Probability and Essential Information Theory

Jan Hajič, Pavel Pecina, Jindřich Helcl, Jindřich Libovický

Feubruary 25, 2026

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Outline

Probability

Outline

Experiments & Sample Spaces

- Experiment, process, test, ...
- Set of possible basic outcomes: **sample space** Ω

Experiment	**Sample Space**
coin toss, die	$\Omega = \{\text{head}, \text{tail}\}$, $\Omega = \{1..6\}$
yes/no opinion poll, quality test	$\Omega = \{0, 1\}$
lottery	$\lvert\Omega\rvert =$ number of combinations
# of traffic accidents per year	$\Omega = \mathbb{N}$
spelling errors	$\Omega = \Sigma^*$, where Σ is an alphabet
missing word	$\lvert\Omega\rvert =$ vocabulary size

Events

Event $A \subseteq \Omega$ is a set of basic outcomes

Ω is the *certain event* $\quad \emptyset$ is the *impossible event*

Example: 3× coin toss

- $\Omega = ?$ $\{\texttt{HHH}, \texttt{HHT}, \texttt{HTH}, \texttt{HTT}, \texttt{THH}, \texttt{THT}, \texttt{TTH}, \texttt{TTT}\}$
- Count cases with exactly two tails: $A = ?$ $\{\texttt{HTT}, \texttt{THT}, \texttt{TTH}\}$
- All heads: $A = ?$ $\{\texttt{HHH}\}$ (elementary event)

...what is a probability of an event?

(Idealized) Probability

- **Repeat** experiment many times, **count** occurrences of event A
- Repeat T times, count c occurrences of A
- Results close to some unknown constant (c/T for large T)
- Call this constant a **probability** of A, denote $p(A)$

This is the **frequentist** view of probability. There are other definitions (e.g., Bayesian), but we will stick to this one.

Estimating Probability

True probability is unknown, we can only **estimate** it:

- From a single series (typical case): set

$$p(A) = \frac{c_1}{T_1}$$

- Otherwise, take the weighted average of all $\frac{c_i}{T_i}$
  (or, if the data allows, simply look at the set of series as if it is a single long series).

This is the **best estimate** (consistent, unbiased, maximum likely, ...).

Estimating Probability: Example

Experiment: 3× coin toss

- $\Omega = \{\texttt{HHH}, \texttt{HHT}, \texttt{HTH}, \texttt{HTT}, \texttt{THH}, \texttt{THT}, \texttt{TTH}, \texttt{TTT}\}$
- Count cases with exactly two tails: $A = \{\texttt{HTT}, \texttt{THT}, \texttt{TTH}\}$

Run experiment 1000 times (i.e., 3000 tosses)

- Counted: 386 cases with two tails
- Estimate: $p(A) = 386/1000 = 0.386$
- Run again: 373, 399, 382, 355, 372, 406, 359
  $p(A) = 0.379$ (weighted average) or simply $3032/8000$
- Uniform distribution assumption: $p(A) = 3/8 = 0.375$

So, it probably not significantly different from the uniform distribution assumption, but we cannot be sure, we would need use statistical tests for that.

Basic Properties of Probability

- $p : 2^{\Omega} \to [0, 1]$
- $p(\Omega) = 1$
- Disjoint events $\Rightarrow p(\bigcup A_i) = \sum_i p(A_i)$

Axiomatic definition of probability: take the above three conditions as axioms.

Immediate consequences:

- $p(\emptyset) = 0$
- $p(\bar{A}) = 1 - p(A)$
- $A \subseteq B \Rightarrow p(A) \leq p(B)$
- $\sum_{a \in \Omega} p(a) = 1$

Can this happen? $A \subset B$ and $p(A) = p(B)$

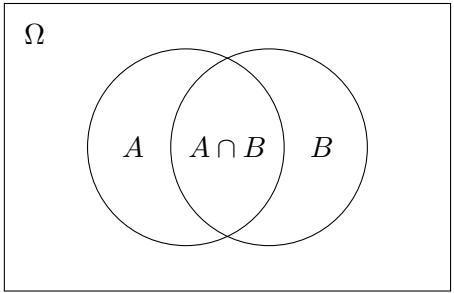
Joint and Conditional Probability

Combining probabilities of **multiple** events:

