LEARNING THE SEMANTICS OF LOGICAL CONTENT

Anonymous authors
Paper under double-blind review

ABSTRACT

A learner that induces logic programs from a template L of N candidate Horn clauses appears to face 2^N hypotheses, but most rule-sets deduce the same facts and are indistinguishable, leading to the notion of semantic information. The true hypothesis class of a recursive template is therefore not the syntax 2^L but the set of realizable deductive closures Ω_L, the distinct "meanings" the template can express. Its log-count $C(L) = \log_2 |\Omega_L|$, the *semantic capacity* of the template, is its learning complexity: when facts arrive online and the learner predicts whether the theory deduces each, Halving makes at most $C(L)$ mistakes in the realizable case, the minimax log-loss (Shtarkov) regret is exactly $C(L)$ once the queries separate the contents, and the agnostic minimax $0/1$ regret is $\Theta(\sqrt{C(L)\,T})$ for both template families we study. We compute $C(L)$ in closed form: $n \log_2 \rho_d$ bits for the bandwidth-d reachability template of n clauses with ρ_d the d-bonacci constant solving $\rho^d = \rho^{d-1} + \cdots + \rho + 1$, and $\frac{1}{4}m^2(1 + o(1))$ bits for nonlinear transitive closure on m constants. In a probabilistic reading a *single* content's size undergoes a sharp phase transition at one base fact per constant. One computable number thus governs counting, coding, and learning, realizing the semantic-information program of Carnap and Bar-Hillel (1952) and Lastras et al. (2026) as a learning rate. Empirically, on Padgett's Florentine families network the 2^{40} clause-sets collapse onto $|\Omega_L| = 4{,}104$ contents and the online code length lands exactly on $C(L) = \log_2 4{,}104 \approx 12$ bits; the same holds on WordNet.

1 INTRODUCTION

Inductive logic programming and, more recently, neuro-symbolic learning ask a learner to choose a set of rules (a hypothesis) from a fixed *template* (language bias) of candidate clauses. It is tempting to measure the difficulty of this choice by the number of syntactic hypotheses, 2^N for a template of N clauses. We show that this count is wrong, and wrong in a way that matters for semantic information, sample complexity, and code length. The reason is denotational. For a definite (Horn) program the model-theoretic, fixpoint, and proof-theoretic semantics coincide (van Emden and Kowalski, 1976; Lloyd, 1987). A program's meaning is its *least Herbrand model* and the set of ground atoms it lets one deduce. Two clause-sets with the same least Herbrand model make identical predictions on every atom, hence they are semantically indistinguishable which is what motivates the notion of semantic information. A learner scored on its deductions can never separate them.

Fix a set of background clauses BK (facts and/or rules) and a finite *template* $L = \{c_1, \ldots, c_N\}$ of definite (Horn) clauses, the rules the language bias makes available. A *hypothesis* is a subset $S \subseteq L$; adding it to BK yields a definite logic program. Such a program has one well-defined meaning, its least Herbrand model $\mathrm{lhm}(\mathrm{BK} \cup S)$: exactly the ground atoms one can *deduce* from it. The object a logic-based learner actually searches is therefore not the syntax 2^L but the set of *realizable deductive closures*

$$\mu : 2^L \to 2^{\mathcal{B}}, \qquad R := \mu(S) = \mathrm{lhm}(\mathrm{BK} \cup S), \qquad \Omega_L := \mathrm{im}\,\mu = \{\mathrm{lhm}(\mathrm{BK} \cup S) : S \subseteq L\} \subseteq 2^{\mathcal{B}},$$

where $\mathcal{B}$ is the (finite) Herbrand base. The map μ is many-to-one; Ω_L is the quotient of syntax by denotational equivalence. An element of Ω_L is a *realizable closure*: a set of ground atoms that some hypothesis actually deduces. Note the two kinds of object: a hypothesis $S \in 2^L$ is a set of *clauses*, its image $R = \mu(S) \in 2^{\mathcal{B}}$ a set of *atoms*; Ω_L collects atom-sets, never rule-sets. We call R the semantic content of S, and $\log_2 |\Omega_L|$ the *semantic information* of the template (the log-number of *distinguishable* contents).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We focus on answering two very different questions, namely: how many contents can the template express (Ω_L capacity), and how much does a random instance deduce ($\mathbb{E}|R|$ extent). The first is a fixed count; the second is a random quantity with a sharp threshold. A template can have enormous capacity yet deduce almost nothing on a typical draw. Treating a large support as though it meant large typical content is exactly the confusion we avoid, and the two facets turn out to be governed by different mathematics: analytic combinatorics for capacity, branching processes for extent.

Our contribution is to isolate the hypothesis class that recursive rule learning actually searches: the realizable closures Ω_L, not the syntax 2^L. Around this object we compute the semantic capacity $C(L) = \log_2 |\Omega_L|$ in closed form and give it three readings: *(i)* A *combinatorial model* (Section 3): Ω_L is exactly countable, and its log-size $C(L)$ is the semantic information of the template. *(ii)* A *probabilistic model* (Section 4): drawing the switchable facts at random turns to the orthogonal *extent* facet, how much a single content deduces, where semantic content undergoes a sharp phase transition. *(iii)* An *online model* (Section 5): $C(L)$ is the learning complexity of the template, at once the realizable Halving mistake bound and the minimax (Shtarkov) log-loss regret; in the agnostic case the minimax $0/1$ regret is $\Theta(\sqrt{C(L)\,T})$, since for the bandwidth family $\mathrm{VC}(\Omega_L) = n - \lfloor n/d \rfloor + O(1)$ and $\mathrm{Ldim}(\Omega_L) = \Theta(C(L))$. The analytic tools are classical; what is new is identifying Ω_L as the object they should be applied to, characterizing it by a nontrivial reachability condition (Proposition 1), and counting it exactly (Theorem 1). For the bandwidth-d reachability template this turns the standard capacity, and learning bounds into concrete rates. We carry one running example, cheating spreading through a class of friends, through all three readings, and confirm on real networks of Florentine families and WordNet (Section 6).

The classical theory of semantic information of Carnap and Bar-Hillel (1952), and its modern philosophical development by Floridi (2011), sought to measure the *content* of a proposition rather than the Shannon bits of a signal. Our $C(L)$ is a concrete, computable realization of that intuition for a recursive template: it counts, in bits, how many distinct *meanings* the template can hold. More directly, this paper can be read as a combinatorial case study within the program of Lastras et al. (2026), which incorporates logical reasoning into Shannon's information theory; there, logic enters the information-theoretic picture in general, whereas here a single recursive family is worked out end to end, and its semantic information is pinned down exactly. The analytic machinery we use are rational generating functions, singularity analysis, and precise minimax redundancy (Szpankowski, 2001; Flajolet and Sedgewick, 2009; Drmota and Szpankowski, 2004; Wu et al., 2026).

Related work This paper sits at the meeting point of five strands, and its contribution is to connect them through the single object Ω_L: **1.** *Semantic information.* The idea that information has a *content* distinct from Shannon's signal bits originates with Carnap and Bar-Hillel (1952) and is developed philosophically by Floridi (2011). That tradition largely measures the content of a single proposition; we instead count, in bits, how many distinct contents a recursive template can hold, and identify that count with a Shannon capacity, a learning rate, and a coding redundancy. The most direct antecedent is the program of Lastras et al. (2026) incorporating logical reasoning into Shannon's theory; our online reading follows Wu et al. (2026). **2.** *Logic-program semantics and (probabilistic) ILP.* That a definite program's meaning is its least Herbrand model — the coincidence of model-theoretic, fixpoint, and proof-theoretic semantics — is classical (van Emden and Kowalski, 1976; Lloyd, 1987). Inductive logic programming (Muggleton, 1991; Muggleton and De Raedt, 1994) learns such programs from examples, its probabilistic extension places a distribution on clauses (Sato, 1995), its differentiable and neural successors optimize a soft 2^L template (Manhaeve et al., 2018), and the field is surveyed by Cropper et al. (2022). Our edge-percolation model of Section 4 is exactly the distribution semantics applied to the transitive-closure template. In the differentiable incarnation of ILP (Evans and Grefenstette, 2018), the soft template is precisely this 2^L parameterization, so the collapse ratio $2^N/|\Omega_L|$ is the slack in such a learner's guarantee (N versus $C(L)$ bits); its realized overpayment is quantified in Appendix C.3. What is new here is not the semantics but the insistence on quotienting syntax by it *before* any capacity: the learner's true hypothesis space is Ω_L, not 2^L. **3.** *Analytic combinatorics and minimax redundancy.* The exact counts come from singularity analysis of rational and infinite-product generating functions (Flajolet and Sedgewick, 2009; Szpankowski, 2001) and the asymptotic enumeration of partial orders of Kleitman and Rothschild (1970; 1975) governing the nonlinear (transitive-closure) variant. The coding-theoretic reading rests on Shtarkov's normalized maximum likelihood (Shtarkov, 1987) and the precise minimax-redundancy asymptotics of Drmota and Szpankowski (2004); Jacquet and Szpankowski (2004); Rissanen (1996); Clarke and Barron (1990); $C(L) = \log_2 |\Omega_L|$ is the Shtarkov sum of the deductive concept class, and its parametric

refinement is a Markov-source redundancy in that line. **4.** *Online learning and concept classes.* Reading contents as classifiers puts the template in the standard mistake-bound and regret framework (Littlestone, 1988; Cesa-Bianchi and Lugosi, 2006; Cover and Thomas, 2006): Halving gives the $\log_2 |\Omega_L|$ mistake bound, online regret, and the Littlestone and VC dimensions bound the achievable rates. Our point is that these familiar bounds become *closed-form rates* in ρ_d and φ once Ω_L is counted, and that a syntactic learner is minimax-suboptimal, by a gap linear in n (Appendix C.3). **5.** *Random graphs and phase transitions.* The extent facet reduces to the transitive closure of a random digraph, whose sharp threshold is Karp's (Karp, 1990), via the branching-process couplings of Athreya and Ney (1972); van der Hofstad (2017). We use that threshold to infer *semantic content*.

2 PRELIMINARIES: REALIZABLE CLOSURES

Fix a finite set of predicate symbols, each with a fixed arity, a set of constant symbols, and a set of variables, and assume there are no function symbols of positive arity. An *atom* is an expression $P(t_1, \ldots, t_r)$, where P is a predicate symbol of arity r and each t_i is either a constant symbol or a variable; it is a *ground atom* if it contains no variables, and sometimes simply called a *fact*. The *Herbrand base* $\mathcal{B}$ is the set of all ground atoms. A *definite clause* is an expression $h \leftarrow b_1, \ldots, b_k$, where $h, b_1, \ldots, b_k$ are atoms; h is the *head* and $b_1, \ldots, b_k$ form the *body*. A *program* is a finite set of definite clauses. The immediate-consequence operator $T_P : 2^{\mathcal{B}} \to 2^{\mathcal{B}}$ of a program P fires, in one step, every rule whose body already holds: $T_P(M) = \{h : (h \leftarrow b_1, \ldots, b_k) \in P,\ \{b_1, \ldots, b_k\} \subseteq M\}$. Its iterates from $\emptyset$ increase to the least model $\mathrm{lhm}(P) = \bigcup_{k \geq 0} T_P{\uparrow}k$, the set of atoms P deduces (van Emden and Kowalski, 1976; Lloyd, 1987).

We work with a set of ground background clauses BK (ground facts and/or rules) and a finite template $L = \{c_1, \ldots, c_N\}$ of definite clauses. A hypothesis is a subset $S \subseteq L$; adjoining it to BK gives a definite program whose meaning is the least Herbrand model $\mathrm{lhm}(\mathrm{BK} \cup S) \subseteq \mathcal{B}$ with $\mathcal{B}$ being the Herbrand base. Throughout, we keep the two kinds of object apart: a *hypothesis* $S \in 2^L$ is a set of clauses (rules the learner switches on); its image $R := \mu(S) \in \Omega_L$ is a set of *atoms* (the facts those rules deduce), $R \subseteq \mathcal{B}$. So μ maps rule-sets to atom-sets, and Ω_L collects the atom-sets, never the rule-sets. We often split BK into a fixed *structural* part and a set E of declared *evidence* (seed) facts.

Switchable rules vs. switchable facts. The switchable elements of a template L may be *rules* or ground *base facts*, and which it is depends on the template family. Both are subsets of the same Horn template: a base fact is simply a Horn clause with empty body, so switching on facts selects a subset of L exactly as switching on rules does, and L is uniformly a set of Horn clauses throughout. In the bandwidth family (Section 3) the switchable elements are rules (edge clauses $r_v \leftarrow r_u$), and we write a hypothesis as $S \subseteq L$. In the transitive-closure family (Theorem 2 and Section 4) the composition rule is fixed in BK and the switchable elements are the base facts; there we write a hypothesis as $F \subseteq L$. The maps μ, $\Omega_L = \mathrm{im}\,\mu$, and $C(L) = \log_2 |\Omega_L|$ are defined identically in both cases; only the type of switched-on object differs. We reserve S for rule-sets and F for fact-sets throughout, so that "S random" always means random rules and "F random" always means random facts.

Everything rests on a clean description of which subsets of $\mathcal{B}$ actually lie in Ω_L, whose log-size is the (semantic) capacity. For the reachability templates below, every clause has the shape $r_v \leftarrow r_u$ (a plain edge $u \to v$).

Proposition 1 (Realizability characterization)**.** *Let the template consist of edge clauses $r_v \leftarrow r_u$ over a directed graph $G = (V, E)$, with* BK *declaring a source set $V_0 \subseteq V$ true. Then $R \subseteq V$ is realizable ($R \in \Omega_L$) if and only if $V_0 \subseteq R$ and every vertex of R is reachable from V_0 along a path lying entirely inside R.*

Example 1 (Cheating in a class)**.** Five students A, B, C, D, F and two predicates, $\mathrm{cheat}/1$ and $\mathrm{friend}/2$, give the Herbrand base $\mathcal{B} = \{\mathrm{cheat}(v) : v \in \{A, B, C, D, F\}\} \cup \{\mathrm{friend}(u, v) : u, v \in \{A, B, C, D, F\}\}$, so that $|\mathcal{B}| = 5 + 25 = 30$, where u, v range over the five students. The background holds a friendship graph (each tie declared in *both* directions, so the relation is symmetric) together with the one student caught cheating — the *evidence* (seed) $E = \{\mathrm{cheat}(B)\}$ — and the template is the single spreading schema $\mathrm{cheat}(Y) \leftarrow \mathrm{cheat}(X), \mathrm{friend}(X, Y)$:

$$\mathrm{BK} = \underbrace{\{\,\mathrm{friend}(X,Y), \mathrm{friend}(Y,X) \;:\; (X,Y) \in \{BA,\ BD,\ AC,\ DC,\ CF\}\,\}}_{\text{structure (each tie, both directions)}} \cup \underbrace{E}_{\text{evidence}},$$

$$L = \{\, \mathrm{cheat}(Y) \leftarrow \mathrm{cheat}(X), \mathrm{friend}(X, Y) \,\} \qquad (N = |L| = 10 \text{ ground clauses}).$$

The 25 friend-atoms are fixed background: every content contains exactly the ten declared ties and no other friend-atom, so contents differ only in their cheat-atoms, the content sub-base on which the learner is queried (a convention we keep throughout the paper). A hypothesis $S \subseteq L$ switches on a set of *conducting* friendships; its meaning $\mu(S) = \mathrm{lhm}(\mathrm{BK} \cup S) \in \Omega_L$ is the set of students it concludes cheat. Taking $S = \{\, \mathrm{cheat}(A) \leftarrow \mathrm{cheat}(B), \mathrm{friend}(B, A);\ \ \mathrm{cheat}(C) \leftarrow \mathrm{cheat}(A), \mathrm{friend}(A, C)\,\} \subseteq L$, each ground body atom $\mathrm{friend}(B, A), \mathrm{friend}(A, C) \in \mathrm{BK}$ holds, so the fixpoint iterates are $\{B\}, \{B, A\}, \{B, A, C\}$ and $R = \mu(S) = \{B, A, C\}$. Two features are the whole point. First, μ is *many-to-one*: adjoining the reverse clause $\mathrm{cheat}(B) \leftarrow \mathrm{cheat}(A), \mathrm{friend}(A, B)$ to S (its body also holds, since $\mathrm{friend}(A, B) \in \mathrm{BK}$) changes no conclusion, so a whole fibre $\mu^{-1}(\{B, A, C\})$ shares one content. Second, not every atom set is a content: $\{B, C\}$ is the least model of no S, since deducing $\mathrm{cheat}(C)$ first needs a cheating friend ($\mathrm{cheat}(A)$ or $\mathrm{cheat}(D)$). By Proposition 1, μ ranges over all $2^{10} = 1024$ hypotheses, but

$$\Omega_L = \big\{\{B\}, \{BA\}, \{BD\}, \{BAC\}, \{BAD\}, \{BCD\}, \{BACD\}, \{BACF\}, \{BCDF\}, \{BACDF\}\big\}$$

so $|\Omega_L| = 10$. So Ω_L collects the deductive closures and only those — a strict, structured subset of the atom sets, here a lattice ($\{BA\}$ and $\{BD\}$ are incomparable: two different outbreaks).

