Notions of Complexity and Information

Matilde Marcolli

Ma148a: Geometry and Physics of Information
Caltech, Fall 2021

Kolmogorov complexity

- Let $T_{\mathcal{U}}$ be a universal Turing machine (a Turing machine that can simulate any other arbitrary Turing machine: reads on tape both the input and the description of the Turing machine it should simulate)
- Given a string w in an alphabet $\mathfrak{A}$, the Kolmogorov complexity

$$\mathcal{K}_{T_{\mathcal{U}}}(w) = \min_{P: T_{\mathcal{U}}(P)=w} \ell(P),$$

minimal length of a program that outputs w

- universality: given any other Turing machine T

$$\mathcal{K}_T(w) = \mathcal{K}_{T_{\mathcal{U}}}(w) + c_T$$

shift by a bounded constant, independent of w; c_T is the Kolmogorov complexity of the program needed to describe T for $T_{\mathcal{U}}$ to simulate it

• conditional Kolmogorov complexity

$$\mathcal{K}_{T_{\mathcal{U}}}(w \,|\, \ell(w)) = \min_{P:\, T_{\mathcal{U}}(P,\ell(w))=w} \ell(P),$$

where the length $\ell(w)$ is given and made available to machine $T_{\mathcal{U}}$

$$\mathcal{K}(w \,|\, \ell(w)) \leq \ell(w) + c,$$

because if know $\ell(w)$ then a possible program is just to write out w: then $\ell(w) + c$ is just number of bits needed for transmission of w plus print instructions

• upper bound

$$\mathcal{K}_{T_{\mathcal{U}}}(w) \leq \mathcal{K}_{T_{\mathcal{U}}}(w \,|\, \ell(w)) + 2 \log \ell(w) + c$$

if don't know a priori $\ell(w)$ need to signal end of description of w (can show for this suffices a "punctuation method" that adds the term $2 \log \ell(w)$)

• any program that produces a description of w is an upper bound on Kolmogorov complexity $\mathcal{K}_{T_{\mathcal{U}}}(w)$

Problems with Kolmogorov complexity

• any program that produces a description of w is an upper bound on Kolmogorov complexity $\mathcal{K}_{T_{\mathcal{U}}}(w)$

• good upper bounds but not lower bounds (non-computability, halting problem)

• $\mathcal{K}$ assigns large complexity to random sequences

• not the heuristic/intuitive notion of "complexity" (interesting patterns)

• are there better notions of complexity?

Kolmogorov Complexity and Entropy

- Independent random variables X_k distributed according to Bernoulli measure $\mathbb{P} = \{p_a\}_{a\in\mathfrak{A}}$ with $p_a = \mathbb{P}(X = a)$
- Shannon entropy $S(X) = -\sum_{a\in\mathfrak{A}} \mathbb{P}(X = a) \log \mathbb{P}(X = a)$
- $\exists c > 0$ such that for all $n \in \mathbb{N}$

$$S(X) \leq \frac{1}{n} \sum_{w\in\mathcal{W}^n} \mathbb{P}(w)\, \mathcal{K}(w \,|\, \ell(w) = n) \leq S(X) + \frac{\#\mathfrak{A}\, \log n}{n} + \frac{c}{n}$$

- expectation value

$$\lim_{n\to\infty} \mathbb{E}(\frac{1}{n}\mathcal{K}(X_1 \cdots X_n \,|\, n)) = S(X)$$

average expected Kolmogorov complexity for length n descriptions approaches Shannon entropy

Kraft inequality for prefix-free codes

• prefix-free codes (prefix codes): code where no code word is a prefix of another code word (self-punctuating codes)

• Kraft inequality for prefix-free codes:
prefix code in an alphabet $\mathfrak{A}$ of size $N = \#\mathfrak{A}$; lengths of code words $\ell(w_1), \ldots \ell(w_m)$

$$\sum_{i=1}^{m} D^{-\ell(w_i)} \leq 1$$

and any such inequality is realized by lengths of code words of some prefix-free code

• Relation between optimal encoding and Shannon entropy

$$S_D(X) \leq \sum_{i=1}^{m} \mathbb{P}(w_i)\ell(w_i) \leq S_D(X) + 1$$

for $D = \#\mathfrak{A}$ and $S_D =$ Shannon entropy with $\log_D$ with $w_1, \ldots, w_m$ code words of optimal lengths for a source X randomly distributed according to Bernoulli $\mathbb{P} = \{p_a\}$

Why Kraft inequality?

