EXPLANATION-ASSISTED LEARNING WITH ADAPTIVE WITNESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

An explanation can tell a learner how to classify examples beyond the one being explained. We study explanation-assisted learning when several explanations are valid and a provider uses the training history to choose which one to reveal. We model explanations as reusable positive witnesses: a positive training input is a bag of instances, and its explanation identifies an instance sufficient for positivity. Test bags have no witnesses. At latent VC dimension one, a learner whose output is independent of witness selection achieves the optimal memoryless sample complexity. At dimension two, a family separates the sample complexities of memoryless explanations, causal explanations, and labels alone; a countable extension separates PAC learnability itself. The lower-bound provider always gives valid explanations, yet its complete witness history carries no information about unseen designated labels beyond the observed labels. Finite Littlestone dimension gives a general causal upper bound, while a witness-span learner for linear subspaces attains the memoryless rate.

1 INTRODUCTION

An explanation supplied with a training label may reveal a rule that also classifies future inputs. Such reuse can reduce the number of labeled examples needed for learning. It underlies classical explanation-based learning (Mitchell et al., 1986; DeJong & Mooney, 1986) and motivates learning from annotator rationales and explanation constraints (Zaidan et al., 2007; Pukdee et al., 2023).

Consider two rules, a and b, each sufficient to justify an input's positive label. An annotator can correctly cite either one. A learner told about b can reuse it on a later input where b fires but a does not. The two explanations are equally valid on the training example, yet provide different information about future examples. A learning guarantee based solely on validity must hold for both choices.

Our model of explanation-assisted learning treats explanations as reusable positive certificates. An input is a *bag* of instances, and its label is positive when it contains an instance from an unknown positive set. A positive training bag comes with one such instance, called a *witness*. That witness certifies the positivity of every future bag containing it. The witness is the extra training supervision; at test time the learner sees only the bag.

If the provider uses a fixed selection rule with independent randomness, the witnessed training records are i.i.d. We call this protocol *memoryless*. Its optimal sample complexity is controlled by the VC dimension of the positive-set class, through a reduction to robust learning (Montasser, 2026). A *causal* provider may instead remember previous records when choosing the next witness. Its witnesses remain valid, and the bags remain i.i.d., but the complete records can be dependent. We ask:

> *When does the learning benefit of valid, reusable witnesses survive history-dependent selection?*

We seek guarantees uniform over all legal providers. The provider cannot change the arriving bags or their labels; it controls only which valid witness accompanies each positive bag.

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.

1.1 Our contributions

Write β for the VC dimension of the positive-set class on individual instances. Our results distinguish $\beta = 1$ from $\beta = 2$.

- **Optimal learning when $\beta = 1$.** Every VC-one class admits a learner with the optimal memoryless sample complexity. The learner replaces different witnesses for the same bag by a common tree constraint, making its output independent of witness selection (Section 3).
- **A three-way separation when $\beta = 2$.** For a family indexed by q, memoryless witnesses, causal witnesses, and labels alone have sample complexities $\Theta((s + \log(1/\delta))/\epsilon)$ with $s = 1, q, q^2$, respectively. A deterministic causal provider hides unseen bag labels along the entire observed history. A countable extension separates PAC learnability itself (Section 4).
- **Upper bounds from online and algebraic structure.** Finite Littlestone dimension gives a general causal upper bound. For linear subspaces, maintaining a witness span achieves the optimal memoryless rate with a compact representation (Section 5).

A learner can ignore witnesses, so causal selection cannot make optimal learning harder than learning from labels alone.

1.2 Related work

Learning from explanations and privileged information. Our prediction task is the OR model of multiple-instance learning (Dietterich et al., 1997; Blum & Kalai, 1998; Sabato & Tishby, 2012). The selected positive instance is privileged training information (Pechyony & Vapnik, 2010; Sharoni & Sabato, 2023). Rationale models, explanation constraints, and explanation-preserving graph perturbations offer other ways to use explanations (Zaidan & Eisner, 2008; Pukdee et al., 2023; Zheng et al., 2024). Here the extra information is an exact sufficient condition, and we study the choice among valid conditions. Arbitrary text rationales need not satisfy this certificate interpretation.

Providers and interaction. Teaching models study information supplied by a cooperative provider (Goldman & Kearns, 1995; Zilles et al., 2011; Doliwa et al., 2014). Strategic labeling can instead slow active learning through abstention (Yan & Zhang, 2024); dynamic benchmarking studies dependence created by successive rounds of data collection and model fitting (Shirali et al., 2023). Our provider must answer every positive bag with a valid witness, while the bag distribution remains fixed.

Learning-theoretic tools. The memoryless benchmark uses robust learning (Montasser et al., 2019; Montasser, 2026). Our causal results use VC-one tree representations (Ben-David, 2015; Chepoi et al., 2020) and online mistake bounds (Littlestone, 1988; Cesa-Bianchi et al., 2004). Appendix G gives further comparisons.

2 Problem formulation

2.1 Explanations as reusable witnesses

An explanation identifies a sufficient condition for a positive label. The learner receives this information during training; test inputs arrive without explanations. We represent the sufficient condition by an instance and the conditions present in an input by a bag.

Let R be a finite instance universe and let $\varnothing \neq \mathcal{G} \subseteq 2^R$ be a known family of possible positive sets. Nature fixes a target $G \in \mathcal{G}$ and a distribution D on bags $A \subseteq R$. The label is

$$h_G(A) = \mathbf{1}\{A \cap G \neq \varnothing\}. \tag{1}$$

The training bags $A_1, \ldots, A_m$ are i.i.d. from D. On a positive bag, the provider returns any $W_t \in A_t \cap G$; on a negative bag it returns $\perp$. The learner observes $\mathcal{T}_m = ((A_t, h_G(A_t), W_t))_{t=1}^m$ and predicts on a fresh bag without a witness. Labels and witnesses are exact; instance identity is the entire explanation message.

For example, a public dictionary of rules $\{\phi_r : r \in R\}$ represents an input x by the rules that fire, $A(x) = \{r : \phi_r(x) = 1\}$. The unknown set G identifies the rules sufficient for positivity. Revealing

$r \in A(x) \cap G$ certifies positivity wherever that rule fires again. We allow arbitrary bag distributions, including dependence among instances within a bag. General learners receive the target incidence table, so the results are information-theoretic; the subspace learner in Section 5.2 has a compact representation.

2.2 Memoryless and causal providers

A *memoryless* provider samples from a fixed kernel $K_{G,D}(\cdot \mid A_t)$ with fresh independent coins. A deterministic pointwise selector is included. For fixed G, D, the condition is

$$\mathcal{L}(W_t \mid A_t, \mathcal{T}_{t-1}, G, D) = K_{G,D}(\cdot \mid A_t). \tag{2}$$

A *causal* provider may use G, D, the current labeled bag, and the transcript through round $t-1$, but cannot see future bags or the learner's private coins. Causal here means non-anticipating. The learner knows neither G, D nor the provider.

We measure the number of records needed to predict with error at most ϵ, with probability at least $1-\delta$, uniformly over targets, distributions, and providers in the specified protocol. Denote these minimax sample complexities by $m^\star_{\mathrm{mem}}$, $m^\star_{\mathrm{cau}}$, and $m^\star_{\mathrm{lab}}$, with the last using labels alone. They satisfy $m^\star_{\mathrm{mem}} \le m^\star_{\mathrm{cau}} \le m^\star_{\mathrm{lab}}$. For a predictor $\widehat{h}$, write $\mathrm{err}_D(\widehat{h}, h_G) = \Pr_{A\sim D,\widehat{h}}[\widehat{h}(A) \neq h_G(A)]$; this averages any fresh prediction randomness. Appendix F records the minimax quantifiers and randomization conventions.

2.3 Complexity measures and benchmark rates

We distinguish the dimensions of the instance class, the bag class, and online prediction:

$$\beta = \mathrm{VC}(\mathcal{G}), \qquad v_{\mathrm{bag}} = \mathrm{VC}(\{h_G : G \in \mathcal{G}\}), \qquad L = \mathrm{Ldim}(\mathcal{G}). \tag{3}$$

The first two dimensions count how many instances or bags can receive all possible binary labelings. Littlestone dimension is the largest depth of a complete binary instance-query tree whose every root-to-leaf labeling is realized by the class. Singleton bags give $\beta \le v_{\mathrm{bag}}$. Standard realizable PAC learning (Hanneke, 2016) gives

$$m^\star_{\mathrm{lab}}(\epsilon, \delta) = \Theta\left(\frac{v_{\mathrm{bag}} + \log(1/\delta)}{\epsilon}\right) \tag{4}$$

for $v_{\mathrm{bag}} \ge 1$, $0 < \epsilon \le 1/8$, and $0 < \delta \le 1/4$. Memoryless witnesses replace the bag dimension by the latent dimension.

Proposition 2.1 (Memoryless benchmark)**.** *For every finite target family with $\beta \ge 1$,*

$$m^\star_{\mathrm{mem}}(\epsilon, \delta) = \Theta\left(\frac{\beta + \log(1/\delta)}{\epsilon}\right) \tag{5}$$

for $0 < \epsilon \le 1/8$ and $0 < \delta \le 1/4$. The upper bound permits a randomized, improper learner. If $\beta = 0$, no samples are needed.

Reduction to robust learning. A positive record certifies one positive instance; a negative record certifies that every member is negative. These constraints reduce witnessed learning to robust learning while preserving β. Memoryless selection supplies the i.i.d. records required by that reduction. Appendix A gives the proof, including a self-contained bound based on Aden-Ali et al. (2024) sufficient for $\beta \le 2$; the general additive dependence follows from Montasser (2026).

The bound is uniform over the fixed kernel: any deterministic pointwise selector of legal witnesses suffices, without optimizing their informativeness. When v_{bag} is much larger than β, the comparison with (4) shows how these explanations reduce the number of training examples needed for the same prediction error.

3 Optimal learning at VC dimension one

At VC dimension one, the learner can remove the effect of witness selection. It keeps a common constraint from every legal witness for a bag, then learns from those constraints at the optimal memoryless rate.

Theorem 3.1 (Causal rate at VC dimension one)**.** *For every finite family with* $\mathrm{VC}(\mathcal{G}) \leq 1$*, there is a deterministic learner with at most one leave-one-out error on every finite bag pool, uniformly over legal witnesses. On* $m \geq 0$ *i.i.d. training bags, against every causal provider,*

$$\Pr[\mathrm{err}_D(\widehat{h}_m, h_G) > \epsilon] \leq (1-\epsilon)^m \leq e^{-m\epsilon}, \quad 0 < \epsilon < 1, \tag{6}$$

$$\mathbb{E}\ \mathrm{err}_D(\widehat{h}_m, h_G) \leq \frac{1}{m+1}. \tag{7}$$

If $\mathrm{VC}(\mathcal{G}) = 1$*, then for* $0 < \epsilon, \delta \leq 1/4$*,*

$$m^\star_{\mathrm{cau}}(\epsilon, \delta) = \Theta\left(\frac{1 + \log(1/\delta)}{\epsilon}\right), \tag{8}$$

and the worst-case expected risk is $\Theta(1/(m+1))$*.*

3.1 A CANONICAL CONSTRAINT FOR EACH WITNESSED BAG

Consider the four targets in Figure 1 and the bag $A = \{a, b\}$. If the true target is G_3, either a or b is a legal witness. Seeing a leaves G_2, G_3 possible, whereas seeing b identifies G_3. Retaining the branch containing G_2, G_3 in both cases removes the effect of the provider's choice. This operation extends to *every* VC-one class.

