The Cost of Complexity

Gerard McCaul

We ask what general cost structure follows when a formal system must be physically realised. Any representation must be stored, manipulated, refreshed, and sometimes erased in matter. Starting from three sparse requirements - that realisation cost respect typicality order, that independent costs add, and that uncertainty be maximal at fixed mean cost - we derive a canonical logarithmic price for maintaining distinctions. That price selects Zipf-like equilibrium distributions over realised distinctions and implies a sharp critical ceiling: in the infinite-vocabulary limit the partition function becomes the Riemann zeta function, and the mean thermodynamic burden of maintaining distinctions diverges at the Zipf point. The same framework extends naturally to lossy representation, placing the information bottleneck in the same thermodynamic family. More broadly, the result suggests that the recurrent power-law organisation seen across language, biology, and machine learning may reflect a generic physical constraint on representation itself, rather than a peculiarity of any one substrate.

I. INTRODUCTION

Many of the most successful theories in science are organised around a balancing act. In mechanics, stationary motion balances kinetic and potential terms. In statistical mechanics, equilibrium balances energy and entropy. In learning theory, useful models balance compression and fidelity, bias and variance, expressivity and overfitting. The particular ontology changes across these examples. The underlying fact does not: persistence and effective description are obtained by holding competing demands in tension.

This suggests a more general question. What is the substrate-independent cost of a physically realised distinction? A representation is not free-floating. Its distinctions must be stored, maintained, manipulated, and sometimes erased in matter. Any general account of representation must therefore answer to thermodynamic cost as well as to descriptive adequacy.

The present paper makes a concrete claim. We show that three assumptions - order-respecting realisation cost, additive composition, and maximum entropy at fixed mean cost - force a canonical logarithmic cost

$$K(r) = \log r, \tag{1}$$

a Zipf equilibrium

$$p(r \mid \lambda) = \frac{r^{-\lambda}}{Z_R(\lambda)}, \tag{2}$$

and a critical cost ceiling at $\lambda = 1$ in the large-vocabulary limit. In that limit the partition function becomes the Riemann zeta function. When fidelity is relaxed, the same structure extends naturally to the information bottleneck.

Large-scale machine learning makes this question difficult to ignore. Resource-constrained intelligence is no longer hypothetical: memory, latency, bandwidth, and power are now explicit design variables in deployed systems. A general theory of representational cost would not replace domain-specific modelling, but it would explain why closely related trade-off structures recur across very different domains.

There is a useful historical analogy. Heat engines were built and optimised before thermodynamics existed in mature form, but thermodynamics clarified which constraints were contingent and which were universal. The present paper aims at an analogous clarification for representation under physical constraint.

Several existing literatures approach parts of this problem. Landauer's principle and the thermodynamics of information tie physically instantiated distinctions to dissipation [1, 2]. Efficient coding studies how representations align their coordinates with structure in the signals that matter [3]. The information bottleneck asks which compressed variables preserve what is relevant to a target [4]. The thermodynamics of prediction relates predictive organisation to dissipation [5]. The recurrence of constrained optimisation, dual variables, and exponential families across these settings is the central clue of the paper.

Our starting point is deliberately sparse. A representation is taken to be a target-relative coarse-graining of possible reality. Before one introduces a probability law, geometry, or dynamics on that coarse-graining, one already has a simpler question: which distinctions must be carried, which of them matter for a target Y, and what does it cost to realise them in matter? We argue that the weakest structure needed to ask this question is an order of typicality on coarse states. From that point onward, the downstream theory is unexpectedly rigid.

We use the term *synthetics* for the broader programme that treats representations themselves as physical objects. The present paper isolates the static grammar of that programme. It does not attempt to describe how representations move between operating points, adapt to new tasks, or change coordinate systems over time. It asks only what equilibrium cost structure is already fixed once a representation is embodied and used efficiently.

The paper proceeds from both ends of the same problem. Sections II and III develop the abstract argument from faithful coarse-graining, typicality order, and additive composition. Section IV re-derives the same logarithmic cost from a concrete operational picture of computation as guess, test, erase. Sections V and VI then

derive the Zipf ensemble and its fidelity-constrained bottleneck extension. The discussion returns to the scope of the static theory and to the dynamical questions it leaves open.

II. COARSE STATES, TYPICALITY, AND EFFICIENT EMBODIMENT

We begin from two structural premises. First, every representation - here meaning a physically embodied description of some target - is physical. Second, physical embodiment is not free. Even the most abstract formalism exists only in the registers that hold it: in marks, glyphs, memory states, or neural structures. In the limiting auto-representational case, the representation is simply the thing itself. We therefore use *theory* and *representation* interchangeably when no confusion arises, but prefer *representation* when we mean the embodied code rather than the target system.

Let Ω denote the space of possible configurations of the universe, and let $Y \subset \Omega$ be the target phenomenon we wish to describe. A representation is specified by three pieces of structure: an abstract alphabet of coarse symbols X, a physical register $R \subset \Omega$ that instantiates those symbols, and a coarse-graining map

$$\pi : \Omega \longrightarrow X \tag{3}$$

that assigns each configuration $\omega \in \Omega$ to a coarse symbol $x \in X$. The symbols X are abstract; the register R is physical. Multiple microstates of R may realise the same symbol, and the equivalence class

$$A_x = \pi^{-1}(x) \tag{4}$$

collects the configurations that the representation treats as indistinguishable.

The coarse-graining is *faithful* relative to Y when X determines Y: if $\pi(\omega) = \pi(\omega')$, then ω and ω' are indistinguishable with respect to Y. In classical statistics this is the requirement that π be a sufficient statistic for predicting Y from Ω [6]; in computational mechanics it is what singles out the causal states [7] as the minimal faithful partition.

Because the symbols are carried by a physical register R, a faithful representation incurs physical cost. The simplest way to see this is through Landauer's principle: erasing information requires at least $kT \ln 2$ of heat dissipation per erased bit [1, 2, 8]. Once symbols must be written, refreshed, and erased in a physical register, representational cost is unavoidable.

A faithful theory cannot be merely a list of names. Its symbols must stand in systematic correspondence with the observations they are meant to capture. That correspondence need not yet endow the symbols with a metric, a probability law, or a dynamics. But it must at least preserve an order: some coarse states are more typical, in the sense of being more prevalent or more readily encountered, than others. This preorder of typicality is the weakest structure sufficient to discuss efficient embodiment at all.

