Probability for Language Modeling

Part II: Sources

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences

Quantitative Cognitive Linguistics Network
25th July 2024

A (not so) rare specimen of a stochastic parrot

The rise of large language models
with their strengths
(fluent relevant grammatical replies)
and weaknesses
(factual hallucinations)
reopens the old question whether
statistical language modeling
makes sense for language science.

Motivating question

Does statistical language modeling make sense for linguistics?

I think one should replace "Does" with "How can".

There is a related question of a theoretical importance:

How much randomness is there in language and speech ...
... and, precisely, how does it interfere with structure?

This question cannot be answered without a certain understanding of mathematical models of randomness.

These slides provide an intro.

1. Overview
2. Markov sources
3. Stationary sources
4. Information theory
5. Sufficient statistic

1 Overview

2 Markov sources

3 Stationary sources

4 Information theory

5 Sufficient statistic

What is a stochastic process?

A process or a source is an infinite sequence of random variables:

$$(X_n)_{n\in\mathbb{N}} := (X_1, X_2, X_3, \ldots)$$

This is a model of sequential data that contain a specified amount of randomness.

To specify a process, it suffices to specify conditional probabilities

$$P(X_{n+1} = x_{n+1} | X_1^n = x_1^n)$$

for all strings $x_1^n := (x_1, x_2, \ldots, x_n)$ and symbols x_{n+1}.

We may also define particular variables as deterministic functions of previously defined random variables:

$$X_{n+1} = f(X_1^n) \iff P(X_{n+1} = f(x_1^n) | X_1^n = x_1^n) = 1.$$

Example 1: IID processes

Independent identically distributed (IID) processes $\supset$ *fair coin.*

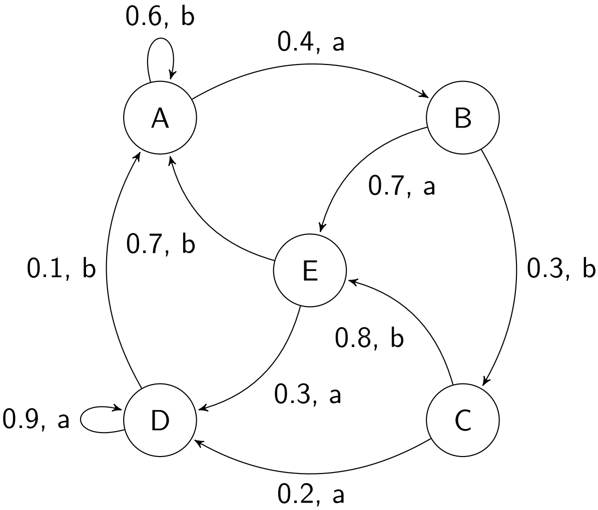
Formally, we have $X_1^n := (X_1, X_2, ..., X_n)$ such that

$$P(X_1^n = x_1^n) = \prod_{i=1}^{n} \pi(x_i), \quad x_1^n \in \mathbb{X}^n.$$

We have $\mathbb{E}\, Z_i = 0$ and $\operatorname{Var} Z_i = 1$ for $Z_i := \frac{1\{X_i = x\} - \pi(x)}{\sqrt{\pi(x)(1-\pi(x))}}$.

1. Law of large numbers (LLN):
   The relative frequencies approach probabilities:
   $\lim_{n\to\infty} \sum_{i=1}^{n} Z_i / n = 0$ with probability 1.
2. Central limit theorem (CLT):
   The distribution of rescaled sample mean $\sum_{i=1}^{n} Z_i / \sqrt{n}$ approaches the Gauss distribution $N(0, 1)$ as $n \to \infty$.
3. Law of the iterated logarithm (LIL):
   $\limsup_{n\to\infty} \left| \sum_{i=1}^{n} Z_i \right| / \sqrt{n \ln \ln n} = 1$ with probability 1.