Target	Positive set	$h_G(\{a, b\})$
G_0	$\varnothing$	0
G_1	$\{c\}$	0
G_2	$\{a\}$	1
G_3	$\{a, b\}$	1

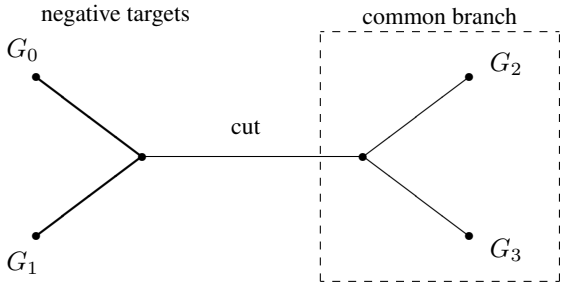

Figure 1: **Different witnesses can be reduced to the same constraint.** For the positive bag $\{a, b\}$ under G_3, witness a permits G_2, G_3, while witness b permits only G_3. Both point to the dashed branch. Keeping that branch discards the extra distinction made by b. The thick subtree joins the targets that label this bag negative.

Tree representation. The tree represents uncertainty about the target: its leaves are candidate sets G, not instances r. Write $X = \mathcal{G}$ for these leaves. VC dimension one means that two instances cannot realize all four membership patterns. This lets us arrange the candidates in a tree T where every nonconstant incidence set is the leaf set on one side of an edge (Ben-David, 2015; Chepoi et al., 2020). For witness r, write H_r for that half-tree and B_r for its candidate leaves:

$$B_r = \{G \in X : r \in G\} = H_r \cap X. \tag{9}$$

In Figure 1, a selects the right branch, while b selects just G_3. Lemma B.1 constructs the tree.

The component determined by a bag. For a bag A, first find the candidates that would label it negative:

$$C_A = \{G \in X : A \cap G = \varnothing\}, \qquad Q_A = \mathrm{hull}_T(C_A). \tag{10}$$

The hull is the smallest connected subtree joining those leaves; in the example, it is the thick left subtree. Assume C_A is nonempty and delete Q_A and its incident edges. For any legal witness r, every candidate in C_A omits r, so Q_A lies on the opposite side of its edge from H_r. Thus H_r avoids Q_A. Since H_r is connected and contains the true target, it must lie in the component containing G. All legal witnesses identify this same component. The learner locates that component using any candidate in B_r, without knowing which candidate is true (Lemma B.2).

Let $J_A(G)$ denote that component. Replace each record by the subtree

$$K_G(A) = \begin{cases} T, & C_A = \varnothing \quad \text{(every target labels the bag positive)}, \\ Q_A, & A \cap G = \varnothing \quad \text{(a negative record)}, \\ J_A(G), & \text{otherwise} \quad \text{(a positive record with a witness)}. \end{cases} \tag{11}$$

We call this replacement *witness canonicalization*. The retained subtree contains G and depends only on A, G, although the learner computes it from the observed record without knowing G.

3.2 THE LEARNING ALGORITHM

Preprocess the known target class into T and fix its reference leaf o.

Training: intersect the constraints. Map each observed record to its subtree using (11). Intersect these subtrees and choose the vertex closest to o:

$$V_m = \bigcap_{i=1}^{m} K_G(A_i), \quad V_0 = T, \qquad p(V_m) = \text{the vertex of } V_m \text{ closest to } o. \tag{12}$$

The intersection is connected and nonempty because every constraint contains G. The vertex $p(V_m)$, called the *gate*, summarizes the training data.

Prediction: test the gate. For a new bag A, compute its negative candidates C_A. Predict positive if C_A is empty; otherwise use

$$\widehat{h}_m(A) = 0 \quad \Longleftrightarrow \quad p(V_m) \in Q_A. \tag{13}$$

The gate may be an internal vertex rather than a candidate target, so this predictor need not belong to the original bag hypothesis class.

Example (continued). Take $o = G_0$ and true target G_3. After the positive bag $\{a, b\}$, either witness leaves the right branch and places the gate at its junction v. For a test bag $\{a\}$, the negative hull joins G_0, G_1 and excludes v, so the learner predicts positive. For $\{b\}$, the negative hull joins G_0, G_1, G_2 and includes v, so it predicts negative. This error is possible even if the first witness was b: canonicalization intentionally discarded that extra information. If $\{b\}$ now appears in training, its only legal witness is b and its retained component is the single leaf G_3. The gate moves from v to G_3, correcting that prediction.

3.3 PROOF SKETCH: LEARNING A THRESHOLD ON THE TARGET PATH

The example's gate moves along the path from G_0 through v to G_3. In general, fix the true target for analysis and follow the path Π from o to G; the learner need not know this path. Every retained subtree contains G, so it meets Π in a final segment beginning at a vertex $\tau_G(A)$:

$$K_G(A) \cap \Pi = [\tau_G(A), G]. \tag{14}$$

Intersecting the constraints moves the gate to the largest threshold seen so far. Once the gate reaches a bag's threshold, it lies inside that bag's negative hull if the bag is negative, or outside that hull if the bag is positive. Consequently, an error requires a threshold beyond the current gate:

$$\widehat{h}_m(A) \neq h_G(A) \quad \Longrightarrow \quad \tau_G(A) >_\Pi p(V_m). \tag{15}$$

Now hold out each bag in a pool in turn and train on the others. Only a bag with the *unique strict largest* threshold can be misclassified, so there is at most one leave-one-out error. The thresholds are functions of the i.i.d. bags, not of the provider's choices. Exchangeability gives the $1/(m+1)$ expected-risk bound; the probability of missing a tail of mass greater than ϵ gives $(1-\epsilon)^m$. Appendix B supplies the proof. This connects canonicalization to one-inclusion prediction (Haussler et al., 1994).

Path classes and countable universes. Root-to-leaf paths on a depth-d binary tree have $\beta = 1$ and $v_{\text{bag}} = L = d$, so witnesses remove the label-only dimension term (Proposition B.5). Query-local finite trees extend the optimal rate to countable universes with finite bags (Corollary B.4).

4 A SEPARATION AT VC DIMENSION TWO

We construct VC-two classes for which the three protocols require different sample sizes.

Theorem 4.1 (Three minimax scales)**.** *For every $q \geq 17$, there is a finite family $\mathcal{G}_q \subseteq 2^{R_q}$ with $\mathrm{VC}(\mathcal{G}_q) = 2$, $\mathrm{Ldim}(\mathcal{G}_q) = q$, and $\log_2 |R_q| = \Theta(q^2)$ such that, for $0 < \epsilon \leq 1/64$ and $0 < \delta \leq 1/8$,*

$$m^\star_{\mathrm{mem}}(\epsilon, \delta) = \Theta((1 + \log(1/\delta))/\epsilon), \tag{16}$$

$$m^\star_{\mathrm{cau}}(\epsilon, \delta) = \Theta((q + \log(1/\delta))/\epsilon), \tag{17}$$

$$m^\star_{\mathrm{lab}}(\epsilon, \delta) = \Theta((q^2 + \log(1/\delta))/\epsilon). \tag{18}$$

The causal lower bound holds for one deterministic, distribution-independent provider defined on the full bag domain.

Table 1 relates these rates to the family's three dimensions.

Training information	Sample complexity	Governing dimension on this family
Memoryless witnesses	$\Theta((1 + \log(1/\delta))/\epsilon)$	Latent VC: 2
Causal witnesses	$\Theta((q + \log(1/\delta))/\epsilon)$	Littlestone: q
Labels alone	$\Theta((q^2 + \log(1/\delta))/\epsilon)$	Bag VC: $\Theta(q^2)$

Table 1: Minimax rates and dimensions on the same family. The worst-case bag distribution may differ across protocols.

4.1 Masking labels across histories

We need q bags $A_1, \ldots, A_q$ with arbitrary labels but latent VC dimension two. One token per bag would fail: arbitrary bag labels would shatter those q tokens. We therefore need many possible witnesses with tightly constrained joint membership.

A shared mask couples the witnesses. Let $S \subseteq [q]$ specify the positive bags and encode S as a slope λ_S over a finite field. An unknown offset u masks this slope in $f(a_i) = u + a_i\lambda_S$, where the a_i are distinct public points. Under a uniform mask, one value has the same distribution for every S. Two values on the same line, however, reveal $\lambda_S = (f(a_i) - f(a_j))/(a_i - a_j)$ and hence every bag label. Giving each bag an independent mask avoids this disclosure, but also lets fixed token memberships vary independently, losing the small VC dimension.

Instead, attach a mask u_t, called a *pad*, to the tth distinct positive bag discovered. One history then evaluates each line $f_t(x) = u_t + x\lambda_S$ at only one public point. Repeated bags receive their original witnesses. For example, with two positive bags:

Arriving bag	Discovery order	Code in the returned witness
A_1	(1)	$(u_1 + a_1\lambda_S)$
A_2	$(1, 2)$	$(u_1 + a_1\lambda_S,\ u_2 + a_2\lambda_S)$
A_1 again	replay (1)	replay $(u_1 + a_1\lambda_S)$

Had A_2 arrived first, its reply would contain $u_1 + a_2\lambda_S$ instead. The two possible first replies determine the slope together, but the provider never shows both in one history. VC dimension examines arbitrary tokens, including replies from different possible histories; the learner sees only replies from one history. This example uses two pads ($q \geq 4$).

4.2 The target family and provider

Fix $q \geq 3$. Both the tokens and the target's pads are fixed before sampling; the provider uses no fresh randomness.

Tokens and bags. A field with 2^q elements has room for all q-bit label vectors. Fix $\mathbb{F} = \mathrm{GF}(2^q)$, an $\mathbb{F}_2$-linear bijection $\Lambda : \mathbb{F}_2^q \to \mathbb{F}$, and distinct $a_1, \ldots, a_q \in \mathbb{F}$. Set $\lambda_S = \Lambda(\mathbf{1}_S)$ and cap the number of pads at $k = q - 2$. For every ordered list $\pi = (i_1, \ldots, i_d)$ of distinct indices and every $z \in \mathbb{F}^{\min\{d,k\}}$, create an instance $e(\pi, z)$ and place it in A_{i_d}. The universe R_q consists of all these tokens. Thus a reply can identify a history and code by selecting an existing instance; it adds no message beyond instance identity.

Targets. Choose a nonempty support S and pads $u \in \mathbb{F}^{r(S)}$, where $r(S) = \min\{|S|, k\}$. The target $G_{S,u}$ contains exactly one token for every ordered list π drawn from S, with code

$$z(\pi, S, u)_t = u_t + a_{i_t}\lambda_S, \qquad 1 \le t \le \min\{d, k\}. \tag{19}$$

Include the empty target as well. Bag A_i is positive precisely when $i \in S$: every target token ends in S, and the list (i) supplies a token whenever $i \in S$.

The cap $k = q - 2$. A list of length at least $q - 1$ already names all but at most one support index. At most two supports can contain it. The cap ensures that its code also fixes *all* pads for either support, so such a token belongs to at most two targets. This prevents these long-history tokens from forming a shattered triple: each member would need to occur in four targets, one for each labeling of the other two tokens.