We therefore introduce a typicality preorder on X:

$$x \succcurlyeq x' \quad \Longleftrightarrow \quad x \text{ is at least as typical as } x'. \tag{5}$$

One may represent this order by a scalar $\tau : X \to \mathbb{R}_{\geq 0}$ satisfying $x \succcurlyeq x' \Leftrightarrow \tau(x) \geq \tau(x')$, unique up to monotone transformation. The primitive object is the order itself, not any particular numerical encoding. Typicality is not here a probability law in disguise; it is the minimal ordinal residue required before any measure or dynamics has been introduced.

Once realisation is costly and coarse states differ in typicality, an efficient register should not make the more typical distinctions systematically more expensive to instantiate. We therefore assign a realisation cost $K(x)$ to each symbol $x \in X$ and require

$$x \succcurlyeq x' \quad \Longrightarrow \quad K(x) \leq K(x'). \tag{6}$$

Equation (6) is intentionally weak. It does not assume an optimal code, only that realisation cost respect the ordinal structure already carried by the symbols [6, 9, 10].

The relation between the target Y and the register R now introduces one further issue: self-reference.

A. Self-reference and imperfect fidelity

Because both the target Y and the register R sit inside Ω, the representation can stand to its target in three basic ways (Fig. 1):

(i) **Disjoint.** The register lies outside the scope of Y. This is the regime of detached observation. Full fidelity is in principle achievable because the representation does not have to describe its own substrate.

(ii) **Coupled.** The register partially overlaps with the target. A thermometer immersed in the bath it measures, or a brain modelling the class of brains, belongs here. Fidelity is then limited by the coupling between register and target.

(iii) **Reflexive.** The register lies inside the scope of Y. Any representation of the whole universe, or of a system that includes its own measuring apparatus, falls into this regime.

The reflexive regime is the crucial one for what follows. Any representation that aspires to universality must compress a target that contains its own register. If π were the identity, the representation would be as complex as the world it describes. If π is non-trivial, the representation must contain a compressed image of itself, and that self-image is necessarily lossy. The analogy is Gödelian: a formal system rich enough to encode its own syntax admits true statements it cannot derive [11]. By the same

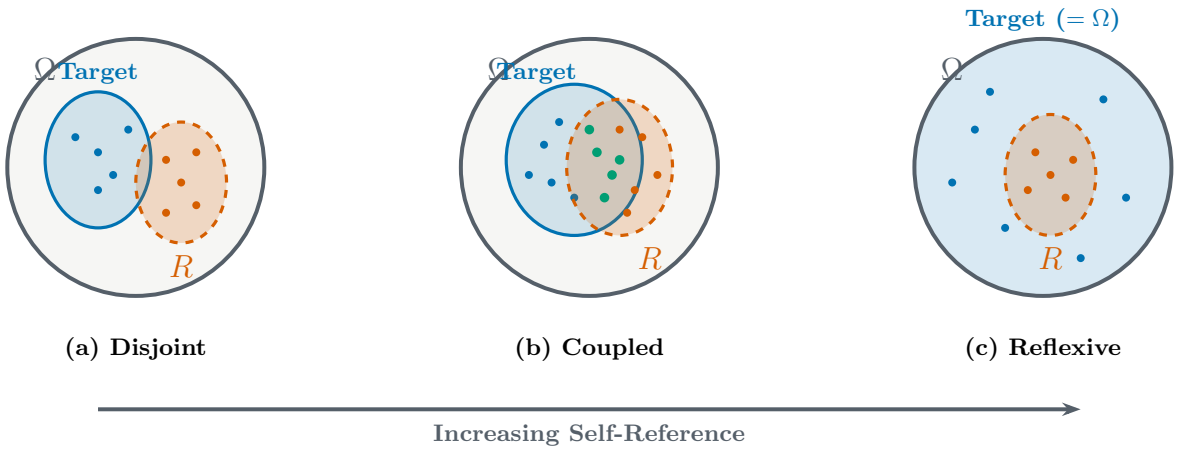


FIG. 1. **Three categories of representation, distinguished by the overlap of the physical register and the target domain within Ω. (a)** Disjoint: the register R (dashed, orange) and the target domain (solid, blue) share no configurations; full fidelity is in principle achievable. **(b)** Coupled: R and the target partially overlap (intersection in green); fidelity is bounded by the coupling constraints. **(c)** Reflexive: the register is strictly embedded entirely within the target domain (Target = Ω); any description must describe itself, thus bounded by incompleteness. Scattered nodes represent configurations within each region.

token, a physical register encoding a non-trivial compression of a universe containing it cannot faithfully distinguish every relevant configuration.

Imperfect fidelity is therefore not an optional complication. In the reflexive case it is structural. The question is not whether distinctions are lost, but which distinctions can be sacrificed at least cost. That is why Section VI will later relax perfect fidelity rather than treat it as the default.

Up to this point we have assumed only a faithful coarse-graining, a physical register carrying the resulting symbols, a preorder of typicality on those symbols, and an efficiency condition requiring realisation cost to respect that order. Since ordinal information is all that survives under admissible relabellings, the only invariant argument available to cost is rank. The next section shows that additive composition then forces the canonical logarithmic form.

III. THE COST OF COMPLEXITY

To this point, we have proposed that a representation assigns a realisation cost $K(x)$ to each of its symbols, and that an efficient representation must order those costs in accordance with the typicality of x. We now show that these requirements already fix the form of the cost function.

For $x \in X$, define its rank by

$$r(x) := \#\{x' \in X : x' \succcurlyeq x\}. \tag{7}$$

Equation (7) counts how many symbols are at least as typical as x. For the present we assume unique typicality values, so that the rank is unambiguous and coincides with ordinal position in the typicality order. This keeps the argument transparent; the general case changes only the bookkeeping.

