

From Complexity to Intelligence

Algorithmic Information Theory and Machine Learning

Nils Holzenberger

October 5, 2026

Outline

- 1 Minimum description length
 - Statistical machine learning
 - Machine learning and compression
 - Minimum description length
- 2 Complexity and neural networks
- 3 Complexity and learning
 - Analogies
 - Anomaly detection
 - Learning with knowledge

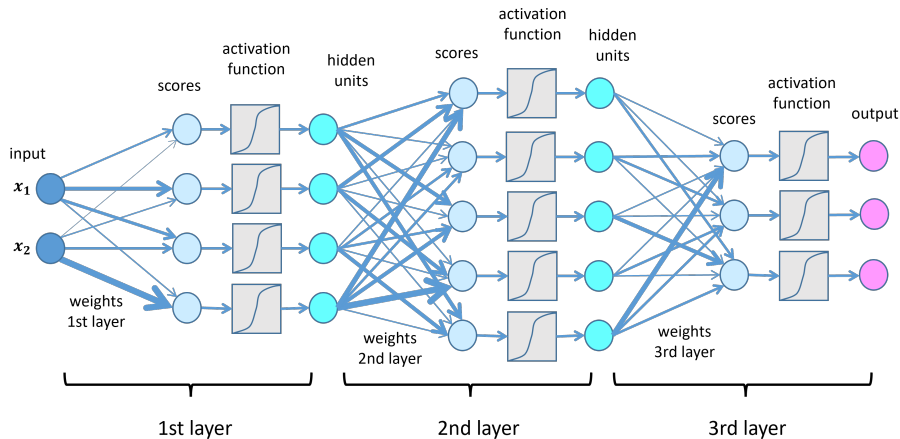
What is this object



Outline

- 1 Minimum description length
 - Statistical machine learning
 - Machine learning and compression
 - Minimum description length
- 2 Complexity and neural networks
- 3 Complexity and learning

Machine Learning



Learning objective in statistical Machine Learning

$$\min_{\theta} \sum_{(x,y) \in \mathcal{D}} -\log p_{\theta}(y|x)$$

- θ are the parameters of the model
- $\mathcal{D}$ is the set of training examples
- $p_{\theta}(y|x)$ is the probability assigned to label y given input x by the parametric model

Maximum Likelihood Estimation

Generative variant:

$$\min_{\theta} \sum_{x \in \mathcal{D}} -\log p_{\theta}(x)$$

The compression objective

This is a compression objective

Shannon's surprise: $I(p) = -\log p$

Cross-entropy: $H(q, p) = -\sum_x q(x) \log p(x)$

With q the empirical distribution, $\forall x, q(x) = \frac{1}{|\mathcal{D}|}$ so

$$H(q, p_\theta) = \sum_{x \in \mathcal{D}} -\frac{1}{|\mathcal{D}|} \log p_\theta(x)$$

Maximum Likelihood Estimation minimizes the cross-entropy between the empirical distribution and the parametric model's distribution. It is equivalent to finding a code with minimal expected length.

A compressionist view of LLMs

Understanding LLM Behaviors via Compression: Data Generation, Knowledge Acquisition and Scaling Laws

Zhixuan Pan^{1,*} Shaowen Wang^{1,*} Pengfei Liao² Jian Li^{1,†}

¹Institute for Interdisciplinary Information Sciences, Tsinghua University

²School of Computer Science and Engineering, Beihang University

panzx24@mails.tsinghua.edu.cn, wangsw23@mails.tsinghua.edu.cn

liaopf22@buaa.edu.cn, lapordge@gmail.com

A compressionist view of LLMs

ANNALS OF TECHNOLOGY

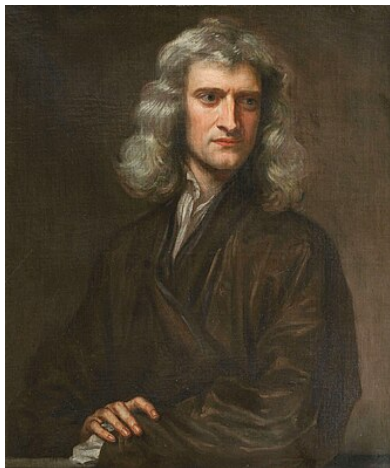
CHATGPT IS A BLURRY JPEG OF THE WEB

OpenAI's chatbot offers paraphrases, whereas Google offers quotes. Which do we prefer?

By **Ted Chiang**

February 9, 2023

Science and compression



- A body remains at rest, or in motion at a constant speed in a straight line, unless it is acted upon by a force.