Finally, we define three quantities, none needing probability yet. This will fix the vocabulary.

Definition 1 (Semantic content, extent, capacity)**.** For a hypothesis $S \subseteq L$: its *semantic content* is the deductive closure $\mu(S) = \mathrm{lhm}(\mathrm{BK} \cup S) \in \Omega_L$; its *semantic extent* is the number of new facts that content adds, $E(S) := |\,\mathrm{lhm}(\mathrm{BK} \cup S)| - |\,\mathrm{lhm}(\mathrm{BK})|$. The *semantic capacity* of the template is $C(L) := \log_2 |\Omega_L|$ bits — the number of bits needed to name which content was produced.

Remark 1 (Why "capacity"). The name is information-theoretic, not merely suggestive. The closure map $\mu : 2^L \to \Omega_L$, $S \mapsto R = \mu(S)$, is a deterministic channel, so any prior P on the clause-sets pushes forward to $\pi = \mu_* P$ on Ω_L with $I(S; R) = H(R)$ and $H(S) = H(R) + H(S \mid R)$; the equivocation $H(S \mid R)$ is exactly the syntactic redundancy μ discards. Maximizing over priors gives $\sup_P I(S; R) = \log_2 |\Omega_L| = C(L)$, uniquely attained by the content-uniform prior, under which the self-information $-\log_2 \pi(R) \equiv C(L)$ is an exact equipartition over realizable theories; a generic prior falls short of capacity by $D(\pi \,\|\, \mathrm{unif}_{\Omega_L})$. Thus $C(L)$ is a genuine Shannon capacity, not just a Hartley count. Appendix D develops this deterministic-channel lift.

3 Combinatorial Model: Counting the Contents

In this section we study a simple template L that allows us to precisely estimate the quantities defined above. The atoms are $r_0, \ldots, r_n$; they *are* the Herbrand base $\mathcal{B}$, and $\mathrm{BK} = \{r_0\}$ declares the source r_0. The bandwidth-d template is

$$L_{n,d} = \{\, c_{i \to j} : r_j \leftarrow r_i,\ 0 \le i < j \le n,\ j - i \le d\,\}, \qquad N = |L_{n,d}| = dn - \tbinom{d}{2}.$$

This is exactly the recursive path/transitive-closure template of ILP, restricted to steps of length $\le d$; d is the reach of one rule, the locality knob of a language bias. In other words, $L_{n,d}$ is a set of edge-clauses: reading $r_j \leftarrow r_i$ as a directed edge $i \to j$ makes $L_{n,d}$ a graph on the atoms. A hypothesis is a subset $S \subseteq L_{n,d}$, a choice of which edges are *switched on*, so S lists the edges that conduct, and there are 2^N such choices. A hypothesis does not stand on its own: one adjoins it to BK and fires the immediate-consequence operator to a fixpoint, starting from the source. The result is the content $R = \mu(S) = \mathrm{lhm}(\mathrm{BK} \cup S) \ \in\ \Omega_L, r_0 \in R$, the set of atoms reachable from r_0 along the switched-on edges of S. So S is a set of *edges* (in $2^{L_{n,d}}$) and R is a set of *atoms* (in $2^{\{0,\ldots,n\}}$).

Example 2 (The cheating class, seated in a row.)**.** Example 1 is an instance: seat the class in a line and let friendship mean "within d seats." Each tie then licenses one ground rule $\mathrm{cheat}(Y) \leftarrow \mathrm{cheat}(X), \mathrm{friend}(X, Y)$ — one edge of G between cheat-atoms — and the abstract family absorbs the fixed tie into that edge, writing it as the plain $r_j \leftarrow r_i$ with r_i in the role of $\mathrm{cheat}(\text{student } i)$.

Write $\Omega_L(n, d) := \Omega_L(L_{n,d})$. We represent a content by the indices of its atoms, $R = \{0 = x_0 < x_1 < \cdots < x_m\} \subseteq \{0, \ldots, n\}$, where x_ℓ is the index of the atom $r_{x_\ell} \in R$; reachability from r_0 along edges of length at most d forces every consecutive gap $x_\ell - x_{\ell-1} \le d$. In view of this, $\Omega_L(n, d)$ is the set of vertex subsets of $\{0, \ldots, n\}$ that contain the source and have no gap exceeding d—not edge-sets, and not arbitrary vertex-sets. Grouping the realizable sets by their top element $x_m = x$ identifies each with a composition of the integer x into parts in $\{1, \ldots, d\}$ (the gaps). This gives $|\Omega_L(n, d)| = \sum_{x=0}^{n} C_d(x)$, with $C_d(x)$ the number of such compositions, so that $\sum_x C_d(x) t^x = (1 - (t + \cdots + t^d))^{-1} = (1 - t)/(1 - 2t + t^{d+1})$ (Flajolet and Sedgewick, 2009; Szpankowski, 2001). This leads to our main result of this section.

Theorem 1 (Exact count and precise asymptotics). *The counts $a_n = |\Omega_L(n,d)|$ have rational generating function*

$$\sum_{n\geq 0} a_n t^n = \frac{1}{1-2t+t^{d+1}}, \qquad \textit{equivalently} \qquad a_n = 2a_{n-1} - a_{n-d-1}\ (n \geq d+1),$$

with $a_k = 2^k$ for $k \leq d$. As $n \to \infty$ with $d \geq 2$ fixed, $a_n = A_d\, \rho_d^n \big(1 + O(\gamma_d^n)\big)$, where $\rho_d \in (1,2)$ is the d-bonacci constant (the unique root in $(1,2)$ of $\rho^d = \rho^{d-1} + \cdots + \rho + 1$), A_d is an explicit constant, and $\gamma_d < 1$. In particular for $d = 2$,

$$\rho_2 = \varphi = \tfrac{1+\sqrt{5}}{2}, \qquad A_2 = \frac{\varphi^3}{\sqrt{5}}, \qquad a_n = F_{n+3} - 1 \sim \frac{\varphi^{n+3}}{\sqrt{5}}.$$

Corollary 1 (Collapse ratio and semantic capacity). *With $N = dn + O(1)$ clauses,*

$$\frac{2^N}{|\Omega_L(n,d)|} = \Big(\frac{2^d}{\rho_d}\Big)^n 2^{O(1)}, \qquad \frac{\log_2 |\Omega_L(n,d)|}{n} \xrightarrow[n\to\infty]{} \log_2 \rho_d \in (0,1).$$

The template carries $\log_2 \rho_d$ bits of content per atom, strictly less than the d bits of syntax per atom; the deficit $d - \log_2 \rho_d$ is the redundancy the language bias builds in by letting several rules compute the same reachability. As $d \to \infty$, $\log_2 \rho_d \uparrow 1$: content saturates at one bit per atom.

The nonlinear jump. The bandwidth family is *linearly* recursive (one recursive atom per body), and its count is a one-dimensional transfer-matrix object with the clean growth constant ρ_d. The canonical ILP target—path, $\mathrm{ancestor}$, transitive closure—is instead *nonlinear*: genuine transitive closure through the doubly-recursive composition rule $\mathrm{tc}(X,Z) \leftarrow \mathrm{tc}(X,Y), \mathrm{tc}(Y,Z)$, whose body carries two recursive atoms. Here the fixed and switchable parts of the program trade places. In the bandwidth family BK is a fact (the source r_0) and L is a set of rules; now BK is the single composition rule above—the recursion is fixed, not learned—and L is the set of ground base facts $L = \{\mathrm{tc}(i,j) : 1 \leq i < j \leq m\}$, $N = |L| = \binom{m}{2}$, each a body-empty Horn clause (cf. Section 2). A hypothesis is a fact-set $F \subseteq L$ (which edges $\mathrm{tc}(i,j)$ are declared; we write F, not S, since the switchable clauses are facts). The content sub-base is $\mathcal{B} = L$, the same set of atoms, and the content is $R = \mu(F) = \mathrm{lhm}(\mathrm{BK} \cup F) = \mathrm{TC}(F)$, obtained by firing the composition rule on F to a fixpoint; there is no source atom, since $\mathrm{TC}(F)$ needs none. Holding the rule fixed and exposing the base relation is exactly the distribution-semantics reading (Sato, 1995) that Section 4 randomizes. Every transitively closed R satisfies $\mathrm{TC}(R) = R$, so $R = \mu(R) \in \Omega_L$; hence Ω_L is precisely the set of transitively closed subrelations of $\{(i,j) : i < j\}$. For a proof of the next result see Appendix A.3.

Theorem 2 (Nonlinear capacity: the transitive-closure jump). *In the ordered reading the contents Ω_L are the strict partial orders on $[m]$ admitting the identity as a linear extension. Their number $T(m) := |\Omega_L|$ satisfies*

$$2^{\lfloor m^2/4 \rfloor} \;\leq\; T(m) \;\leq\; 2^{\,m^2/4 + o(m^2)}, \qquad \textit{so} \qquad \kappa := \lim_{m\to\infty} \frac{\log_2 T(m)}{m^2} = \frac{1}{4},$$

and $C(L) = \log_2 T(m) = \frac{1}{4} m^2 (1 + o(1))$ bits. The same $\kappa = \frac{1}{4}$ governs the unordered variant, where Ω_L is the set of all *transitive relations on $[m]$. Thus nonlinear recursion has quadratic capacity $\Theta(m^2)$, against the linear $\Theta(n)$ of the bandwidth family.*

4 Probabilistic Model: A phase transition in semantic content

We now switch to a probabilistic model and turn to the semantic extent, that is, how *much* a single content deduces. The natural random model is the distribution semantics of probabilistic logic programs (De Raedt and Kimmig, 2015): fix the recursive rule as background, select each candidate base fact independently, and recompute the least model. The random object is thus the *graph* of selected base facts, never the rule. Throughout this section BK and L are fixed; the random object is the hypothesis. In the transitive-closure template (Theorem 3) BK is the composition rule, L is the set of all candidate base facts $\mathrm{tc}(i,j)$, $i \neq j$, over m constants, and $F \subseteq L$ switches each fact on independently with probability p.

With each base fact selected with probability p (Erdős–Rényi model), the selected facts F form a random digraph $\mathcal{D}(m,p)$, and the content $R = \mu(F) = \mathrm{TC}(F)$ is its transitive closure. The semantic

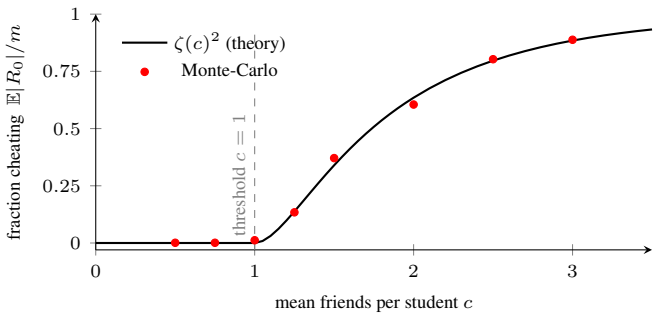


Figure 1: The cheating epidemic. Fraction of an m-student class implicated after one student is caught, $\mathbb{E}|R_0|/m$, against the mean number of friends per student c, for random friendships $G(m, c/m)$. The order parameter is $\zeta(c)^2$, with ζ solving $1 - \zeta = e^{-c\zeta}$ (solid). Monte-Carlo ($m = 3000$, 250 trials per point).

extent of the draw (Definition 1) is $E(F) = |R| = \#\{(i, j) : i \rightsquigarrow j\}$, and $\mathbb{E}|R|$ is the average semantic content.

By symmetry $\mathbb{E}|R| = m \cdot \mathbb{E}|R_0|$ up to the diagonal, where $R_0 = \{v : 0 \rightsquigarrow v\}$ is the set reachable from one constant. Exploring R_0 by breadth-first search, each newly reached vertex has $\mathrm{Binomial}(m-1, p)$ out-neighbours; in the sparse regime $p = c/m$ this is asymptotically $\mathrm{Poisson}(c)$, so R_0 is the total progeny of a Poisson(c) branching process (Karp, 1990; Athreya and Ney, 1972). Branching processes have a threshold at mean offspring $c = 1$, and it transfers verbatim to the deductive closure. Detailed proof is in Appendix A.4.

Theorem 3 (Phase transition of semantic content)**.** *Let the base facts be $\mathcal{D}(m, c/m)$ and let $\zeta = \zeta(c)$ be the unique positive root of $1 - \zeta = e^{-c\zeta}$ (so $\zeta = 0$ for $c \le 1$ and $\zeta > 0$ for $c > 1$). As $m \to \infty$,*

$$\mathbb{E}|R_0| \longrightarrow \begin{cases} \dfrac{1}{1-c}, & c < 1, \\ \zeta(c)^2\, m\,(1 + o(1)), & c > 1, \end{cases} \qquad \mathbb{E}|R| = \begin{cases} \Theta(m), & c < 1, \\ \zeta(c)^2\, m^2\,(1 + o(1)), & c > 1. \end{cases}$$

The average semantic content jumps from linear to quadratic in m at the critical density $c = 1$: below it a random program deduces a bounded number of facts per constant, and its total content is $\Theta(m)$; above it, the content is a fixed positive fraction $\zeta(c)^2$ of all m^2 possible facts; at $c = 1$ exactly, $\mathbb{E}|R| = \Theta(m^{4/3})$.

The bandwidth-d template of Section 3 has no such transition (proof in Appendix A.5).

Corollary 2 (Bounded reach cannot percolate)**.** *For the bandwidth-d template of Section 3 (each atom has at most d in-neighbours, edges on with probability $\theta < 1$), the reachable set is a.s. finite and $\mathbb{E}|R_0| = O_{d,\theta}(1)$, uniformly in n, for every fixed d. There is no critical $\theta < 1$.*

Corollary 2 is the correct reading of the "bounded content" phenomenon: a bounded-reach recursive template is a quasi-one-dimensional reachability process, which in one dimension cannot percolate, so its semantic content stays bounded no matter how reliable the rules. The phase transition appears only once rules reach across an unbounded number of constants (the mean-field regime). The two results together locate the boundary exactly: *semantic content is a critical phenomenon*, with critical density one base fact per constant, and bounded-arity language biases sit permanently on its trivial side. This is an extent-axis statement and must not be confused with the capacity of Section 3: a template can have exponential capacity (many possible contents) while its *typical* content, under an independent-clause prior, is $O(1)$ per source.