Provider. On each new positive designated bag, extend the discovery list to π and return $e(\pi, z(\pi, S, u))$. On a repetition, replay that bag's first token. After k discoveries, longer lists carry the same completed code. Every reply belongs to both the arriving bag and the fixed target. To define one provider on *all* bags, return the first legal token in a fixed ordering on any other positive bag, without changing the discovery list; negative bags receive $\perp$. The rule is deterministic and causal.

4.3 Sample complexity and dimensions

For the lower bound, draw a random support and uniform pads, keeping the provider deterministic. Witnesses then leave the support posterior unchanged:

Proposition 4.2 (No additional information about the designated labels)**.** *Fix the provider above. Draw S from any prior on subsets of $[q]$ and, conditional on nonempty S, draw u uniformly from $\mathbb{F}^{r(S)}$. Let $I \in [q]^m$ be a fixed sequence of designated-bag indices, or a random sequence independent of the target. If Y, W are its labels and complete witness transcript, then*

$$\mathcal{L}(S \mid I, Y, W) = \mathcal{L}(S \mid I, Y). \tag{20}$$

To see the identity, fix I, Y. They determine the discovery order and the number t_0 of exposed pads. For any compatible S, adding the fixed vector $(a_{i_s}\lambda_S)_{s=1}^{t_0}$ to uniform independent pads gives a uniform vector in $\mathbb{F}^{t_0}$. All replies copy prefixes of this vector; repetitions and discoveries beyond the cap add no further code values. Thus the likelihood of the *entire* transcript is the same for every compatible S, not merely the likelihood of each reply considered separately.

Causal sample lower bound. To turn the identity into a sample lower bound, keep one designated bag negative and put total probability 32ϵ on the other $q - 1$ bags, with independent fair labels. Before order q/ϵ samples, too many of these rare bags remain unseen. Even after reading *all* witnesses, their labels remain independent fair bits, so enough prediction error remains with constant probability. A two-target rare-bag argument supplies the $\log(1/\delta)$ term. Appendix C gives the probability calculation.

VC dimension. The pad cap excludes tokens of length at least $q - 1$ from a shattered triple. Order three remaining tokens by length. If the two longest use the same list with different codes, they cannot both be positive. If one list extends the other and both tokens are positive, every target containing the longer also contains the shorter. Either case rules out a required labeling. Otherwise, their lists first differ at some position t. In every target containing both tokens, subtracting their code equations gives

$$\lambda_S = \frac{z^1_t - z^2_t}{a_{i^1_t} - a_{i^2_t}}.$$

This fixes S, and the longer token fixes all pads needed to check the third. The third token's membership is therefore fixed: labelings 111 and 110 cannot both occur. Two depth-one tokens, in contrast, can realize all four labelings by using supports containing both indices, either one, or neither, and choosing the pads accordingly. Hence $\mathrm{VC}(\mathcal{G}_q) = 2$, giving the constant-dimensional memoryless rate.

Littlestone dimension. Consider an instance learner that predicts positive only when every remaining target contains the queried token. It never makes a false positive. As long as its mistaken positive tokens lie on one prefix chain, each mistake must lengthen that chain: a longer positive token already forces every consistent shorter one. There are at most q positions to add. At the first divergence within the coded positions, the subtraction above fixes the support. Each later mistake must reveal a previously unknown pad; at most $q-2$ pads exist. Beyond the cap, diverging lists together name all q indices, fixing the support and all pads. These two phases give $O(q)$ mistakes, enough for the causal upper rate by Section 5. Appendix C sharpens the count to q and constructs a matching query tree, proving $L = q$.

Bag VC dimension. Fix full support $S = [q]$ and a full permutation $\pi_\star$. A token $e(\pi_\star, z)$ belongs to exactly one target: its list fixes the support, and its code fixes all pads. The code has $q(q-2)$ binary coordinates. For each coordinate j, form a bag B_j of these private tokens whose jth bit is one. A full-support target meets B_j exactly when its private token has that bit set. Varying the code realizes every labeling of these bags, so $v_{\mathrm{bag}} \geq q(q-2)$; counting targets gives a matching upper bound.

On A_i, the learner predicts support membership; on B_j, it predicts a code bit. The posterior identity concerns the support labels, not the code bits. The q causal lower bound and the q^2 label-only lower bound therefore use different bag distributions on the same family. Theorem C.1 in Appendix C gives the exact dimension and size bounds. In particular, with maximum bag size $b = |R_q|$, the three rate numerators scale as $1, \sqrt{\log b}, \log b$.

A countable extension. A countable disjoint union preserves VC dimension two, but a large enough block defeats any proposed causal sample size. This gives one class that is PAC learnable under every memoryless provider and not distribution-free PAC learnable under deterministic causal providers, even with finite bags (Corollary C.2).

5 Positive guarantees beyond VC dimension one

Online mistake bounds and witness spans give two further causal guarantees.

5.1 A general bound from Littlestone dimension

Littlestone dimension L, defined in Section 2, characterizes the optimal worst-case number of mistakes in realizable online learning (Littlestone, 1988). A witness lets us use this instance-level guarantee for bag prediction: every mistaken bag prediction supplies an instance on which the online predictor was wrong.

Theorem 5.1 (A general causal upper bound)**.** *For every nonempty finite family with $L = \mathrm{Ldim}(\mathcal{G})$, every $m \geq 1$ and $0 < \delta < 1$, a deterministic, generally improper learner satisfies, uniformly over causal providers, with probability at least $1-\delta$,*

$$\mathrm{err}_D(\widehat{h}_m, h_G) \leq \frac{2(L + \log(1/\delta))}{(1-e^{-1})m}. \tag{21}$$

Consequently, $O((L + \log(1/\delta))/\epsilon)$ samples suffice for $0 < \epsilon, \delta < 1$. If $L = 0$, no samples are needed.

The learning rule. Run the standard optimal instance predictor. Before each record arrives, save its induced bag predictor, which labels a bag positive if any member is predicted positive. On a mistaken training bag, update the instance predictor using one correction:

- A false negative means every member was predicted negative. The returned positive witness identifies an instance on which the predictor was wrong.
- A false positive means some member was predicted positive. Since the bag is actually negative, any such member supplies a negative instance correction.

At the end, predict by a majority vote of the m saved bag predictors, with fixed tie-breaking. No test witness is required.

Proof idea. Each correction reduces the remaining Littlestone dimension, allowing at most L bag mistakes. Since the next bag is independent, the conditional mistake probability equals the current predictor's population risk. An online-to-batch argument bounds the sum of these risks; majority voting then bounds the final risk (Cesa-Bianchi et al., 2004); see Appendix D.

The combined numerator $\min\{L, v_{\mathrm{bag}}\} + \log(1/\delta)$ is tight on the affine family. Finite L is not necessary: a disjoint union of binary-path families with unbounded depth has infinite L but retains the VC-one guarantee.

5.2 Linear subspaces: an optimal witness-span learner

Proposition 5.2 (Subspace witnesses)**.** *Let $R = \mathbb{K}^n$ over a finite field, and let $\mathcal{G}$ be all linear subspaces of dimension at most d, with $1 \le d \le n$. Then $\mathrm{VC}(\mathcal{G}) = d$ and, for $0 < \epsilon \le 1/8$ and $0 < \delta \le 1/4$,*

$$m^{\star}_{\mathrm{cau}}(\epsilon, \delta) = \Theta((d + \log(1/\delta))/\epsilon).$$

Maintaining a witness span attains the upper bound without enumerating $\mathcal{G}$.

Maintaining a witness span. Start with $V = \{0\}$. Predict a bag positive exactly when it intersects V, and add each observed positive witness to the maintained span. Since every witness is in the target, V stays inside the target and never causes a false positive. A false-negative bag misses the current span entirely. Every legal witness from that bag therefore adds an independent direction, regardless of the provider's choice. For example, if $G = \mathrm{span}\{e_1, e_2\}$ and the first witness is e_1, the next false-negative bag cannot contain a vector in $\mathrm{span}\{e_1\}$. Any witness from it completes the target span.

The learner makes at most d errors, and adding witnesses can only shrink its remaining error region. Appendix E uses these properties to bound the risk of the final span and proves a matching lower bound on singleton bags. Storing a basis requires at most nd field elements.

6 Conclusion and discussion

Fixed witness selection. Proposition 2.1 guarantees optimal memoryless sample complexity for any fixed rule choosing a legal witness from the current bag. The rule may rank witnesses by a score $s(A, r)$ with fixed tie-breaking, or sample from a fixed bag-conditioned distribution with fresh independent coins. For annotation protocols satisfying our reusable-certificate model, this allows choosing the shortest, first in a fixed order, or most salient valid explanation. All criteria must be fixed functions of the current bag, without reference to earlier annotations. Such rules incur no extra sample cost; the guarantee does not address annotation time or computation.

The VC-two separation shows how history dependence can break this guarantee. Its provider caches each bag's first witness, but that choice depends on which other bags arrived earlier. Caching therefore does not enforce (2). The fixed-rule condition is sufficient, not necessary: the VC-one and subspace results tolerate every causal provider.

Scope and limitations. We assume a known target family and exact positive certificates reusable across inputs; a witness that fails on another bag containing it violates this model. The affine construction requires very large bags. General learners are information-theoretic; the subspace learner has a compact representation.

Open questions. At latent VC dimension two with bag size at most b, the worst-case numerator lies between $\Omega(\sqrt{\log b})$ and $O(\log b)$. Can this gap be closed? Which restrictions on history dependence preserve the memoryless rate, and how do the bounds change when witnesses reveal information about unseen labels?

AI use statement

We used LLMs to assist with proving mathematical claims and with retrieving and identifying related work. It also assisted with writing: revising the manuscript's organization and exposition using related

theory papers as structural references, drafting worked examples from the stated constructions, and improving clarity and reducing repetitive phrasing. Additional assistance covered LaTeX preparation and layout checks. The authors are responsible for the mathematical claims, proofs, citations, and final text.

REFERENCES

Ishaq Aden-Ali, Mikael Møller Høgsgaard, Kasper Green Larsen, and Nikita Zhivotovskiy. Majority-of-three: The simplest optimal learner? In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of the Thirty-Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 22–45. PMLR, 2024. URL https://proceedings.mlr.press/v247/aden-ali24a.html.

Patrick Assouad. Densité et dimension. *Annales de l'Institut Fourier*, 33(3):233–282, 1983. doi: 10.5802/aif.938. URL https://www.numdam.org/articles/10.5802/aif.938/.

Shai Ben-David. 2 Notes on Classes with Vapnik–Chervonenkis Dimension 1, 2015. URL https://arxiv.org/abs/1507.05307.

Avrim Blum and Adam Kalai. A note on learning from multiple-instance examples. *Machine Learning*, 30(1):23–29, 1998. doi: 10.1023/A:1007402410823. URL https://doi.org/10.1023/A:1007402410823.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004. doi: 10.1109/TIT.2004.833339. URL https://doi.org/10.1109/TIT.2004.833339.

Victor Chepoi, Kolja Knauer, and Manon Philibert. Two-dimensional partial cubes. *The Electronic Journal of Combinatorics*, 27(3):P3.29, 2020. doi: 10.37236/8934. URL https://doi.org/10.37236/8934.

Gerald DeJong and Raymond Mooney. Explanation-based learning: An alternative view. *Machine Learning*, 1(2):145–176, 1986. doi: 10.1007/BF00114116. URL https://doi.org/10.1007/BF00114116.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71, 1997. doi: 10.1016/S0004-3702(96)00034-3. URL https://doi.org/10.1016/S0004-3702(96)00034-3.