- $$\frac{d\vec{p}}{dt} = \sum \vec{F}$$

- $$\vec{F}_{A \rightarrow B} = -\vec{F}_{B \rightarrow A}$$

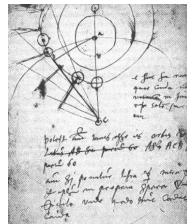
Science and compression



- The orbit of a planet is an ellipse with the Sun at one of the two foci.
- A line segment joining a planet and the Sun sweeps out equal areas during equal intervals of time.
- The square of a planet's orbital period is proportional to the cube of the length of the semi-major axis of its orbit.

$$\frac{T^2}{a^3} = \text{constant}$$

Science and compression



Science and compression

COMPRESSION

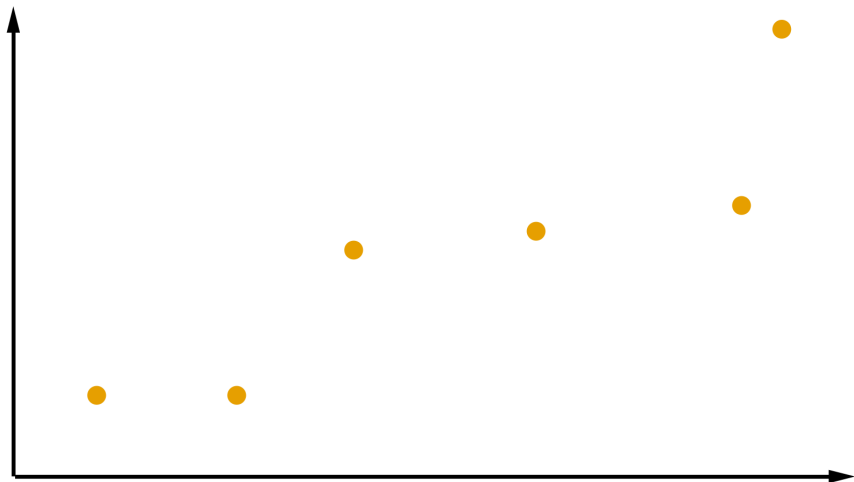
Reality → Observations → Empirical laws → Principles