- Joint probability: A and B at the same time
  $p(A, B) = p(A \cap B)$
- Conditional probability: A given B
  $p(A \mid B) = p(A \cap B)/p(B)$

Estimating from counts:

$$p(A \mid B) = \frac{p(A, B)}{p(B)} = \frac{c(A \cap B)/T}{c(B)/T} = \frac{c(A \cap B)}{c(B)}$$



Bayes Rule

$p(A, B) = p(B, A)$ since $p(A \cap B) = p(B \cap A)$
Therefore, $p(A \mid B) \cdot p(B) = p(B \mid A) \cdot p(A)$

And therefore:

$$p(A \mid B) = \frac{p(B \mid A) \cdot p(A)}{p(B)}$$

Independence

Can we compute $p(A, B)$ from $p(A)$ and $p(B)$?

Recall from previous slide:

$$\begin{aligned} p(A \mid B) &= p(B \mid A) \cdot p(A)/p(B) \\ p(A \mid B) \cdot p(B) &= p(B \mid A) \cdot p(A) \\ p(A, B) &= p(B \mid A) \cdot p(A) \end{aligned}$$

We're almost there: how does $p(B \mid A)$ relate to $p(B)$?

$$p(B \mid A) = p(B) \quad \text{iff } A \text{ and } B \text{ are } \textbf{independent}$$

Examples: two coin tosses; weather today and weather on April 7th 1348

Probability Chain Rule

$$p(A_1, A_2, A_3, A_4, \ldots, A_n) =$$

$$p(A_1 \mid A_2, A_3, A_4, \ldots, A_n) \cdot p(A_2 \mid A_3, A_4, \ldots, A_n)$$
$$\cdot p(A_3 \mid A_4, \ldots, A_n) \cdots p(A_{n-1} \mid A_n) \cdot p(A_n)$$

This is a direct consequence of the Bayes rule.

The Golden Rule (of Classic Statistical NLP)

Conditional probability $p(A \mid B)$ in NLP application

	Speech recognition	Machine translation	Language Modeling
A	transcription	target language sentence	next word
B	audio signal	source language sentence	previous word

Goal is to find A that maximizes $p(A \mid B)$

$$\begin{aligned}\arg\max_A p(A \mid B) &= \arg\max_A \frac{p(B \mid A) \cdot p(A)}{p(B)\cancel{p(B)}} \\ &= \arg\max_A p(B \mid A) \cdot p(A)\end{aligned}$$

...as $p(B)$ is constant when changing As.

Random Variable

- A **random variable** is a function $X : \Omega \to Q$
  - In general: $Q = \mathbb{R}^n$, typically $\mathbb{R}$
  - Easier to handle real numbers than real-world events
- Random variable is **discrete** if Q is countable (i.e., also if finite)
- **Example:** die: natural "numbering" $[1, 6]$, coin: $\{0, 1\}$
- **Probability distribution:**

$$p_X(x) = p(X = x) \stackrel{\text{def}}{=} p(A_x) \text{ where } A_x = \{a \in \Omega : X(a) = x\}$$

- Often just $p(x)$ if it is clear from context what X is

Expectation & Joint/Conditional Distributions

Expectation is a mean of a random variable (weighted average):

$$E(X) = \sum_{x \in X(\Omega)} x \cdot p_X(x)$$

Example: one six-sided die: 3.5, two dice (sum): 7

Properties of joint and conditional RVs analogous to events

- Analogous to probability of events
- Bayes: $p_{X|Y}(x, y) = p(x \mid y) = \frac{p(y|x) \cdot p(x)}{p(y)}$
- Chain rule: $p(w, x, y, z) = p(z) \cdot p(y \mid z) \cdot p(x \mid y, z) \cdot p(w \mid x, y, z)$

Essential Information Theory

Outline

The Notion of Entropy

Entropy $\sim$ "chaos", fuzziness, opposite of order, ...