Thorsten Doliwa, Gaojian Fan, Hans Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, VC-dimension and sample compression. *Journal of Machine Learning Research*, 15(89):3107–3131, 2014. URL https://www.jmlr.org/papers/v15/doliwa14a.html.

Sally A. Goldman and Michael J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995. doi: 10.1006/jcss.1995.1003. URL https://doi.org/10.1006/jcss.1995.1003.

Sally A. Goldman and H. David Mathias. Teaching a smarter learner. *Journal of Computer and System Sciences*, 52(2):255–267, 1996. doi: 10.1006/jcss.1996.0020. URL https://doi.org/10.1006/jcss.1996.0020.

Steve Hanneke. The optimal sample complexity of PAC learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016. URL https://www.jmlr.org/papers/v17/15-389.html.

David Haussler, Nick Littlestone, and Manfred K. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994. doi: 10.1006/inco.1994.1097. URL https://doi.org/10.1006/inco.1994.1097.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988. doi: 10.1007/BF00116827. URL https://doi.org/10.1007/BF00116827.

Philip M. Long and Lei Tan. PAC learning axis-aligned rectangles with respect to product distributions from multiple-instance examples. *Machine Learning*, 30(1):7–21, 1998. doi: 10.1023/A:1007450326753. URL https://doi.org/10.1023/A:1007450326753.

Tom M. Mitchell, Richard M. Keller, and Smadar T. Kedar-Cabelli. Explanation-based generalization: A unifying view. *Machine Learning*, 1(1):47–80, 1986. doi: 10.1023/A:1022691120807. URL https://doi.org/10.1023/A:1022691120807.

Omar Montasser. Bagging robustly learns VC classes with linear sample complexity, 2026. URL https://arxiv.org/abs/2608.13514.

Omar Montasser, Steve Hanneke, and Nathan Srebro. VC classes are adversarially robustly learnable, but only improperly. In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2512–2530. PMLR, 2019. URL https://proceedings.mlr.press/v99/montasser19a.html.

Dmitry Pechyony and Vladimir Vapnik. On the theory of learning with privileged information. In *Advances in Neural Information Processing Systems 23*, pp. 1894–1902. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/hash/c73dfe6c630edb4c1692db67c510f65c-Abstract.html.

Rattana Pukdee, Dylan Sam, J. Zico Kolter, Maria-Florina Balcan, and Pradeep Ravikumar. Learning with explanation constraints. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9c537882044c8b5352c363e840872ddb-Abstract.html.

Sivan Sabato and Naftali Tishby. Multi-instance learning with any hypothesis class. *Journal of Machine Learning Research*, 13(97):2999–3039, 2012. URL https://www.jmlr.org/papers/v13/sabato12a.html.

Michal Sharoni and Sivan Sabato. On the capacity limits of privileged ERM. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 523–534. PMLR, 2023. URL https://proceedings.mlr.press/v206/sharoni23a.html.

Ali Shirali, Rediet Abebe, and Moritz Hardt. A theory of dynamic benchmarks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=i8L9qoeZOS.

Tom Yan and Chicheng Zhang. The Human-AI substitution game: Active learning from a strategic labeler. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6cf4eecbba3f10850d487ef6bf3dabaf-Paper-Conference.pdf.

Omar Zaidan and Jason Eisner. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 31–40. Association for Computational Linguistics, 2008. URL https://aclanthology.org/D08-1004/.

Omar Zaidan, Jason Eisner, and Christine Piatko. Using "Annotator Rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 260–267. Association for Computational Linguistics, 2007. URL https://aclanthology.org/N07-1033/.

Xu Zheng, Farhad Shirani, Tianchun Wang, Shouwei Gao, Wenqian Dong, Wei Cheng, and Dongsheng Luo. PAC learnability under explanation-preserving graph perturbations, 2024. URL https://arxiv.org/abs/2402.05039.

Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Models of cooperative teaching and learning. *Journal of Machine Learning Research*, 12(11):349–384, 2011. URL https://www.jmlr.org/papers/v12/zilles11a.html.

SUPPLEMENTARY MATERIAL

Appendix A proves the memoryless benchmark. Appendices B and C prove the two main results and their extensions. Appendices D and E prove the positive guarantees. Additional model conventions and related work appear in Appendices F and G.

A THE MEMORYLESS REDUCTION

For a distribution P on labeled points and a perturbation map $\mathcal{U}$, write

$$R_{\mathcal{U}}(g; P) = \Pr_{(x,y)\sim P}[\exists u \in \mathcal{U}(x) : g(u) \neq y] .$$

A distribution P is robustly realizable by $\mathcal{F}$ if $R_{\mathcal{U}}(f^\star; P) = 0$ for some $f^\star \in \mathcal{F}$.

A.1 FROM WITNESSED BAGS TO ROBUST LEARNING

The reduction has two parts: encode each enriched record as a robustly labeled center, then lift the resulting instance predictor back to bags. The auxiliary bootstrap bound used below is proved in Section A.2.

Proof of Proposition 2.1. We verify the reduction, apply the robust bound, and prove the lower bound.

Center reduction. Take the disjoint union of R and one formal center c_z for every possible enriched record $z = (A, y, w)$. On this enlarged space define

$$\widetilde{f}_G(r) = \mathbf{1}\{r \in G\}, \qquad \widetilde{f}_G(c_{(A,y,w)}) = y.$$

The center labels are fixed across the target class: the bit y is part of the token name. Hence $\widetilde{\mathcal{F}} = \{\widetilde{f}_G : G \in \mathcal{G}\}$ has VC dimension β; only the coordinates in R vary, and their traces are exactly those of $\mathcal{G}$.

Define the public perturbation map by

$$\mathcal{U}(c_{(A,y,w)}) = \begin{cases} \{c_{(A,1,w)}, w\}, & y = 1, \\ \{c_{(A,0,\perp)}\} \cup A, & y = 0, \end{cases} \qquad \mathcal{U}(r) = \{r\}.$$

Every perturbation set contains its center and is nonempty, including when a negative bag is empty. If z is generated under target G, then $\widetilde{f}_G$ has value y throughout $\mathcal{U}(c_z)$. The induced distribution on (c_z, y) is robustly realizable by $\widetilde{\mathcal{F}}$. Under a memoryless provider it is also i.i.d.

Let P_z denote this labeled-center distribution and abbreviate $R_{\mathcal{U}}(g) = R_{\mathcal{U}}(g; P_z)$.

General latent dimension. We use the following realizable form of Theorem 1 of Montasser (2026). Let $\mathcal{F}$ be a binary class of VC dimension d and dual VC dimension $d^\star$, let $\mathcal{U}$ be any nonempty-valued perturbation map, and fix a deterministic robust empirical-risk minimization oracle for $\mathcal{F}$. If a distribution P is robustly realizable by $\mathcal{F}$, then, for $m \geq 4$, bagging $N = O(d^\star + \log(1/\delta))$ oracle outputs produces a randomized improper majority vote g such that, with probability at least $1 - \delta$ over the sample and the bootstrap draws,

$$R_{\mathcal{U}}(g; P) \leq C \frac{d + \log(1/\delta)}{m} \tag{22}$$

for a universal constant C. The dual dimension controls the number of oracle calls, not the risk bound.

Verification of the hypotheses. Because the class is finite, fix an ordering and let the oracle return the first empirical robust-risk minimizer. Every bootstrap sample drawn from the realizable center

distribution has a zero-loss member of $\widetilde{\mathcal{F}}$, so the rule is deterministic and robustly consistent. The perturbation map is finite-valued, so Corollary A.2 with $d = \beta$ gives

$$O\Big(\frac{\beta(1+\log(1/\delta))}{\epsilon}\Big)$$

examples directly. When $\beta \le 2$, this proves the upper bound in (5). For arbitrary β, Assouad's inequality $d^\star \le 2^{\beta+1} - 1$ makes the ensemble finite (Assouad, 1983); setting $P = P_z$ and $d = \beta$ in (22) gives the claimed general bound.

Lift to bags. Lift the resulting instance predictor g to bags by

$$\widehat{h}(A) = \mathbf{1}\{g(r) = 1 \text{ for some } r \in A\}. \tag{23}$$

For the analysis only, couple a fresh positive bag with an independent witness drawn from the provider's fixed kernel conditional on that bag; the predictor receives no test-time witness. Robust correctness implies $g(w) = 1$, so the lifted bag prediction is positive. On a negative record, it implies $g(r) = 0$ for every $r \in A$, so the bag prediction is negative. Pointwise in the enriched test record,

$$\mathbf{1}\{\widehat{h}(A) \ne h_G(A)\} \le \sup_{u \in \mathcal{U}(c_z)} \mathbf{1}\{g(u) \ne y\}.$$

Taking expectations and applying the corresponding robust bound proves the upper bound in (5).

Lower bound. Let $r_1, \dots, r_\beta$ be shattered by $\mathcal{G}$ and restrict the bag distribution to their singleton bags together with the empty bag. On a positive singleton the witness is forced to be its unique instance, and on a negative bag the witness field is $\perp$. The enriched record is a deterministic recoding of the ordinary labeled example. The realizable VC construction gives $\Omega(\beta/\epsilon)$ samples at constant confidence.

For the confidence term, choose a shattered instance r and targets G_0, G_1 that omit and contain it, respectively. Put probability 3ϵ on $\{r\}$ and the remaining mass on $\varnothing$. If the rare singleton is absent from the sample, the complete enriched transcripts under G_0 and G_1 coincide. For each realized output, let p be its probability of predicting one on $\{r\}$. The rare-bag contributions to its two risks are $3\epsilon p$ and $3\epsilon(1-p)$, one of which is greater than ϵ. Averaging over the learner's output and the uniform prior on the two targets gives conditional failure probability at least $1/2$. Hence one target satisfies

$$P_m^{\mathrm{mem}}(\epsilon) \ge \tfrac{1}{2}(1 - 3\epsilon)^m \tag{24}$$

for one fixed pointwise provider. This gives $\Omega(\log(1/\delta)/\epsilon)$; here $-\log(1-3\epsilon) \le 6\epsilon$ for $0 < \epsilon \le 1/4$. The maximum of the two lower bounds is within a factor two of their sum.

If $\mathrm{VC}(\mathcal{G}) = 0$, every coordinate $r \in R$ is constant across the distinct members of $\mathcal{G}$. Hence all members coincide and the target is known. □

A.2 A DIRECT CONFIDENCE BOUND FOR MEMORYLESS WITNESSES

Following the full-bootstrap leave-one-out construction of Montasser (2026), we derive the needed estimate from the consistent-rule second-moment bound of Aden-Ali et al. (2024, Lemma 5).