Example 2: A unifilar finite-state process



Example 3: Unifilar processes (definition)

A process $(X_n)_{n\in\mathbb{N}}$ is called unifilar if there is another process $(Y_n)_{n\in\mathbb{N}}$, called the underlying process, such that each symbol X_n is a stochastic function of the corresponding state Y_n,

$$P(X_n|Y_1^n, X_1^{n-1}) = \varepsilon(Y_n, X_n), \qquad \text{(emission probability)}$$

and the next state Y_{n+1} is a deterministic function of the previous state Y_n and symbol X_n,

$$Y_{n+1} = \delta(Y_n, X_n). \qquad \text{(transition function)}$$

Examples:

- Higher order Markov chains: $Y_n = X_{n-k}^{n-1}$ for a fixed k.
- Recurrent neural networks: Y_n — hidden state of network.
- Any process: $Y_n = X_1^{n-1}$.

Thus, we often put restrictions on $(Y_n)_{n\in\mathbb{N}}$ (finite alphabet, etc.).

Example 4: Santa Fe processes

- A Santa Fe process is a stochastic process $(X_i)_{i\in\mathbb{Z}}$ where individual variables can be decomposed as pairs

$$X_i = (K_i, Z_{K_i})$$

  with two processes $(K_i)_{i\in\mathbb{Z}}$ and $(Z_k)_{k\in\mathbb{N}}$.
- The narration $(K_i)_{i\in\mathbb{Z}}$ consists of topics $K_i : \Omega \to \mathbb{N}$.
- The knowledge $(Z_k)_{k\in\mathbb{N}}$ consists of facts $Z_k : \Omega \to \{-1, 1\}$.
- Process $(X_i)_{i\in\mathbb{Z}}$ is a simple model of a non-contradictory text: Whenever a certain topic is discussed again ($K_i = K_j$), the same fact is reported ($Z_{K_i} = Z_{K_j}$).
- That said, we may assume narration $(K_i)_{i\in\mathbb{Z}}$ and knowledge $(Z_k)_{k\in\mathbb{N}}$ to be pretty arbitrary processes and investigate consequences of our particular choices.

Example 5: Large language models

Large language models are also certain stochastic sources.

In language models based on transformers, probabilities $P(X_t|X_{t-M}^{t-1})$ are computed by stacking two mechanisms:

- embeddings — vectors x_t corresponding to words/concepts,
- attention — a nonlinear operation on embeddings

$$y_t = \sum_{s=t-M}^{t-1} \frac{\exp(x_t \cdot x_s)}{\sum_{r=t-M}^{t-1} \exp(x_t \cdot x_r)} x_s.$$

The GPT-3 language model:

- Number of parameters: $N = 175$ billions (800 GB RAM).
- Context length: $M = 2048$ words.
- Training data: Common Crawl (410 bln, 60%), WebText2 (19 bln, 22%), books (67 bln, 16%), Wikipedia (3 bln, 3%).

Mathematical landscape

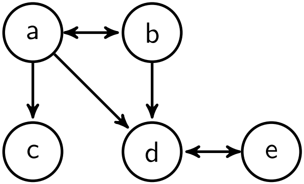
Hierarchy of stochastic processes:

- Fair-coin process
- IID processes
- Markov processes
- Hidden Markov processes
- Stationary processes
- Asymptotically mean stationary (AMS) processes
- Non-stationary processes

1 Overview
2 Markov sources
3 Stationary sources
4 Information theory
5 Sufficient statistic

Markov chains

	a	b	c	d	e
a	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0
b	$\frac{1}{6}$	0	0	$\frac{5}{6}$	0
c	0	0	1	0	0
d	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$
e	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$



A stochastic process $(X_i)_{i \in \mathbb{N}}$ over a countable alphabet $\mathbb{X}$ is called a Markov process if for all $n \in \mathbb{N}$, we have

$$P(X_1^n = x_1^n) = \pi(x_1) \prod_{i=2}^n \tau(x_{i-1}, x_i)$$

for some vector $\pi : \mathbb{X} \to [0,1]$, called the initial distribution, and some matrix $\tau : \mathbb{X} \times \mathbb{X} \to [0,1]$, called the transition matrix.