Theorem 1 (Rank reduction)**.** *Any cost function that respects only ordinal typicality can depend on a coarse state only through its rank in that order. Equivalently, there exists a nondecreasing function $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$ such that*

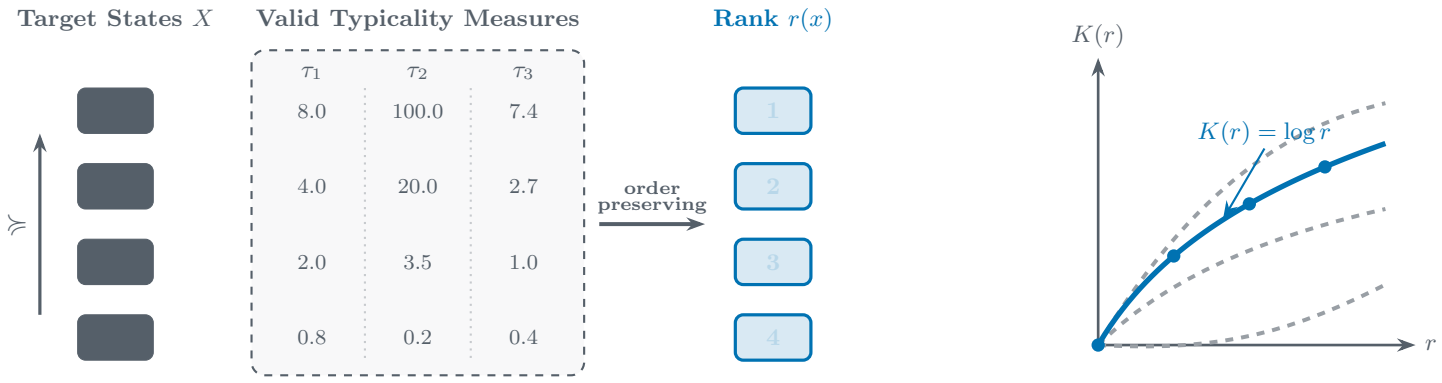
$$K(x) = f\big(r(x)\big) \tag{8}$$

for every symbol x.

Proof. Suppose $\varphi : X \to X'$ is an order-preserving relabelling of the typicality preorder. Such a relabelling preserves everything that is admissible in the present setup, because the preorder is the only structure assumed on X. If two coarse states have the same ordinal position in that preorder, then exchanging them by an admissible relabelling cannot change any legitimate cost assignment. Hence the cost of a state can depend only on its ordinal position, namely its rank $r(x)$. Monotonicity of K along the preorder then implies that the resulting function of rank is nondecreasing. □

There is one further property that any efficient representation must satisfy: *independence of composition*. The cost of jointly representing two independent subsystems should equal the sum of the individual costs. This is the standard additivity property familiar from entropy, Kolmogorov complexity, and thermodynamic free energy [1, 6, 9].

While the costs of independent systems are additive, the state spaces they enumerate are *multiplicative*. Consider two independently realised subsystems with coarse-state sets X and Y. Let $x \in X$ and $y \in Y$ have ranks



(a) Unique Ordinal Invariant **(b) Canonical Cost Schedule**

FIG. 2. **Rank as the ordinal invariant. (a)** Three arbitrary typicality scalars τ compatible with the same preorder $\succcurlyeq$ collapse under order-preserving relabelling to the unique monotonic invariant: rank $r(x)$. **(b)** Any nondecreasing function of rank yields an admissible representation cost (grey dashed curves); however, imposing independence of composition singles out $K(r) = \log r$ as the unique canonical schedule (blue).

$r(x)$ and $r(y)$. Under independent composition the composite outcome is the pair (x, y). Because typicality order is respected factor-wise, the number of composite states at least as typical as (x, y) equals the product of the individual counts:

$$r(x, y) = r(x)\, r(y). \tag{9}$$

Equation (9) makes the multiplicative structure explicit: each composite state inherits the product of the two marginal ranks, so equal-rank contours on the composition grid are hyperbolae of equal cost.

This is a counting statement, not a measure-theoretic one. Under independent factor-wise ordering, the composite states at least as typical as (x, y) are exactly the pairs formed from states at least as typical as x and states at least as typical as y, so the counts multiply.

Rank is therefore multiplicative under independent composition, whereas realisation cost is additive. Accordingly, the function f in Eq. (8) must satisfy

$$f(rs) = f(r) + f(s) \tag{10}$$

for positive integers r and s. This discrete Cauchy equation fixes the cost to be logarithmic in rank.

Theorem 2 (Logarithmic cost)**.** *Assume $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$ is nondecreasing and satisfies Eq.* (10)*. Then there exists a constant $\kappa \geq 0$ such that*

$$f(r) = \kappa \log r. \tag{11}$$

Proof. By induction on Eq. (10), $f(r^m) = m f(r)$ for every positive integer m. Writing $r = \prod_{i=1}^{k} p_i^{a_i}$ in its prime factorisation, repeated application of Eq. (10) gives

$$f(r) = \sum_{i=1}^{k} a_i\, f(p_i). \tag{12}$$

Thus f is completely determined by its values on primes.

Now fix primes p and q, and choose positive integers m, n such that $p^m \leq q^n < p^{m+1}$. Monotonicity gives $m f(p) \leq n f(q) < (m+1) f(p)$. Dividing by $n f(p)$ and using the logarithmic form of the numerical bracket:

$$\frac{m}{n} \leq \frac{f(q)}{f(p)} < \frac{m+1}{n}, \qquad \frac{m}{n} \leq \frac{\log q}{\log p} < \frac{m+1}{n}. \tag{13}$$

Both $f(q)/f(p)$ and $\log q / \log p$ are trapped in the same interval of width $1/n$. Since n can be made arbitrarily large, the two ratios must agree:

$$\frac{f(p)}{\log p} = \frac{f(q)}{\log q} =: \kappa \tag{14}$$

for every pair of primes p, q.

Substituting $f(p_i) = \kappa \log p_i$ into Eq. (12) yields $f(r) = \kappa \sum_i a_i \log p_i = \kappa \log \prod_i p_i^{a_i} = \kappa \log r$. □

Finally, the constant κ in Eq. (11) fixes only the unit of cost. We may therefore absorb this into the units of later derived parameters, and present a *canonical* cost function

$$K(r) = \log r. \tag{15}$$

The functional form of K already appears in many figures of merit across physics and information theory. The same algebraic fact underlies the entropy-code-length correspondence [6, 9] and the Huffman optimality result [10]. In Kolmogorov complexity, the nth shortest program has complexity $\lfloor \log n \rfloor + O(\log \log n)$ [12]. It is also identical to the per-state *rank surprisal* familiar from free-energy arguments in neuroscience and biology. The important point here is not the reappearance of the logarithm, but the fact that it is derived from first principles using only compositionality and efficiency.

