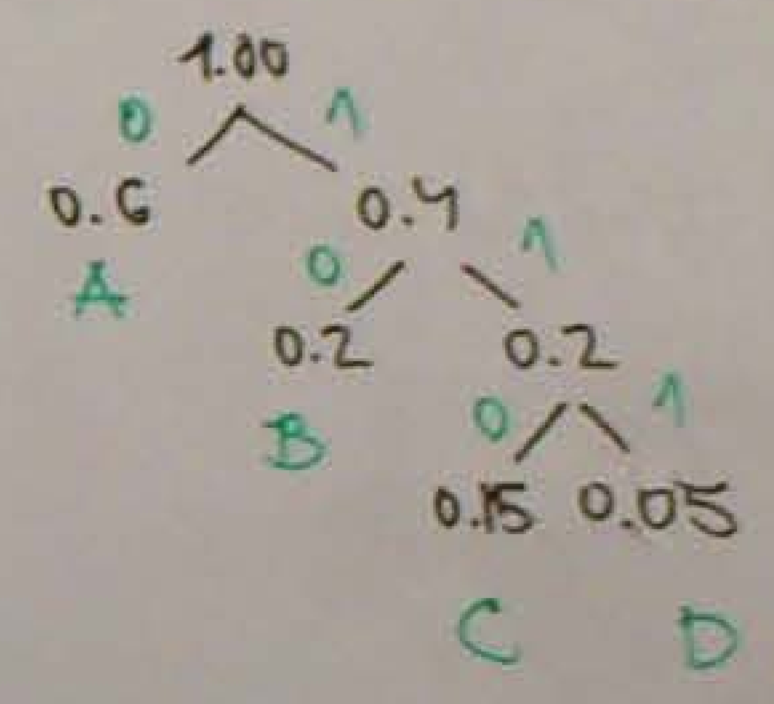
Př: $\mathcal{A} = \{A, B, C, D\}$

A .... 0.60 — 0.60 — 0.60 > 1.00
B .... 0.20 — 0.20 > 0.40
C .... 0.15 > 0.20
D 0.05

p_i

A ... 0.30 > 0.59 > 1.00
B ... 0.29
C ... 0.28 > 0.41
D ... 0.13

→ všechny symboly mají 2 bity



A → 0
B → 10
C → 110
D → 111