

# Essential Information Theory

## PA154 Jazykové modelování (1.3)

Pavel Rychlý

pary@fi.muni.cz

February 23, 2017

**Source:** Introduction to Natural Language Processing (600.465)  
Jan Hajič, CS Dept., Johns Hopkins Univ.  
[www.cs.jhu.edu/~hajic](http://www.cs.jhu.edu/~hajic)

# The Notion of Entropy

- Entropy – “chaos” , fuzziness, opposite of order, . . .
  - ▶ you know it
    - ▶ it is much easier to create “mess” than to tidy things up . . .
- Comes from physics:
  - ▶ Entropy does not go down unless energy is used
- Measure of uncertainty:
  - ▶ if low . . . low uncertainty

## Entropy

The higher the entropy, the higher uncertainty, but the higher “surprise” (information) we can get out of experiment.

# The Formula

- Let  $p_x(x)$  be a distribution of random variable  $X$
- Basic outcomes (alphabet)  $\Omega$

## Entropy

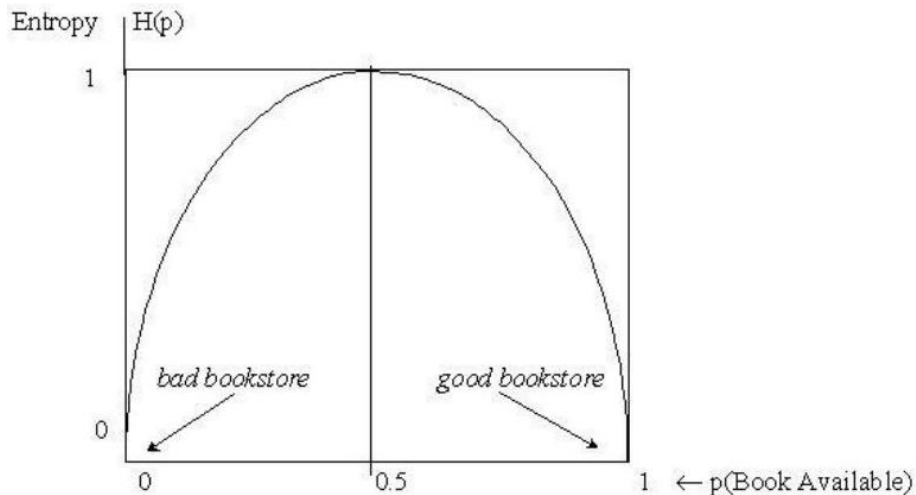
$$H(X) = - \sum_{x \in \Omega} p(x) \log_2 p(x)$$

- Unit: bits ( $\log_{10}$ : nats)
- Notation:  $H(X) = H_p(X) = H(p) = H_X(p) = H(p_X)$

## Using the Formula: Example

- Toss a fair coin:  $\Omega = \{head, tail\}$ 
  - ▶  $p(head) = .5, p(tail) = .5$
  - ▶  $H(p) = -0.5 \log_2(0.5) + (-0.5 \log_2(0.5)) = 2 \times ((-0.5) \times (-1)) = 2 \times 0.5 = 1$
- Take fair, 32-sided die:  $p(x) = \frac{1}{32}$  for every side  $x$ 
  - ▶  $H(p) = -\sum_{i=1 \dots 32} p(x_i) \log_2 p(x_i) = -32(p(x_1) \log_2 p(x_1))$   
(since for all  $i$   $p(x_i) = p(x_1) = \frac{1}{32}$ )  
 $= -32 \times (\frac{1}{32} \times (-5)) = 5$  (now you see why it's called **bits**?)
- Unfair coin:
  - ▶  $p(head) = .2 \dots \mathbf{H(p) = .722}$
  - ▶  $p(head) = .1 \dots \mathbf{H(p) = .081}$

## Example: Book Availability



## ■ When $H(p) = 0$ ?

- ▶ if a result of an experiment is **known** ahead of time:
- ▶ necessarily:

$$\exists x \in \Omega; p(x) = 1 \& \forall y \in \Omega; y \neq x \Rightarrow p(y) = 0$$

## ■ Upper bound?

- ▶ none in general
- ▶ for  $|\Omega| = n : H(p) \leq \log_2 n$ 
  - ▶ nothing can be more uncertain than the uniform distribution

- Recall:

- ▶  $E(X) = \sum_{x \in X(\Omega)} p_x(x) \times x$

- Then:

$$E \left( \log_2 \left( \frac{1}{p(x)} \right) \right) = \sum_{x \in X(\Omega)} p_x(x) \log_2 \left( \frac{1}{p_x(x)} \right) =$$
$$- \sum_{x \in X(\Omega)} p_x(x) \log_2 p_x(x) = H(p_x) =_{\text{notation}} H(p)$$

## ■ Recall:

- ▶ 2 equiprobable outcomes:  $H(p) = 1$  bit
- ▶ 32 equiprobable outcomes:  $H(p) = 5$  bits
- ▶ 4.3 billion equiprobable outcomes:  $H(p) \cong 32$  bits

## ■ What if the outcomes are not equiprobable?

- ▶ 32 outcomes, 2 equiprobable at 0.5, rest impossible:
  - ▶  $H(p) = 1$  bit
- ▶ any measure for comparing the entropy (i.e. uncertainty/difficulty of prediction) (also) for random variables with *different number of outcomes?*



- Perplexity:
  - ▶  $G(p) = 2^{H(p)}$
- ... so we are back at 32 (for 32 eqp. outcomes), 2 for fair coins, etc.
- it is easier to imagine:
  - ▶ NLP example: vocabulary size of a vocabulary with uniform distribution, which is equally hard to predict
- the “wilder” (biased) distribution, the better:
  - ▶ lower entropy, lower perplexity

# Joint Entropy and Conditional Entropy

- Two random variables:  $X$  (space  $\Omega$ ),  $Y$  ( $\Psi$ )
- Joint entropy:
  - ▶ no big deal:  $((X, Y)$  considered a single event):

$$H(X, Y) = - \sum_{x \in \Omega} \sum_{y \in \Psi} p(x, y) \log_2 p(x, y)$$

- Conditional entropy:

$$H(Y|X) = - \sum_{x \in \Omega} \sum_{y \in \Psi} p(x, y) \log_2 p(y|x)$$

recall that  $H(X) = E \left( \log_2 \frac{1}{p_x(x)} \right)$

(weighted “average”, and weights are not conditional)

# Conditional Entropy (Using the Calculus)

- other definition:

$$\begin{aligned} H(Y|X) &= \sum_{x \in \Omega} p(x) H(Y|X = x) = \\ &\quad \text{for } H(Y|X = x), \text{ we can use} \\ &\quad \text{the single-variable definition (} x \sim \text{constant)} \\ &= \sum_{x \in \Omega} p(x) \left( - \sum_{y \in \Psi} p(y|x) \log_2 p(y|x) \right) = \\ &= - \sum_{x \in \Omega} \sum_{y \in \Psi} p(y|x) p(x) \log_2 p(y|x) = \\ &= - \sum_{x \in \Omega} \sum_{y \in \Psi} p(x, y) \log_2 p(y|x) \end{aligned}$$

# Properties of Entropy I

## ■ Entropy is non-negative:

- ▶  $H(X) \geq 0$
- ▶ proof: (recall:  $H(X) = -\sum_{x \in \Omega} p(x) \log_2 p(x)$ )
  - ▶  $\log_2(p(x))$  is negative or zero for  $x \leq 1$ ,
  - ▶  $p(x)$  is non-negative; their product  $p(x) \log(p(x))$  is thus negative,
  - ▶ sum of negative numbers is negative,
  - ▶ and  $-f$  is positive for negative  $f$

## ■ Chain rule:

- ▶  $H(X, Y) = H(Y|X) + H(X)$ , as well as
- ▶  $H(X, Y) = H(X|Y) + H(Y)$  (since  $H(Y, X) = H(X, Y)$ )

# Properties of Entropy II

- Conditional Entropy is better (than unconditional):
  - ▶  $H(Y|X) \leq H(Y)$
- $H(X, Y) \leq H(X) + H(Y)$  (follows from the previous (in)equalities)
  - ▶ equality iff  $X, Y$  independent
  - ▶ (recall:  $X, Y$  independent iff  $p(X, Y) = p(X)p(Y)$ )
- $H(p)$  is concave (remember the book availability graph?)
  - ▶ concave function  $f$  over an interval  $(a, b)$ :  
 $\forall x, y \in (a, b), \forall \lambda \in [0, 1] :$   
 $f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y)$
  - ▶ function  $f$  is convex if  $-f$  is concave
- for proofs and generalizations, see Cover/Thomas