- You know it: it is much easier to create "mess" than to tidy things up...
- In physics, entropy does not go down unless energy is applied

Measure of uncertainty:

- Low value $\Rightarrow$ low uncertainty
- The higher the entropy, the higher uncertainty, but the higher "surprise" (information) we can get out of an experiment

The Formula

Let $p_X(x)$ be a distribution of random variable X

Basic outcomes (alphabet) Σ:

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

- **Unit:** bits ($\log_e$: nats)
- **Notation:** $H(X) = H_p(X) = H(p) = H_X(p) = H(p_X)$

Using the Formula: Example

Toss a fair coin: $\Sigma = \{\text{head}, \text{tail}\}$

- $p(\text{head}) = 0.5$, $p(\text{tail}) = 0.5$
- $H(p) = -0.5 \log_2(0.5) + (-0.5 \log_2(0.5)) = 2 \times 0.5 = 1$

Fair 32-sided die: $p(x) = 1/32$ for every side x

- $H(p) = -32 \times (1/32) \times (-5) = 5$ bits

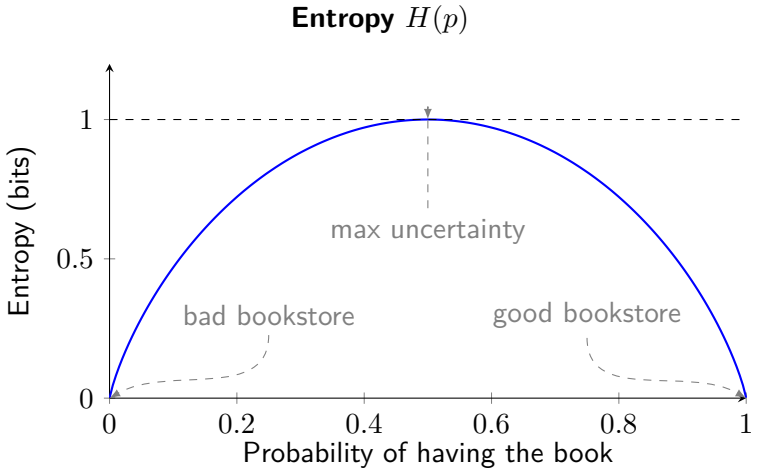
Unfair coin:

- $p(\text{head}) = 0.2$ ...$H(p) = 0.722$
- $p(\text{head}) = 0.01$ ...$H(p) = 0.081$

Interactive Demo

https://ufallab.ms.mff.cuni.cz/~libovicky/entropy.html

Example: Book Availability



The Limits

When $H(p) = 0$?

- If the result of an experiment is known ahead of time
- Necessarily: $\exists x \in \Sigma; p(x) = 1$ and $\forall y \neq x : p(y) = 0$

Upper bound?

- None in general
- For $|\Sigma| = n$: $H(p) \leq \log_2 n$
- Nothing can be more uncertain than the uniform distribution

(Proof in NPFL129 using Lagrange multipliers.)

Perplexity: Motivation

Recall:

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

What if the outcomes are not equiprobable?

- 32 outcomes, 2 equiprobable at 0.5, rest impossible: $H(p) = ?$ 1 bit

Perplexity

Perplexity:

$$\boxed{G(p) = 2^{H(p)}}$$

...so we are back at 32 (for 32 equiprobable 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 Ω), Y (space Γ)

Joint entropy ((X, Y) considered a single event):

$$H(X, Y) = -\sum_x \sum_y p(x, y) \log_2 p(x, y)$$

In NPFL129, this $H(X, Y)$ denotes cross-entropy.

Conditional entropy: (Still weighting with joint probability!)

$$H(Y \mid X) = -\sum_x \sum_y p(x, y) \log_2 p(y \mid x)$$

Properties of Entropy I

Entropy is non-negative:

- $H(X) \geq 0$
- Proof: $\log(p(x))$ is negative or zero for $x \leq 1$; $p(x)$ is non-negative; their product is negative; sum of negatives is negative; and $-f$ is positive for negative f

Chain rule:

- $H(X, Y) = H(Y \mid X) + H(X)$
- $H(X, Y) = H(X \mid Y) + H(Y)$ (since $H(Y, X) = H(X, Y)$)

Properties of Entropy II

Conditional entropy is better (than unconditional):

$$H(Y \mid X) \leq H(Y)$$

Joint entropy bound:

$$H(X, Y) \leq H(X) + H(Y)$$

- Equality iff X, Y independent
- Recall: X, Y independent iff $p(X, Y) = p(X)p(Y)$

$H(p)$ **is concave:**

$$\forall x, y \in (a, b), \lambda \in [0, 1] : f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y)$$

“Coding” Interpretation of Entropy

The least (average) number of bits needed to encode a message (string, sequence, series, ...) (each element being a result of a random process with some distribution p): $= H(p)$

Remember various compressing algorithms?

