Natural Language and Strong Nonergodicity

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences

Statistics of Languages: Theories and Experiments
19th–22nd July 2017

1 Introduction

2 Ergodic processes

3 Strongly nonergodic processes

4 Theorem about facts and words

5 Conclusion

6 Additional clarifying remarks

My background and interests

Background:

1. Master's in Physics.
2. Programming work in Computational Linguistics.
3. PhD and Dr. Habil. in Information Theory.

Current interests:

What kind of a stochastic process may model the process of generation of texts in natural language?

1. Statistical laws of language.
2. Probability theory.
3. Information theory (also algorithmic information theory).
4. Computational linguistics.

Plan of the talk

Natural Language and Strong Nonergodicity:

1. **What is a stationary ergodic process?**
   — *A process is ergodic if all empirical frequencies in the long run converge to the probabilities.*
2. **A linguistic interpretation of nonergodic processes.**
   — *Different texts concern different topics. Hence the frequencies of keywords in a randomly selected text are random variables depending on the random text topic.*
3. **Theorem about facts and words.**
   — *If the stochastic process of text generation is sufficiently strongly nonergodic, then the number of "words" detected in the text by the PPM algorithm must be sufficiently large.*

1 Introduction

2 Ergodic processes

3 Strongly nonergodic processes

4 Theorem about facts and words

5 Conclusion

6 Additional clarifying remarks

Ergodic theorem and ergodic processes

1. Consider a discrete process $(X_i)_{i=1}^{\infty} = (X_1, X_2, X_3, \ldots)$.
2. For a string $w = (x_1, \ldots, x_n)$ define random variable

$$Y_i^w := \begin{cases} 1 & \text{if } X_i = x_1, \ldots, X_{i+n-1} = x_n, \\ 0 & \text{else.} \end{cases}$$

3. Process $(X_i)_{i=1}^{\infty}$ is called stationary if expectations $\mathbb{E}\, Y_i^w$ do not depend on i for all w.

Theorem (ergodic theorem)

For any stationary process $(X_i)_{i=1}^{\infty}$, there exist random limits

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i^w = Y^w \text{ almost surely.}$$

4. Process $(X_i)_{i=1}^{\infty}$ is called ergodic if Y^w are constant for all w.

Examples of stationary ergodic processes

1. Process $(X_i)_{i=1}^{\infty}$ is called IID (independent identically distributed) if

$$P(X_1 = x_1, ..., X_n = x_n) = \pi(x_1)...\pi(x_n).$$

— IID processes are ergodic.

2. Process $(X_i)_{i=1}^{\infty}$ is called Markov if

$$P(X_1 = x_1, ..., X_n = x_n) = \pi(x_1)p(x_2|x_1)...p(x_n|x_{n-1}).$$

— A Markov process is ergodic if $p(x_i|x_{i-1}) > c > 0$.

3. Process $(Y_i)_{i=1}^{\infty}$ is called hidden Markov if $Y_i = f(X_i)$ for a certain Markov process $(X_i)_{i=1}^{\infty}$.

— A hidden Markov process is ergodic if the underlying Markov process is ergodic.

1 Introduction

2 Ergodic processes

3 Strongly nonergodic processes

4 Theorem about facts and words

5 Conclusion

6 Additional clarifying remarks

Is natural language ergodic or not?

1. A process is ergodic when frequencies of strings in a sample in the long run converge to constants.
2. Suppose now, we choose at random a text in natural language.
3. **Imagine selecting a random book from a library.**
4. Imagine counting the frequencies of a keyword, such as *bijection* for a text in maths, *fossil* for a text in paleontology.
5. We expect that the frequencies of keywords are random variables with values depending on the random text topic.
6. Since keywords are some strings, the stochastic process that models natural language should be not ergodic = nonergodic.

By counting keywords, we can infer the random text topic.

Nonergodic processes—another perspective

Intuition: Process is nonergodic $\iff$ there exist $\geq$ two topics.