Example 3 (The cheating epidemic)**.** The same threshold governs cheating at scale, now in the reading of this section. In Example 1 the friendships were fixed background and the spreading rules were switchable; here the roles swap: BK holds the spreading rule $\mathrm{cheat}(Y) \leftarrow \mathrm{cheat}(X), \mathrm{friend}(X, Y)$ and the seed $\mathrm{cheat}(B)$; L is the set of all base facts $\mathrm{friend}(u, v)$ over m students; and the hypothesis $F \subseteq L$ is drawn at random, each pair present with probability c/m (mean degree c). The friendship edges play the role of the random base facts F of Theorem 3. Two things differ: the graph is undirected, $G(m, c/m)$ rather than $D(m, c/m)$, and the content is the connected component R_0 of the seed rather than the transitive closure of all pairs. The argument is unchanged: the breadth-first

exploration of R_0 is the same $\mathrm{Poisson}(c)$ branching process, so the same threshold and the same $\zeta(c)^2$ apply (Figure 1). Cheating is a critical phenomenon with threshold one friend per student: below it contained ($O(1)$ others), above it an epidemic reaching a fraction $\zeta(c)^2$ of the class.

5 Adversarial Model: Learning and minimax regret

Section 4 was the semantic extent: how much a single content deduces. We now return to the capacity. We observe that Ω_L is, literally, the hypothesis class (in ML it is often denoted as $\mathcal{H}$) a learner searches, and $C(L)$ *is* the learning complexity. Each content $R \in \Omega_L$ predicts, on every ground atom, whether it is deducible: $h_R(\gamma) = \mathbf{1}[\gamma \in R]$. So $\Omega_L \subseteq \{0,1\}^{\mathcal{B}}$ is a finite concept class, one concept per realizable theory. The two errors contributing to the loss are logic's own: a false positive is a deduced falsehood (over-general rule), a false negative a missed truth (incomplete rule).

Mismatch loss. We score a learner by the $0/1$ (mismatch) loss against the best content in hindsight. As a warm-up we deal first with realizable case.

Theorem 4 (Realizable mistake bound)**.** *In the realizable online setting, the Halving algorithm (Littlestone, 1988; Cesa-Bianchi and Lugosi, 2006) over Ω_L makes at most $\log_2 |\Omega_L| = C(L)$ mistakes. The exact mistake-bound value is the Littlestone dimension $\mathrm{Ldim}(\Omega_L)$, and $\mathrm{VC}(\Omega_L) \le \mathrm{Ldim}(\Omega_L) \le \log_2 |\Omega_L| = C(L)$ (Littlestone, 1988) (see detailed proof in Appendix A.6).*

Now drop realizability: the label sequence $y_{1:T} \in \{0,1\}^T$ is arbitrary (adversarial), and we measure $0/1$ regret against the best content in hindsight. The regret becomes

$$\mathrm{Regret}_T^{0/1} \;=\; \sum_{t=1}^{T} \mathbf{1}[\hat{y}_t \neq y_t] \;-\; \min_{R \in \Omega_L} \sum_{t=1}^{T} \mathbf{1}[h_R(\gamma_t) \neq y_t],$$

where $\hat{y}_t \in \{0,1\}$ is the learner's predicted label for γ_t.

Theorem 5 (Agnostic minimax rate)**.** *For every finite content class Ω_L and every $T \ge \mathrm{Ldim}(\Omega_L)$, the minimax expected $0/1$ regret over T rounds satisfies*

$$c_1 \sqrt{\mathrm{Ldim}(\Omega_L)\, T} \;\le\; \min_{\text{learner}} \max_{y_{1:T}} \mathbb{E}\, \mathrm{Regret}_T^{0/1} \;\le\; \sqrt{\tfrac{\ln 2}{2}\, C(L)\, T},$$

with $c_1 = \frac{1}{4}$ (see Algorithm 2 in Appendix B).

Proof. Upper bound. Run Hedge (a version of EWA algorithm) over the $|\Omega_L|$ content-experts, the R-th predicting $h_R(\gamma_t)$. Tuned Hedge has $0/1$ regret at most $\sqrt{(T/2)\ln M}$ against the best of M experts (Cesa-Bianchi and Lugosi, 2006); here $M = |\Omega_L|$ and $\ln M = (\ln 2)\, C(L)$, giving $\sqrt{(\ln 2/2)\, C(L)\, T}$. *Lower bound.* By the agnostic online lower bound (Ben-David et al., 2009), any class $\mathcal{H}$ with $\mathrm{Ldim}(\mathcal{H}) = D$ forces minimax regret at least $c_1\sqrt{DT}$ (one may take $c_1 = \frac{1}{4}$ for $T \ge D$): a Littlestone mistake tree of depth D embeds a sequential Rademacher game whose expected regret against the best root-to-leaf comparator is $\Theta(\sqrt{DT})$. Apply with $\mathcal{H} = \Omega_L$ and $D = \mathrm{Ldim}(\Omega_L)$. Full proof in Appendix A.7. □

The two bounds "match" *iff* $\mathrm{Ldim}(\Omega_L) = \Theta(C(L))$: the Hedge upper bound holds for any finite class, but without a lower bound on Ldim the matching regret lower bound could be as small as $\sqrt{\mathrm{Ldim}\; T}$ with $\mathrm{Ldim} \ll C(L)$. For the bandwidth family the gap closes as stated below and proved in Appendix A.8.

Corollary 3 (Bandwidth-d agnostic rate)**.** *For the bandwidth-d class, $\mathrm{Ldim}(\Omega_L(n,d)) = \Theta(C(L))$, we have $\mathrm{VC}(\Omega_L(n,d)) = n - \lfloor n/d \rfloor + O(1)$, and the sandwich $\mathrm{VC} \le \mathrm{Ldim} \le \log_2 |\Omega_L|$ of Theorem 4 makes both ends linear in n. Hence both sides of Theorem 5 are $\Theta\big(\sqrt{C(L)\, T}\big)$, so the minimax agnostic $0/1$ regret is $\Theta\big(\sqrt{C(L)\, T}\big)$.*

Logarithmic Loss. Under logarithmic loss each content is the deterministic predictor $p_t^R(y) = \mathbf{1}[y = \mathbf{1}[\gamma_t \in R]]$, and a strategy $\hat{p} = (\hat{p}_t)$ pays per round the log loss $\ell_t^{\log}(\hat{p}_t, y_t) \;:=\; -\log_2 \hat{p}_t(y_t)$, so that its cumulative loss is the code length $-\log_2 \hat{P}(y_{1:T})$ with $\hat{P}(y_{1:T}) = \prod_t \hat{p}_t(y_t)$. The minimax regret over a query sequence is

$$\mathrm{Regret}_T^{\star}(\gamma_{1:T}) = \min_{\hat{p}} \max_{y} [\sum_t \ell_t^{\log}(\hat{p}_t, y_t) - \min_R \sum_t \ell_t^{\log}(p_t^R, y_t)].$$

Algorithm 1: Universal deductive predictor (version-space mixture)

Input: content class Ω_L; query stream $\gamma_1, \gamma_2, \ldots$; prior weights $w(R) > 0$ (default $w \equiv 1$)
$V \leftarrow \Omega_L$; `// version space: contents consistent with labels`
for $t = 1, 2, \ldots, T$ **do**
  receive query γ_t;
  $\hat{p}_t \leftarrow \dfrac{\sum_{R \in V : \gamma_t \in R} w(R)}{\sum_{R \in V} w(R)}$; `// mixture probability that` γ_t `is entailed`
  predict $\hat{p}_t \in [0,1]$ (log loss), or $\hat{y}_t \leftarrow \mathbf{1}[\hat{p}_t \geq \frac{1}{2}]$ (0/1 loss);
  observe $y_t = \mathbf{1}[\gamma_t \in R^*]$;
  $V \leftarrow \{R \in V : \mathbf{1}[\gamma_t \in R] = y_t\}$; `// prune inconsistent contents`

Theorem 6 (Minimax log-loss regret is the Shtarkov sum — and it is the capacity)**.** *Write* $\Omega_L|_{\gamma_{1:T}} = \{(\mathbf{1}[\gamma_1 \in R], \ldots, \mathbf{1}[\gamma_T \in R]) : R \in \Omega_L\}$ *for the label patterns the contents induce on the queries. For every query sequence the minimax regret is the log Shtarkov sum,*

$$\mathrm{Regret}_T^\star(\gamma_{1:T}) = \log_2 \sum_{y \in \{0,1\}^T} \max_{R \in \Omega_L} P_R(y) = \log_2 \big|\Omega_L|_{\gamma_{1:T}}\big|,$$

attained uniquely by the normalized-maximum-likelihood (Shtarkov) strategy (Shtarkov, 1987). Consequently $\mathrm{Regret}_T^\star \leq \log_2 |\Omega_L| = C(L)$*, with equality once the queries realize all* $|\Omega_L|$ *patterns, and by Sauer–Shelah* $\mathrm{Regret}_T^\star \leq \min\{C(L),\ \mathrm{VC}(\Omega_L) \log_2(T+1)\}$.

Proof sketch. On outcome y a strategy with sequence likelihood $\hat{P}$ has regret $\log_2(M(y)/\hat{P}(y))$ with $M(y) = \max_R P_R(y)$; the NML distribution $q^\star(y) = M(y)/S$ makes the ratio constant at $S = \sum_y M(y)$, and shows no strategy beats $\log_2 S$. Here $P_R(y) \in \{0,1\}$ is the indicator that y is R's pattern, so S counts distinct patterns, $S = |\Omega_L|_{\gamma_{1:T}}|$; this is at most $|\Omega_L|$, with equality once the queried atoms separate the contents. The precise minimax form is that of Drmota and Szpankowski (2004). Detailed proof is in Appendix A.9. □

Evaluation of Algorithm 1. The bandwidth family of Section 3 turns Theorem 6 into a concrete rate: querying the prefix $r_0, \ldots, r_{T-1}$ realizes exactly the smaller template's contents, so $\big|\Omega_L|_{\gamma_{1:T}}\big| = |\Omega_L(T-1, d)| = a_{T-1} \sim A_d \rho_d^{T-1}$ and $\mathrm{Regret}_T^\star = (T-1)\log_2 \rho_d + \log_2 A_d + o(1)$: the redundancy grows linearly, at $\log_2 \rho_d$ bits per informative query, until it saturates at $C(L) = \log_2 a_n$ once the full base $r_0, \ldots, r_n$ is queried. Against this value we place the efficient, horizon-free Bayes mixture under the uniform prior on Ω_L — the version-space predictor of Algorithm 1, whose hard-label version is exactly Halving. Theorem 7 (Appendix A.10) evaluates it for an *arbitrary* template and query sequence, not only the bandwidth family.

Theorem 7 (Exact regret of Algorithm 1)**.** *Fix a query sequence* $\gamma_{1:T}$ *and a realizable outcome, let* $w : \Omega_L \to \mathbb{R}_{>0}$ *be the prior weight function parameterizing Algorithm 1 (default* $w \equiv 1$*), extended to sets by* $w(A) = \sum_{R \in A} w(R)$*; and let* V_{T+1} *be the final version space.* (i) *(Exact regret.) Under logarithmic loss the cumulative regret of Algorithm 1 is*

$$\mathrm{Regret}_T = \log_2 \frac{w(\Omega_L)}{w(V_{T+1})} \ \leq\ \log_2 \frac{w(\Omega_L)}{w(R^\star)},$$

The right-hand side is the code length of the true theory under w*. Under the uniform prior* $w \equiv 1$ *this reads* $\mathrm{Regret}_T = \log_2 \frac{|\Omega_L|}{|V_{T+1}|} \leq \log_2 |\Omega_L| = C(L)$*: it pays only for the contents it has not yet ruled out.* (ii) *(Matching lower bound.) For the same query sequence, every sequential predictor has worst-case regret at least the minimax value* $\mathrm{Regret}_T^\star = \log_2 \big|\Omega_L|_{\gamma_{1:T}}\big|$ *of Theorem 6. Hence when the queries separate the contents (*$\big|\Omega_L|_{\gamma_{1:T}}\big| = |\Omega_L|$*, so* $|V_{T+1}| = 1$*), the uniform Algorithm 1 attains* $\log_2 |\Omega_L| = C(L)$ *and is minimax-optimal.*

Remark 2 (Mistake rates by recursion type). Halving is the hard-label version of Algorithm 1, so in the realizable setting identifying a recursive theory costs at most $C(L)$ mistakes: $O(n)$ for the linear bandwidth-d template, where $C(L) = n \log_2 \rho_d\,(1 + o(1))$ by Theorem 1, and $O(m^2)$ for nonlinear transitive closure, where $C(L) = \frac{1}{4} m^2 (1 + o(1))$ by Theorem 2. The nonlinear rate cannot

be improved: the atoms $\{\mathrm{tc}(a,b) : a \in A,\, b \in B\}$ of a balanced bipartition $[m] = A \sqcup B$ are shattered (Appendix A.3), so $\mathrm{VC}(\Omega_L) = \Theta(m^2)$ and *every* learner needs $\Theta(m^2)$ examples.

6 Empirical validation

We validate the two operational claims of the framework: that syntax collapses onto a small semantic content class, and that the online cost of Algorithm 1 is exactly the semantic information $C(L) = \log_2 |\Omega_L|$, on a real relational dataset, Padgett's Florentine families marriage network (15 families, 20 ties). We use the template with the seed $E = \{\mathsf{marked}(\mathsf{Medici})\}$ and one clause $\mathsf{marked}(Y) \leftarrow \mathsf{marked}(X), \mathsf{tie}(X, Y)$ per orientation of each tie, where $\mathsf{marked}(X)$ reads "family X has been reached from the seed". Thus with $N = 40$ ground clauses, the content sub-base is the 15 marked-atoms, one per family (the ties are fixed background). Appendix C gives details and scales the same experiment to the WordNet taxonomy.

Semantic collapse. Enumerating the least Herbrand model of all $2^N = 2^{40} \approx 1.1 \times 10^{12}$ clause-sets, the syntax collapses onto exactly $|\Omega_L| = 4{,}104$ realizable contents — the connected atom-sets containing the seed, as Proposition 1 predicts. The collapse ratio is $2^N/|\Omega_L| \approx 2.7 \times 10^8$: a marriage network of forty switchable rules expresses only $C(L) = \log_2 4{,}104 = 12.00$ bits of semantic content, not the 40 syntactic bits a clause-set learner would budget for. The ≈ 28-bit gap is the slack between the syntactic and semantic guarantees.

Algorithm 1 on real data. We run the version-space mixture online over Ω_L (uniform prior): the 15 atoms are revealed in random order, each labeled by whether a hidden target content $R^* \in \Omega_L$ deduces it, and the learner predicts the entailment bit and pays log-loss. Figure 2 reports both losses. Its cumulative log-loss reaches $C(L) = 12.00$ bits *exactly* for every target and every query order (right endpoint of every trajectory), matching Theorem 7: once the Herbrand base separates the contents the version space collapses to R^* and the paid code length is precisely the semantic information. Under $0/1$ loss the same run identifies the target in at most $C(L)$ mistakes — mean 5.3 and never more than 10 over 300 real targets, comfortably below the $\log_2 |\Omega_L| \approx 12$ bound of Theorem 4. The identity persists at scale: on the WordNet carnivore taxonomy (366 concepts, 2^{365} clause-sets, $|\Omega_L| = 10^{90.9}$) Algorithm 1, run implicitly by an $O(n)$ tree dynamic program, pays exactly $C(L) = 302$ bits (Table 2). Thus on real data the single number $C(L)$ governs both the coding regret (attained) and the mistake bound.

