

# Bayesian multimodeling: minimum description length

MIPT

2022

# Occam's razor



- William of Ockham: «Plurality must never be posited without necessity».

# Occam's razor



- William of Ockham: «Plurality must never be posited without necessity».
- Modern interpretation: entities should not be multiplied beyond necessity.
- Paul Dirac: «A theory with mathematical beauty is more likely to be correct than an ugly one that fits some experimental data.»
- Albert Einstein: «Everything Should Be Made as Simple as Possible, But Not Simpler»

# When Occam's razor does not work

Occam's razor is an empirical rule for sorting hypothesis during research.

It can be wrong:

- Ernst Mach: molecules are a fictitious construct, as they are not observable.

# Minimum description length

## Task

Given a string: 001011001011001011... 001011, where the pattern 001011 repeats 100500 times.

How can we describe this string?

- `s == "001011...001011001011001011")`
- `s == (''.join('001011' for _ in range(100500)))`

# Kolmogorov complexity

## Definition

Given a computable partially defined mapping from a set of binary words into itself:

$$T : \{0, 1\}^* \rightarrow \{0, 1\}^*.$$

Kolmogorov complexity of the binary string  $x$  is a minimal description length w.r.t.  $T$ :

$$K_T(x) = \min_{f \in \{0, 1\}^*} \{|f| : T(f) = x\},$$

# Kolmogorov complexity

Generally, Kolmogorov complexity is uncomputable.

## Definition

Given a computable partially defined mapping from a set of binary words into itself:

$$T : \{0, 1\}^* \rightarrow \{0, 1\}^*.$$

Kolmogorov complexity of the binary string  $x$  is a minimal description length w.r.t.  $T$ :

$$K_T(x) = \min_{f \in \{0, 1\}^*} \{|f| : T(f) = x\},$$

# Discrete distribution entropy

## Definition

Given a discrete variable  $x$  with probability  $p$  and values  $x_1, \dots, x_n$ , An entropy of  $x$  is:

$$H(x) = - \sum_{i=1}^n p(x = x_i) \log p(x = x_i).$$



# Discrete distribution entropy

## Definition

Given a discrete variable  $x$  with probability  $p$  and values  $x_1, \dots, x_n$ , An entropy of  $x$  is:

$$H(x) = - \sum_{i=1}^n p(x = x_i) \log p(x = x_i).$$

- interpretation: measure of disorder in the distribution;
- maximal at uniform distribution;
- minimal at a distribution with concentration only at one event ( $x_i = 1, x_j = 0, i \neq j$ ).

# Discrete distribution entropy

## Definition

Given a discrete variable  $x$  with probability  $p$  and values  $x_1, \dots, x_n$ , An entropy of  $x$  is:

$$H(x) = - \sum_{i=1}^n p(x = x_i) \log p(x = x_i).$$

- interpretation: measure of disorder in the distribution;
- maximal at uniform distribution;
- minimal at a distribution with concentration only at one event ( $x_i = 1, x_j = 0, i \neq j$ ).
- **relation to Kolmogorov complexity:**

$$K(x) \leq H(x) + O(\log n)$$

for binary string with length  $n$ .

# Minimum description length principle

$$\text{MDL}(f, \mathcal{D}) = L(f) + L(\mathcal{D}|f),$$

where  $f$  is a model,  $\mathcal{D}$  is a dataset,  $L$  is a description length in bits.

$$\text{MDL}(f, \mathcal{D}) \sim L(f) + L(\mathbf{w}^*|f) + L(\mathcal{D}|\mathbf{w}^*, f),$$

$\mathbf{w}^*$  — optimal parameters.

|                |                   |                                  |   |
|----------------|-------------------|----------------------------------|---|
| $\mathbf{f}_1$ | $L(\mathbf{f}_1)$ | $L(\mathbf{w}_1^* \mathbf{f}_1)$ | $L(\mathcal{D} \mathbf{w}_1^*, \mathbf{f}_1)$ |
| $\mathbf{f}_2$ | $L(\mathbf{f}_2)$ | $L(\mathbf{w}_2^* \mathbf{f}_2)$ | $L(\mathcal{D} \mathbf{w}_2^*, \mathbf{f}_2)$ |
| $\mathbf{f}_3$ | $L(\mathbf{f}_3)$ | $L(\mathbf{w}_3^* \mathbf{f}_3)$ | $L(\mathcal{D} \mathbf{w}_3^*, \mathbf{f}_3)$ |

# MDL: example

## Задача

Given a string: 001011001011001011... 001011, where the pattern 001011 repeats 100500 times.

How can we describe this string?