IV. THE GUESSING MACHINE

We now re-derive the same logarithmic cost from a concrete operational picture. The point is not to enlarge the ontology of the theory, but to show that the same cost law is forced by the physics of search and erasure. The previous section reached $K(r) = \log r$ from ordinal typicality and additive composition. Here we reach it from the guess–test–erase cycle of physical computation.

A. Computation as Diophantine search

The bridge between abstract representation and concrete mechanism is provided by the Davis–Putnam–Robinson–Matiyasevich characterisation. A set is *recursively enumerable* if membership can be confirmed by an algorithm, even if non-membership need not be decided in finite time [13]. DPRM states that every recursively enumerable set $R \subseteq \mathbb{N}^m$ can be characterised by the existence of integer solutions to a polynomial equation:

$$u \in R \quad \Longleftrightarrow \quad \exists\, x \in \mathbb{N}^n : D(u, x) = 0, \tag{16}$$

where $D : \mathbb{N}^{m+n} \to \mathbb{Z}$ is a polynomial with integer coefficients [14, 15]. Here u is an m-tuple of *parameters* - the input to the computation - and x is an n-tuple of *witnesses*. The terminology is borrowed from proof theory: a witness is a candidate solution whose existence certifies a positive answer, much as an explicit counterexample certifies a conjecture false. The witness is hard to find but easy to check - one need only evaluate $D(u, x)$ and test whether the result is zero.

Any computation whose positive outcomes can be verified in finite time can therefore be cast as the search for integer witnesses to a polynomial equation. The specific polynomial D encodes the architecture - logic gates, biochemical pathways, neural circuitry. The search form itself is universal: guess a witness, test the constraint, discard failures, and continue.

B. Guess, test, erase

This universal search structure motivates the following abstraction. A *Non-Deterministic Diophantine Machine* (NDDM) is specified by a parameter space $U \subseteq \mathbb{N}^m$, a witness space $X \subseteq \mathbb{N}^n$, an integer-coefficient polynomial $D : U \times X \to \mathbb{Z}$, and a family of guess distributions $\{p_u\}_{u \in U}$ over X. Given input u, the machine performs a single computational step:

1. **Guess.** Sample a candidate witness $x \sim p_u$.
2. **Test.** Evaluate the constraint $D(u, x) = 0$.
3. **Erase.** If $D(u, x) \neq 0$, reset the witness register to a fixed reference state and return to step 1. If $D(u, x) = 0$, halt.

The NDDM is not a special architecture but a universal coarse-grained description of the search implicit in any code-bearing computation (Fig. 3). This universality is guaranteed by DPRM: any system capable of general-purpose computation - digital hardware, a genome under selection, a neural circuit - admits a Diophantine representation in which the computational steps reduce to the guess–test–erase cycle above [14].

The connection to the framework of Section II is now immediate. The witness register is the physical register $R \subset \Omega$; the integer witnesses are the coarse states X; and the polynomial D encodes the representation's correspondence with its target Y. The Diophantine representation is a concrete instantiation of the symbol map $\pi : \Omega \to X$ from Eq. (3) - one guaranteed to exist for any system capable of general computation. What Section II called a faithful representation, the NDDM realises as a polynomial whose zero set captures exactly the admissible witness–parameter pairs.

C. The cost of a wrong guess

Each time the machine rejects a witness, the register must be erased. By Landauer's principle, erasing a register whose state x_k indexes x_k effectively distinguishable microstates dissipates at least $k_B T \ln x_k$ of heat [1, 8]. For a witness tuple $x = (x_1, \ldots, x_n)$, the total minimal erasure cost is

$$Q_{\min}(x) \;\geq\; k_B T \sum_{k=1}^{n} \ln x_k. \tag{17}$$

The sum $\sum_k \ln x_k$ is the *log-volume* of the witness: the minimal information, in nats, that must be irreversibly discarded when the machine erases a failed guess. Without loss of generality we may encode witness tuples as a single integer via a computable *Gödelisation* $g : \mathbb{N}^n \to \mathbb{N}$ - any fixed, invertible mapping from tuples to positive integers (e.g. the Cantor pairing function [12]) - reducing the log-volume to $\ln g(x)$. Since g is fixed and computable, the distinction is immaterial for the asymptotic behaviour that concerns us, and we write simply $\ln x$ for a single encoded witness.

With witnesses encoded as positive integers and ordered by size, the integer label *is* the rank: smaller integers are more readily encountered as witnesses (more typical, in the language of Section II), and the ordinal position of witness x in the size ordering is precisely the rank $r(x)$ defined in Eq. (7). Once witnesses are Gödelised and ordered by size, the witness label is not merely analogous to rank; it *is* rank.

This identification makes the convergence of the two derivations explicit. In Section III, compositionality and efficiency implied $K(r) = \log r$ through the functional equation $f(rs) = f(r) + f(s)$. Here, the physics of erasure gives $Q_{\min} \propto \ln x$. The functional form is the same because the underlying constraint is the same: the cost