Baseline: learning over clause-sets. The uniform Bayes mixture over the 2^N clause-sets predicts only through μ, so it *is* Algorithm 1 with prior $w(R) = |\mu^{-1}(R)|$, and Theorem 7(i) gives its regret exactly, $\log_2(2^N/|\mu^{-1}(R^\star)|)$. On the Florentine network (all 4,104 fibres counted exactly) it pays 6 to 20 bits — above $C(L)$ on 79% of the contents, 14.35 on average — against 12.00 on every target; on the bandwidth-2 template it pays $1.01\,n$ bits on average over contents and $\frac{4}{3}n$ in the worst case against $C(L) = 0.694\,n$, a gap linear in n. Uniform-on-syntax is a skewed prior on meanings: it wins only on huge-fibre targets (random clause-sets, where it pays $H(\pi) = O(1)$ by Corollary 2), while uniform-on-contents is minimax (Theorem 7(ii)). The realized gap is a constant fraction of the collapse ratio, which is the slack between the guarantees N and $C(L)$ (Appendix C.3).

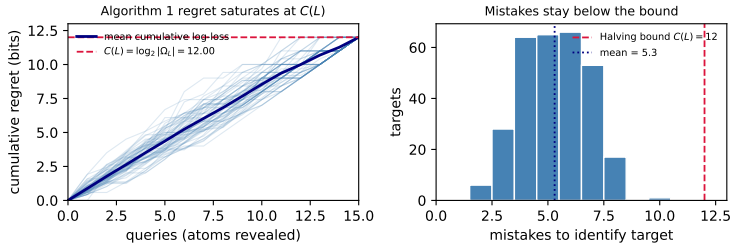


Figure 2: Algorithm 1 on the Florentine network ($|\Omega_L| = 4{,}104$). **Left:** cumulative log-loss over 60 random runs (light) and their mean (dark) saturates at $C(L)$ (dashed), as Theorem 7 predicts. **Right:** mistakes over 300 runs, mean 5.3, all below the Halving bound $C(L) \approx 12$.

AI USE STATEMENT

Generative AI tools were used in this work for the following purposes: writing and debugging experiment code, generating figures, drafting and editing parts of the text, and checking some of the proofs. The problem formulation, the main results, and the overall structure of the paper are the authors' own. Every proof was independently verified by the authors, and all AI-assisted text, code, and figures were reviewed and edited by us. The authors take full responsibility for the content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Krishna B. Athreya and Peter E. Ney. *Branching Processes*. Springer, 1972.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, 2009.

Rudolf Carnap and Yehoshua Bar-Hillel. An outline of a theory of semantic information. Technical Report 247, Research Laboratory of Electronics, MIT, 1952.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Bertrand S. Clarke and Andrew R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

Andrew Cropper, Sebastijan Dumančić, Richard Evans, and Stephen H. Muggleton. Inductive logic programming at 30. *Machine Learning*, 111(1):147–172, 2022.

Luc De Raedt and Angelika Kimmig. Probabilistic (logic) programming concepts. *Machine Learning*, 100:5–47, 2015.

Michael Drmota and Wojciech Szpankowski. Precise minimax redundancy and regret. *IEEE Transactions on Information Theory*, 50(11):2686–2707, 2004.

Richard Evans and Edward Grefenstette. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61:1–64, 2018.

Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

Luciano Floridi. *The Philosophy of Information*. Oxford University Press, 2011.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

Philippe Jacquet and Wojciech Szpankowski. Markov types and minimax redundancy for Markov sources. *IEEE Transactions on Information Theory*, 50(7):1393–1402, 2004.

Richard M. Karp. The transitive closure of a random digraph. *Random Structures & Algorithms*, 1 (1):73–93, 1990.

Daniel J. Kleitman and Bruce L. Rothschild. The number of finite topologies. *Proceedings of the American Mathematical Society*, 25(2):276–282, 1970.

Daniel J. Kleitman and Bruce L. Rothschild. Asymptotic enumeration of partial orders on a finite set. *Transactions of the American Mathematical Society*, 205:205–220, 1975.

Luis A. Lastras, Jonathan Lenchner, Barry M. Trager, Wojciech Szpankowski, Mark S. Squillante, Chai Wah Wu, Ronald Fagin, and Alexander Gray. Fundamental limits incorporating logical reasoning into Shannon's information theory. *Proceedings of the National Academy of Sciences (PNAS)*, 123, 2026.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988.

John W. Lloyd. *Foundations of Logic Programming*. Springer, 2 edition, 1987.

Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. DeepProbLog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

George A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

Stephen Muggleton. Inductive logic programming. *New Generation Computing*, 8(4):295–318, 1991.

Stephen Muggleton and Luc De Raedt. Inductive logic programming: Theory and methods. *Journal of Logic Programming*, 19–20:629–679, 1994.

Asaf Nachmias and Yuval Peres. The critical random graph, with martingales. *Israel Journal of Mathematics*, 176:29–41, 2010.

John F. Padgett and Christopher K. Ansell. Robust action and the rise of the Medici, 1400–1434. *American Journal of Sociology*, 98(6):1259–1319, 1993.

Jorma Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, 1996.

Taisuke Sato. A statistical learning method for logic programs with distribution semantics. In *Proceedings of the 12th International Conference on Logic Programming (ICLP)*, pages 715–729. MIT Press, 1995.

Yuri M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3–17, 1987.

Wojciech Szpankowski. *Average Case Analysis of Algorithms on Sequences*. Wiley, 2001.

Remco van der Hofstad. *Random Graphs and Complex Networks*, volume 1. Cambridge University Press, 2017.

Maarten H. van Emden and Robert A. Kowalski. The semantics of predicate logic as a programming language. *Journal of the ACM*, 23(4):733–742, 1976.

Changlong Wu, Ananth Grama, and Wojciech Szpankowski. Online universal learning from information-theoretic perspective. *Foundations and Trends in Communications and Information Theory*, 23:225–443, 2026.

A FULL PROOFS

This appendix gives the complete arguments for the results whose proofs were sketched in the body. Throughout, $\mu(S) = \mathrm{lhm}(\mathrm{BK} \cup S)$, $\Omega_L = \mathrm{im}\,\mu$, and for the bandwidth family $a_n = |\Omega_L(n, d)|$.

A.1 REALIZABILITY (PROPOSITION 1)

Proof. Let the template consist of edge clauses $r_v \leftarrow r_u$ over $G = (V, E)$ with source set V_0 declared by BK.

(⇒). Suppose $R = \mu(S) = \mathrm{lhm}(\mathrm{BK} \cup S)$ for some $S \subseteq L$, and let $v \in R \setminus V_0$. Since v is not a background source, it entered the least model at some finite stage: $v \in T_P{\uparrow}(k+1) \setminus T_P{\uparrow}k$ for some $k \geq 0$, where $P = \mathrm{BK} \cup S$. By definition of T_P, v was produced by firing a clause $r_v \leftarrow r_u \in S$ whose body r_u already held, i.e. $u \in T_P{\uparrow}k \subseteq R$ and $u \to v \in E$. Thus v has an in-neighbour $u \in R$ that entered strictly earlier; iterating on u terminates at V_0 and exhibits a path from V_0 to v lying entirely in R, the stated reachability condition. (When G is acyclic, "every v has an in-neighbour in R" already forces this, since backward walks cannot cycle.)

(⇐). Suppose $V_0 \subseteq R$ and every $v \in R$ is reachable from V_0 inside R. Take $S := \{(u \to v) \in E : u, v \in R\}$, the edges internal to R. First, $\mathrm{lhm}(\mathrm{BK} \cup S) \subseteq R$: every clause of S has its head in R by construction, and $V_0 \subseteq R$, so no atom outside R is ever produced. Second, $R \subseteq \mathrm{lhm}(\mathrm{BK} \cup S)$: define the within-$R$ distance $\delta(v)$ of $v \in R$ from V_0 along paths lying in R (finite by hypothesis). Induct on δ. If $\delta(v) = 0$ then $v \in V_0$, so v holds. If $\delta(v) = \ell + 1$, pick an in-R-neighbour u with $\delta(u) = \ell$; by induction $u \in \mathrm{lhm}(\mathrm{BK} \cup S)$, and $(u \to v) \in S$, so firing it adds v. Hence $R = \mathrm{lhm}(\mathrm{BK} \cup S)$ and $R \in \Omega_L$. □

A.2 EXACT COUNT AND ASYMPTOTICS (THEOREM 1)

Proof. Generating function. The compositions with parts in $\{1, \ldots, d\}$ have the ordinary generating function

$$\sum_{x \geq 0} C_d(x)\, t^x = \frac{1}{1 - (t + t^2 + \cdots + t^d)} = \frac{1 - t}{1 - 2t + t^{d+1}},$$

the last step by $t + \cdots + t^d = t(1 - t^d)/(1 - t)$ and simplification. Hence $a_n = \sum_{x=0}^{n} C_d(x)$ is the partial-sum sequence, and the partial-sum operator multiplies the generating function by $(1 - t)^{-1}$:

$$\sum_{n \geq 0} a_n t^n = \frac{1}{1 - t} \cdot \frac{1 - t}{1 - 2t + t^{d+1}} = \frac{1}{1 - 2t + t^{d+1}}.$$

Clearing denominators, $\sum_n a_n t^n = (1 - 2t + t^{d+1})^{-1}$, and reading off the coefficient of t^n for $n \geq d + 1$ gives $a_n - 2a_{n-1} + a_{n-d-1} = 0$, i.e. the stated recurrence; for $k \leq d$ every subset of $\{1, \ldots, k\}$ has all gaps $\leq d$, so $a_k = 2^k$.

Dominant root. Write $Q(t) = 1 - 2t + t^{d+1}$. Substituting $\rho = 1/t$ and multiplying by ρ^{d+1} gives $\rho^{d+1} - 2\rho^d + 1 = 0$, which factors as

$$\rho^{d+1} - 2\rho^d + 1 = (\rho - 1)\big(\rho^d - \rho^{d-1} - \cdots - \rho - 1\big).$$

The factor $\rho - 1$ contributes the root $\rho = 1$ (equivalently $t = 1$); the remaining roots solve the *d-bonacci equation* $\rho^d = \rho^{d-1} + \cdots + \rho + 1$, whose unique root in $(1, 2)$ is the d-bonacci constant ρ_d. By the Perron–Frobenius theory of the companion matrix of this positive recurrence, ρ_d strictly dominates all other roots in modulus, and $\rho_d > 1$ dominates the spurious root at 1. Hence $t_d := 1/\rho_d \in (0, \frac{1}{2})$ is the unique pole of $\sum a_n t^n$ of smallest modulus, and it is simple. *Singularity analysis.* For a rational generating function with a simple dominant pole at t_d, $a_n = \dfrac{-1}{Q'(t_d)\, t_d}\, \rho_d^n \big(1 + O(\gamma_d^n)\big)$ with $Q'(t) = -2 + (d+1)t^d$ and $\gamma_d < 1$ the ratio of the smallest to the next-smallest singularity modulus (Flajolet and Sedgewick, 2009, Ch. IV). This gives $A_d = \big((2 - (d+1)t_d^d)\, t_d\big)^{-1}$. □

A.3 NONLINEAR CAPACITY IS QUADRATIC WITH CONSTANT $\kappa = \frac{1}{4}$ (THEOREM 2)

Proof. Fix m constants and the composition rule $\mathrm{tc}(X, Z) \leftarrow \mathrm{tc}(X, Y), \mathrm{tc}(Y, Z)$ as background; the switchable clauses are the base facts. A hypothesis F selects a set of base facts, and its content

is $\mu(F) = \mathrm{TC}(F)$, the transitive closure. Since every transitive relation is its own closure, Ω_L is exactly the set of realizable transitive relations. We first treat the *ordered* template (base facts (i, j) with $i < j$, the acyclic/DAG language bias), for which a content is a transitively closed subrelation of the ground order $\{(i, j) : i < j\}$; the unordered variant is handled at the end. Write $T(m) = |\Omega_L|$.

Lower bound $T(m) \geq 2^{\lfloor m^2/4 \rfloor}$. Split $[m] = A \sqcup B$ with $A = \{1, \ldots, \lfloor m/2 \rfloor\}$ and $B = \{\lfloor m/2 \rfloor + 1, \ldots, m\}$, and let R be an arbitrary subset of $A \times B$. Every pair $(a, b) \in R$ has $a \in A$, $b \in B$, hence $a < b$, so R respects the ground order. R is transitive: a violation would need $(x, y), (y, z) \in R$ with $(x, z) \notin R$, but then y is at once a head (so $y \in B$) and a tail (so $y \in A$), which is impossible. Hence every such R is a realizable content, and distinct subsets give distinct relations, so $T(m) \geq 2^{|A|\,|B|} = 2^{\lfloor m/2 \rfloor \lceil m/2 \rceil} = 2^{\lfloor m^2/4 \rfloor}$.

Upper bound $T(m) \leq 2^{\,m^2/4 + o(m^2)}$. Each content R is a *strict partial order* on $[m]$: it is transitive by membership in Ω_L; it is irreflexive and antisymmetric because $R \subseteq \{(i, j) : i < j\}$ forbids every (i, i) and forbids (i, j) and (j, i) together. Distinct contents are therefore distinct labeled strict partial orders, so $T(m) \leq P(m)$, the number of partial orders on m labeled points. By the Kleitman–Rothschild enumeration (Kleitman and Rothschild, 1975), $\log_2 P(m) = \frac{m^2}{4} + o(m^2)$: almost every labeled poset is a three-layer order with layer sizes $\frac{m}{4}, \frac{m}{2}, \frac{m}{4}$ and about $\frac{3}{16}m^2$ comparable pairs, and these dominate the count. Hence $\log_2 T(m) \leq \frac{1}{4}m^2 + o(m^2)$.

Conclusion (ordered case). Combining, $\lfloor m^2/4 \rfloor \leq \log_2 T(m) \leq \frac{1}{4}m^2 + o(m^2)$; dividing by m^2 and letting $m \to \infty$ gives $\kappa = \lim_m \log_2 T(m)/m^2 = \frac{1}{4}$, and $C(L) = \log_2 T(m) = \frac{1}{4}m^2(1 + o(1))$.

Unordered variant. If the base facts are all ordered pairs (i, j) with $i \neq j$ (both directions switchable), then Ω_L is the set of *all* transitive relations on $[m]$; write $\mathrm{Tr}(m) = |\Omega_L|$. The same bipartite family still lies in Ω_L, so $\mathrm{Tr}(m) \geq 2^{\lfloor m^2/4 \rfloor}$. For the upper bound, send a transitive relation R to the preorder $\bar{R} := R \cup \Delta$, where $\Delta = \{(i, i) : i \in [m]\}$. Then $\bar{R}$ is reflexive and still transitive, so $\bar{R}$ is a preorder. This map is at most 2^m-to-one, because R and $\bar{R}$ coincide off the diagonal. Hence $\mathrm{Tr}(m) \leq 2^m\, Q(m)$, where $Q(m)$ is the number of preorders on $[m]$ — equivalently the number of finite topologies — for which $\log_2 Q(m) = \frac{m^2}{4} + o(m^2)$ (Kleitman and Rothschild, 1970). Since $2^m = 2^{o(m^2)}$, this gives $\log_2 \mathrm{Tr}(m) \leq \frac{1}{4}m^2 + o(m^2)$, so $\kappa = \frac{1}{4}$ holds. □

A.4 Phase transition of semantic content (Theorem 3)

Proof. Write $p = c/m$ and fix the source 0. By vertex-symmetry of $\mathcal{D}(m, p)$, $\mathbb{E}|R| = \sum_i \mathbb{E}|\{j : i \rightsquigarrow j\}| = m\,\mathbb{E}|R_0|$, the diagonal $i = j$ contributing only $O(m) = o(m^2)$; so it suffices to analyse the forward reachable set $R_0 = \{v : 0 \rightsquigarrow v\}$.