• Main observation: a set of prefix-free binary code words corresponds to a binary tree and oriented paths from the root to one of the leaves ($0 =$ turn right, $1 =$ turn left at the next node)

• for simplest tree with only one step equality $\frac{1}{2} + \frac{1}{2} = 1$

• for other binary trees, Kraft inequality proved inductively over subtrees: isolating root and first subsequent nodes

• Shannon entropy estimate from Kraft inequality

$$S(X) - \sum_{i=1}^{m} \mathbb{P}(w_i)\ell(w_i) \leq \sum_i \mathbb{P}(w_i) \log_2(\frac{2^{-\ell(w_i)}}{\mathbb{P}(w_i)})$$

$$= \log_2(e) \sum_i \mathbb{P}(w_i) \log(\frac{2^{-\ell(w_i)}}{\mathbb{P}(w_i)}) \leq \log_2(e) \sum_i \mathbb{P}(w_i)(\frac{2^{-\ell(w_i)}}{\mathbb{P}(w_i)} - 1) \leq 0$$

using $\log(x) \leq x - 1$ and Kraft inequality

Kraft inequality for Turing machines

- prefix-free Turing machine: programs on which it halts are prefix-free codes (unidirectional input/output tapes, bidirectional work tapes...)
- universal prefix-free Turing machine $T_{\mathcal{U}}$
- encode programs P using a prefix-free (binary) code
- Kraft inequality

$$\sum_{P\,:\,T_{\mathcal{U}}(P)\text{ halts}} 2^{-\ell(P)} \leq 1$$

- Universal (Sub)Probability

$$\mathbb{P}_{T_{\mathcal{U}}}(w) = \sum_{P\,:\,T_{\mathcal{U}}(P)=w} 2^{-\ell(P)} = \mathbb{P}(T_{\mathcal{U}}(P) = w)$$

over an ensemble of randomly drawn programs (expressed by binary codes) most don't halt (or crash) but some halt and output w

Levin's Probability Distribution

- prefix-free Kolmogorov complexity

$$\mathcal{KP}_{T_{\mathcal{U}}}(x) = \min_{P\,:\,T_{\mathcal{U}}(P)=x} \ell(P)$$

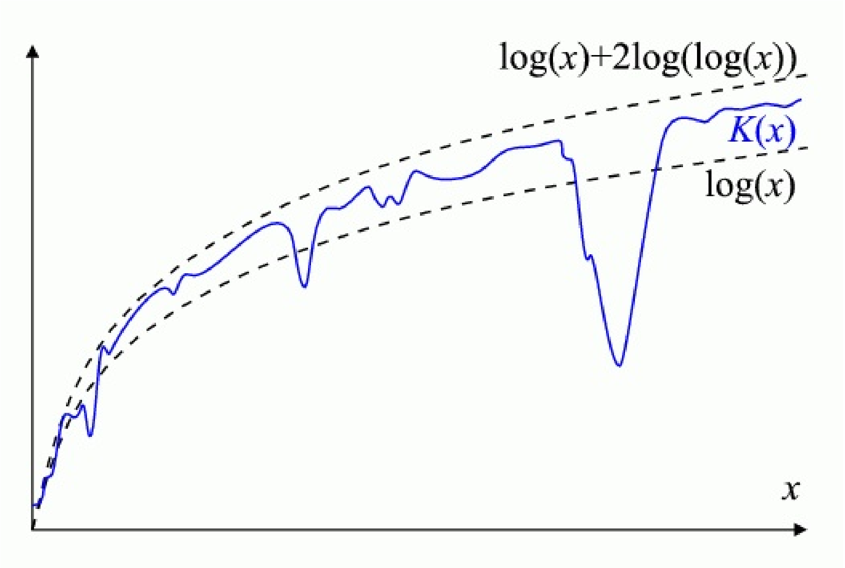
$T_{\mathcal{U}} =$ universal prefix-free Turing machine

- Relation of universal measure to Kolmogorov complexity:

$$\mathbb{P}_{T_{\mathcal{U}}}(w) \sim 2^{-\mathcal{KP}_{T_{\mathcal{U}}}(w)}$$

- dominance of shortest program
  - L.A. Levin, *Various measures of complexity for finite objects (axiomatic description)*, Soviet Math. Dokl., Vol.17 (1976) N.2, 522–526.
  - A.K. Zvonkin, L.A. Levin, *Complexity of finite objects and the development of the concept of information and randomness by means of the theory of algorithms*, Uspehi Mat. Nauk, Vol.25 (1970) no. 6(156), 85–127.

behavior of prefix-free Kolmogorov complexity



Gell-Mann Effective Complexity

- unlike Kolmogorov complexity does not measure description length of whole object
- based on description length of "regularities" (structured patterns) contained in the object
- a completely random sequence has maximal Kolmogorov complexity but zero effective complexity (it contains no structured patterns)
- distinguish system complexity from structural complexity

Gell-Mann Effective Complexity

• Nihat Ay, Markus Mueller, Arleta Szkola, *Effective complexity and its relation to logical depth*, IEEE Trans. Inf. Th., Vol. 56/9 (2010) 4593–4607. [arXiv:0810.5663]

- total information: combination of Kolmogorov complexity and Shannon entropy

$$\mathcal{T}(x, \mathbb{E}) := \mathcal{K}(x|\mathbb{E}) + H(\mathbb{E})$$

  with $\mathbb{E}$ a statistical ensemble and x a datum
- Kolmogorov complexity term $\mathcal{K}(x|\mathbb{E})$ measures algorithmic complexity of computing x assuming it belongs to the statistical ensemble $\mathbb{E}$
- $H(\mathbb{E})$ computes the Shannon entropy of the ensemble