Lemma A.1 (Full-bootstrap robust vote)**.** *Let $\mathcal{X}$ be finite or countable, let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ have VC dimension d, and let $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}} \setminus \{\varnothing\}$ be finite-valued. Suppose that A is a deterministic rule which, on every robustly realizable finite sample, returns a member of $\mathcal{F}$ with zero empirical robust loss. For $n \ge 1$, a sample $S = (Z_1, \dots, Z_n)$, and $I = (i_1, \dots, i_n) \in [n]^n$, let $S_I = (Z_{i_1}, \dots, Z_{i_n})$ and define*

$$b_S(u) = \mathbb{E}_{I \sim \mathrm{Unif}([n]^n)}[2\mathsf{A}(S_I)(u) - 1], \qquad g_S(u) = \mathbf{1}\{b_S(u) > 0\}.$$

The expectation is an exact finite average over the n^n bootstrap sequences, so g_S is a well-defined predictor in the information-theoretic model of Section 2. There is a universal constant C_0 such that every distribution P robustly realizable by $\mathcal{F}$ satisfies

$$\mathbb{E}_{S \sim P^n} R_{\mathcal{U}}(g_S; P) \le C_0 \frac{d}{n}. \tag{25}$$

Proof. We use the following equivalent form of Aden-Ali et al. (2024, Lemma 5). Let $\mathcal{C}$ have VC dimension at most d, let $c^\star \in \mathcal{C}$, and let $\widehat{c}$ be any deterministic consistent rule. If $J \sim Q^m$ and

$$p(x) = \Pr\{\widehat{c}_J(x) \neq c^\star(x)\},$$

then, for a universal constant c_0,

$$\mathbb{E}_{x\sim Q} p(x)^2 \leq c_0 \frac{d}{m}. \tag{26}$$

Indeed, the square is the probability that two independent runs of the same consistent rule both err at x.

Fix a robustly realizable sequence $T = (Z_1, \ldots, Z_N)$, where $Z_i = (x_i, y_i)$, and let Q be uniform on its N positions. For a bootstrap index sequence $J \sim Q^m$, put

$$p_i = \max_{u\in\mathcal{U}(x_i)} \Pr\{\mathsf{A}(T_J)(u) \neq y_i\}.$$

Choose a maximizer $\phi(i) \in \mathcal{U}(x_i)$ and consider the class on $[N]$

$$\mathcal{C}_\phi = \{i \longmapsto \mathbf{1}\{f(\phi(i)) \neq y_i\} : f \in \mathcal{F}\}.$$

It has VC dimension at most d. On any shattered set of positions, ϕ is injective and $\mathcal{F}$ shatters its image after the fixed label flips. Robust realizability puts the zero function in $\mathcal{C}_\phi$, while

$$J \longmapsto (i \mapsto \mathbf{1}\{\mathsf{A}(T_J)(\phi(i)) \neq y_i\})$$

is a consistent rule for this class. Applying (26) gives

$$\frac{1}{N}\sum_{i=1}^{N} p_i^2 \leq c_0 \frac{d}{m}. \tag{27}$$

Set $N = n + 1$ and $m = n$. For each i, let $b_{T_{-i}}$ be the full-bootstrap vote trained on the sequence with position i removed, and put

$$L_i = \sup_{u\in\mathcal{U}(x_i)} \mathbf{1}\{(2y_i - 1)b_{T_{-i}}(u) \leq 0\}.$$

If $L_i = 1$, some $u \in \mathcal{U}(x_i)$ is misclassified by a random bootstrap hypothesis trained on T_{-i} with probability at least $1/2$. A length-n bootstrap from all N positions omits i with probability

$$q_n = \left(1 - \frac{1}{n+1}\right)^n \geq e^{-1}.$$

Conditional on omission it is a bootstrap from T_{-i}; if position i is present, robust consistency forces correctness at every $u \in \mathcal{U}(x_i)$. Thus $L_i = 1$ implies $p_i \geq q_n/2 \geq 1/(2e)$. Equation (27) now yields

$$\frac{1}{N}\sum_{i=1}^{N} L_i \leq 4e^2 c_0 \frac{d}{n}.$$

Finally draw $T \sim P^{n+1}$. Each T_{-i} has law P^n and is independent of Z_i. Exchangeability gives

$$\mathbb{E}_{S\sim P^n} R_{\mathcal{U}}(g_S; P) \leq \mathbb{E}_T \frac{1}{N}\sum_{i=1}^{N} L_i \leq 4e^2 c_0 \frac{d}{n},$$

which proves the claim with $C_0 = 4e^2 c_0$. □

Corollary A.2 (Direct confidence bound)**.** *Let the assumptions of Lemma A.1 hold with* $1 \leq d < \infty$. *For* $0 < \epsilon \leq 1/2$ *and* $0 < \delta \leq 1/4$, *there is a generally improper learner using*

$$O\left(\frac{d(1 + \log(1/\delta))}{\epsilon}\right)$$

examples whose robust risk is at most ϵ *with probability at least* $1 - \delta$. *When* $1 \leq d \leq 2$, *this is* $O((1 + \log(1/\delta))/\epsilon)$. *If* $d = 0$, *no examples are needed.*

Proof. If $d = 0$, all members of $\mathcal{F}$ agree pointwise, so robust realizability makes any fixed member robustly correct. Assume $d \geq 1$. Let C_0 be the constant in Lemma A.1 and take

$$n_0 = \left\lceil \frac{16 C_0 d}{\epsilon} \right\rceil.$$

A full-bootstrap candidate trained on n_0 examples has expected robust risk at most $\epsilon/16$. Markov's inequality shows that its risk is at most $\epsilon/4$ with probability at least $3/4$.

Train

$$k = \left\lceil \frac{\log(2/\delta)}{\log 4} \right\rceil$$

candidates on independent blocks. Except on an event of probability $4^{-k} \leq \delta/2$, at least one has risk at most $\epsilon/4$. On an independent validation block of size

$$v = \left\lceil \frac{12}{\epsilon} \log \frac{2(k+1)}{\delta} \right\rceil,$$

select a candidate of smallest empirical robust loss. This loss is observable because $\mathcal{U}$ is finite-valued. Conditional on the candidates, Chernoff bounds give

$$\Pr\{\widehat{R} > \epsilon/2 \mid R \leq \epsilon/4\} \leq e^{-\epsilon v/12}, \qquad \Pr\{\widehat{R} \leq \epsilon/2 \mid R > \epsilon\} \leq e^{-\epsilon v/8}.$$

When a good candidate exists, fix one of them. By the choice of v, a union bound over its first tail event and the second tail event for each bad candidate is at most $(k+1)e^{-\epsilon v/12} \leq \delta/2$. Outside this union, the fixed good candidate has empirical loss at most $\epsilon/2$, while every candidate of risk greater than ϵ has empirical loss greater than $\epsilon/2$. Together with the candidate-existence bound, an empirical minimizer has risk at most ϵ with probability at least $1 - \delta$. Finally, $kn_0 + v = O(d(1 + \log(1/\delta))/\epsilon)$ in the stated range. □

B PROOFS FOR LATENT VC DIMENSION ONE

Lemma B.1 (Compatible-split tree)**.** *Fix a reference target $o \in X$. There is a finite graph tree T with leaf set X such that every nonconstant incidence set B_r is the leaf set of one of the two graph half-trees obtained by deleting an edge. Denote that half-tree by H_r, so*

$$H_r \cap X = B_r. \tag{28}$$

Different instances may be represented by the same edge.

Proof of Lemma B.1. For each nonconstant instance r, orient its incidence split away from o:

$$L_r = \begin{cases} B_r, & o \notin B_r, \\ X \setminus B_r, & o \in B_r. \end{cases}$$

The nonempty sets L_r form a laminar family. Indeed, if all four quadrants of the two incidence splits for r and s were nonempty, four targets would realize every trace on $\{r, s\}$, contradicting $\mathrm{VC}(\mathcal{G}) \leq 1$. Complementing either side preserves the four-quadrant condition. Since $o \in (X \setminus L_r) \cap (X \setminus L_s)$, one of the other three quadrants is empty. Equivalently, L_r, L_s are nested or disjoint.

Add X and every target singleton to this laminar family, identifying duplicates. Join each proper member to its least strict superset. The resulting inclusion tree has leaves indexed by X. The edge above L_r separates L_r from $X \setminus L_r$, so the appropriate side of that edge has leaf set B_r. Constant-zero instances are never returned. A bag containing a constant-one instance is positive for every target and has empty negative cell. If $|X| = 1$, take the one-vertex tree. □

Lemma B.2 (Canonical target-side component)**.** *Let $G \in X$ and suppose A is positive under G, with $C_A \neq \varnothing$. For every legal witness $r \in A \cap G$, the half-tree H_r lies in the same component of $T \setminus Q_A$: the component containing the leaf G. This component can be recovered from (A, r) and does not depend on which legal witness is returned.*

Proof of Lemma B.2. For a legal witness r, the target leaf G belongs to B_r. Every target in C_A omits every instance in A, and in particular omits r. Thus all leaves of C_A lie in the graph half-tree opposite H_r. Their hull Q_A also lies on that side, so $H_r \cap Q_A = \varnothing$. Since H_r is connected and contains G, it lies in the component of $T \setminus Q_A$ containing G. This component depends only on (G, A).

The learner locates it from the returned witness. The public nonempty leaf set B_r is known, and the required component is the one containing any leaf of B_r, equivalently the full half-tree H_r. □

Fix the true target G and let Π be the path from o to G, ordered toward G. Every canonical subtree contains G, so its intersection with Π is a terminal subpath. The closest point $p(V)$ of any nonempty intersection V also lies on Π. This defines $\tau_G(A)$ in (14). If $K_1, \ldots, K_m$ are the canonicalized records, the point of $\cap_i K_i$ closest to o is

$$\max_{\Pi}\{o, \tau_G(A_1), \ldots, \tau_G(A_m)\}. \tag{29}$$

This follows from convexity of subtrees in a graph tree, or by intersecting their terminal subpaths on Π.

Lemma B.3 (A mistake lies beyond the gate)**.** *Let $p \in \Pi$ be used in the gate rule* (13)*. If that rule misclassifies a bag A, then $\tau_G(A) >_\Pi p$.*

Proof. Suppose first that A is negative. Then $K_G(A) = Q_A$, and $Q_A \cap \Pi = [\tau_G(A), G]$. A mistaken positive prediction says $p \notin Q_A$, hence $p <_\Pi \tau_G(A)$.

Now suppose A is positive and $C_A \neq \varnothing$. Its canonical subtree is the component of $T \setminus Q_A$ containing G. If Q_A meets Π, all its vertices on Π lie strictly before the terminal subpath $[\tau_G(A), G]$. A mistaken negative prediction has $p \in Q_A$, and again $p <_\Pi \tau_G(A)$. If $Q_A \cap \Pi = \varnothing$, a negative prediction is impossible. A bag with $C_A = \varnothing$ is declared positive and is always correct. □

Proof of Theorem 3.1. Fix a tagged bag pool $A_1, \ldots, A_n$. In every counterfactual branch, Lemma B.2 maps the enriched record at coordinate j to the same target-dependent subtree $K_G(A_j)$, regardless of the witness chosen on that branch. With coordinate i held out, the gate is

$$p_{-i} = \max_{\Pi}\big(\{o\} \cup \{\tau_G(A_j) : j \neq i\}\big).$$

By Lemma B.3, an error at i requires $\tau_G(A_i) >_\Pi p_{-i}$. Hence $\tau_G(A_i)$ is the unique strict maximum of the pool thresholds. At most one coordinate can satisfy this condition. This argument is pointwise in all provider choices.