Step 1 (exploration process). Discover R_0 by breadth-first search: colour 0 active and all others neutral; repeatedly take an active vertex u, reveal its arcs to the still-neutral vertices (each present independently with probability p), recolour every newly hit vertex active, and retire u. Let D_t be the number discovered after t retirements ($D_0 = 1$). The number of children of the t-th retired vertex is $\xi_t \sim \mathrm{Binomial}(m - D_t,\, p)$, independent across t given the history.

Step 2 (branching coupling). Since $m - D_t \leq m - 1$, each ξ_t is stochastically dominated by $\mathrm{Binomial}(m - 1, p)$, and while $D_t = o(m)$ it also dominates $\mathrm{Binomial}(m - 1 - o(m), p)$; both bounds converge in law to $\mathrm{Poisson}(c)$ as $m \to \infty$. A standard coupling (van der Hofstad, 2017; Athreya and Ney, 1972) matches the exploration, up to its first εm discoveries, with a Galton–Watson process T_c of offspring law $\mathrm{Poisson}(c)$.

Step 3 (subcritical $c < 1$). For T_c with mean offspring $c < 1$, generation k has expected size c^k, so the expected total progeny is $\sum_{k \geq 0} c^k = 1/(1 - c) < \infty$; extinction is a.s. with exponentially decaying total-progeny tails. The exploration is dominated above by a $\mathrm{Binomial}(m - 1, p)$ branching process, itself subcritical, so $\{|R_0|\}_m$ is uniformly integrable; combined with $|R_0| \xrightarrow{d} |T_c|$ this gives $\mathbb{E}|R_0| \to \frac{1}{1-c}$, hence $\mathbb{E}|R| = m\,\mathbb{E}|R_0| = \Theta(m)$.

Step 4 (supercritical $c > 1$). The extinction probability q of T_c is the smallest root of $q = e^{-c(1-q)}$; writing $\zeta := 1 - q$ recasts this as $1 - \zeta = e^{-c\zeta}$, whose unique positive root exists exactly for $c > 1$. The exploration splits along the two fates of T_c: with probability $\to q = 1 - \zeta$ it dies after $O_P(1)$

discoveries; with probability $\to \zeta$ it survives and, by Karp's analysis of $\mathcal{D}(m, c/m)$ (Karp, 1990), fills the giant out-component, whose relative size equals the survival probability, so $|R_0| = \zeta m(1 + o(1))$ w.h.p. Taking expectations, $\mathbb{E}|R_0| = \zeta \cdot \zeta m(1 + o(1)) + (1 - \zeta) \cdot O(1) = \zeta^2 m(1 + o(1))$, hence $\mathbb{E}|R| = \zeta^2 m^2(1 + o(1))$.

Step 5 (the critical point $c = 1$). Because the exploration of Step 1 only ever reveals arcs into still-neutral vertices, $|R_0|$ has exactly the law of the component of a vertex in the undirected graph $G(m, p)$, so $\mathbb{E}|R_0|$ is the Erdős–Rényi susceptibility. At $p = 1/m$ the limiting Galton–Watson process is critical ($\zeta(1) = 0$), with total-progeny tail $\Pr[|T_1| \geq k] = \Theta(k^{-1/2})$. The exploration inherits this tail only up to $k \asymp m^{2/3}$: after D_t discoveries the mean offspring is $1 - D_t/m$, and this drift becomes visible exactly at $D_t \asymp m^{2/3}$, so $\Pr[|R_0| \geq k] = \Theta(k^{-1/2})$ for $k \leq m^{2/3}$ (van der Hofstad, 2017, Ch. 5), while beyond that scale $\Pr[|R_0| > A\, m^{2/3}] \leq 4m^{-1/3} e^{-A^2(A-4)/32}$ for $A > 8$ (Nachmias and Peres, 2010, Thm. 1). Summing the tail,

$$\mathbb{E}|R_0| = \sum_{k \geq 1} \Pr[|R_0| \geq k] = \sum_{k \leq m^{2/3}} \Theta(k^{-1/2}) \; + \; O\big(m^{2/3} \cdot m^{-1/3}\big) = \Theta(m^{1/3}),$$

the well-known $\Theta(m^{1/3})$ critical susceptibility. Hence $\mathbb{E}|R_0| = \Theta(m^{1/3}) = o(m)$ and $\mathbb{E}|R| = m\,\mathbb{E}|R_0| = \Theta(m^{4/3}) = o(m^2)$: the order of $\mathbb{E}|R|$ passes from $\Theta(m)$ through $\Theta(m^{4/3})$ to $\Theta(m^2)$ as c crosses 1, and the transition is sharp. □

A.5 Proof of Corollary 2

Proof. Fix d and $\theta < 1$. On the bandwidth-d template the atoms are $r_0, \ldots, r_n$ with source r_0, and the only clauses into r_j are the $\leq d$ edges $r_j \leftarrow r_i$ with $1 \leq j - i \leq d$, each present independently with probability θ. Whether r_j is reachable from the source therefore depends on the history only through the reachability indicators of the preceding d atoms and the (independent) presence of those $\leq d$ edges.

A finite-window Markov chain. Encode the state at position j as the reachability pattern of the length-d window ending at j,

$$W_j \;=\; \big(\mathbf{1}[r_{j-d+1} \text{ reachable}], \ldots, \mathbf{1}[r_j \text{ reachable}]\big) \in \{0,1\}^d.$$

Sliding the window from j to $j + 1$ drops the oldest coordinate and appends $\mathbf{1}[r_{j+1}$ reachable$]$, and r_{j+1} is reachable iff at least one of the reachable atoms among $r_{j+1-d}, \ldots, r_j$ has its edge to r_{j+1} switched on. Given W_j, this new bit is an independent Bernoulli whose parameter depends only on W_j and θ. Hence $(W_j)_{j \geq 0}$ is a time-homogeneous Markov chain on the 2^d window-states, with transition matrix $M = M(d, \theta)$.

The all-zero window absorbs. Let $\mathbf{0} = (0, \ldots, 0)$ be the all-unreachable window. If $W_j = \mathbf{0}$ then none of the atoms feeding r_{j+1} is reachable, so r_{j+1} is unreachable and $W_{j+1} = \mathbf{0}$: the state $\mathbf{0}$ is absorbing. Moreover $\mathbf{0}$ is reachable from every state with positive probability, since from any window the next d edges are all absent with probability $(1 - \theta)^{\leq d} \geq (1 - \theta)^d > 0$, which switches off every incoming edge for d consecutive steps and empties the window. Thus every non-absorbing state is transient, and the substochastic matrix $\widehat{M}$ obtained by restricting M to the non-zero states has spectral radius $\rho(\widehat{M}) < 1$ (Perron–Frobenius, since $\mathbf{0}$ is reached from every state with positive probability).

Exponential tail and the bound. Let $Z = |R_0| - 1 = \sum_{j \geq 1} \mathbf{1}[r_j \text{ reachable}]$ count the reachable atoms past the source. A reachable atom at position j requires $W_j \neq \mathbf{0}$, so

$$\Pr[r_j \text{ reachable}] \;\leq\; \Pr[W_j \neq \mathbf{0}] \;\leq\; \big\|\widehat{M}^{\,j}\big\|_1 \;\leq\; C_d\, \rho(\widehat{M})^{\,j},$$

for a constant C_d depending only on d. Summing the geometric series,

$$\mathbb{E}\,|R_0| \;=\; 1 + \sum_{j \geq 1} \Pr[r_j \text{ reachable}] \;\leq\; 1 + \frac{C_d}{1 - \rho(\widehat{M})} \;=\; O_{d,\theta}(1),$$

uniformly in n. In particular $|R_0|$ is a.s. finite and has exponentially decaying tails. Since $\rho(\widehat{M}) < 1$ for *every* $\theta < 1$, there is no $\theta < 1$ at which the mean diverges: the bandwidth-d process is permanently subcritical, with no critical point. □

A.6 THE REALIZABLE MISTAKE BOUND AND THE LITTLESTONE SANDWICH: THEOREM 4

We prove Theorem 4: Halving over Ω_L makes at most $C(L)$ mistakes, the exact mistake-bound value is $\mathrm{Ldim}(\Omega_L)$, and $\mathrm{VC}(\Omega_L) \le \mathrm{Ldim}(\Omega_L) \le \log_2 |\Omega_L| = C(L)$. All three statements hold for any finite concept class; we state them for Ω_L.

Proof of Theorem 4. Halving mistake bound. In the realizable setting the target $R^* \in \Omega_L$ labels every atom by $y = \mathbf{1}[\gamma \in R^*]$. Halving maintains the version space $V \subseteq \Omega_L$ of contents consistent with all labels seen and predicts on a queried atom γ by majority vote, $\hat{y} = \mathbf{1}[\,|\{R \in V : \gamma \in R\}| \ge |V|/2\,]$. On a mistake the true label disagrees with the majority, so at least half of V is inconsistent with that label and is pruned; hence $|V|$ at least halves on every mistake. Since R^* is never pruned, $|V| \ge 1$ throughout, and $|V| = |\Omega_L|$ initially, so the number of mistakes is at most $\log_2 |\Omega_L| = C(L)$.

Ldim *is the exact mistake bound.* The optimal (adversarial) mistake bound of the realizable online problem equals the Littlestone dimension $\mathrm{Ldim}(\Omega_L)$ (Littlestone, 1988): an adversary drives a mistake tree of depth $\mathrm{Ldim}(\Omega_L)$ and no learner can be forced to err more often.

The sandwich $\mathrm{VC}(\Omega_L) \le \mathrm{Ldim}(\Omega_L) \le \log_2 |\Omega_L|$. For the upper bound, Halving witnesses a learner with mistake bound $\log_2 |\Omega_L|$, and $\mathrm{Ldim}(\Omega_L)$ is the *minimal* achievable mistake bound, so $\mathrm{Ldim}(\Omega_L) \le \log_2 |\Omega_L|$. For the lower bound, let $T \subseteq \mathcal{B}$ be shattered by Ω_L, $|T| = \mathrm{VC}(\Omega_L)$. Ordering T arbitrarily and branching on the two labels at each atom yields a complete binary mistake tree of depth $|T|$: every root-to-leaf path is a label pattern on T, and shattering supplies a content realizing it, so the tree is Ω_L-realizable. Hence $\mathrm{Ldim}(\Omega_L) \ge \mathrm{VC}(\Omega_L)$ (Littlestone, 1988). □

A.7 THE AGNOSTIC MINIMAX RATE: PROOF OF THEOREM 5

We prove the two-sided bound of Theorem 5: for every finite content class Ω_L and every $T \ge \mathrm{Ldim}(\Omega_L)$,

$$c_1 \sqrt{\mathrm{Ldim}(\Omega_L)\, T} \;\le\; \min_{\text{learner}} \max_{y_{1:T}} \mathbb{E}\,\mathrm{Regret}_T^{0/1} \;\le\; \sqrt{\tfrac{\ln 2}{2}\, C(L)\, T}, \qquad c_1 = \tfrac{1}{4}.$$

Throughout, each content $R \in \Omega_L$ acts as the expert predicting $h_R(\gamma_t) = \mathbf{1}[\gamma_t \in R]$, the loss is $0/1$, and the comparator is the best content in hindsight.

Upper bound (Hedge over the content-experts). Run the exponential-weights forecaster over the $M = |\Omega_L|$ content-experts with learning rate $\eta > 0$: maintain weights $w_t(R) \propto \exp(-\eta L_{t-1}(R))$, where $L_{t-1}(R)$ is the cumulative $0/1$ loss of R through round $t-1$, and predict $\hat{y}_t = 1$ with probability $\sum_R w_t(R)\, h_R(\gamma_t)$. With losses in $[0,1]$, the standard exponential-weights inequality (Cesa-Bianchi and Lugosi, 2006, Thm. 2.2) gives, against the best expert $R^\star$,

$$\mathbb{E}\,\mathrm{Regret}_T^{0/1} \;\le\; \frac{\ln M}{\eta} \;+\; \frac{\eta T}{8}.$$

Optimizing $\eta = \sqrt{8 \ln M / T}$ makes the two terms equal and yields $\mathbb{E}\,\mathrm{Regret}_T^{0/1} \le \sqrt{(T/2)\ln M}$. Since $M = |\Omega_L|$ and $\ln M = (\ln 2)\log_2 |\Omega_L| = (\ln 2) C(L)$, this is $\sqrt{(\ln 2/2)\, C(L)\, T}$, holding for *any* finite class regardless of its Littlestone dimension. (A doubling trick over η removes the known-horizon assumption at the cost of a constant.)

Lower bound (a shattered mistake tree, played in phases). Let $d = \mathrm{Ldim}(\Omega_L)$ and let $\mathcal{T}$ be a complete binary mistake tree of depth d shattered by Ω_L: to each internal node v is attached a query atom $\gamma_v \in \mathcal{B}$, and for every root-to-leaf path there is a content $R \in \Omega_L$ whose labels on the path's d atoms are exactly the branch labels taken along the path. Assume $T \ge d$ and set $L = \lfloor T/d \rfloor \ge 1$; we use the first $dL \le T$ rounds and on the remaining rounds present an arbitrary atom with fair-coin labels, which contributes zero expected regret.

We describe a *randomized* adversary; since the minimax regret dominates the value against any fixed adversary (Yao), the resulting expected regret is a lower bound. Play in d phases, starting at the root v_1. In phase i we sit at a node v_i and, for L consecutive rounds, present the *same* atom γ_{v_i} with i.i.d. labels $y \sim \mathrm{Unif}\{0,1\}$. Let $\mathrm{maj}_i \in \{0,1\}$ be the majority label of phase i; move to the child of v_i

selected by maj_i, obtaining v_{i+1}. This traces one root-to-leaf path $v_1, \dots, v_d$ and a node-labelling $(\mathrm{maj}_1, \dots, \mathrm{maj}_d)$; by shattering there is a content $R^\star \in \Omega_L$ with $h_{R^\star}(\gamma_{v_i}) = \mathrm{maj}_i$ for every i.

Learner's loss. Within phase i the atom γ_{v_i} is fixed and the L labels are fresh fair coins independent of everything the learner has observed. Hence every prediction is correct with probability exactly $\frac{1}{2}$, and the learner's expected loss in phase i is $L/2$, for a total of $\mathbb{E}[\text{learner}] = dL/2$.

Comparator's loss. The content $R^\star$ predicts maj_i throughout phase i, so its phase-i loss is the minority count $\frac{L}{2} - \frac{1}{2}|S_i|$, where $S_i = \sum_{s=1}^{L}(2y_{i,s} - 1)$ is a sum of L Rademacher signs. By Khintchine's inequality (sharp L^1 constant), $\mathbb{E}|S_i| \geq L^{1/2}/\sqrt{2}$, so

$$\mathbb{E}[R^\star] \;=\; \sum_{i=1}^{d}\Big(\tfrac{L}{2} - \tfrac{1}{2}\mathbb{E}|S_i|\Big) \;\leq\; \frac{dL}{2} - \frac{d}{2}\cdot\frac{\sqrt{L}}{\sqrt{2}}.$$

Regret. Subtracting, and using $\min_R L_T(R) \leq L_T(R^\star)$,

$$\begin{aligned}\min_{\text{learner}}\max_{y_{1:T}} \mathbb{E}\,\mathrm{Regret}_T^{0/1} \;&\geq\; \mathbb{E}[\text{learner}] - \mathbb{E}[R^\star] \;\geq\; \frac{d\sqrt{L}}{2\sqrt{2}} \\ &=\; \frac{1}{\sqrt{8}}\sqrt{d\,(dL)} \;\geq\; \frac{1}{\sqrt{8}}\sqrt{\tfrac{1}{2}\,d\,T} \;=\; \frac{1}{4}\sqrt{\mathrm{Ldim}(\Omega_L)\,T},\end{aligned}$$

using $dL \geq T/2$ for $T \geq d$. This is the claimed $c_1\sqrt{\mathrm{Ldim}(\Omega_L)\,T}$ with $c_1 = \frac{1}{4}$.