Theorem

Process $(X_i)_{i=1}^{\infty}$ is nonergodic if and only if there exists a function $f(x_1, \dots, x_n)$ of a sequence of symbols and a binary random variable Z such that $0 < P(Z = 0) < 1$ and

$$\lim_{n\to\infty} P(f(X_{t+1}, \dots, X_{t+n}) = Z) = 1 \tag{1}$$

for any position t.

Definition

A binary variable Z satisfying (1) will be called a random fact.

Thus, a process is nonergodic if there exists $\geq$ one random fact.
A random fact tells which of two topics the random text is about.

Santa Fe process—an example of a nonergodic process

1. Let $(Z_k)_{k=1}^{\infty}$ be an IID process with $Z_k \in \{0, 1\}$ and
$$P(Z_k = 0) = P(Z_k = 1) = 1/2.$$
2. Let $(K_i)_{i=1}^{\infty}$ be an IID process with $K_i \in \{1, 2, 3, ...\}$ and
$$P(K_i = k) \propto \frac{1}{k^{\alpha}}, \quad \alpha > 1.$$
3. The Santa Fe process is $(X_i)_{i=1}^{\infty}$, where
$$X_i = (K_i, Z_{K_i}).$$
4. The Santa Fe process is nonergodic since all Z_k are probabilistically independent random facts.

Strong nonergodicity

Intuition: Santa Fe process is strongly nonergodic since there exist infinitely many probabilistically independent random facts.

Definition

Process $(X_i)_{i=1}^{\infty}$ is called strongly nonergodic if there exist functions $f_k(x_1, ..., x_n)$ of a sequence of symbols and a binary IID process $(Z_k)_{k=1}^{\infty}$ such that $P(Z_k = 0) = 1/2$ and

$$\lim_{n\to\infty} P(f_k(X_{t+1}, ..., X_{t+n}) = Z_k) = 1$$

for any position t and any $k = 1, 2, 3, ...$.

(number of persistent topics) $\approx 2^{\text{(number of independent random facts)}}$

strong nonergodicity $\iff$ continuum of topics

Nonergodicity via continuum of topics

1. We have an intuition that:
   - different texts concern different topics,
   - and the topic of a text can be inferred from the text.
2. Natural language would be strongly nonergodic if:

   continuum of topics: the exact topic of an infinitely long text had to be described by an infinitely long sequence of independent binary random variables $Z_1, Z_2, Z_3, \ldots$,

   persistence of topics: there existed fixed binary functions $f_1, f_2, f_3, \ldots$ which would allow to infer $Z_1, Z_2, Z_3, \ldots$ from any sufficiently long finite portion of the infinitely long text.
3. In the above reasoning we assume some idealization:
   - Texts are assumed infinitely long (real ones are finite!).
   - Some topics do not change in a given text
     (is there any persistent topic of an infinitely long text?)

Nonergodicity via frozen randomness of the environment

Consider the physical and cultural environment we live in, which we try to describe and control.

We may suppose that this environment contains some amount of frozen randomness, which accumulates over time.

Natural language is strongly nonergodic if the environment:

1. settles down on a random one of a continuum of possibilities,
2. and is ultimately described by all sufficiently long texts.

topic of all texts $\Longleftrightarrow$ frozen randomness of the environment

1 Introduction

2 Ergodic processes

3 Strongly nonergodic processes

4 Theorem about facts and words

5 Conclusion

6 Additional clarifying remarks

Towards the theorem about facts and words

1. Our considerations may seem pure philosophy...

   ... without any measurable consequences.
2. We will show that the opposite is true.
3. There is the following proposition (informal statement):

Theorem (facts and words)

Suppose we have a finite text drawn from a stationary process. Then the number of distinct PPM words detectable in the text must be roughly greater than the number of independent random facts inferrable from the text.

Intuition: Rich vocabulary of a text is a necessary consequence of a complex world described in the text.