For the i.i.d. tail, abbreviate $\tau_i = \tau_G(A_i)$ and let $p_m = \max_\Pi\{o, \tau_1, \ldots, \tau_m\}$. The τ_i are i.i.d. because they are functions only of the target and the bags. For $p \in \Pi$, put

$$\psi(p) = \Pr_{A \sim D}[\tau_G(A) >_\Pi p].$$

Lemma B.3 gives $\operatorname{err}_D(\widehat{h}_m, h_G) \leq \psi(p_m)$. If the finite path Π has a vertex with $\psi(p) > \epsilon$, let p_ϵ be the last such vertex. The event $\psi(p_m) > \epsilon$ implies $p_m \leq_\Pi p_\epsilon$, and hence $\tau_i \leq_\Pi p_\epsilon$ for every i. It has probability at most

$$\Pr(\tau_1 \leq_\Pi p_\epsilon)^m = (1 - \psi(p_\epsilon))^m < (1 - \epsilon)^m.$$

If there is no such vertex, the event is empty. This proves (6) without a continuity assumption on the threshold distribution.

For the expectation bound, append one independent bag A_{m+1}. An error on it implies that τ_{m+1} is the unique strict maximum among the $m + 1$ exchangeable thresholds. Each coordinate has the same probability of being that maximum, and the events are disjoint, so

$$\mathbb{E}\operatorname{err}_D(\widehat{h}_m, h_G) = \Pr[\widehat{h}_m(A_{m+1}) \neq h_G(A_{m+1})] \leq \frac{1}{m+1}.$$

For the lower bound, $\mathrm{VC}(\mathcal{G}) = 1$ supplies an instance r and two targets G_0, G_1 that omit and contain r. Put mass 3ϵ on the singleton bag $\{r\}$ and the remaining mass on the empty bag. If the singleton is never sampled, the enriched transcripts under G_0 and G_1 coincide, while the two targets disagree on a set of mass 3ϵ. For a realized output that predicts one there with probability p, the rare-bag

contributions to the two risks are $3\epsilon p$ and $3\epsilon(1-p)$. At least one exceeds ϵ, so averaging over outputs and the uniform target prior gives conditional failure probability at least $1/2$. Thus one target satisfies

$$\Pr[\mathrm{err}_D(\widehat{h}_m, h_G) > \epsilon] \geq \tfrac{1}{2}(1-3\epsilon)^m. \tag{30}$$

The witness on the positive singleton is forced, so this is a causal robust lower bound. The inequality $-\log(1-3\epsilon) \leq 6\epsilon$ for $0 < \epsilon \leq 1/4$, together with Equations (6) and (30) yield (8).

For expected risk, the case $m = 0$ follows directly from the same two-target prior. For $m \geq 1$, give $\{r\}$ mass $1/(m+1)$. On the event that it is unseen, the average conditional error of the two targets on this bag is at least one half. Hence one target has expected risk at least

$$\frac{1}{2(m+1)}\left(1 - \frac{1}{m+1}\right)^m \geq \frac{1}{2e(m+1)}.$$

Together with (7), this proves the expected-risk claim. □

B.1 Countable universes

Corollary B.4 (Countable VC-one classes)**.** *Let R be countable, let the bag domain consist of the finite subsets of R, and let $\mathcal{G} \subseteq 2^R$ be a known class with $\mathrm{VC}(\mathcal{G}) \leq 1$. For every $m \geq 0$, there is a query-local tree gate such that, against every causal provider,*

$$\mathbb{E}\ \mathrm{err}_D(\widehat{h}_m, h_G) \leq \frac{1}{m+1}. \tag{31}$$

For $0 < \epsilon, \delta \leq 1/2$, a deterministic confidence amplification has sample complexity

$$O\left(\frac{1+\log(1/\delta)}{\epsilon}\right). \tag{32}$$

For each fixed training transcript, the construction is a single predictor on all finite bags and receives no unlabeled test pool. If $\mathrm{VC}(\mathcal{G}) = 1$, this order is optimal for $0 < \epsilon, \delta \leq 1/4$.

Proof of Corollary B.4. Fix an enumeration of R. Given a training transcript $\mathcal{T}_m$ and a finite query bag A, form

$$U = A \cup \bigcup_{i=1}^{m} A_i, \qquad X_U = \{G \cap U : G \in \mathcal{G}\}.$$

The set U is finite and X_U is a finite class of VC dimension at most one. Choose all reference traces and tree-construction ties by the fixed enumeration, and run the finite tree gate on X_U. Its local target is $G \cap U$; since every training bag, query bag, and returned witness lies in U, the observed labels and witness legality are unchanged by this restriction.

For fixed bags and target, Lemma B.2 makes every canonical constraint independent of the legal witness returned. The whole query-local prediction is therefore independent of the provider's history. Now fix a pool $A_1, \ldots, A_n$. When coordinate i is held out and used as the query, its local universe is

$$A_i \cup \bigcup_{j \neq i} A_j = \bigcup_{j=1}^{n} A_j,$$

the same finite set for every i. All leave-one-out predictions consequently use the same finite trace class and tree. Theorem 3.1 gives at most one leave-one-out error, pointwise over the legal witnesses.

Apply this to $m+1$ i.i.d. bags, tagging each positive bag by its first legal instance in the fixed enumeration for the probabilistic argument. The leave-one-out errors are exchangeable and sum to at most one. Each coordinate therefore has expected error at most $1/(m+1)$. Witness independence lets us replace the canonical tags by the actual causal provider, proving (31).

We record the confidence amplification. It is enough to take $0 < \epsilon, \delta \leq 1/2$. Put

$$n = \left\lceil \frac{16}{\epsilon} \right\rceil, \qquad K = \left\lceil \frac{\log(2/\delta)}{\log 4} \right\rceil.$$

Train K query-local gates on disjoint consecutive blocks of n bags. Canonicalization holds simultaneously for every finite query bag, so each entire candidate predictor is a function only of its own bag block and the fixed target, not of witnesses chosen in earlier blocks. The candidate risks are therefore independent. Markov's inequality gives risk at most $\epsilon/4$ with probability at least $3/4$ for each candidate. Thus, with probability at least $1 - \delta/2$, some candidate has risk at most $\epsilon/4$.

On a fresh label-only validation block of size

$$v = \left\lceil \frac{12}{\epsilon} \log \frac{4(K+1)}{\delta} \right\rceil,$$

return a candidate of least empirical error. Conditional on the candidates, Chernoff bounds show that a candidate of risk at most $\epsilon/4$ has empirical error above $\epsilon/2$ with probability at most $e^{-v\epsilon/12}$, while a candidate of risk greater than ϵ has empirical error at most $\epsilon/2$ with probability at most $e^{-v\epsilon/8}$. A union bound makes the total validation failure at most $\delta/2$. Since $Kn + v$ has the order in (32), this proves the upper bound. The singleton rare-bag lower bound in the proof of Theorem 3.1 applies unchanged when $\mathrm{VC}(\mathcal{G}) = 1$. □

B.2 Binary paths

Proposition B.5 (Binary-path separation)**.** *For every $d \geq 1$, let R be the edges of a complete rooted binary tree of depth d, and let $\mathcal{G}$ consist of its 2^d root-to-leaf paths. Then*

$$\mathrm{VC}(\mathcal{G}) = 1, \qquad \mathrm{Ldim}(\mathcal{G}) = d, \qquad v_{\mathrm{bag}} = d. \tag{33}$$

Hence, in the range of Theorem 3.1,

$$m^{\star}_{\mathrm{lab}}(\epsilon, \delta) = \Theta\left(\frac{d + \log(1/\delta)}{\epsilon} \right), \tag{34}$$

$$m^{\star}_{\mathrm{cau}}(\epsilon, \delta) = \Theta\left(\frac{1 + \log(1/\delta)}{\epsilon} \right). \tag{35}$$

Proof of Proposition B.5. Every edge has a nonconstant trace on the path family. Given two edges, either they are comparable in the ancestor order, in which case a path cannot contain the lower edge without the upper one, or they lie in different branches, in which case no path contains both. No pair is shattered, and $\mathrm{VC}(\mathcal{G}) = 1$.

For $j \in [d]$, let A_j be the set of all right-child edges between levels $j-1$ and j. Encode a path by its left–right word $b \in \{0,1\}^d$. Then $h_{G_b}(A_j) = b_j$, so the d bags are shattered. There are exactly 2^d target classifiers, giving the reverse cardinality bound $v_{\mathrm{bag}} \leq d$.

For Littlestone dimension, query the right-child edge at the root of the underlying tree. An answer of one restricts the version space to paths in the right subtree, while an answer of zero restricts it to paths in the left subtree. Repeating this construction recursively gives a complete instance tree of depth d, so $\mathrm{Ldim}(\mathcal{G}) \geq d$. The cardinality bound $\mathrm{Ldim}(\mathcal{G}) \leq \log_2 |\mathcal{G}| = d$ gives equality. The two sample-complexity statements follow from (4) and Theorem 3.1. □

C Proofs for the affine prefix block

Throughout this section, use the field, public points, tokens, and targets of Section 4, with $k = q - 2$. We first verify witness legality and the posterior identity, then prove the three dimension bounds, and finally combine them to obtain the minimax rates.

Theorem C.1 (Affine dimensions and size)**.** *For every $q \geq 3$, the constructed class has*

$$\mathrm{VC}(\mathcal{G}_q) = 2, \qquad \mathrm{Ldim}(\mathcal{G}_q) = q, \tag{36}$$

$$q(q-2) \leq v_{\mathrm{bag}},$$

$$v_{\mathrm{bag}} \leq q(q-2) + \log_2\left(\binom{q}{2} + q + 2 \right). \tag{37}$$

The designated bags are pairwise disjoint and shattered. Moreover,

$$\log_2 |R_q| = \Theta(q^2), \qquad \log_2 |A_i| = \Theta(q^2). \tag{38}$$

C.1 WITNESS LEGALITY AND THE POSTERIOR IDENTITY

Proof of Proposition 4.2. Fix $q \geq 3$, put $k = q - 2$, and use the field, public points, tokens, and targets defined in Section 4. The map $S \mapsto \lambda_S$ is injective, and $a_i - a_j \neq 0$ for $i \neq j$.

Construction and provider. If $S \neq S'$, then the corresponding targets have different label vectors on the designated bags. If $S = S'$ but $u \neq u'$, choose a coordinate on which the pads differ and an ordered tuple from S long enough to expose it in (19). The targets are therefore distinct. Moreover,

$$A_i \cap G_{S,u} \neq \varnothing \quad \Longleftrightarrow \quad i \in S, \tag{39}$$

because the one-term tuple (i) occurs precisely when $i \in S$. The designated bags realize the full Boolean cube.

The provider records the ordered list of distinct positive designated bags as they first appear. When a new positive bag extends this list to π, it returns $e(\pi, z(\pi, S, u))$. On a repeated positive bag it replays the token used on that bag's first occurrence; a negative bag receives $\perp$. Fix an order on R_q to extend this rule to the whole bag domain. On any other positive bag the provider returns the first element of its intersection with the target, without changing the positive-discovery list. Every response is legal, and the rule is deterministic and causal on all of 2^{R_q}.