- They do well on data with repeating ($=$ easily predictable $=$ low entropy) patterns
- Their results have high entropy $\Rightarrow$ compressing compressed data does nothing

Kullback-Leibler Divergence (Relative Entropy)

Remember: long series of experiments...c_i/T_i oscillates around some number...we can only estimate it...to get a distribution q.

So we get a distribution q; the true distribution is p. How big error are we making?

Kullback-Leibler distance:

$$D(p||q) = \sum_{x \in \Sigma} p(x) \log_2 \frac{p(x)}{q(x)} = E_p \log_2 \frac{p(x)}{q(x)}$$

Comments on Relative Entropy

Conventions:

- $0 \log 0 = 0$
- $p \log(p/0) = \infty$ (for $p > 0$)

Distance? (less "misleading": Divergence)

- Not quite:
  - Not symmetric: $D(p||q) \neq D(q||p)$
  - Does not satisfy the triangle inequality
- But useful to look at it that way

$H(p) + D(p||q)$: bits needed for encoding p if q is used

Mutual Information (MI) in Terms of Relative Entropy

Random variables X, Y; $p_{XY}(x, y)$, $p_X(x)$, $p_Y(y)$

Mutual information (between two random variables X, Y):

$$I(X, Y) = D(p(x, y) \| p(x)p(y))$$

$I(X, Y)$ measures how much (our knowledge of) Y contributes (on average) to easing the prediction of X

Or, how $p(x, y)$ deviates from (independent) $p(x)p(y)$

Mutual Information: the Formula

Rewrite the definition:

$$\begin{aligned} I(X,Y) &= D(p(x,y)\|p(x)p(y)) \\ &= \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

Measured in bits (what else? :-)

From Mutual Information to Entropy

By how many bits the knowledge of Y lowers the entropy $H(X)$:

$$
\begin{aligned}
I(X,Y) &= \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(y)p(x)} \\
&= \sum_x \sum_y p(x,y) \log_2 \frac{p(x \mid y)}{p(x)} \quad \text{(using } p(x,y)/p(y) = p(x \mid y)\text{)} \\
&= \sum_x \sum_y p(x,y) \log_2 p(x \mid y) - \sum_x \sum_y p(x,y) \log_2 p(x) \\
&= -H(X \mid Y) + (-\sum_x p(x) \log_2 p(x)) \\
&= H(X) - H(X \mid Y)
\end{aligned}
$$

Properties of MI vs. Entropy

- $I(X,Y) = H(X) - H(X \mid Y)$
  Number of bits the knowledge of Y lowers the entropy of X
- By symmetry also $I(X,Y) = H(Y) - H(Y \mid X)$ (by symmetry)
- $I(X,Y) = H(X) + H(Y) - H(X,Y)$

Is it a metric then? No.

- $I(X,Y) = I(Y,X)$ (symmetry)
- $I(X,Y) \geq 0$ (since $D(p||q) \geq 0$)
- $I(X,X) = H(X)$ (since $H(X \mid X) = 0$, so it is not reflexive)

Cross-Entropy

Typical case: we've got a series of observations $T = \{t_1, t_2, t_3, \ldots, t_n\}$

Simple estimate: $\forall y \in \Sigma : \tilde{p}(y) = c(y)/|T|$, where $c(y) = |\{t : t = y\}|$

...but the true p is unknown; every sample is too small!

Natural question: how well do we do using $\tilde{p}$ instead of p?