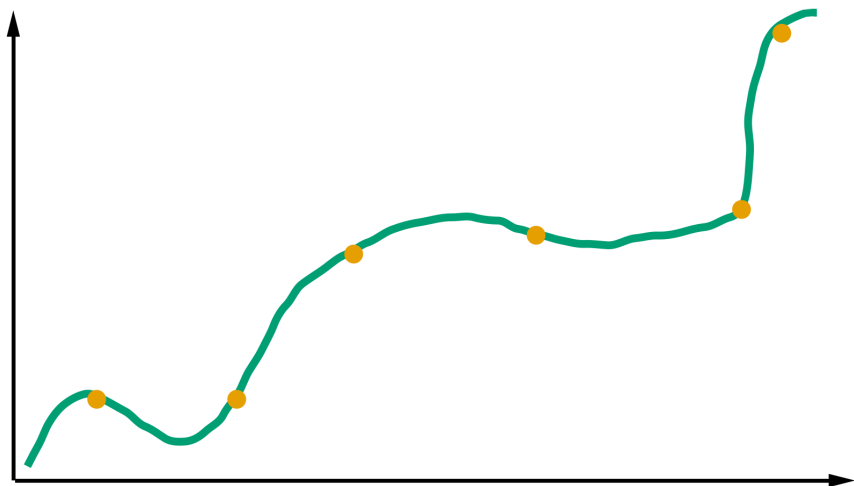
$$\frac{T^2}{a^3} = \text{constante}$$

$$\frac{d\vec{p}}{dt} = \sum_i \vec{F}_i$$

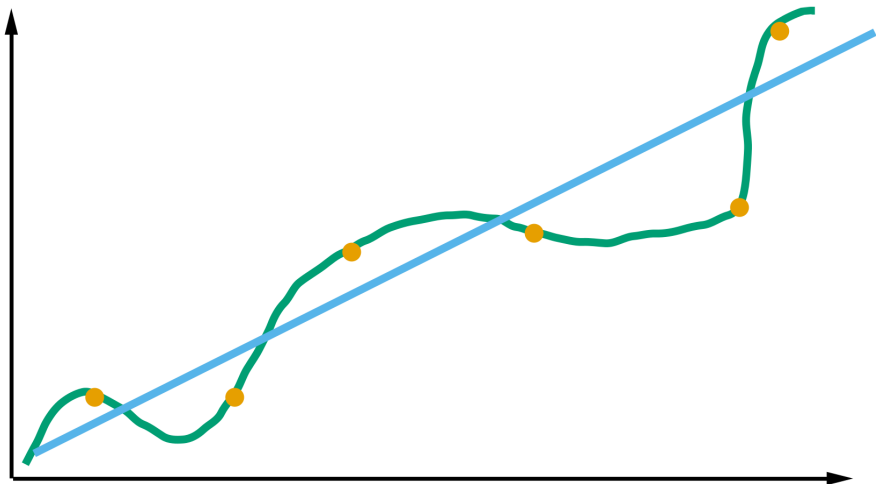
Induction is compression



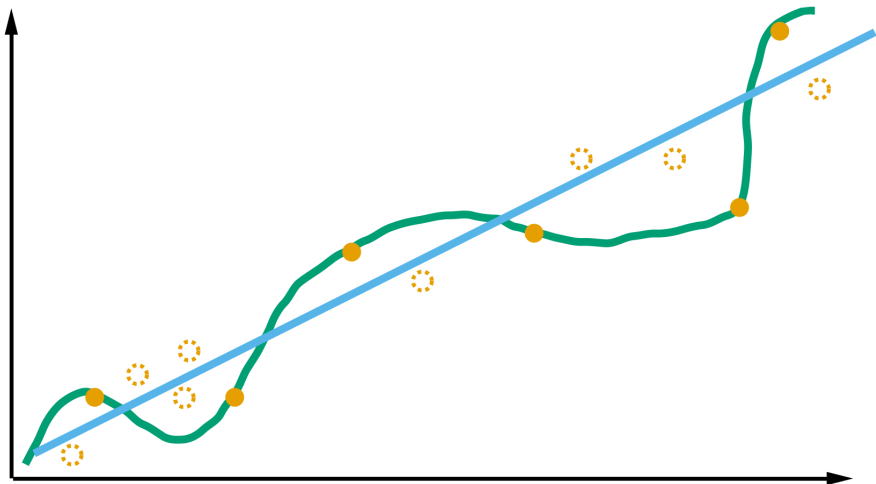
Induction



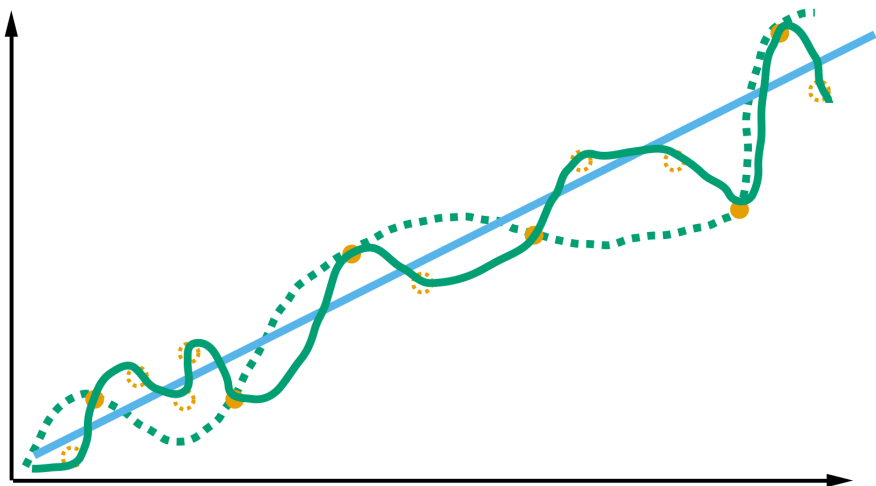
Induction is compression



Induction is compression



Induction is compression



Solomonoff's induction

A FORMAL THEORY OF INDUCTIVE INFERENCE, Part I^{*†}

Ray J. Solomonoff

Visiting Professor, Computer Learning Research Center

Royal Holloway, University of London

Mailing Address: P.O.B. 400404, Cambridge, Ma. 02140, U.S.A.

Information and Control, Volume 7, No. 1, Pp. 1-22, March 1964 Copyright
by Academic Press Inc.

The MDL formula

$$C(\mathcal{D}) \leq C(\mathcal{D}|M) + C(M)$$

→ MDL objective $\min_M \sum_{x \in \mathcal{D}} C(x|M) + C(M)$

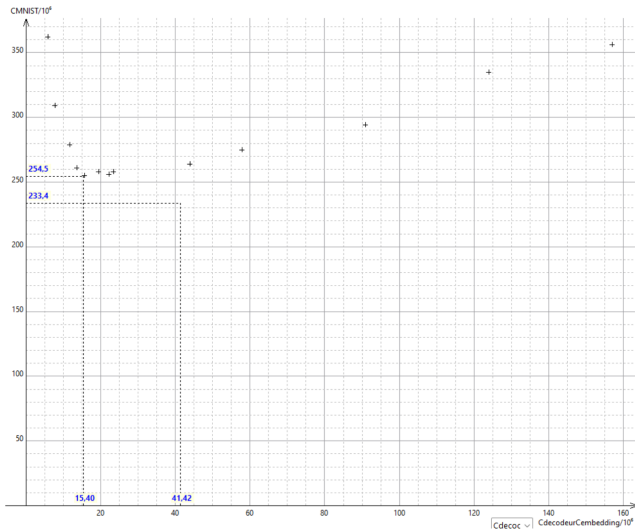
C only works with lossless compression.