A.8 The bandwidth-d rate: VC, Littlestone: Corollary 3

We prove Corollary 3: for the bandwidth-d family $\mathrm{Ldim}(\Omega_L(n,d)) = \Theta(C(L))$, so both sides of Theorem 5 are $\Theta(\sqrt{C(L)\,T})$. We identify a content with its indicator on the atoms $r_1, \dots, r_n$; the source r_0 lies in every content and carries no shattering information, so we drop it and write $[n] = \{1, \dots, n\}$. Now, $R \in \Omega_L(n,d)$ iff $0 \in R$ and, listing $R = \{0 = x_0 < x_1 < \cdots < x_m\}$, every gap satisfies $x_\ell - x_{\ell-1} \leq d$.

Definition 2 (d-gap-free sets)**.** A set $T \subseteq [n]$ is *d-gap-free* if it contains no d consecutive integers, i.e. there is no t with $\{t, \dots, t+d-1\} \subseteq T$. Let $g(n,d) := \max\{|T| : T \subseteq [n] \text{ is } d\text{-gap-free}\}$.

Lemma 1 (VC dimension of the bandwidth family)**.** *For all $n \geq 1$, $d \geq 1$,*

$$g(n,d) \;\leq\; \mathrm{VC}(\Omega_L(n,d)) \;\leq\; g(n,d) + d, \qquad g(n,d) = n - \Big\lfloor \tfrac{n}{d} \Big\rfloor.$$

In particular $\mathrm{VC}(\Omega_L(n,d)) = \big(1 - \frac{1}{d}\big)n + O(1)$.

Proof. Every d-gap-free T is shattered. Fix such a T with $|T| = k$ and a target $\sigma \in \{0,1\}^T$; write $P = \sigma^{-1}(1)$. We build $R \in \Omega_L$ with $R \cap T = P$ and $R \cap \sigma^{-1}(0) = \emptyset$, free on $[n] \setminus T$. Start from $R_0 := \{0\} \cup P$, which already has the right labels on T, and repair realizability without touching $\sigma^{-1}(0)$. For $v \in R_0$ with $v > 0$, the window $I_v := \{v-d, \dots, v-1\} \cap \{0, \dots, n\}$ must contain an allowed atom: otherwise all d integers of I_v lie in $\sigma^{-1}(0) \subseteq T$, making $T \supseteq I_v$ a block of d consecutive integers, contradicting d-gap-freeness (when $v \leq d$ the atom $0 \in I_v$ is always allowed). Switch on the largest allowed $u < v$ and recurse on it; the index strictly decreases, so the recursion halts at 0. The resulting R has $0 \in R$ and every $v \in R \setminus \{0\}$ an in-R-neighbour within d, so $R \in \Omega_L$ and $R \cap T = P$. As σ was arbitrary, T is shattered; thus $\mathrm{VC}(\Omega_L(n,d)) \geq g(n,d)$.

A shattered T is d-gap-free up to a top block. Suppose $T \supseteq B := \{t, \dots, t+d-1\}$ and some $w \in T$ exceeds $t+d-1$. Take σ labelling all of B with 0 and w with 1. If $R \in \Omega_L$ realized σ then $w \in R$ while $B \cap R = \emptyset$; the elements of R below w form a chain $0 = y_0 < \cdots < y_s$ with all gaps $\leq d$, and $w - y_s \leq d$ forces $y_s \geq w - d \geq t$. Descending, a single step of size $\leq d$ must land the chain in $[t, t+d-1] = B$, exhibiting an element of $B \cap R = \emptyset$ — a contradiction. So σ is unrealizable and T is not shattered: any block a shattered T contains sits at its very top, contributing at most $d-1$ extra indices, whence $\mathrm{VC}(\Omega_L(n,d)) \leq g(n,d) + d$.

Evaluation of $g(n,d)$. Partitioning $[n]$ into $\lfloor n/d \rfloor$ length-d blocks (and a shorter tail), a d-gap-free set omits at least one integer per block, so $|T| \leq n - \lfloor n/d \rfloor$. Conversely $T^\star = [n] \setminus \{d, 2d, \dots, d\lfloor n/d \rfloor\}$ contains no d consecutive integers (every length-d window holds a multiple of d) and has size $n - \lfloor n/d \rfloor$. Hence $g(n,d) = n - \lfloor n/d \rfloor$. □

Proof of Corollary 3. The sandwich $\mathrm{VC}(\Omega_L) \le \mathrm{Ldim}(\Omega_L) \le \log_2 |\Omega_L| = C(L)$ of Theorem 4, applied to $\Omega_L = \Omega_L(n,d)$, has left side $\big(1-\frac{1}{d}\big)n+O(1)$ by Lemma 1 and right side $(\log_2 \rho_d)\, n+O(1)$ by Theorem 1. Both are linear in n with strictly positive slope ($1-\frac{1}{d} > 0$ and $\log_2 \rho_d > 0$ for $d \ge 2$), so $\mathrm{Ldim}(\Omega_L(n,d)) = \Theta(n) = \Theta(C(L))$. Substituting into Theorem 5, the lower bound $c_1\sqrt{\mathrm{Ldim}\ T}$ and the upper bound $\sqrt{(\ln 2/2)\, C(L)\, T}$ are both $\Theta(\sqrt{C(L)\, T})$, so the minimax agnostic $0/1$ regret is $\Theta(\sqrt{C(L)\, T})$. □

Remark 3 (The trapped constant). Corollary 3 settles the order but not $\lim_n \mathrm{Ldim}(\Omega_L(n,d))/n$, which is confined to $[\,1-\frac{1}{d},\ \log_2 \rho_d\,]$ ($= [0.5,\ 0.6942]$ for $d=2$). The interval is nonempty and strictly inside $(0,1)$, and its width $\log_2 \rho_d - (1-\frac{1}{d}) \to 0$ as $d \to \infty$, so the constant is asymptotically pinned for large reach.

A.9 Minimax log-loss regret is the Shtarkov sum (Theorem 6)

Proof. Fix the query sequence $\gamma_{1:T}$. For an outcome $y \in \{0,1\}^T$ write $M(y) = \max_{R\in\Omega_L} P_R(y)$ and $S = \sum_y M(y)$ (the Shtarkov sum). A strategy with sequence likelihood $\hat{P}$ suffers cumulative log loss $-\log_2 \hat{P}(y)$, while the best expert suffers $\min_R(-\log_2 P_R(y)) = -\log_2 M(y)$; so its regret on y is $\log_2\big(M(y)/\hat{P}(y)\big)$ and its worst-case regret is $\max_y \log_2\big(M(y)/\hat{P}(y)\big)$.

Achievability. The normalized-maximum-likelihood distribution $q^\star(y) = M(y)/S$ (Shtarkov, 1987) is a probability distribution on $\{0,1\}^T$ and makes the ratio constant, $M(y)/q^\star(y) \equiv S$, so its worst-case regret is exactly $\log_2 S$.

Optimality. For any $\hat{P}$ with $\sum_y \hat{P}(y) \le 1$, the mediant inequality gives $\max_y M(y)/\hat{P}(y) \ge \sum_y M(y)/\sum_y \hat{P}(y) \ge S$. Hence $\min_{\hat{P}} \max_y \log_2(M(y)/\hat{P}(y)) = \log_2 S$, attained uniquely by $q^\star$.

Evaluating S. Here each expert is deterministic: $P_R(y) = \prod_t \mathbf{1}[y_t = \mathbf{1}[\gamma_t \in R]] \in \{0,1\}$ is the indicator that y is R's label pattern on $\gamma_{1:T}$. Thus $M(y) \in \{0,1\}$ equals 1 iff y is some content's pattern, and $S = \sum_y M(y)$ counts the distinct patterns, $S = |\Omega_L|_{\gamma_{1:T}}|$. Restriction can only merge contents, so $S \le |\Omega_L|$, with equality once the queried atoms separate all contents; and by Sauer–Shelah $S \le \sum_{i\le \mathrm{VC}(\Omega_L)} \binom{T}{i} \le (T+1)^{\mathrm{VC}(\Omega_L)}$, giving $\mathrm{Regret}_T^\star \le \min\{C(L), \mathrm{VC}(\Omega_L)\log_2(T+1)\}$. The precise minimax evaluation is that of Drmota and Szpankowski (2004). □

A.10 Proof of Theorem 7

Proof of Theorem 7. Write $V_t = \{R \in \Omega_L :\ R \text{ consistent with } y_{<t}\}$ for the version space entering round t, so $V_1 = \Omega_L$ and $V_{t+1} = \{R \in V_t : \mathbf{1}[\gamma_t \in R] = y_t\}$. Since the outcome is realizable, the generating target $R^\star \in \Omega_L$ is consistent with every label, hence $R^\star \in V_t$ for all t; as $w(R^\star) > 0$ this gives $w(V_t) \ge w(R^\star) > 0$, so every prediction is well defined and every per-round loss is finite.

(i) Exact regret. The algorithm predicts $\hat{p}_t = w(\{R \in V_t : \gamma_t \in R\})/w(V_t)$ for label 1. If $y_t = 1$ then $V_{t+1} = \{R \in V_t : \gamma_t \in R\}$ and $\hat{p}_t(y_t) = \hat{p}_t = w(V_{t+1})/w(V_t)$; if $y_t = 0$ then $V_{t+1} = \{R \in V_t : \gamma_t \notin R\}$ and $\hat{p}_t(y_t) = 1 - \hat{p}_t = w(V_{t+1})/w(V_t)$. Either way

$$\hat{p}_t(y_t) = \frac{w(V_{t+1})}{w(V_t)}.$$

The sequential mass assigned to y therefore telescopes,

$$\widehat{P}(y) = \prod_{t=1}^{T} \hat{p}_t(y_t) = \prod_{t=1}^{T} \frac{w(V_{t+1})}{w(V_t)} = \frac{w(V_{T+1})}{w(V_1)} = \frac{w(V_{T+1})}{w(\Omega_L)}.$$

In the realizable case the best content incurs zero log-loss: any R consistent with the whole outcome (e.g. $R^\star$) predicts each observed label with probability 1, so $\min_{R\in\Omega_L} \sum_t \ell_t^{\log}(p_t^R, y_t) = 0$. Hence the cumulative regret on y is

$$\mathrm{Regret}_T = -\log_2 \widehat{P}(y) = \log_2 \frac{w(\Omega_L)}{w(V_{T+1})} \ \le\ \log_2 \frac{w(\Omega_L)}{w(R^\star)},$$

the inequality because $R^\star \in V_{T+1}$. Under $w \equiv 1$ this reads $\mathrm{Regret}_T = \log_2(|\Omega_L|/|V_{T+1}|) \le \log_2 |\Omega_L| = C(L)$, since $|V_{T+1}| \ge 1$.

(ii) Lower bound and optimality. Let $\hat{q} = (\hat{q}_t)$ be any sequential log-loss predictor. Its conditionals satisfy $\hat{q}_t(0 \mid y_{<t}) + \hat{q}_t(1 \mid y_{<t}) = 1$, so $\widehat{Q}(y) := \prod_t \hat{q}_t(y_t \mid y_{<t})$ defines a (sub-)probability on $\{0,1\}^T$ with $\sum_y \widehat{Q}(y) \le 1$. On a realizable outcome the best content again has zero loss, so the predictor's regret there is $-\log_2 \widehat{Q}(y)$, and the realizable outcomes are exactly the $S_T = |\Omega_L|_{\gamma_{1:T}}|$ distinct label patterns (Theorem 6). If its worst-case regret over realizable outcomes were $< \log_2 S_T$, then $\widehat{Q}(y) > 1/S_T$ for each of the S_T patterns; summing over them gives $\sum_{y \text{ pattern}} \widehat{Q}(y) > 1 \ge \sum_y \widehat{Q}(y)$, a contradiction. Hence every predictor has worst-case regret at least $\log_2 S_T = \mathrm{Regret}_T^\star$, the minimax value of Theorem 6.

Now compare Algorithm 1 (uniform w) to this floor. The map sending each $R \in \Omega_L$ to its pattern on $\gamma_{1:T}$ partitions Ω_L into the S_T nonempty fibres $\{V_{T+1}(y) : y \text{ realizable}\}$, so $\sum_{y \text{ pattern}} |V_{T+1}(y)| = |\Omega_L|$. When the queries separate the contents, $S_T = |\Omega_L|$ and every fibre is a singleton: $|V_{T+1}(y)| = 1$ for all realizable y, so by (i) the algorithm's regret is $\log_2 |\Omega_L| = \log_2 S_T = \mathrm{Regret}_T^\star$ on *every* outcome; its worst case meets the floor and it is minimax-optimal. In general (i) gives the data-dependent regret $\log_2(|\Omega_L|/|V_{T+1}(y)|)$, which differs from the constant minimax value $\log_2 S_T$ by

$$\log_2 \frac{|\Omega_L|}{S_T\,|V_{T+1}(y)|},$$

positive on small fibres and negative on large ones (the fibre sizes average $|\Omega_L|/S_T$). Thus Algorithm 1 is anytime and horizon-free but equalizes the regret across outcomes only when it separates the contents; the constant value $\log_2 S_T$ on every outcome is attained by the horizon-dependent NML strategy of Theorem 6 alone. □

The only per-step primitive is a weighted count over the version space. For the bandwidth-d family this is a single $2^d \times 2^d$ transfer-matrix pass, $O(2^d n)$ per step; on a tree (Appendix C.2) it is one $O(n)$ dynamic program, so the mixture over $|\Omega_L|$ contents is evaluated without ever enumerating Ω_L. A non-uniform prior w changes the guarantee to $\log_2\big(w(\Omega_L)/w(R^\star)\big)$, the code length of the true theory under w: a good language bias buys smaller redundancy, exactly as in universal coding (Grünwald, 2007).

B Algorithm for the Mismatch Loss

Algorithm 1 prunes: on each label it discards the contents inconsistent with it. Pruning is correct in the realizable log-loss setting, where the target is never discarded, but it is wrong for the agnostic mismatch loss, where a content that errs once may still be the best comparator in hindsight. The fix is to reweight rather than prune: each content carries a weight that decays geometrically in its accumulated mistakes, so a single error is penalised but not fatal. This is the exponential-weights forecaster over the content class Ω_L, made explicit with a prediction rule and an efficient implementation. The only difference from Algorithm 1 is the last line: Algorithm 1 sets the weight of an inconsistent content to 0 (pruning), whereas Algorithm 2 multiplies it by β (discounting); taking $\beta \to 0$ with the majority-vote prediction recovers Algorithm 1 and, exactly, Halving.