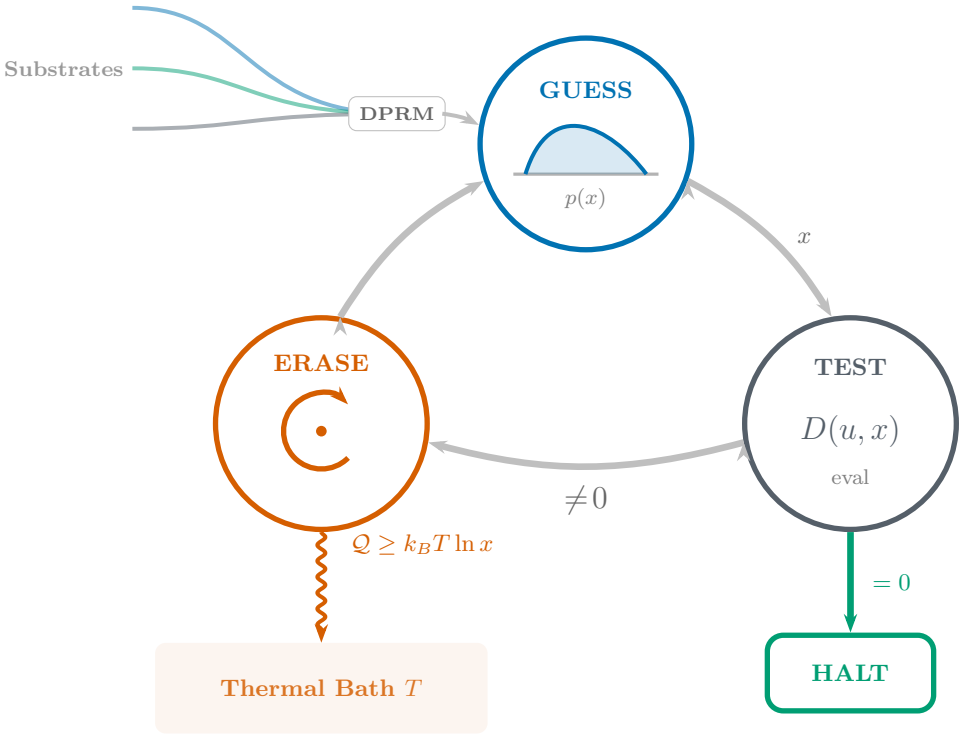


FIG. 3. **The guessing machine.** Any code-bearing system admits, via the DPRM characterisation, a universal coarse-grained description as a Non-Deterministic Diophantine Machine (NDDM). The machine samples integer witnesses x from a guess distribution p, tests the polynomial constraint $D(u, x) = 0$, and erases the witness register on failure. Each erasure dissipates at least $k_B T \ln x$ of heat into the thermal bath (Landauer's principle). The mean dissipation per step is bounded below by the rank surprisal $\ell = \mathbb{E}_p[\ln X]$. The specific polynomial D encodes the architecture; the guess–test–erase cycle is universal.

of independent subsystems must add while their state spaces multiply, and the logarithm is the unique monotone function satisfying this [6]. Section III deduced this requirement from structural assumptions; here it is read off directly from the thermodynamics of the register.

D. Rank surprisal

Let p denote the stationary distribution of witnesses encountered at the moment of erasure. The *rank surprisal* of the computation is

$$\ell(p) \;:=\; \mathbb{E}_p[\ln X] \;=\; \sum_{x \in X} p(x)\, \ln x, \tag{18}$$

the mean log-volume of erased witnesses. This quantity is universal in two senses. First, it depends only on the coarse-grained witness statistics p, not on the mechanics of how D is evaluated or the substrate in which it is realised. Second, by DPRM, any code-bearing system admits such a representation, so ℓ is defined for any system capable of general computation.

Combining Eq. (17) with the definition of ℓ gives a direct thermodynamic bound: the minimal average heat dissipated per computational step satisfies

$$\frac{Q_{\min}(\text{step})}{k_B T} \;\geq\; \ell(p). \tag{19}$$

The rank surprisal is a lower bound on the thermodynamic cost of computation, measured in natural units. It holds for any witness distribution with finite ℓ, independently of the architecture.

Two independent derivations have now identified the same structure: a logarithmic cost function over a countable set of representational states, with the rank surprisal ℓ quantifying the irreducible thermodynamic price of maintaining the representation. The remaining question is which distribution of witnesses is maximally non-committal at a given mean cost - and what structure that distribution must take.

V. THE UNIVERSAL COST CURVE

However a representation is realised - whether in silicon, in nucleotides, or in synaptic weights - its use of representational states defines a probability distribution. A state visited often receives high probability; a state visited rarely receives low probability. Write $p(r)$ for the probability that the representation occupies the coarse

state of rank r. The preceding sections have fixed the cost of each state to be $K(r) = \log r$ and identified the mean cost

$$\ell \;=\; \sum_r p(r)\,\log r \tag{20}$$

with the rank surprisal - the irreducible thermodynamic price of maintaining the representation. The question is now: if the mean cost ℓ is all we know about p, what form must the distribution take?

A. Maximum entropy and the logic of honest inference

The answer is supplied by Jaynes' principle of maximum entropy [16]. Among all distributions consistent with a set of constraints, the one with the greatest Shannon entropy

$$S[p] \;=\; -\sum_{r=1}^{R} p(r)\,\log p(r) \tag{21}$$

is the unique distribution that introduces no information beyond the constraints themselves. Any other choice would implicitly assume structural knowledge we do not possess. Shore and Johnson showed that maximum entropy is the only inference rule consistent with the requirements of subset independence, coordinate invariance, and system independence [17] - the informational analogues of the physical principles (compositionality, efficiency) that led us to $K(r) = \log r$ in the first place.

In our setting the constraints are normalisation and a fixed mean cost:

$$\sum_{r=1}^{R} p(r) = 1, \qquad \sum_{r=1}^{R} p(r)\,\log r = \ell. \tag{22}$$

Here R is a finite cutoff on the number of coarse states (the vocabulary size of the representation). Introducing Lagrange multipliers α for normalisation and λ conjugate to mean cost, the variational problem

$$\frac{\partial}{\partial p(r)}\Big[S[p] - \alpha\big(\textstyle\sum p(r) - 1\big) - \lambda\big(\textstyle\sum p(r)\log r - \ell\big)\Big] = 0 \tag{23}$$

yields the stationary condition $-\log p(r) - 1 - \alpha - \lambda \log r = 0$, so that

$$p(r \mid \lambda) \;=\; \frac{r^{-\lambda}}{Z_R(\lambda)}, \tag{24}$$

with partition function

$$Z_R(\lambda) \;=\; \sum_{r=1}^{R} r^{-\lambda}. \tag{25}$$

Proposition 1 (Zipf equilibrium)**.** *Let $K(r) = \log r$ be the canonical cost on ranks $\{1, \dots, R\}$. The maximum-entropy distribution at fixed mean cost $\ell = \langle K \rangle$ is the finite-cutoff Zipf family $p(r \mid \lambda) = r^{-\lambda}/Z_R(\lambda)$. The multiplier λ is uniquely determined by ℓ through the equation of state $\ell(\lambda) = -\partial_\lambda \log Z_R(\lambda)$.*

Under the stated constraints, this is the unique maximum-entropy ensemble on the rank line. Related power-law families appear in Mandelbrot's informational equilibria [18], but here the Zipf distribution follows directly from the logarithmic cost and the requirement of honest inference.