Caution: Converse is not true! Rich vocabulary does not imply high complexity of the described world.

Numbers of random facts and PPM words

1. Consider a random text $X_1^n := (X_1, ..., X_n)$.
2. The set of independent random facts inferrable from X_1^n is:

$$U(X_1^n) := \left\{ l \in \{1, 2, ...\} : f_k(X_1^n) = Z_k \text{ for all } k \le l \right\}.$$

3. The set of all substrings of length m in X_1^n is:

$$V(m|X_1^n) := \left\{ x_1^m : X_{t+1}^{t+m} = x_1^m \text{ for some } 0 \le t \le n - m \right\}.$$

4. Let $G_{\text{PPM}}(X_1^n)$ be the PPM order of X_1^n, i.e,
   the order of the adaptive Markov approximation of the text which yields the best compression rate of the text.
5. The set of distinct PPM words detectable in X_1^n is:

$$V_{\text{PPM}}(X_1^n) := V(G_{\text{PPM}}(X_1^n)|X_1^n).$$

Theorem about facts and words

$$H(X_1^n) := -\sum_{x_1^n} P(X_1^n = x_1^n) \log P(X_1^n = x_1^n) \text{ — entropy}$$

$$I(X_1^n; X_{n+1}^{2n}) := H(X_1^n) + H(X_{n+1}^{2n}) - H(X_1^{2n}) \text{ — mutual information}$$

$$\operatorname*{hilb}_{n\to\infty} a_n := \limsup_{n\to\infty} \frac{\log^+ a_n}{\log n} \text{ — Hilberg exponent: } \operatorname*{hilb}_{n\to\infty} n^\beta = \beta$$

Theorem (facts and words)

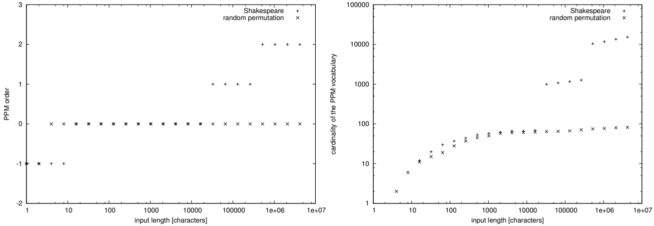
We have inequalities

$$\begin{aligned}\operatorname*{hilb}_{n\to\infty} \mathbb{E} \operatorname{card} U(X_1^n) &\le \operatorname*{hilb}_{n\to\infty} I(X_1^n; X_{n+1}^{2n}) \\ &\le \operatorname*{hilb}_{n\to\infty} \mathbb{E} \left[G_{\text{PPM}}(X_1^n) + \operatorname{card} V_{\text{PPM}}(X_1^n)\right].\end{aligned}$$

For Santa Fe processes, we have an exact power law

$$\operatorname*{hilb}_{n\to\infty} \mathbb{E} \operatorname{card} U(X_1^n) = \operatorname*{hilb}_{n\to\infty} I(X_1^n; X_{n+1}^{2n}) = \beta, \quad \beta = 1/\alpha \in (0,1).$$

Some empirical data for Shakespeare



$G_{\mathrm{PPM}}(x_1^n)$ and **card** $V_{\mathrm{PPM}}(x_1^n)$ versus the input length n for 35 plays by Shakespeare and a text shuffled by characters.

For IID and Markov processes over a finite alphabet, we have

$$\operatorname*{hilb}_{n\to\infty} \mathbb{E}\left[G_{\mathrm{PPM}}(X_1^n) + \mathbf{card}\, V_{\mathrm{PPM}}(X_1^n)\right] = 0.$$

For natural language, we seem to have a stepwise power law

$$\operatorname*{hilb}_{n\to\infty} \mathbb{E}\left[G_{\mathrm{PPM}}(X_1^n) + \mathbf{card}\, V_{\mathrm{PPM}}(X_1^n)\right] = \beta, \quad \beta \in (0,1).$$

Zipf's law and strong nonergodicity

1. Zipf's law — a power law for the distribution of words (words as given by spelling rules).
2. Herdan's law — power law growth of the number of distinct words vs. the text length, an integrated version of Zipf's law.
3. In natural language we seem to have not only Herdan's law for orthographic words but also for PPM words.
4. By the theorem about facts and words, we cannot exclude that natural language is strongly nonergodic.
5. We may suppose that some sort of Zipf's law for PPM words holds for some strongly nonergodic processes more generally.