Theorem 8 (Mismatch guarantee)**.** *Run Algorithm 2 over Ω_L, and write $M = |\Omega_L|$.*

(i) (Realizable limit.) *With $\beta \to 0$ and majority-vote prediction it is Halving, so in the realizable case it makes at most $\log_2 M = C(L)$ mistakes.*

(ii) (Agnostic regret.) *With the randomized prediction and $\eta = \sqrt{8 \ln M/T}$, against an arbitrary label sequence,*

$$\mathbb{E}\Big[\mathrm{Regret}_T^{0/1}\Big] = \mathbb{E}\Big[\textstyle\sum_t \mathbf{1}[\hat{y}_t \ne y_t]\Big] - \min_{R \in \Omega_L} \sum_t \mathbf{1}[h_R(\gamma_t) \ne y_t] \;\le\; \sqrt{\tfrac{\ln 2}{2}\, C(L)\, T}.$$

Proof. Each content R is the expert predicting $h_R(\gamma_t) = \mathbf{1}[\gamma_t \in R]$, with 0/1 loss $\ell_t(R) = \mathbf{1}[h_R(\gamma_t) \ne y_t]$ and mistake count $m_t(R) = \sum_{s<t} \ell_s(R)$. The update multiplies by $\beta = e^{-\eta}$ on each mistake, so $w_t(R) = \beta^{m_t(R)} = e^{-\eta\, m_t(R)}$; write $W_t = \sum_R w_t(R)$ and $p_t = s_1/W_t$.

Algorithm 2: Exponential-weights deductive predictor (mismatch loss)

Input: content class Ω_L; query stream $\gamma_1, \gamma_2, \ldots$; discount $\beta = e^{-\eta} \in [0, 1)$ (default $\eta = \sqrt{8 \ln |\Omega_L| / T}$; $\beta \to 0$ gives Halving)

$w(R) \leftarrow 1$ for all $R \in \Omega_L$ `// weight` $= \beta^{\#\texttt{mistakes so far}}$

for $t = 1, 2, \ldots$ **do**

receive query γ_t

$s_1 \leftarrow \sum_{R \in \Omega_L : \gamma_t \in R} w(R), \quad s_0 \leftarrow \sum_{R \in \Omega_L : \gamma_t \notin R} w(R)$ `// weighted vote for/against entailment`

predict $\hat{y}_t \leftarrow \mathbf{1}[s_1 \geq s_0]$ (deterministic), or $\hat{y}_t \leftarrow 1$ w.p. $s_1/(s_1 + s_0)$ (randomized)

observe $y_t = \mathbf{1}[\gamma_t \in R^*]$

$w(R) \leftarrow \beta\, w(R)$ for every R with $\mathbf{1}[\gamma_t \in R] \neq y_t$ `// discount inconsistent contents`

(i) With $\beta = 0$ a content that errs is zeroed permanently, so at every round the positive-weight set is exactly the version space of contents consistent with the labels seen, and the rule $\mathbf{1}[s_1 \geq s_0]$ is the majority vote over it: this is Halving over Ω_L. The target is never removed, so $W_t > 0$ throughout, and each mistake eliminates at least half the version space; hence at most $\log_2 M = C(L)$ mistakes.

(ii) Loss identity. For the randomized rule, if $y_t = 1$ the prediction errs with probability s_0/W_t, and if $y_t = 0$ with probability s_1/W_t; in both cases this equals $\widehat{\ell}_t := \sum_R w_t(R)\, \ell_t(R)/W_t$, the weight-average expert loss, since the erring contents are precisely $\{R : \gamma_t \notin R\}$ when $y_t = 1$ and $\{R : \gamma_t \in R\}$ when $y_t = 0$. *Potential argument.* By Hoeffding's lemma ($\ell_t \in [0, 1]$),

$$\frac{W_{t+1}}{W_t} = \mathbb{E}_{R \sim p_t}\big[e^{-\eta\, \ell_t(R)}\big] \leq \exp\big(-\eta\, \widehat{\ell}_t + \eta^2/8\big).$$

Taking logarithms, summing over t, and using $W_1 = M$ together with $W_{T+1} \geq w_{T+1}(R) = e^{-\eta\, L_T(R)}$ for any fixed R (where $L_T(R) = \sum_t \ell_t(R)$) gives $\sum_t \widehat{\ell}_t - \min_R L_T(R) \leq \ln M/\eta + \eta T/8$. The left side is $\mathbb{E}[\mathrm{Regret}_T^{0/1}]$ by the loss identity. Optimising $\eta = \sqrt{8 \ln M/T}$ yields $\sqrt{(T/2) \ln M} = \sqrt{(\ln 2/2)\, C(L)\, T}$, using $\ln M = (\ln 2) C(L)$. (A time-varying $\eta_t = \sqrt{8 \ln M/t}$, or a doubling trick, removes the dependence on the horizon T at the cost of an absolute constant.) □

C Further empirical results on Real Data

This appendix supplements the Florentine experiment of Section 6 (Figure 2): Appendix C.1 probes its robustness and the mechanism behind the exact regret, and Appendix C.2 carries the same two laws to a real taxonomy three orders of magnitude larger, and Appendix C.3 compares Algorithm 1 with the same learner run over clause-sets. Neither restates the collapse count or the exact-regret result already reported in Section 6. All figures are reproduced by the accompanying code.

We first spell out the two Florentine experiments: (i) the *semantic collapse* of syntactic rule-sets onto the content class Ω_L, and (ii) the behaviour of the online predictor, Algorithm 1.

Data. Padgett's Florentine families (Padgett and Ansell, 1993) is a real 15-node network whose edges are marriage ties between Renaissance families; it is a standard benchmark in social-network analysis. We use it because our running example — influence spreading through a network of ties — is exactly this kind of object.

Template. Fix one family as the *seed* (default: Medici). The single rule schema

$$\mathsf{marked}(Y) \;\leftarrow\; \mathsf{marked}(X),\ \mathsf{tie}(X, Y)$$

is instantiated over both directions of every marriage tie, giving $N = 2 \times 20 = 40$ ground clauses. Here $\mathsf{marked}(X)$ is the unary predicate carrying the spread: $\mathsf{marked}(f) = 1$ means family f has been reached from the seed, i.e. f lies in the current content R; it is the domain-neutral reading of the cheat predicate of Example 1, with the marriage tie $\mathsf{tie}(X, Y)$ as fixed background (the analogue of friend). A *hypothesis* is a subset S of these 40 clauses — a choice of which ties actually conduct the marking. There are $2^N = 2^{40} \approx 1.1 \times 10^{12}$ hypotheses.

Meaning. A hypothesis S does not stand on its own: its *meaning* is the set of families it ends up marking, i.e. the least Herbrand model $\mu(S) = \mathrm{lhm}(\mathrm{BK} \cup S)$ obtained by starting from the seed and firing the switched-on ties to a fixpoint. This is a set of *atoms* (marked families), not a set of clauses. The object of interest is the set of all meanings the template can produce,

$$\Omega_L \;=\; \{\, \mu(S) : S \subseteq L \,\} \;\subseteq\; 2^{\{15 \text{ families}\}}.$$

Which sets are meanings? Not every set of families is realizable. By the realizability characterization (Proposition 1), for a symmetric tie relation $R \in \Omega_L$ if and only if R contains the seed and is *connected* in the network. Enumerating these gives exactly $|\Omega_L| = 4{,}104$ meanings.

Why the map is many-to-one (a worked case). Take the meaning $R = \{\text{Medici}, \text{Salviati}, \text{Pazzi}\}$, a path in the network. Every hypothesis that (a) switches on the forward ties Medici $\to$ Salviati $\to$ Pazzi, (b) switches *off* any tie from {Medici, Salviati} to a family outside R (otherwise the marking would spread further), and (c) does *anything at all* on the reverse ties and on ties whose source is never reached — produces exactly this R. A whole "fibre" of distinct rule-sets collapses to the one meaning. Summing the fibre sizes over the $4{,}104$ meanings recovers all 2^{40} hypotheses.

C.1 Florentine families: robustness, regret build-up, and the prior

Robustness across the seed. We repeat the construction with each of the 15 families as the seed E (Table 1, Figure 3 left). The content count $|\Omega_L|$, and hence $C(L) = \log_2 |\Omega_L|$, varies with how central the seed is — Medici, the most central family, induces the largest content class (12.00 bits), the peripheral Pazzi the smallest (10.42) — so the semantic information of the same template is a property of where the evidence is injected. The two laws are nonetheless invariant across all 15 seeds: the code length of Algorithm 1 equals $C(L)$ exactly, and the mistake count never exceeds it (Theorems 7 and 4).

How the regret is paid. The exact-regret identity of Section 6 is not an end-of-run coincidence; it is accrued query by query as the data separate contents. After k atoms the minimax value (Theorem 6) is $\log_2 S_k$, where $S_k = |\Omega_L|_{\gamma_{1:k}}|$ is the number of distinct label-patterns the first k atoms induce on Ω_L. Averaged over query orders, $S_k = 2, 4, 7, 13, 25, 46, 82, \ldots, 1780, 2740, 4104$ for $k = 1, \ldots, 15$, so $\log_2 S_k$ climbs monotonically from 0.9 to $C(L) = 12.0$ (Figure 3, right). This is the mechanism behind the identity: the regret is exactly the log-count of contents the queries can tell apart, and the full content sub-base $\mathcal{B}$ separates all $4{,}104$, so the paid code length is $\log_2 |\Omega_L| = C(L)$.

Effect of the prior. Under the uniform prior over contents the code length is the constant $C(L)$ for every target — the minimax-optimal behaviour of Theorem 7. Replacing it with a size prior $w(R) \propto 2^{-|R|}$ makes the code length target-dependent (mean 13.30 bits over the $4{,}104$ targets, shorter than $C(L)$ on small contents and longer on large ones): a matched language bias buys shorter codes on typical targets at the cost of worst-case optimality, exactly the universal-coding trade-off (Grünwald, 2007). Consistent with Proposition 2, the capacity-achieving default is uniform on *contents*, not on the 2^L clause-sets.

C.2 The WordNet taxonomy at scale

Section 6 enumerates Ω_L outright, which is feasible only for small instances — computing $|\Omega_L|$ is #P-hard in general. On a *tree*, however, the version space factorizes and Algorithm 1 runs *implicitly*, with no enumeration: an observed label propagates (a concept marked forces its ancestors marked; a concept unmarked forces its whole subtree unmarked), and the number of consistent contents together with the mixture weight $\hat{p}_t$ of the next query are computed by a single $O(n)$ dynamic program over the tree. The mixture over $|\Omega_L|$ meanings is thus evaluated at every step, however large $|\Omega_L|$ is.

Formal setup (WordNet). Fix a category and let its descendants in the WordNet (Miller, 1995) is-a (hypernymy) taxonomy form a rooted tree $\mathcal{T} = (V, \text{parent})$ with root r and $|V| = n$ concepts ($n = 366$ for CARNIVORE). The constants are the concepts; there are no function symbols, so the Herbrand universe is V itself. A single unary predicate under carries the content: $\mathsf{under}(c)$ reads "concept c is included in the selected sub-taxonomy." The content sub-base — the atoms the learner

seed family	$\lvert\Omega_L\rvert$	$C(L)$ (bits)	code length	mistakes (mean / max)
Medici	4104	12.00	$C(L)$	5.3 / 10
Guadagni	3294	11.69	$C(L)$	5.1 / 9
Albizzi	3206	11.65	$C(L)$	5.3 / 10
Salviati	2738	11.42	$C(L)$	5.0 / 8
Ridolfi	2648	11.37	$C(L)$	5.1 / 8
Bischeri	2614	11.35	$C(L)$	5.1 / 8
Strozzi	2616	11.35	$C(L)$	4.9 / 8
Barbadori	2590	11.34	$C(L)$	5.0 / 8
Tornabuoni	2450	11.26	$C(L)$	4.7 / 8
Castellani	2370	11.21	$C(L)$	4.8 / 8
Peruzzi	2136	11.06	$C(L)$	4.8 / 9
Acciaiuoli	2053	11.00	$C(L)$	4.9 / 9
Lamberteschi	1648	10.69	$C(L)$	4.7 / 8
Ginori	1604	10.65	$C(L)$	4.6 / 7
Pazzi	1370	10.42	$C(L)$	4.6 / 8

Table 1: Algorithm 1 on the Florentine network with each family as seed. The content count and $C(L)$ track the seed's centrality; across all seeds the code length equals $C(L)$ exactly and the mistakes stay below it.

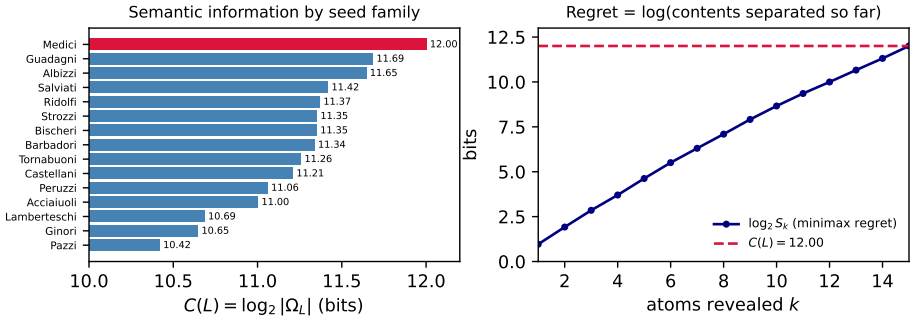


Figure 3: **Left:** semantic information $C(L)$ for each seed family (Table 1); Medici (red) is largest. **Right:** regret build-up — $\log_2 S_k$, the minimax value after k queries, rises to $C(L)$ as the queries separate the $4{,}104$ contents (Medici seed).

is queried on — is

$$\mathcal{B} = \{\mathsf{under}(c) : c \in V\}, \qquad |\mathcal{B}| = n,$$

the parent atoms being fixed background. The parent relation is fixed structural background (the analogue of tie in the Florentine case) and is never predicted; the background BK additionally declares the root, $\mathsf{under}(r)$, as the seed ($V_0 = \{r\}$).

Template and hypotheses. The reachability schema $\mathsf{under}(v) \leftarrow \mathsf{under}(\mathrm{parent}(v))$ grounds to one edge clause per non-root concept, so

$$L = \big\{\, \mathsf{under}(v) \leftarrow \mathsf{under}(u) : v \in V \setminus \{r\},\ u = \mathrm{parent}(v) \,\big\}, \qquad N = |L| = n - 1,$$

a plain edge clause $r_v \leftarrow r_u$ in the form of Proposition 1 (here $N = 365$, so there are $2^{365} \approx 10^{110}$ clause-sets). A hypothesis $S \subseteq L$ switches on a set of parent→child edges; its meaning is $\mu(S) = \mathrm{lhm}(BK \cup S) \subseteq \mathcal{B}$, obtained by firing the switched-on edges from the root to a fixpoint.