B. Thermodynamic structure

The partition function $Z_R(\lambda)$ generates the full thermodynamics of the representation. Differentiating $\log Z_R(\lambda)$ gives the mean cost

$$C(\lambda) \;=\; -\frac{d}{d\lambda}\log Z_R(\lambda), \tag{26}$$

while substituting Eq. (24) into Eq. (21) yields the entropy

$$S(\lambda) \;=\; \lambda\, C(\lambda) + \log Z_R(\lambda). \tag{27}$$

Writing λ^{-1} for the *complexity temperature*, the *free complexity* follows by Legendre transform:

$$F(\lambda) \;:=\; C(\lambda) - \lambda^{-1} S(\lambda) \;=\; -\lambda^{-1}\log Z_R(\lambda). \tag{28}$$

The multiplier λ prices canonical cost: large λ concentrates mass on low-rank (cheap) states; small λ spreads mass across the vocabulary. At $\lambda = 1$ the distribution is standard Zipf's law, $p(r) \propto 1/r$, and $Z_R(1) = H_R \sim \ln R$ diverges as $R \to \infty$. For $\lambda > 1$ the partition function converges as a Dirichlet series; for $\lambda \le 1$ it diverges. In the infinite-cutoff limit,

$$Z_\infty(\lambda) \;:=\; \lim_{R\to\infty} Z_R(\lambda) \;=\; \sum_{r=1}^{\infty} r^{-\lambda} \;=\; \zeta(\lambda), \qquad \lambda > 1. \tag{29}$$

The universal cost curve therefore reduces, in the infinite-vocabulary limit, to the Riemann zeta function. The critical point $\lambda = 1$ separates regimes in which an unbounded vocabulary can and cannot be maintained at finite mean cost.

Differentiating once more gives the susceptibility

$$\chi(\lambda) \;:=\; -\frac{dC}{d\lambda} \;=\; \mathrm{Var}_\lambda(\log r), \tag{30}$$

the variance of cost across the rank line - the analogue of a heat capacity (Fig. 4). At the critical point $\chi \to \infty$: the system becomes maximally sensitive to changes in the cost budget.

The Legendre structure - $C(\lambda)$, $S(\lambda)$, $F(\lambda)$ - is inherited from standard statistical mechanics once $K(r) =$

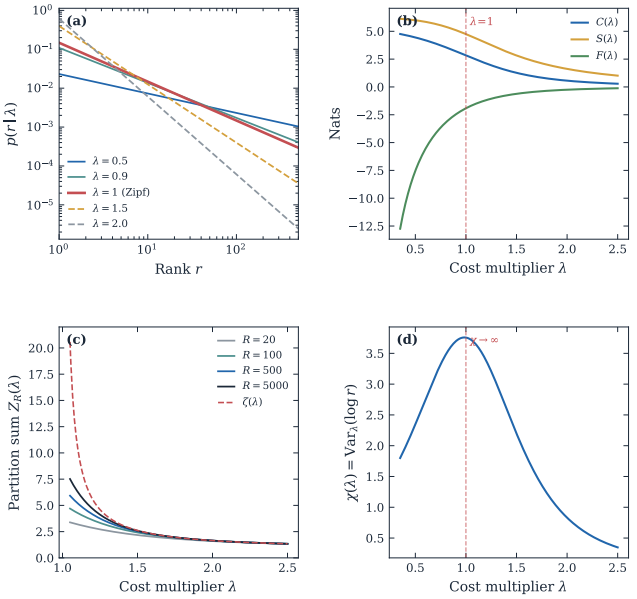


FIG. 4. **Finite-cutoff Zipf family and critical structure.** **(a)** Representative rank distributions $p(r \mid \lambda)$ on log–log axes for several values of the cost multiplier λ; the $\lambda = 1$ curve (red) is standard Zipf's law. **(b)** Response functions: mean cost $C(\lambda)$, entropy $S(\lambda)$, and free complexity $F(\lambda)$ (all in nats, $R = 500$). **(c)** Finite-cutoff partition sums $Z_R(\lambda)$ for increasing vocabulary sizes R, approaching the infinite-cutoff limit $Z_\infty(\lambda) = \zeta(\lambda)$ for $\lambda > 1$. **(d)** Susceptibility $\chi(\lambda) = \mathrm{Var}_\lambda(\log r)$, diverging at the critical point.

$\log r$ is fixed. The singular behaviour at $\lambda = 1$ is the central quantitative consequence of the theory: systems operating near it can spread probability across many ranks, but only at the price of diverging susceptibility. The rank surprisal ℓ diverges as $\lambda \to 1^+$, which by the Landauer bound (Eq. 19) implies unbounded heat dissipation per computational step. Richer codes therefore require greater thermodynamic expenditure to maintain.

VI. IMPERFECT FIDELITY AND THE INFORMATION BOTTLENECK

The distribution derived in the previous section answers a specific question: what does the representation look like when the only constraint is the mean cost it can afford? Implicitly, this assumes *perfect fidelity* - every coarse state captures exactly those features of the target it is supposed to represent. Section II established that no non-trivial self-representation can achieve this at finite cost: any representation with finite resources is necessarily lossy. We now ask what changes when we take that imperfection into account.

This relaxation is not optional in the reflexive regime introduced in Section II. Any non-trivial self-embedded representation must be lossy at finite cost. The relevant question is therefore not whether fidelity is imperfect, but how fidelity trades against representational cost.

A. Measuring fidelity

To proceed, we need a measure of what the representation preserves. Let X denote the coarse states (equivalently, the ranked witnesses of Section IV) and Y the target observations. The question is: how much does knowing X tell us about Y?

The axiomatic foundations of information theory constrain the answer severely. Any measure of statistical dependence between X and Y that satisfies non-negativity, vanishes if and only if X and Y are independent, is invariant under invertible transformations, and decomposes additively for independent pairs, must take the form of a mutual information [6, 19]:

$$I(X;Y) \;=\; \sum_{x,y} p(x,y)\,\log \frac{p(x,y)}{p(x)\,p(y)}. \tag{31}$$

Equivalently, $I(X;Y) = D_{\mathrm{KL}}\big(p(x,y)\,\|\,p(x)p(y)\big)$: the Kullback–Leibler divergence from statistical independence. Under these axioms, mutual information is the canonical choice - the same kind of uniqueness result that singled out the logarithm in Section III. Mutual information is to fidelity what $K(r) = \log r$ is to cost: the unique functional form compatible with the structural axioms.