Compressing MNIST

Minimize $\sum_{x \in \mathcal{D}} C(x|M) + C(M)$ where

- $\mathcal{D}$ is MNIST, x is a black-and-white image of a digit, $32 \times 32 = 1024$ pixels
- M is an autoencoder, with varying size d in bottleneck layer
- $C(x|M)$ is the sum of
 - the size of the hidden state of the encoder: $|\mathcal{D}| \cdot d \cdot 16$ bits,
 - the decoding error: $\sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{1024} C([x_i - \hat{x}_i]_j) = \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{1024} \log(1 + [x_i - \hat{x}_i]_j)$
- $C(M)$ is the size of the decoder: $|\theta| \cdot 16$ bits

Compressing MNIST



Outline

- 1 Minimum description length
- 2 Complexity and neural networks
- 3 Complexity and learning

Probabilities and MDL

- Assume probabilistic framework, i.e. samples from $\mathcal{D}$ with p.d.f. p drawn i.i.d.
- minimize $K(p_\theta) + \sum_{x \in \mathcal{D}} K(x|p_\theta)$
- typically, $K(x|p_\theta) = -\log p_\theta(x)$
- every x has the same code length

A COMPLEXITY-BASED THEORY OF COMPOSITIONALITY

Eric Elmoznino^{*,1,2}, Thomas Jiralerspong^{*,1,2}, Yoshua Bengio^{1,2}, Guillaume Lajoie^{1,2}

¹Mila – Quebec AI Institute, ²Université de Montréal

Variable-length compression

- Sample a variable-length sequence of symbols $w_1, \dots, w_n$ from a set of symbols $\mathcal{V}$: $K(p_w) + K(w_1, \dots, w_n | p_w)$
- Turn each symbol w_i into a vector x_i with a mapping f :
 $\sum_i K(x_i | f(w_i)) + K(f)$
- The model is f , p_w and the $(w_j^{(i)})_j$ for each x_i (denote by matrix W)
- $\min_{f, p_w, W} \sum_i K(x_i | f(W_i)) + K(W | p_w) + K(p_w) + K(f)$

$K(W|p_w)$

- Use the distribution to describe the samples:

$$K(W|p_w) = \sum_i \sum_j -\log p_w(W_{i,j})$$

- This is an approximation of the true complexity
- In principle the error requires an infinite number of bits to be encoded. Assume a fixed-size precision.

$K(p_w)$ and $K(f)$

- p_w and f are typically large models with many parameters
- How to measure their description length?
 - Encode each parameter on fixed-size precision, or
 - Use a prequential code¹

¹Léonard Blier and Yann Ollivier, *The Description Length of Deep Learning models*, NeurIPS 2018

$$K(x_i | f(W_i))$$

- Decode the sequence $w_1, \dots, w_n$ into x
- f can be parameterized as a Transformer model or Recurrent Neural Network

How to find the minimum?

- In practice, train a discrete auto-encoder
- $W = \text{encoder}(X)$
- Loss function: $-\log p_w(\text{encoder}(X)) - \log p(W|\text{decoder}(\text{encoder}(X)))$
- $\triangle!$ The above only minimizes $K(W|p_w)$ and $K(x_i|f(W_i))$

Outline

- 1 Minimum description length
- 2 Complexity and neural networks
- 3 Complexity and learning**
 - Analogies
 - Anomaly detection
 - Learning with knowledge

ghi $\rightarrow$ ghj, uuvvww $\rightarrow$?

ghi	$\rightarrow$	ghj
uuvvww	$\rightarrow$	uuvvxx
		uuvvjj
		uuvvwx
		ghj
		uuwvwx
		uuvvj
		uuvvww
		uuvvwj
		error
		ghjjjj
		ghjvww
		uujuvww
		uuw
		uuwvww

Simple for a simple reason

INVERTED JENNY



The World's Most Famous Stamp

HISTORY OF FIRST U.S. GOVERNMENT AIRMAIL
 PRODUCTION OF FIRST U.S. AIRMAIL STAMP
 ROBEY'S DISCOVERY AT THE POST OFFICE
 SALE RECORDS OF THE 100 INVERTED JENNYS
 OWNERS' BIOGRAPHIES AND STORIES

Society of Inverted Jenny Owners



Membership requires documentation of
 past or present ownership of a genuine
 Inverted Jenny.

MEMBER
SIGN IN

Not a member?
REGISTER HERE



2.000€



2.000.000€

Simple for a simple reason: $C(s) \leq C(s|f) + C(f)$

What is this object