Realizable contents. By Proposition 1, on a tree each node's only in-neighbour is its parent, so $R \in \Omega_L$ iff $r \in R$ and R is ancestor-closed—a concept is included only if its parent is:

$$\Omega_L = \{R \subseteq V :\ r \in R,\ \mathrm{parent}(v) \in R \text{ for all } v \in R \setminus \{r\}\},$$

i.e. the subtrees of $\mathcal{T}$ hanging from the root (the is-a reading of the reachability family; the Florentine instance is its social-network reading). The count factorizes: writing $f(v)$ for the number of valid

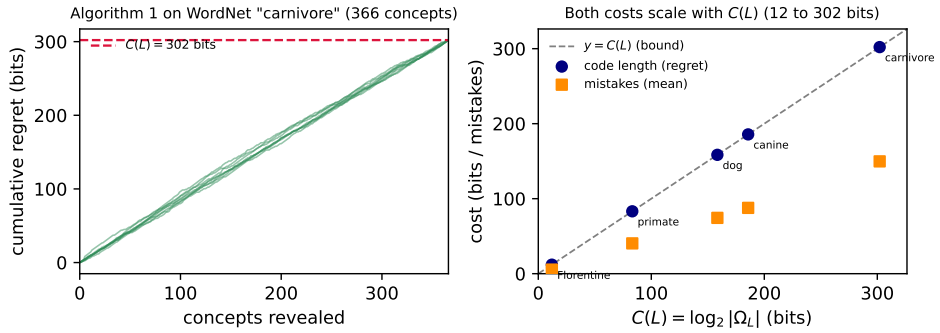


Figure 4: **Left:** Algorithm 1 on the carnivore taxonomy (366 concepts, $|\Omega_L| = 10^{90.9}$); the cumulative regret climbs to $C(L) = 302$ bits. **Right:** across the five instances of Table 2, the code length lands exactly on the line $y = C(L)$ from 12 to 302 bits, while the mistakes grow with $C(L)$ but stay below it.

category (root)	concepts	$\lvert\Omega_L\rvert$	$C(L)$ (bits)	code length	mistakes (mean / max)
Florentine (reference)	15	4.1×10^3	12.0	$C(L)$	5.3 / 10
primate	106	$10^{25.0}$	83.2	$C(L)$	40.4 / 46
dog	190	$10^{47.7}$	158.5	$C(L)$	74.4 / 81
canine	224	$10^{55.9}$	185.7	$C(L)$	87.8 / 99
carnivore	366	$10^{90.9}$	302.1	$C(L)$	149.8 / 162

Table 2: Algorithm 1 on the WordNet is-a taxonomy, run implicitly by dynamic programming. The code length equals $C(L)$ exactly at every scale, even for $|\Omega_L| = 10^{90}$; mistakes stay under $C(L)$.

configurations of the subtree at v given $v \in R$,

$$f(v) = \prod_{c:\ \mathrm{parent}(c)=v} \big(1 + f(c)\big), \qquad |\Omega_L| = f(r),$$

each child being dropped or kept-with-$f(c)$-options. For CARNIVORE this gives $|\Omega_L| = 10^{90.9}$, hence the semantic capacity $C(L) = \log_2 |\Omega_L| = 302.1$ bits, against the 2^{365} clause-sets—collapse ratio $2^N/|\Omega_L| \approx 10^{18.9}$. The product above is the $O(n)$ dynamic program by which Algorithm 1 evaluates the version-space mixture without ever enumerating Ω_L.

Results. Table 2 runs Algorithm 1 on categories spanning 83 to 302 bits of semantic information, i.e. content classes from 10^{25} to 10^{90} meanings. In every case the cumulative code length equals $C(L)$ *exactly* (to 13 decimal places; the residual is floating-point), and the mistakes stay below $C(L)$, at mean $\approx C(L)/2$. For carnivore (366 concepts) the $2^{365} \approx 10^{110}$ clause-sets collapse onto $|\Omega_L| = 10^{90.9}$ contents (collapse ratio $10^{18.9}$), and Algorithm 1 pays exactly $C(L) = 302$ bits without ever materializing a content (Figure 4). The single number $C(L)$ therefore governs the online cost on real data across more than an order of magnitude of scale — and, since the tree case runs in linear time, the #P-hardness of counting Ω_L is not an obstacle to *using* it.

C.3 BASELINE: THE SAME LEARNER RUN OVER CLAUSE-SETS

What does a learner lose by searching the syntax 2^L instead of the contents Ω_L? The natural baseline is the same Bayes mixture as Algorithm 1, run over the 2^N clause-sets under the uniform prior. A clause-set predicts only through its content $\mu(S)$, so this learner *is* Algorithm 1 with the non-uniform prior $w(R) = |\mu^{-1}(R)|$, the fibre size, and Theorem 7(i) gives its regret exactly: once the queries separate the contents it pays

$$\mathrm{Regret}_T^{\mathrm{syn}}(R^\star) \;=\; \log_2 \frac{2^N}{|\mu^{-1}(R^\star)|} \qquad \text{against} \qquad \mathrm{Regret}_T^{\mathrm{sem}}(R^\star) \;=\; \log_2 |\Omega_L| = C(L).$$

The two learners therefore differ in one thing only — uniform on *syntax* versus uniform on *meanings* — and the comparison is exact, with no tuning. Averaged over contents the clause-set learner pays $C(L) + D(\mathrm{unif}_{\Omega_L} \,\|\, \pi)$ with $\pi = \mu_* \mathrm{unif}_{2^L}$ (Proposition 2); averaged over its own prior it pays $H(\pi) \le C(L)$.

Florentine network. We compute all $4{,}104$ fibre sizes exactly (they sum to 2^{40}). Algorithm 1 pays $C(L) = 12.00$ bits on every target. The clause-set mixture pays between 6.00 and 20.00 bits: more than $C(L)$ on 79% of the contents, 14.35 bits on average over contents ($D(\mathrm{unif}\|\pi) = 2.35$ bits), and 8 bits (67%) more in the worst case (Figure 5, left). Under $0/1$ loss the corresponding weighted-majority learner makes 5.58 mistakes on average and up to 12, against 5.29 and 10 for Halving over Ω_L. The ordering reverses when the target is itself the content of a uniformly random clause-set: those targets have large fibres, and the clause-set mixture pays $H(\pi) = 9.81$ bits (3.75 mistakes) against 12.00 (4.84).

Bandwidth template. For $L_{n,d}$ the fibre mass factorizes along the line, $-\log_2 \pi(R) = \sum_{j=1}^{n} c_j$ with $c_j = k_j$ if $r_j \notin R$ and $c_j = -\log_2(1 - 2^{-k_j})$ if $r_j \in R$, where k_j is the number of atoms of R among $r_{j-d}, \ldots, r_{j-1}$; a transfer pass gives the worst case and the mean exactly for any n (Figure 5, right; Table 3). The overpayment is *linear in* n: for $d = 2$ the clause-set mixture pays $1.01\,n$ bits on average over contents and $\frac{4}{3}n$ in the worst case, against $C(L) = 0.694\,n$. It stays well below the syntactic guarantee $N \approx dn$: the collapse ratio $2^N/|\Omega_L|$ is the gap between the two *guarantees*, of which the realized regret gap is a constant fraction (about one quarter on average and one half in the worst case for $d = 2$). On typical random programs the clause-set prior is again the better code: by Corollary 2 random reachability dies out, so $H(\pi) = O(1)$ (7.25 bits for $d = 2$, uniformly in n) while $C(L) = \Theta(n)$.

Reading. Uniform-on-syntax is a strongly non-uniform prior on meanings: it concentrates on the few contents with huge fibres (small outbreaks, or fully saturated ones) and starves the exponentially many thin-fibre contents. The content-uniform Algorithm 1 is the minimax choice (Theorem 7(ii)): it pays exactly $C(L)$ on every target, whereas the clause-set learner's cost ranges over $[\,O(1),\ \Theta(N)\,]$ depending on how syntactically redundant the truth happens to be.

	$C(L)$ (Alg. 1, every target)	clause-set, mean over Ω_L	clause-set, worst	guarantee N
Florentine ($N = 40$)	12.00	14.35	20.00	40
bandwidth $d = 2$, $n = 20$	14.81	19.76	27.00	39
bandwidth $d = 2$, per atom	0.694	1.009	1.333	2
bandwidth $d = 3$, per atom	0.879	1.129	1.500	3

Table 3: Log-loss regret (bits) of the uniform mixture over clause-sets versus Algorithm 1. All entries are exact (fibre counts / transfer pass); per-atom rows are the $n \to \infty$ rates.

D Semantic Information: the deterministic-channel lift

The quantities of Section 2 are combinatorial including a count $|R|$ and a Hartley $\log_2 |\Omega_L|$. They are the deterministic shadows of information-theoretic objects that appear the moment a prior is placed on hypotheses. Because the closure map μ is deterministic, the lift is exact and forced. Let P be a law on 2^L; through the deterministic channel $S \mapsto R = \mu(S)$ it induces a random content with pushforward $\pi = \mu_* P$ on Ω_L. That is, $\pi = \mu_* P$ is the distribution of the content R (on Ω_L) induced by pushing the clause-set prior P (on 2^L) through μ. S carries the input prior; R carries π.

Definition 3 (Probabilistic semantic quantities)**.** For a prior P on 2^L with pushforward $\pi = \mu_* P$: the *semantic entropy* is $H(\Omega_L) := H(R)$; the *semantic information* is $I_P := I(S; R)$ across the generation channel; the *semantic capacity* is $C(L) := \sup_P I(S; R)$.

Proposition 2 (The deterministic channel collapses the triple)**.** *For $R = \mu(S)$ and any prior P:*

(i) $I(S; R) = H(R)$ and $H(S) = H(R) + H(S \mid R)$; the equivocation $H(S \mid R)$ is the semantic information μ discards*, and it is 0 iff μ is injective;*

(ii) $C(L) = \sup_P I(S; R) = \log_2 |\Omega_L|$, attained exactly when π is uniform on Ω_L; this recovers the combinatorial capacity of Definition 1;

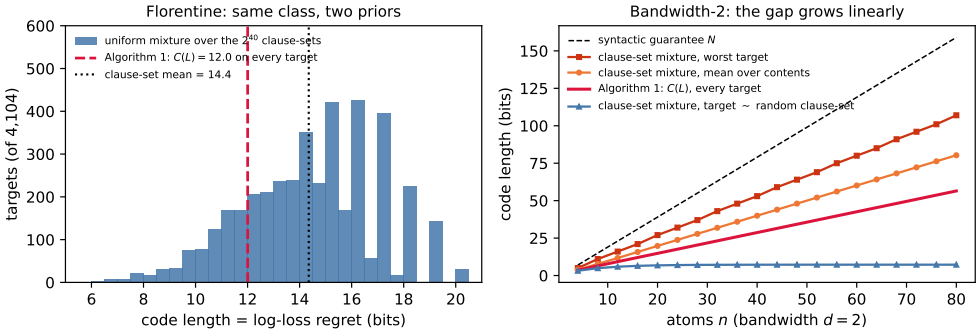


Figure 5: **Left:** Florentine network. Code length of the uniform mixture over the 2^{40} clause-sets on each of the 4,104 targets (histogram; mean 14.35, max 20), against the constant $C(L) = 12.0$ of Algorithm 1. **Right:** bandwidth-2 template. The clause-set mixture's worst-case and content-averaged code lengths grow linearly in n, above $C(L)$ and below the syntactic guarantee N; on targets drawn from random clause-sets it pays only $H(\pi) = O(1)$.

(iii) under that capacity-achieving prior the self-information $-\log_2 \pi(R) \equiv \log_2 |\Omega_L|$ *is constant — an exact equipartition for realizable theories — while a generic prior falls short of capacity by* $\log_2 |\Omega_L| - H(R) = D(\pi \,\|\, \mathrm{unif}_{\Omega_L})$.

Proof sketch. $R = \mu(S)$ is deterministic, so $H(R \mid S) = 0$ and $I(S;R) = H(R)$; the chain rule gives the decomposition. As P ranges over laws on 2^L, $\pi = \mu_* P$ ranges over all laws on Ω_L, so $\sup_P H(R) = \log_2 |\Omega_L|$ at the uniform π; the deficit is the KL divergence to uniform. □

This fixes the dictionary: semantic content, after taking logs, is entropy $H(R)$; semantic information is $I(S;R)$, here equal to $H(R)$ because generation is noiseless; and semantic capacity is the Shannon capacity $\sup_P I(S;R)$, here equal to $\log_2 |\Omega_L|$ because the channel is deterministic.

Example 4 (The cheating class, made probabilistic)**.** In reality, not every friend of a cheated person is a cheater. This is the *dual* of the epidemic in Example 3: there the friendship *graph* was random and the spreading rule fixed; here we keep the fixed ten-tie graph of Example 1 and randomize the *rules* instead ("S random", a bond percolation on the fixed graph). To model it, we put a prior on the template of Example 1 by letting each friendship conduct cheating independently with probability θ (Bernoulli on clauses). The pushforward π_θ gives each content the total mass of its fibre; the fibre sizes $|\mu^{-1}(R)|$ over the ten contents are $256, 128, 128, 32, 64, 32, 160, 32, 32, 160$ (summing to 2^{10}). At $\theta = \frac{1}{2}$ (uniform 1024 clause-sets),

$$H(S) = N = 10, \qquad H(R) = I(S;R) = 2.962, \qquad H(S \mid R) = 7.038 \text{ bits},$$

so μ discards 7.038 of the 10 hypothesis bits — the semantic information lost to the quotient. The capacity is prior-free, $C(L) = \log_2 10 = 3.322$ bits, attained only by the content-uniform prior; the Bernoulli($\frac{1}{2}$) prior falls 0.360 bit short, exactly $D(\pi_{1/2} \,\|\, \mathrm{unif}_{\Omega_L})$, because its pushforward is uneven (fibres range from 32 to 256). Equivalently, in the friendship-percolation reading (each friendship conducts w.p. θ), the middle student C — reachable through A or through D — cheats with the *reliability* probability $1 - (1-\theta^2)^2$, a redundancy no chain can express. Capacity, entropy, and extent are three different bits; the mean extent is the subject of Section 4.

E Concluding Remarks

For recursive Horn templates the right object is the semantic content class Ω_L. Its semantic information $C(L) = \log_2 |\Omega_L|$ is a single invariant, and counting, coding, and learning all measure it. Syntax over-counts by the collapse ratio $2^N/|\Omega_L|$. So any capacity, sample-complexity, or code-length claim stated over clause-sets, rather than contents, is inflated by exactly that factor. The template poses two very different questions. Capacity asks how many contents exist; semantic extent asks how much one content deduces. The two are governed by different mathematics, and conflating them

is the error the framework is built to prevent. The running example makes this concrete. Catch one cheater, and the possible outbreaks form a 10-element lattice: $C(L) = \log_2 10$ bits, not 2^{10} clause-sets. The probabilistic rule reproduces the capacity/entropy/gap triad, and at scale a sharp epidemic threshold at one friend per student. And Halving learns who cheats in $C(L)$ mistakes. This is the concrete, combinatorial face of incorporating logical reasoning into information theory (Lastras et al., 2026). Open directions include the precise second-order asymptotics of the nonlinear count, the exact constant in $\mathrm{Ldim}(\Omega_L) = \Theta(C(L))$, a noisy semantic capacity, and the #P-hardness of computing $|\Omega_L|$ for general DAGs.

F THE MACHINE-LEARNING ↔ LOGIC DICTIONARY

The learning reading of Section 5 rests on a fixed correspondence between the vocabulary of statistical/online learning and the logic-programming objects of Section 2. The input space is the Herbrand base itself, so an "example" is a ground atom γ paired with the bit recording whether the target theory deduces it; and the hypothesis class is the set of contents Ω_L, never the 2^L clause-sets.

What is learned is the target content R^*; a prediction is a guess of the entailment bit of the queried atom; and learning is nearest-point identification in the metric space (Ω_L, d_D) with $d_D(R, R') = D(R \triangle R')$.

machine learning	**logic / this paper**
input (instance) space $\mathcal{X}$	Herbrand base $\mathcal{B}$ (queried on the content sub-base)
instance $x \in \mathcal{X}$	a ground atom $\gamma \in \mathcal{B}$ (a variable-free fact)
label $y \in \{0, 1\}$	deduction bit $y = \mathbf{1}[\gamma \in R^*]$
hypothesis / concept class $\mathcal{H}$	realizable closures $\Omega_L \subseteq \{0, 1\}^{\mathcal{B}}$ (not 2^L)
hypothesis $h \in \mathcal{H}$	a content $R \in \Omega_L$, read as $h_R(\gamma) = \mathbf{1}[\gamma \in R]$
prediction $\hat{y}$	$\hat{y} = \mathbf{1}[\gamma \in \hat{R}]$ (0/1), or belief $\hat{p} \in [0, 1]$ (log loss)
target / ground truth h^*	the target theory R^*; realizable iff $R^* \in \Omega_L$
raw parameterization of $\mathcal{H}$	clause-sets $S \in 2^L$, decoded by many-to-one μ

Table 4: Objects. The many-to-one $\mu : 2^L \to \Omega_L$ is the parameter-to-classifier map; semantically equal S are one h.