The faithfulness loss introduced in Section II - how much Ω is not captured by the coarse states - can likewise be written as a conditional mutual information $I(Y;\Omega \mid X)$, itself a KL divergence. Maximum fidelity corresponds to $I(Y;\Omega \mid X) = 0$; all of Y is encoded in X. The finite-cost constraint of the previous section is the regime where this is assumed to hold.

B. Two constraints, one Lagrangian

We now have two independently motivated quantities: the mean representational cost $\ell = \sum p(r) \log r$, derived from the physics of erasure, and the fidelity $I(X;Y)$, derived from the axioms of information theory. Both are constraints on the representation. The natural question is: what is the maximum-entropy representation subject to *both*?

One useful way to relate this section to the previous one is to separate the universal rank law from the task-dependent weighting. Schematically,

$$p(r) \;\propto\; g(r)\, r^{-\lambda}, \tag{32}$$

where $r^{-\lambda}$ is the universal baseline set by the log-rank cost and $g(r)$ collects whatever additional weighting is needed to preserve fidelity. The pure cost curve of Section V is the special case $g(r) = 1$. Fidelity therefore enters not by replacing the canonical rank law, but by supplying the extra structure beyond bare expected cost that entropy and free complexity then register.

The variational problem is

$$\max_{p(X|\Omega)} S[p] \quad \text{subject to} \quad \langle \log r \rangle = \ell, \quad I(X;Y) \geq I_0, \tag{33}$$

where $S[p]$ is the entropy of the representation and I_0 is the minimum acceptable fidelity. The standard Lagrangian relaxation introduces two multipliers: λ conjugate to mean cost and β conjugate to fidelity:

$$\mathcal{L} \;=\; S[p] \;-\; \lambda \left(\langle \log r \rangle - \ell \right) \;-\; \beta \left(I_0 - I(X;Y) \right). \tag{34}$$

In the deterministic-encoder setting, $I(\Omega;X) = H(X)$, so the entropy term together with the fixed rank-cost constraint can be absorbed into a single penalty on representation complexity. One then recovers the standard information-bottleneck functional

$$\mathcal{L}_{\mathrm{IB}} \;=\; I(\Omega;X) \;-\; \beta\, I(X;Y), \tag{35}$$

introduced by Tishby, Pereira, and Bialek [4]. This clarifies what is universal and what is task-specific: λ fixes the baseline thermodynamic price of rank, whereas β prices fidelity relative to that baseline.

Each term in Eq. (34) has a direct physical interpretation:

- $S[p]$: the entropy of the representation - the uncertainty in which coarse state the representation occupies;
- $\langle \log r \rangle = \ell$: the rank surprisal - the thermodynamic cost of erasure per computational step;
- $I(X;Y)$: the faithfulness of the representation - how much of the target is captured;
- λ: the canonical cost multiplier; its reciprocal λ^{-1} is the complexity temperature;
- β: the fidelity premium, pricing faithfulness against compression.

C. The bottleneck frontier

Varying β at fixed λ traces out a one-parameter family of representations: the *bottleneck frontier*. At $\beta = 0$ the fidelity constraint is switched off, and we recover the $g(r) = 1$ baseline of Section V - the pure cost curve at maximum entropy. As β increases, the representation must preserve more information about Y, at the expense of either higher cost or lower entropy.

The critical point $\lambda = 1$ persists. It marks the thermodynamic ceiling on representational cost: whatever the fidelity target, the mean rank surprisal ℓ cannot remain finite once the partition sum crosses into the divergent regime. The bottleneck adds a second dimension (fidelity), but the cost ceiling is set by the same zeta divergence as before (Fig. 5).

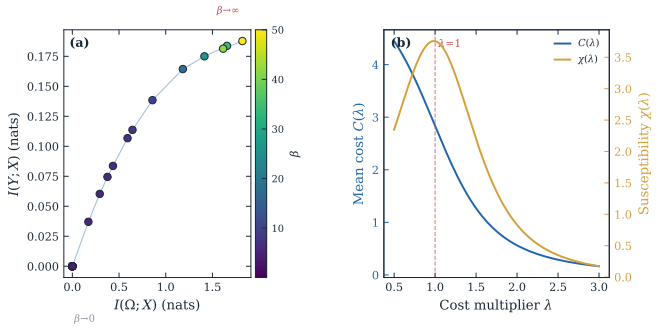


FIG. 5. **The bottleneck frontier and thermodynamic response.** Left: the information-bottleneck trade-off for a structured discrete source, coloured by the fidelity premium β. The frontier runs from the trivial representation at small β to the highest-fidelity codes at large β. Right: mean cost $C(\lambda)$ and susceptibility $\chi(\lambda)$ as functions of the cost multiplier λ for the rank-line ensemble ($R = 500$). The susceptibility diverges at $\lambda = 1$ (vertical dashed line) in the large-R limit, marking the thermodynamic ceiling on representational complexity.

The equation of state

$$C(\lambda) = -\frac{d}{d\lambda} \log Z_R(\lambda) \tag{36}$$

parameterises the frontier in thermodynamic variables. The susceptibility $\chi(\lambda) = \mathrm{Var}_\lambda(\log r)$ diverges as $\lambda \to 1^+$ in the large-R limit: the system becomes infinitely sensitive to changes in the cost budget at the critical point.

D. The bottleneck from physical principles

From this perspective, the standard bottleneck Lagrangian emerges as the natural two-constraint extension once representational cost is fixed by embodiment and fidelity is measured by mutual information. In the standard treatment, the bottleneck is postulated as a compression objective and the trade-off parameter β is introduced by hand [4]. Here the same structure appears because:

1. The cost function $K(r) = \log r$ follows from compositionality and efficiency (Section III) or equivalently from erasure physics (Section IV).
2. The fidelity measure $I(X;Y)$ is singled out by the axioms of information theory, given that we seek a measure of statistical dependence.
3. The MaxEnt procedure itself is the unique inference rule consistent with subset independence, coordinate invariance, and compositionality [17].

The bottleneck is therefore not a separate formalism but the natural extension of the cost-curve analysis to imperfect fidelity. We have not proved a literal closed-form

identification between β and λ^{-1}; rather, λ fixes the universal thermodynamic price of rank while β prices fidelity relative to that same baseline. The weighting form of Eq. (32) is the local statement of that relation.