Communicating classes

For a Markov process with a given transition matrix τ, we say that x leads to y and write it as $x \to y$ if

$$P(X_n = y \text{ for some } n \in \mathbb{N} | X_1 = x) > 0.$$

We also write $x \leftrightarrow y$ if $x \to y$ and $y \to x$.

Relation $\leftrightarrow$ is an equivalence relation on $\mathbb{X}$, i.e., it is

- reflexive: $x \leftrightarrow x$,
- symmetric: $x \leftrightarrow y$ if and only if $y \leftrightarrow x$,
- transitive: $x \leftrightarrow y$ and $y \leftrightarrow z$ implies $x \leftrightarrow z$.

The communicating class of x is defined as

$$[x] := \{y \in \mathbb{X} : x \leftrightarrow y\}.$$

Communicating classes are disjoint and partition space $\mathbb{X}$.

Irreducible, finite, and stationary chains

A transition matrix τ or the respective Markov process are called irreducible if space $\mathbb{X}$ is the single communicating class.

A Markov process is called (in)finite if space $\mathbb{X}$ is (in)finite.

A distribution $\bar{\pi}$ is called invariant for a given transition matrix τ if

$$\sum_{y \in \mathbb{X}} \bar{\pi}(y)\tau(y, x) = \bar{\pi}(x) \quad \text{for all } x \in \mathbb{X}.$$

Theorem

- *Let $(X_i)_{i \in \mathbb{N}}$ be a finite Markov process. Then an invariant distribution exists but need not be unique.*
- *Let $(X_i)_{i \in \mathbb{N}}$ be an irreducible Markov process. Then the invariant distribution is unique if it exists.*

Ergodic theorem

We inductively define random variables called passage times

$$T_0^x := 0, \qquad T_n^x := \inf\left\{n \in \mathbb{N} : n > T_{n-1}^x, X_n = x\right\}.$$

The successive recurrence times are $R_n^x := T_{n+1}^x - T_n^x$.

Random variables $R_1^x, R_2^x, R_3^x, \dots$ form an IID process.

Theorem (ergodic theorem)

Let $(X_i)_{i\in\mathbb{N}}$ be an irreducible Markov process such that the invariant distribution $\bar{\pi}$ exists. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} 1\{X_i = x\} = \frac{1}{\mathbb{E}\, R_i^x} = \bar{\pi}(x) \text{ a.s.}$$

1 Overview

2 Markov sources

3 Stationary sources

4 Information theory

5 Sufficient statistic

Stationary processes

A stochastic process $(X_i)_{i\in\mathbb{Z}}$ is called stationary if for all $t \in \mathbb{Z}$, all $k \in \mathbb{N}$ and all strings x_1^k, we have

$$P(X_{t+1}^{t+k} = x_1^k) = P(X_1^k = x_1^k).$$

Example: Markov sources with an invariant initial distribution.

Theorem (Birkhoff ergodic theorem)

For any stationary process $(X_i)_{i\in\mathbb{Z}}$, all $k \in \mathbb{N}$, and all strings x_1^k, there exist limits

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1\left\{X_{i+1}^{i+k} = x_1^k\right\} \ a.s.$$

$$\lim_{n\to\infty} a_n := a \iff \inf_{n\in\mathbb{N}} \sup_{k\geq n} a_k = a = \sup_{n\in\mathbb{N}} \inf_{k\geq n} a_k$$

Ergodic processes

A stationary process $(X_i)_{i\in\mathbb{Z}}$ is called ergodic if for all $k \in \mathbb{N}$ and all strings x_1^k, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} 1\left\{X_{i+1}^{i+k} = x_1^k\right\} = P(X_1^k = x_1^k) \text{ a.s.}$$

Examples: Markov sources with an invariant initial distribution and an irreducible transition matrix; IID processes.