- for a datum x, one looks for a choice of $\mathbb{E}$ that minimizes the total information: $\mathbb{E}$ is a best fitting statistical model for x
- one also wants a choice of $\mathbb{E}$ with the property that x is "typical" in the statistics determined by $\mathbb{E}$ $\Rightarrow$ probability $\mathbb{E}(x)$ of x in the statistics $\mathbb{E}$ not much smaller than predicted by Shannon entropy $2^{-H(\mathbb{E})}$
- these conditions identify a set $\mathcal{M}_x$ of candidates $\mathbb{E}$: good statistical models explaining the datum x
- effective complexity of datum x is minimal value of Kolmogorov complexity $\mathcal{K}(\mathbb{E})$ over candidate models $\mathbb{E}$

$$\mathcal{E}(x) = \min_{\mathbb{E}\in\mathcal{M}_x} \mathcal{K}(\mathbb{E})$$

- completely random patterns have *small* effective complexity

Logical Depth

1. Charles H. Bennett, *Logical Depth and Physical Complexity*, in "The Universal Turing Machine – a Half-Century Survey" (Ed. Rolf Herken), Oxford University Press, 1988.
2. Charles H. Bennett, Peter Gács, Ming Li, Paul M.B. Vitányi, Wojciech H. Zurek, *Information distance*, IEEE Transactions on Information Theory, 44(1998) N.4, 1407–1423.

- Bennett's notion of logical depth is another variant of complexity using execution time of a nearly-minimal program rather than length of minimal program

$$\mathcal{D}_\alpha(x) = \min_P \{\tau(P) \,|\, \ell(P) - \mathcal{K}(x) \leq \alpha, \ T_U(P) = x\}$$

- computing minimum time of execution of a program P that outputs x, whose length equals (or just slightly larger than) minumum one (whose length is $\mathcal{K}(x)$)
- allowed discrepancy measured by parameter α

- from minimal to nearly-minimal: avoid problem that some slightly longer programs may have shorter execution time
- it seems small change from from length of a program to its execution time but significant effect in reducing role of randomness in high complexity patterns
- how $\mathcal{D}_\alpha(x)$ changes compared to effective complexity $\mathcal{E}(x)$?
- phase transition: for small values of $\mathcal{E}(x)$ also $\mathcal{D}_\alpha(x)$ takes small values; when effective complexity crosses a threshold value (which depends on Kolmogorov complexity) logical depth suddenly jumps to extremely large values (Ay–Mueller–Szkola)
- so effective complexity $\mathcal{E}(x)$ considered a more stable notion of complexity

Integrated Information (an idea from neuroscience – Tononi)

1. G. Tononi G (2008) *Consciousness as integrated information: A provisional manifesto*, Biol. Bull. 215 (2008) N.3, 216–242.
2. M. Oizumi, N. Tsuchiya, S. Amari, *Unified framework for information integration based on information geometry*, PNAS, Vol. 113 (2016) N. 51, 14817–14822.

- want to measure amount of informational complexity in a system that is not separately reducible to its individual parts
- possibilities of causal relatedness among different parts of the system

Computing integrated information

- consider all possible ways of splitting a given system into subsystems
- the state of the system at a given time t is described by a set of observables X_t and the state at a near-future time by X_{t+1}
- partition λ into N subsystems $\Rightarrow$ splitting of these variables $X_t = \{X_{t,1}, \ldots, X_{t,N}\}$ and $X_{t+1} = \{X_{t+1,1}, \ldots, X_{t+1,N}\}$ into variables describing the subsystems
- all causal relations among the $X_{t,i}$ or among the $X_{t+1,i}$, also causal influence of the $X_{t,i}$ on the $X_{t+1,j}$ through time evolution captured (statistically) by joint probability distribution $\mathbb{P}(X_{t+1}, X_t)$
- compare information content of this joint distribution with distribution where only causal dependencies between X_{t+1} and X_t through evolution within separate subsystem not across subsystems

- set $\mathcal{M}_\lambda$ of probability distributions $\mathbb{Q}(X_{t+1}, X_t)$ with property that
$$\mathbb{Q}(X_{t+1,i}|X_t) = \mathbb{Q}(X_{t+1,i}|X_{t,i})$$
for each subset $i = 1, \dots, N$ of the partition λ
- minimize Kullback-Leibler divergence between actual system and its best approximation in $\mathcal{M}_\lambda$ over choice of partition λ
- integrated information
$$\Phi = \min_\lambda \min_{\mathbb{Q}\in\mathcal{M}_\lambda} \mathrm{KL}(\mathbb{P}(X_{t+1}, X_t)||\mathbb{Q}(X_{t+1}, X_t))$$
- value Φ represents additional information in the whole system not reducible to smaller parts

Question: a Complexity version of integrated information based on Gell-Mann effective complexity?