Two caveats. First, the shape of the bottleneck frontier depends on the joint distribution $p(\Omega, Y)$, which is domain-specific. What is universal is the parameterisation and, within this rank-cost construction, the critical point at $\lambda = 1$. Domain structure determines where on the phase diagram a system operates; it does not move the phase boundary. Second, for deterministic encoders $I(\Omega; X) = H(X)$ is a monotone function of C at fixed λ; the general stochastic case, in which the encoder itself introduces noise, remains an open question.

VII. DISCUSSION

A. Summary of results

Starting from two premises - that every representation is physically embodied, and that no embodiment is free - we derived a chain of results with no additional domain-specific assumptions.

1. **Log-rank cost.** If a representation's cost respects typicality order and composes additively for independent subsystems, then the canonical cost law is $K(r) = \log r$ (Sections III and IV).

2. **Zipf equilibrium.** Maximum entropy at fixed mean cost yields the finite-cutoff Zipf family $p(r \mid \lambda) = r^{-\lambda}/Z_R(\lambda)$ (Section V). In the infinite-cutoff limit the partition function becomes $Z_\infty(\lambda) = \zeta(\lambda)$, with a critical point at $\lambda = 1$.

3. **Bottleneck extension.** When perfect fidelity is relaxed, mutual information is the canonical fidelity measure, and the standard bottleneck Lagrangian emerges as the natural two-constraint extension of the same framework (Section VI).

4. **Landauer–zeta bound.** The rank surprisal $\ell = \mathbb{E}[\ln X]$ lower-bounds the heat dissipated per computational step. Because ℓ diverges as $\lambda \to 1^+$, Zipf-like operation near the critical point requires unbounded thermodynamic resources in the infinite-vocabulary limit.

Under the stated assumptions, each link follows from a separate structural principle: the cost function from compositionality, the distribution from maximum entropy, the fidelity measure from the axioms of information theory, and the thermodynamic lower bound from Landauer's principle. What remains domain-specific is the typicality order, the cutoff R, and the joint distribution $p(\Omega, Y)$ that shapes the bottleneck frontier.

B. Scope: a static theory of representations

The scope of the paper should be stated plainly. Everything derived here concerns the *stationary* distribution of a representation. The typicality order is taken as given, the cost budget is fixed, and $p(r \mid \lambda)$ is an equilibrium distribution rather than a trajectory. There is no learning dynamics here, no adaptation rule, and no mechanism by which a representation moves across the phase diagram.

This is a deliberate restriction. The present paper fixes the static grammar of embodied representation: the form of the cost law, the associated equilibrium ensemble, the critical ceiling, and the fidelity-constrained extension. Any dynamical theory must be built on top of these constraints rather than in place of them.

C. Power laws and the critical regime

Zipf's law appears in word frequencies [20], neural activity [21], genomic sequences [22], urban populations [23], and many other empirical domains. The present derivation does not compete with the mechanism-specific accounts proposed in each case. It identifies the equilibrium family that such mechanisms can populate once representational cost is logarithmic in rank.

Within this framework, $\lambda = 1$ is the critical point at which the susceptibility $\chi(\lambda) = \mathrm{Var}_\lambda(\log r)$ diverges. That makes the vicinity of Zipf's law a natural regime in which to look for systems balancing rich effective vocabularies against finite thermodynamic budgets. Our illustrative fits suggest this possibility, but a systematic empirical test lies beyond the present paper.

D. Outlook: from statics to dynamics

The static theory points directly to three dynamical questions:

- how representations move on the (λ, β) phase diagram under learning, control, or selection;
- whether the critical point $\lambda = 1$ acts as an attractor, and if so, with what rates and universality classes;
- how the Landauer–zeta bound constrains not only equilibrium structure but also the speed at which representations can be reorganised.

These questions connect the present framework to work on thermodynamic computation [5, 24], information-theoretic learning dynamics [25], and the statistical mechanics of evolving codes. The present paper fixes the static grammar. Any dynamical theory of representation must now explain how real systems move on this phase diagram without violating the same cost ceiling.

[1] R. Landauer, IBM Journal of Research and Development **5**, 183 (1961).

[2] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa, Nature Physics **11**, 131 (2015).

[3] H. B. Barlow, in *Sensory Communication* (MIT Press, 2012).

[4] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, arXiv:physics/0004057 (2000).

[5] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks, Physical Review Letters **109**, 120604 (2012).

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley-Interscience, Hoboken, 2006).

[7] C. R. Shalizi and J. P. Crutchfield, Journal of Statistical Physics **104**, 817 (2001).

[8] C. H. Bennett, IBM Journal of Research and Development **17**, 525 (1973).

[9] C. E. Shannon, Bell System Technical Journal **27**, 379 (1948).

[10] D. Huffman, Proceedings of the IRE **40**, 1098 (1952).

[11] K. Gödel, Monatshefte für Mathematik und Physik **38**, 173 (1931).

[12] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 3rd ed. (Springer, New York, 2008).

[13] M. Sipser, *Introduction to the Theory of Computation*, 3rd ed. (Cengage Learning, Boston, 2013).

[14] Y. V. Matiyasevich, *Hilbert's Tenth Problem* (MIT Press, Cambridge, MA, 1993).

[15] M. Davis, H. Putnam, and J. Robinson, Annals of Mathematics **74**, 425 (1961).

[16] E. T. Jaynes, Physical Review **106**, 620 (1957).

[17] J. E. Shore and R. W. Johnson, IEEE Transactions on Information Theory **26**, 26 (1980).

[18] B. Mandelbrot, in *Communication Theory* (Academic Press, 1953) pp. 486–502.

[19] I. Csiszár, The Annals of Probability **3**, 146 (1975).

[20] G. K. Zipf, *The Psycho-Biology of Language* (Houghton Mifflin, Boston, 1935).

[21] T. Mora and W. Bialek, Journal of Statistical Physics **144**, 268 (2011).

[22] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Physical Review Letters **73**, 3169 (1994).

[23] X. Gabaix, Quarterly Journal of Economics **114**, 739 (1999).

[24] D. H. Wolpert, Journal of Physics A: Mathematical and Theoretical **52**, 193001 (2019).

[25] R. Shwartz-Ziv and N. Tishby, Opening the black box of deep neural networks via information, arXiv:1703.00810 (2017).