Theorem (ergodicity criterion)

A stationary process $(X_i)_{i\in\mathbb{Z}}$ is ergodic if for all $k \in \mathbb{N}$ and all strings x_1^k, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} P(X_1^k = x_1^k, X_{i+1}^{i+k} = x_1^k) = P(X_1^k = x_1^k)^2.$$

Ergodic decomposition and prior distributions

- Just like any stationary Markov process can be decomposed into irreducible Markov processes, any stationary process can be decomposed into ergodic processes.
- The important difference is that a stationary Markov process can decompose into countably many ergodic components, whereas a general stationary process can decompose into uncountably many ergodic components.
- Non-ergodic processes are Bayesian mixtures of ergodic processes, where the prior distribution can be arbitrary.
- For example, non-ergodic Santa Fe process $X_i = (K_i, Z_{K_i})$ where $(Z_k)_{k\in\mathbb{N}}$ is an IID process decomposes into uncountably many IID Santa Fe processes $X_i = (K_i, z_{K_i})$ where $(z_k)_{k\in\mathbb{N}}$ are realizations of process $(Z_k)_{k\in\mathbb{N}}$.

1 Overview

2 Markov sources

3 Stationary sources

4 Information theory

5 Sufficient statistic

Block entropy

The block entropy a stationary process $(X_i)_{i\in\mathbb{Z}}$ is

$$H(n) := H(X_1^n) = H(X_1, ..., X_n) = H(X_{i+1}, ..., X_{i+n}).$$

For convenience, we also put $H(0) = 0$.

We have

$$\Delta H(n) := H(n) - H(n-1) = H(X_n | X_1^{n-1}),$$
$$\Delta^2 H(n) := H(n) - 2H(n-1) + H(n-2) = -I(X_1; X_n | X_2^{n-1}).$$

Remark: Block entropy $H(n)$ is non-negative ($H(n) \geq 0$), non-decreasing ($\Delta H(n) \geq 0$) and concave ($\Delta^2 H(n) \leq 0$).

Entropy rate

The entropy rate of a stationary process $(X_i)_{i\in\mathbb{Z}}$ is

$$h = \lim_{n\to\infty} \Delta H(n) = H(1) + \sum_{n=2}^{\infty} \Delta^2 H(n) = \lim_{n\to\infty} \frac{H(n)}{n}.$$

We have $0 \leq h \leq H(1)$.

For a Markov process with invariant distribution $\bar{\pi}$ and matrix τ,

$$h = \sum_{x} \bar{\pi}(x) \left[-\sum_{x'} \tau(x,x') \log \tau(x,x') \right].$$

Theorem (Shannon-McMillan-Breiman theorem)

For any stationary ergodic process $(X_i)_{i\in\mathbb{Z}}$, we have

$$\lim_{n\to\infty} \frac{1}{n} [-\log P(X_1^n)] = h \text{ a.s.}$$

Excess entropy

The excess entropy of a stationary process $(X_i)_{i\in\mathbb{Z}}$ is

$$E = \lim_{n\to\infty} I(X_{-n+1}^0; X_1^n) = \lim_{n\to\infty} [2H(n) - H(2n)]$$
$$= \lim_{n\to\infty} [H(n) - nh] = \sum_{n=1}^{\infty} [\Delta H(n) - h] = -\sum_{n=2}^{\infty} (n-1)\Delta^2 H(n).$$

For a Markov process with invariant distribution $\bar{\pi}$ and matrix τ,

$$E = H(1) - h = \sum_x \bar{\pi}(x) \left[\sum_{x'} \tau(x, x') \log \frac{\tau(x, x')}{\bar{\pi}(x')} \right].$$

Theorem (ergodic decomposition of excess entropy)