Transcript opacity. We next prove the posterior identity. Let ν be any prior on $S \subseteq [q]$. Conditional on a nonempty S, draw U uniformly from $\mathbb{F}^{r(S)}$; for the empty set use the unique empty pad. Fix a word $I \in [q]^m$ of designated-bag indices independently of (S, U), allowing repetitions, and condition on its labels. Its ordered list of distinct positive discoveries $\pi = (i_1, \ldots, i_t)$ is now fixed. Put $t_0 = \min\{t, q-2\}$. Every S consistent with the observed labels has at least t elements, so $U_1, \ldots, U_{t_0}$ exist. All nonnull witness tokens are deterministic functions of the public prefixes and

$$Z_s = U_s + a_{i_s}\lambda_S, \qquad 1 \leq s \leq t_0.$$

For every such S, the vector $(Z_1, \ldots, Z_{t_0})$ is uniform on $\mathbb{F}^{t_0}$. Discoveries beyond depth $q-2$ attach the same completed code to a longer public prefix, and repetitions replay earlier tokens. The conditional channel from S to W is therefore constant on all label sets consistent with (I, Y). Bayes' rule proves (20).

□

C.2 THE THREE DIMENSIONS

Latent VC dimension in Theorem C.1. For the VC upper bound, call a token deep if its tuple length is at least $q-1$. A deep tuple is contained in at most two label sets S. For each one, the $q-2$ displayed code equations determine every coordinate of the pad, so a deep token belongs to at most two targets. Each member of a shattered triple would occur in four distinct targets, hence no deep token can occur in such a triple.

Suppose three shallow tokens are shattered. Write them as $e(\pi^j, z^j)$ and order their lengths $d_1 \geq d_2 \geq d_3$. Shattering supplies a target containing all three. If $\pi^1 = \pi^2$, the tokens are distinct and hence $z^1 \neq z^2$, but a target contains only one code word for a fixed public tuple. This contradicts the all-one trace.

If π^2 is a strict prefix of π^1, consistency in the common target says that z^2 is the corresponding prefix of z^1. Every target containing the first token then contains the second, so the trace in which only the first of these two is present is impossible.

In the remaining case π^1 and π^2 first differ at a depth $s \leq d_2$. In every target containing both tokens,

$$z^1_s - z^2_s = (a_{i^1_s} - a_{i^2_s})\lambda_S,$$

where $i^1_s \neq i^2_s$ are the entries at their first disagreement. The coefficient is nonzero because the public points a_i are distinct. The equation determines λ_S, and the injective assignment $S \mapsto \lambda_S$ determines S. The first token then fixes $u_1, \ldots, u_{d_1}$. The third tuple has length at most d_1, so both its containment in S and all of its code equations are fixed among targets containing the first two tokens. The all-one trace shows that its fixed value is one, making the trace 110 impossible. All three cases contradict shattering.

For the matching lower bound, choose distinct $i, j \in [q]$. Pick a label set containing them and a pad so that the corresponding depth-one tokens are

$$x = e((i), u_1 + a_i\lambda_S), \qquad y = e((j), u_1 + a_j\lambda_S).$$

The target $G_{S,u}$ gives trace 11 on $\{x, y\}$. The empty target gives 00. A target in the fiber over $\{i\}$ can choose its first pad coordinate to contain x and gives 10; similarly a target over $\{j\}$ gives 01. Thus $\mathrm{VC}(\mathcal{G}_q) = 2$.

□

Littlestone dimension in Theorem C.1. For the Littlestone lower bound, let $J = \{1, \dots, k\}$ with ordered tuple $\rho = (1, \dots, k)$, and let $p_0 = q - 1$, $p_1 = q$. Choose distinct $\xi_0, \xi_1 \in \mathbb{F}$. At a node whose preceding branch labels are $b = (b_1, \dots, b_{t-1})$, $1 \le t \le k$, query

$$e\big(\rho_{\le t}, (\xi_{b_1}, \dots, \xi_{b_{t-1}}, \xi_1)\big). \tag{40}$$

For every full branch $b \in \{0, 1\}^k$ and every support

$$S \in \{J,\ J \cup \{p_0\},\ J \cup \{p_1\},\ [q]\},$$

choose the target with pads

$$u_t^{S,b} = \xi_{b_t} - a_t\lambda_S, \qquad 1 \le t \le k. \tag{41}$$

It follows from (19) that this target gives label b_t to the query at depth t. Thus all four supports remain below every leaf of the first k levels.

At such a leaf, write $\xi_b = (\xi_{b_1}, \dots, \xi_{b_k})$ and query $e((\rho, p_0), \xi_b)$. This separates the supports according to membership of p_0. On its zero branch query $e((\rho, p_1), \xi_b)$; on its one branch query $e((\rho, p_0, p_1), \xi_b)$. The last query separates the two remaining supports according to membership of p_1. We have constructed a complete instance tree of depth $k + 2 = q$, so $\mathrm{Ldim}(\mathcal{G}_q) \ge q$.

For the reverse inequality, maintain the version space consistent with all token labels observed so far. Predict a token positive only when every remaining target contains it. Since the true target remains in the version space, this rule makes no false positives; correct negative observations may shrink the version space but cannot create additional mistakes. Each mistake is therefore a positive token not forced by the preceding mistaken positive tokens.

Two token prefixes are comparable here when their public ordered tuples are comparable in the prefix order. Every mistaken token is positive under the true target, so the code word on a shorter mistaken prefix agrees with the corresponding prefix of a longer one.

Before the first pair of incomparable token prefixes, the mistaken prefixes form a strict chain. A shorter prefix cannot cause a new mistake after a longer one, because a target containing the longer token also contains its consistent ancestors. If no fork occurs, the chain has length at most q.

Suppose the first fork occurs when a new mistaken token becomes incomparable with the preceding chain. It is then incomparable with the chain's longest prefix; otherwise it would be comparable with every prefix in the chain. Let ℓ be the length of that longest prefix, and let $d_\vee$ be the larger depth of it and the new token. Let t be their first disagreement. In particular, $d_\vee \ge \ell$. A depth-q token fixes $S = [q]$ and all k pads, hence a unique target; if a fork occurs, the preceding chain therefore has $\ell \le q - 1$. If $t \le k$, the affine difference fixes λ_S and hence S, while the longer token fixes the first $\min\{d_\vee, k\}$ pads. If $d_\vee \le k$, every later mistake must expose a new pad, so the total is at most

$$\ell + 1 + (k - d_\vee) \le k + 1 = q - 1.$$

If $d_\vee > k$, the support and all pads are already fixed at the fork. It is the last mistake, and $\ell + 1 \le q$.

Two distinct ordered tuples cannot first disagree at position q, since a common prefix of length $q - 1$ has a unique unused index. Thus $t > k = q - 2$ implies $t = q - 1$. If $t > k$, the two prefixes share their first $q - 2$ entries and differ at entry $q - 1$. Those entries jointly cover $[q]$, so $S = [q]$; either token also fixes every pad. The target is again unique after the fork, and the total is at most q. Thus the conservative rule has mistake bound q. The online mistake-tree characterization of Littlestone dimension gives $\mathrm{Ldim}(\mathcal{G}_q) \le q$.

□

Bag VC dimension and size in Theorem C.1. We next calculate the bag VC dimension. The number of targets is

$$|\mathcal{G}_q| = 1 + \sum_{s=1}^{q} \binom{q}{s} |\mathbb{F}|^{\min\{s,k\}}$$
$$= 1 + \sum_{s=1}^{k-1} \binom{q}{s} |\mathbb{F}|^s + \left(\binom{q}{2} + q + 1 \right) |\mathbb{F}|^k. \tag{42}$$

Put $Q = |\mathbb{F}| = 2^q$. Since $\sum_{s=1}^{k-1} \binom{q}{s} \le Q - 1$ and $Q^s \le Q^{k-1}$ throughout the sum,

$$1 + \sum_{s=1}^{k-1} \binom{q}{s} Q^s \le 1 + (Q-1)Q^{k-1} \le Q^k.$$

This also covers $q = 3$, when the sum is empty. The cardinality bound therefore gives

$$v_{\text{bag}} \le \log_2 |\mathcal{G}_q| \le qk + \log_2 \left(\binom{q}{2} + q + 2 \right). \tag{43}$$

For the reverse inequality, fix the full permutation $\pi_\star = (1, \dots, q)$. Each token $r_z = e(\pi_\star, z)$ belongs to exactly one target: its prefix forces $S = [q]$, and its code fixes all k pads. Identify $z \in \mathbb{F}^k$ with a bit string of length qk. For every bit coordinate $j \in [qk]$, form the bag

$$B_j = \{r_z : \text{the } j\text{th bit of } z \text{ is one}\}.$$

A full-support target meets B_j exactly when the corresponding private token has bit j equal to one. These targets realize all 2^{qk} label vectors on $B_1, \dots, B_{qk}$, proving $v_{\text{bag}} \ge qk$ and hence (37).

For completeness, writing $(q)_d = q!/(q-d)!$, direct counting gives

$$|R_q| = \sum_{d=1}^{q} (q)_d \, 2^{q \min\{d,k\}}, \tag{44}$$

$$|A_i| = \sum_{d=1}^{q} (q-1)_{d-1} \, 2^{q \min\{d,k\}} = \frac{|R_q|}{q}. \tag{45}$$

The depth-q term gives a lower bound of order $q! \, 2^{q(q-2)}$ for $|R_q|$, while the whole sum is at most $q^{q+1} 2^{q(q-2)}$. The analogous bounds hold for $|A_i|$. Thus (38) follows.

□

C.3 From the hidden labels to sample complexity

Proof of Theorem 4.1. It remains to derive the i.i.d. lower bound. Put $n_{\text{rare}} = q - 1$ and $p = 32\epsilon/n_{\text{rare}}$. Give each A_i, $2 \le i \le q$, probability p, and give A_1 the remaining probability $1 - 32\epsilon$. Draw the labels on $A_2, \dots, A_q$ as independent fair bits, keep A_1 negative, and draw a uniform pad conditional on the resulting label set. This is a prior on deterministic target–provider worlds.

Suppose $m \le n_{\text{rare}}/(256\epsilon)$. If N_{seen} is the number of distinct rare bags observed, then

$$\mathbb{E}N_{\text{seen}} \le n_{\text{rare}} mp \le n_{\text{rare}}/8, \qquad \Pr(N_{\text{seen}} > n_{\text{rare}}/2) \le 1/4.$$

Condition further on the complete index word. On the event $N_{\text{seen}} \le n_{\text{rare}}/2$, put $N_{\text{unseen}} = n_{\text{rare}} - N_{\text{seen}} \ge n_{\text{rare}}/2$. Let $\mathcal{R}$ denote the learner's construction coins.