1. Introduction
2. Ergodic processes
3. Strongly nonergodic processes
4. Theorem about facts and words
5. Conclusion
6. Additional clarifying remarks

Conclusion

Is then natural language strongly nonergodic?

We cannot be sure, but we cannot exclude it, since:

1. We probably live in a world full of frozen randomness, found both in culture and in nature.
2. We try to describe this randomness using natural language.
3. We use surprisingly many distinct words, satisfying Zipf's law, which suggests that this randomness is practically infinite.

1 Introduction

2 Ergodic processes

3 Strongly nonergodic processes

4 Theorem about facts and words

5 Conclusion

6 Additional clarifying remarks

Ergodic theorem revisited

For a string $w = x_1^n = (x_1, \dots, x_n)$, we define

$$Y_i^w := \begin{cases} 1 & \text{if } X_i^{i+n-1} = w, \\ 0 & \text{else.} \end{cases}$$

Theorem (ergodic theorem)

For any stationary process $(X_i)_{i=1}^\infty$, there exist random limits

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n Y_i^w = Y^w \text{ almost surely.}$$

Distribution $\phi(X_1^n = w) := Y^w$ is ergodic almost surely.

If we adopt a frequentist interpretation of probability,
we can see only ergodic processes (provided they are stationary).

Ergodic decomposition

Nonergodic processes arise when we adopt a Bayesian interpretation of probability.

Theorem (ergodic decomposition)

Any stationary distribution $\boldsymbol{P}$ can be represented as

$$\boldsymbol{P}(\boldsymbol{X}_1^n = \boldsymbol{x}_1^n) = \int \phi(\boldsymbol{X}_1^n = \boldsymbol{x}_1^n) d\nu(\phi),$$

where $\boldsymbol{\nu}$ is a unique distribution on stationary ergodic distributions.

Stationary ergodic distributions are some building blocks from which we can construct any stationary distribution.

Ergodic decomposition and computability

Ergodic Bernoulli(θ) process distribution:

$$\phi(X_1^n = x_1^n|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}, \quad x_i \in \{0,1\}.$$

Nonergodic mixture Bernoulli process distribution:

$$P(X_1^n = x_1^n) = \int \phi(X_1^n = x_1^n|\theta) w(\theta) d\theta, \quad w(\theta) \text{ is a prior.}$$

1. Suppose that parameter θ is not a computable real number. Then distribution $\phi(X_1^n = x_1^n|\theta)$ is not computable.
2. Suppose that prior w is a computable distribution. Then distribution $P(X_1^n = x_1^n)$ is computable.

Although ergodic distributions are some building blocks from which we can construct any stationary distribution, some nonergodic distributions are computationally simpler than their ergodic components (i.e., their building blocks).

Algorithmic randomness and nonergodicity

Kolmogorov complexity $K(x_1^n)$ is the length of the shortest self-delimiting program that prints out string x_1^n.

An infinite sequence of data $x_1, x_2, \ldots$ is called Martin-Löf algorithmically random w.r.t. a computable distribution P when

$$\inf_{n>0} \left[K(x_1^n) + \log P(X_1^n = x_1^n) \right] > -\infty.$$

The set of algorithmically random sequences has full measure P.

When we are given an infinite sequence of data, we may entertain a hypothesis that the sequence is algorithmically random with respect to some distribution. This distribution need not be ergodic.