For any process $(X_i)_{i\in\mathbb{Z}}$ and parameter $\Theta = f(X_{-\infty}^t) = g(X_{t+1}^\infty)$,

$$E = I(X_{-\infty}^t; X_{t+1}^\infty) = H(\Theta) + I(X_{-\infty}^t; X_{t+1}^\infty|\Theta).$$

Hilberg exponent

For a stationary process, we have the Hilberg exponent

$$\beta := \operatorname*{hilb}_{n\to\infty} \left(H(X_1^n) - nh\right) = \operatorname*{hilb}_{n\to\infty} I(X_1^n; X_{n+1}^{2n}) \in [0,1],$$

where to measure power-law growth, we introduce the operator

$$\operatorname*{hilb}_{n\to\infty} S(n) := \left[\limsup_{n\to\infty} \frac{\log S(n)}{\log n}\right]_+ .$$

In particular, we obtain

$$\operatorname*{hilb}_{n\to\infty} n^\beta = \beta \text{ if } \beta \geq 0.$$

Theorem (excess bound)

If $\lim_{n\to\infty} S(n)/n = s$ *and* $S(n) \geq ns$ *then*

$$\operatorname*{hilb}_{n\to\infty} \left(S(n) - ns\right) = \operatorname*{hilb}_{n\to\infty} \left(2S(n) - S(2n)\right).$$

1 Overview

2 Markov sources

3 Stationary sources

4 Information theory

5 Sufficient statistic

Hilberg's law

- $\beta := \text{hilb}_{n\to\infty} (H(X_1^n) - nh) = \text{hilb}_{n\to\infty} I(X_1^n; X_{n+1}^{2n}) \in [0, 1]$.
- Finite unifilar processes have $\beta = 0$.
- Santa Fe processes $X_i = (K_i, Z_{K_i})$ with Zipfian narration

$$(K_i)_{i\in\mathbb{Z}} \sim \text{IID}, \quad P(K_i = k) \sim \frac{1}{k^\alpha}, \quad \alpha > 1,$$

and uniformly distributed knowledge

$$(Z_k)_{k\in\mathbb{N}} \sim \text{IID}, \quad P(Z_k = \pm 1) = \frac{1}{2},$$

are strongly nonergodic, we have $E = \infty$ and $\beta = 1/\alpha$.

Relationship $\beta > 0$ is called Hilberg's law.

Sufficient statistics (summaries)

A sufficient statistic $T_{X\to Y}$ is a function of variable X such that variables X and Y are independent given $T_{X\to Y}$.

We observe

$$\begin{aligned} I(X;Y) &= I(T_{X\to Y};Y) + \underbrace{I(X;Y|T_{X\to Y})}_{0} \\ &= I(T_{X\to Y};T_{Y\to X}) + \underbrace{I(T_{X\to Y};Y|T_{Y\to X})}_{0} \\ &= I(T_{X\to Y};T_{Y\to X}) \le \min\{H(T_{X\to Y}), H(T_{Y\to X})\}. \end{aligned}$$

The minimal sufficient statistic yields an insight into the divergence of excess entropy E and a bound for the Hilberg exponent β.

Example 1: Finite-state process

Consider a unifilar process $(X_i)_{i\in\mathbb{Z}}$ such that
the underlying process $(Y_i)_{i\in\mathbb{Z}}$ has exactly K distinct states.

We have

$$\begin{aligned} I(X_1^n; X_{n+1}^{2n}) &\le I(X_1^n, Y_{n+1}; X_{n+1}^{2n}) \\ &= I(Y_{n+1}; X_{n+1}^{2n}) + \underbrace{I(X_1^n; X_{n+1}^{2n}|Y_{n+1})}_{0} \\ &\le H(Y_{n+1}) \le \log K, \end{aligned}$$

since samples X_1^n and X_{n+1}^{2n} are indpendent given state Y_{n+1}.