Idea: simulate actual p by using a different T' (or rather: by using different observation we simulate the insufficiency of T vs. some other data)

Cross Entropy: The Formula

$$H_{p'}(\tilde{p}) = H(p') + D(p'||\tilde{p})$$

$$\boxed{H_{p'}(\tilde{p}) = -\sum_{x \in \Sigma} p'(x) \log_2 \tilde{p}(x)}$$

- p' is certainly not the true p, but we can consider it the "real world" distribution against which we test $\tilde{p}$
- **(Cross) Perplexity:** $G_{p'}(p) = G_{T'}(p) = 2^{H_{p'}(\tilde{p})}$

Conditional Cross Entropy

So far: "unconditional" distribution(s) $p(x)$, $p'(x)$...

In practice: virtually always conditioning on context

Interested in: sample space Γ, r.v. Y, $y \in \Gamma$; context: sample space Σ, r.v. X, $x \in \Sigma$

"Our" distribution $p(y \mid x)$, test against $p'(y, x)$:

$$H_{p'}(p) = -\sum_{y} \sum_{x} p'(y, x) \log_2 p(y \mid x)$$

Sample Space vs. Data

In practice, it is often inconvenient to sum over the sample space(s)

Use the following formula:

$$H_{p'}(p) = -\sum_{y}\sum_{x} p'(y,x)\log_2 p(y \mid x) = -\frac{1}{|T'|}\sum_{i=1}^{|T'|}\log_2 p(y_i \mid x_i)$$

This is in fact the normalized log probability of the "test" data:

$$H_{p'}(p) = -\frac{1}{|T'|}\log_2 \prod_{i=1}^{|T'|} p(y_i \mid x_i)$$

Computation Example

Alphabet $\Sigma = \{a, b, \ldots, z\}$
Distribution: $p(a) = 0.25$, $p(b) = 0.5$, $p(\cdot) = 1/64$ for $\{c \ldots r\}$, $= 0$ for rest

Data (test): barb $\Rightarrow p'(a) = p'(r) = 0.25$, $p'(b) = 0.5$

Sum over Σ:

$$-\sum_{\sigma} p'(\sigma) \log_2 p(\sigma) = 0.5 + 0.5 + 0 + \ldots + 1.5 + 0 + \ldots = 2.5$$

Sum over data:

i / s_i	1/b	2/a	3/r	4/b	
$\frac{1}{\lvert T' \rvert}(-\log_2 p(s_i))$	1	2	6	1	$= 10 \times \frac{1}{4} = 2.5$

Cross Entropy: Some Observations

$H(p) \lessgtr H_{p'}(p)$: ALL possibilities!

Previous example: $H(p) = 2.5$ bits $= H_{p'}(p)$ (barb)

Other data:

- **probable:** $(1/8)(6+6+6+1+2+1+6+6) = 4.25$
  - $H(p) < 4.25$ bits $= H_{p'}(p)$
- **abba:** $(1/4)(2+1+1+2) = 1.5$
  - $H(p) > 1.5$ bits $= H_{p'}(p)$
- **baby:** $-p'(\text{'y'}) \log_2 p(\text{'y'}) = -0.25 \log_2 0 = \infty$ (??)

Cross Entropy: Usage

Comparing data?

NO!

Rather: **comparing distributions** (vs. real data)

Have (got) 2 distributions: p and q (on some Σ, X)

- Which is better?
- Better: has lower cross-entropy (perplexity) on real data S

Perplexity of language models on WikiText-2 Benchmark

Model	PPL
Unigram	≈ 1050
Bigram	≈ 600
4-gram	≈ 240
5-gram	≈ 220
GPT-2 Medium (345M)	22.7
GPT-2 Large (774M)	19.9
GPT-2 XL (1.5B)	18.3
Qwen2-7B	7.9
Gemma-2-9B-IT	7.3
LLaMA-2-7B	5.5
Qwen2-72B	5.6
Mistral-7B	5.3
LLaMA-2-13B	4.9
LLaMA-2-70B	3.3

Summary

1. **Probability basics:** sample spaces, events, Bayes rule, chain rule, independence
2. **Random variables:** distributions, expectations, conditional distributions
3. **The Golden Rule:**
   $\arg\max_A p(A \mid B) = \arg\max_A p(B \mid A) \cdot p(A)$
4. **Information theory:** entropy, perplexity, KL divergence, mutual information, cross-entropy

https://ufal.mff.cuni.cz/courses/npfl124