- `s == "001011...001011001011001011"`
- `s == (''.join('001011' for _ in range(100500)))`
- `import re; re.match('(001011){100500}')`
- $L(f_1) = 0, L(\mathcal{D}|f_1) = 100505;$
- $L(f_2) = 0, L(\mathcal{D}|f_2) = 45;$
- $L(f_3) \gg 0, L(\mathcal{D}|f_3) = 38;$

# MDL and Kolmogorov complexity

**Kolmogorov complexity** is a minimum description length for the dataset, described by a given programming language.

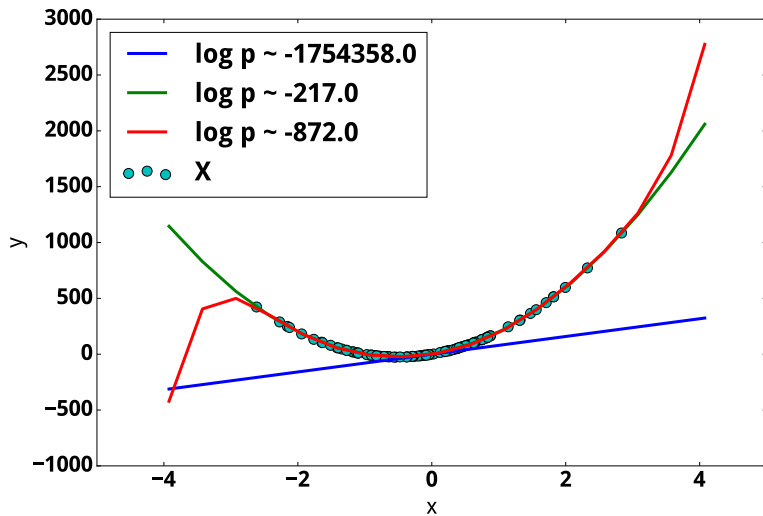
## **Invariance theorem**

Given any description language  $L$ , the optimal description language is at least as efficient as  $L$ , with some constant overhead.

## **Difference from MDL:**

- Incomputable.
- Code length depends on the language. For small sample size the invariance theorem can give poor results.

# Probability coding



# MDL + probability

Model selection problem can be viewed as a problem of information transmission, from encoder to decoder.

Given a dataset  $X, x \in X$ .

- Encoder encodes information about the dataset  $X$  with some code  $f$  and transmits it to the decoder.
- Decoder decodes the information  $f(X)$  and restores the data  $X$  (with loss).
- Problem is to find optimal coding method for  $x$
- Code length:  $-\log p(x)$

Quality criterion is *Regret*:

$$R(x) = -\log P(x) + \min_{f \in \mathfrak{F}} (\log P(x|f)).$$

It shows the difference between the length of the real code  $\log P(x)$  for  $x$  and the best code from the set of codes  $\mathfrak{F}$ .

Regret for datasets with parameterized distribution:

$$R(X) = \max_{x \in X} (-\log P(x) + \min_w (\log P(x|w))).$$

# MDL and Evidence

## Statement

Let the likelihood function  $p(X|w, f)$  be from an exponential family of distributions:

$$p(x|w, f) = h(x)g(\boldsymbol{\eta})\exp(\boldsymbol{\eta} \cdot T(x)),$$

where  $h, g, T$  are some functions,  $\boldsymbol{\eta}$  is a distribution parameter.

Let prior be Jeffreys prior:

$$p(w|f) = \frac{\sqrt{I(w)}}{\int_w \sqrt{I(w)}},$$

Then regret and evidence differ only by some constant when  $n \rightarrow \infty$ :

$$\lim_{n \rightarrow \infty} \left( R(X) - \int_w p(X|w, f)p(w|f)dw \right) = \text{Const.}$$



# MDL vs. Evidence

| <b>Evidence</b>                 | <b>MDL</b>                   |
|---------------------------------|------------------------------|
| Uses prior knowledge            | Independent from prior       |
| Uses data generation hypothesis | Minimizes description length |

# References

- MacKay D. J. C., Mac Kay D. J. C. Information theory, inference and learning algorithms. – Cambridge university press, 2003.
- Grunwald P. A tutorial introduction to the minimum description length principle //arXiv preprint math/0406077. – 2004.
- Успенский В., Шень А., Верещагин Н. Колмогоровская сложность и алгоритмическая случайность. – Litres, 2017
- Grunwald P., Vitányi P. Shannon information and Kolmogorov complexity //arXiv preprint cs/0410002. – 2004.
- Vereshchagin N. K., Vitányi P. M. B. Kolmogorov's structure functions and model selection //IEEE Transactions on Information Theory. – 2004. – Т. 50. – №. 12. – С. 3265-3290.
- Штарьков Ю. М. Универсальное последовательное кодирование отдельных сообщений //Проблемы передачи информации. – 1987. – Т. 23. – №. 3. – С. 3-17.
- When Occam's razor does not work:  
<https://hsm.stackexchange.com/questions/26/was-occam-s-razor-ever-wrong>