This process has $E \le \log K$ and $\beta = 0$.

Example 2: Biased coin with a prior

Consider a process where we first draw a probability $\theta \in [0, 1]$ and then we repeatedly toss a biased coin—formally, Bernoulli(θ):

$$P(X_1^n = x_1^n | \Theta = \theta) = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}.$$

We have the Markov chain

$$X_1^n \to \sum_{i=1}^n X_i \to \Theta \to \sum_{i=n+1}^{2n} X_i \to X_{n+1}^{2n}.$$

Hence

$$I(X_1^n; X_{n+1}^{2n}) = I\left(\sum_{i=1}^n X_i; \sum_{i=n+1}^{2n} X_i\right) \le H\left(\sum_{i=1}^n X_i\right) \le \log(n+1).$$

This process is strongly nonergodic, has $E = \infty$ but $\beta = 0$.

Example 3: Santa Fe process

Santa Fe process $X_i = (K_i, Z_{K_i})$ with Zipfian narration

$$(K_i)_{i\in\mathbb{Z}} \sim \text{IID}, \quad P(K_i = k) \sim \frac{1}{k^\alpha}, \quad \alpha > 1,$$

and uniformly distributed knowledge

$$(Z_k)_{k\in\mathbb{N}} \sim \text{IID}, \quad P(Z_k = \pm 1) = \frac{1}{2}.$$

Denote $\{X_1^n\} := \{X_i : 1 \le i \le n\}$. We have the Markov chain

$$X_1^n \to \{X_1^n\} \to (Z_k)_{k\in\mathbb{N}} \to \{X_{n+1}^{2n}\} \to X_{n+1}^{2n}.$$

Hence

$$\mathbb{E} \min \mathbb{N} \setminus \{X_1^n\} \cap \{X_{n+1}^{2n}\} \lesssim I(X_1^n; X_{n+1}^{2n}) \lesssim \mathbb{E} \# \{X_1^n\} \log \max \{X_1^n\}.$$

We have Herdan-Heaps' law $\# \{X_1^n\} \sim n^{1/\alpha}$.

This process is strongly nonergodic, has $E = \infty$ and $\beta = 1/\alpha$.

Minimal sufficient statistic for natural language?

- What is the minimal sufficient statistic for natural language?
  - Is it closer to $\sum_{i=1}^{n} X_i$ or to $\{X_i : 1 \le i \le n\}$? *(data aggregation vs. data memorization)*
  - Think of embeddings at the levels of word, sentence, paragraph, chapter, book, etc.
  - Think also of a mental representation of factual knowledge. If it is unbounded, we likely have Hilberg's law.
  - How to model the pressure to forget useless facts?
  - Different people remember different things.—Purposeful randomization of memories or different initial conditions?
- A memory theory (refinement of unifilar processes) is in need:
  - Memories have a non-trivial structure (symbols, vecs, freqs).
  - Memories operate at various time scales.
  - Memories operate in parallel and in complexes.
  - Memories are unbounded but finite and prone to forgetting.
  - Given memories, generation of texts may be simple.
- We need an appropriate level of abstraction—somewhere between Markov chains and general (AMS) processes!

The Road to Wisdom by Piet Hein

The road to wisdom? Well, it's plain
And simple to express:
Err
and err
and err again,
but less
and less
and less.

Further reading

P. Billingsley. *Ergodic Theory and Information*. Wiley & Sons, 1965.

T. M. Cover and J. A. Thomas. *Elements of Information Theory, 2nd ed.* Wiley & Sons, 2006.

J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: The entropy convergence hierarchy. *Chaos*, 15:25–54, 2003.

Ł. Dębowski. *Information Theory Meets Power Laws: Stochastic Processes and Language Models*. Wiley & Sons, 2021.

Ł. Dębowski. A short course in universal coding. Book manuscript in preparation, 2024.

J. R. Norris. *Markov Chains*. Cambridge University Press, 1997.