By (20), conditional on the complete enriched sample and $\mathcal{R}$, the unseen labels remain independent fair bits. For each unseen coordinate i, let α_i be the probability that the returned test-time predictor labels A_i positive, where the probability is over its fresh prediction randomization. Its conditional error is the random variable

$$E_i = \begin{cases} \alpha_i, & h_G(A_i) = 0, \\ 1 - \alpha_i, & h_G(A_i) = 1. \end{cases}$$

The variables E_i are independent, take values in $[0,1]$, and have mean $1/2$. Put $E_{\mathrm{tot}} = \sum_i E_i$, whose mean is $\mu = N_{\mathrm{unseen}}/2$. The bounded-variable Chernoff inequality

$$\Pr(E_{\mathrm{tot}} \le (1-\eta)\mu) \le \exp(-\eta^2\mu/2).$$

holds for independent summands in $[0,1]$. Writing $\Pr_\star$ for the conditional probability above and taking $\eta = 7/8$ gives

$$\begin{aligned}\Pr_\star\Big(E_{\mathrm{tot}} \le \frac{n_{\mathrm{rare}}}{32}\Big) &\le \Pr_\star\Big(E_{\mathrm{tot}} \le \frac{N_{\mathrm{unseen}}}{16}\Big) \\ &\le \exp\Big(-\frac{49N_{\mathrm{unseen}}}{256}\Big) \le \exp\Big(-\frac{49n_{\mathrm{rare}}}{512}\Big) < \frac{1}{4}.\end{aligned}$$

Averaging removes the conditioning and the construction coins. Whenever $E_{\mathrm{tot}} > n_{\mathrm{rare}}/32$, the rare bags alone contribute risk $pE_{\mathrm{tot}} > \epsilon$. Every learner therefore has average failure probability greater than $(3/4)^2 > 1/2$ under this prior. Some fixed deterministic target–provider world in its support has at least this failure probability. For $\delta \le 1/8$ this proves

$$m^\star_{\mathrm{cau}}(\epsilon,\delta) > \frac{q-1}{256\epsilon}. \tag{46}$$

For the confidence term, put mass 4ϵ on A_1 and the remaining mass on A_2. Compare the empty target with a fixed target in the fiber over $S = \{1\}$. Whenever A_1 is absent, their complete enriched transcripts coincide, while the two bag classifiers disagree on mass 4ϵ. Under the uniform prior on the two worlds, the average failure probability is at least

$$\frac{1}{2}(1-4\epsilon)^m.$$

Since $-\log(1-4\epsilon) \le 8\epsilon$ for $\epsilon \le 1/8$, this quantity exceeds δ when

$$m \le \frac{\log(1/(4\delta))}{8\epsilon}. \tag{47}$$

One fixed world attains at least the average failure probability.

For $q \ge 17$ and $\delta \le 1/8$,

$$\frac{q-1}{256} \ge \frac{q}{512}, \qquad \frac{1}{8}\log\frac{1}{4\delta} \ge \frac{1}{24}\log\frac{1}{\delta}.$$

The maximum of (46) and (47) is at least half their sum, which gives the lower half of (17). Its upper half follows from Theorem 5.1 and $\mathrm{Ldim}(\mathcal{G}_q) = q$. The memoryless rate has an upper bound from the center reduction and Corollary A.2; the two-target rare-bag argument in the proof of Proposition 2.1 gives the matching lower bound. The label-only rate follows from (4) and (37). This completes the proof of Theorem 4.1. □

C.4 A COUNTABLE CLASS THAT IS NOT CAUSALLY LEARNABLE

Corollary C.2 (A countable separation)**.** *On a countable universe with finite bags, there is a countable class of latent VC dimension two that is PAC learnable under every memoryless provider but not distribution-free PAC learnable under deterministic causal providers. For* $0 < \epsilon \le 1/8$ *and* $0 < \delta \le 1/4$*, its memoryless complexity is*

$$m^\star_{\mathrm{mem}}(\epsilon,\delta) = \Theta((1+\log(1/\delta))/\epsilon). \tag{48}$$

Proof of Corollary C.2. Let the global witness-token universe be the disjoint union of the universes R_q, $q = 17, 18, \ldots$. The global class consists of the empty target and every nonempty target from one block, extended by emptiness on all other blocks. A triple contained in one block is not shattered by Theorem C.1; a triple meeting two blocks has no all-one trace. A pair in any one block is shattered, so the global VC dimension is two.

For a target in block q, run its affine prefix provider on the designated bags of that block. On every other positive bag, return the first legal token in a fixed enumeration of the countable universe, without

changing the affine-prefix state; negative bags receive $\perp$. The lower-bound distributions only visit the designated bags of one block.

Fix any proposed sample size m. Choose $q \geq 17$ with $q > 4m + 1$, and support the bag distribution on block q. At $\epsilon = 1/64$ and $\delta = 1/8$, equation (46) gives a deterministic target–provider pair on which the learner fails with probability greater than δ. No sample size works uniformly over targets and distributions, so the countable class is not distribution-free PAC learnable.

It remains to consider memoryless providers. On the countable alphabet of finite enriched records, repeat the center-and-perturbation construction from the proof of Proposition 2.1. The resulting instance class has VC dimension two: all center labels are fixed, and the only varying coordinates are those of the original class. Under a memoryless provider, the labeled centers are i.i.d. and the induced distribution is robustly realizable by this class.

Fix an enumeration of the target class. On each robustly realizable finite sample, return the first target with zero empirical robust loss. Such a target exists; each candidate can be checked because every perturbation set is finite, so the enumeration eventually reaches one. This is a deterministic robustly consistent rule of the kind used in Lemma A.1. Corollary A.2, applied with $d = 2$, therefore gives a generally improper predictor g with robust risk at most ϵ from $O((1 + \log(1/\delta))/\epsilon)$ examples. The OR lift has bag error only when the center record has robust error, exactly as in the finite proof. This proves the upper half of (48). Its lower half already holds on any two-target subproblem of one block by the rare-bag argument in the proof of Proposition 2.1. □

D PROOF OF THE GENERAL CAUSAL BOUND

Proof of Theorem 5.1. For a nonempty version space $V \subseteq \mathcal{G}$, an instance $r \in R$, and $b \in \{0, 1\}$, write

$$V_{r,b} = \{H \in V : \mathbf{1}\{r \in H\} = b\},$$

and set $\mathrm{Ldim}(\varnothing) = -1$. Fix an ordering of R and a deterministic rule for ties. The standard optimal prediction at r is

$$a_V(r) \in \arg\max_{b \in \{0,1\}} \mathrm{Ldim}(V_{r,b}). \tag{49}$$

Starting from $V_1 = \mathcal{G}$, the predictor used before round i is

$$g_i(A) = \max_{r \in A} a_{V_i}(r), \qquad g_i(\varnothing) = 0. \tag{50}$$

The version space changes only after a bag mistake. If g_i makes a false negative, the displayed witness $w_i \in A_i \cap G$ satisfies $a_{V_i}(w_i) = 0$; set $V_{i+1} = (V_i)_{w_i,1}$. If g_i makes a false positive, choose the first instance $r_i \in A_i$ with $a_{V_i}(r_i) = 1$. The bag is negative, so $r_i \notin G$; set $V_{i+1} = (V_i)_{r_i,0}$. On a correct prediction, leave the version space unchanged.

The target remains in every V_i. At each update the new version space is the branch opposite the prediction in (49), whose Littlestone dimension is strictly below that of V_i: the predicted branch has dimension at least that of the retained branch, and the two branches form the children of an instance tree rooted at r_i or w_i. Thus, for every realized sequence of legal witnesses,

$$\sum_{i=1}^{m} M_i \leq L, \qquad M_i = \mathbf{1}\{g_i(A_i) \neq h_G(A_i)\}. \tag{51}$$

Let $\mathcal{F}_{i-1}$ contain the transcript and all provider randomness revealed before bag A_i is drawn. The predictor g_i is $\mathcal{F}_{i-1}$-measurable, whereas A_i is a fresh draw from D. Consequently, with

$$R_i = \mathrm{err}_D(g_i, h_G),$$

we have

$$\mathbb{E}[M_i \mid \mathcal{F}_{i-1}] = R_i. \tag{52}$$

Set $\gamma = 1 - e^{-1}$ and

$$Z_t = \exp\left(\gamma \sum_{i=1}^{t} R_i - \sum_{i=1}^{t} M_i\right).$$

Since M_i is conditionally Bernoulli with mean R_i,

$$\mathbb{E}[e^{-M_i} \mid \mathcal{F}_{i-1}] = 1 - R_i + e^{-1}R_i = 1 - \gamma R_i \le e^{-\gamma R_i}.$$

Therefore $(Z_t)_{t=0}^m$ is a nonnegative supermartingale with $Z_0 = 1$. Markov's inequality and (51) show that, with probability at least $1 - \delta$,

$$\sum_{i=1}^m R_i \le \frac{L + \log(1/\delta)}{1 - e^{-1}}. \tag{53}$$

Return the majority vote of $g_1, \dots, g_m$. Whenever it is wrong on a bag, at least half of the component predictors are wrong on that bag. Hence

$$\mathrm{err}_D(\widehat{h}_m, h_G) \le \frac{2}{m} \sum_{i=1}^m R_i.$$

Together with (53), this proves (21). In particular, for $0 < \epsilon, \delta < 1$,

$$m^\star_{\mathrm{cau}}(\epsilon, \delta) \le \left\lceil \frac{2e}{e-1} \frac{L + \log(1/\delta)}{\epsilon} \right\rceil. \tag{54}$$

If $L = 0$, no instance has different labels in two targets; all targets therefore induce the same bag classifier. □

E Proof for linear-subspace explanations

Proof of Proposition 5.2. Let $e_1, \dots, e_d$ be linearly independent vectors. For each $J \subseteq [d]$, the subspace $\mathrm{span}\{e_j : j \in J\}$ has dimension at most d and has trace $\{e_j : j \in J\}$ on these vectors. Thus $\mathrm{VC}(\mathcal{G}) \ge d$. Conversely, a shattered set cannot contain the zero vector, which belongs to every subspace. It cannot be linearly dependent: in a minimal dependence, one vector belongs to the span of the others, and a subspace containing the others cannot omit it. A shattered set is therefore independent. Its all-one trace must be contained in a target of dimension at most d, so it has size at most d. Hence $\mathrm{VC}(\mathcal{G}) = d$.

Let $V_1 = \{0\}$, and use $g_i(A) = \mathbf{1}\{A \cap V_i \neq \varnothing\}$ before round i. If g_i makes a false negative, update to $V_{i+1} = \mathrm{span}(V_i \cup \{W_i\})$; otherwise leave the state unchanged. By induction, $V_i \subseteq G$ at every time. Thus g_i never makes a false positive. On a false negative $A_i \cap V_i = \varnothing$, so every legal witness $W_i \in A_i \cap G$ lies outside V_i. Each mistake raises $\dim V_i$ by one, and $\sum_{i=1}^m M_i \le d$ for every sequence of legal explanations.

Write $R_i = \mathrm{err}_D(g_i, h_G)$. Prefix causality and the fresh i.i.d. bag give $\mathbb{E}[M_i \mid \mathcal{F}_{i-1}] = R_i$, using the filtration in Appendix D. The same exponential supermartingale gives, with probability at least $1 - \delta$,

$$\sum_{i=1}^m R_i \le \frac{d + \log(1/\delta)}{1 - e^{-1}}.$$

Since V_i only grows within G, its false-negative region only shrinks. Consequently $R_{m+1} \le R_i$ for every $i \le m$, and the final, proper predictor induced by V_{m+1} satisfies

$$\mathrm{err}_D(g_{m+1}, h_G) \le \frac{1}{m} \sum_{i=1}^m R_i \le \frac{d + \log(1/\delta)}{(1 - e^{-1})m}.$$

All operations are subspace-membership checks and basis updates over the finite field; a basis stores at most nd field elements. There is no need to enumerate the family of subspaces.

For the lower bound, restrict the distribution to the singleton bags of $e_1, \dots, e_d$ and the empty bag. Each positive explanation is forced, so the usual realizable VC lower bound yields $\Omega(d/\epsilon)$ at constant confidence. The two targets $\{0\}$ and $\mathrm{span}\{e_1\}$, with probability 3ϵ on $\{e_1\}$ and the remaining mass on the empty bag, give failure probability at least $\frac{1}{2}(1 - 3\epsilon)^m$ for one target, exactly as in the proof of Proposition 2.1. This supplies $\Omega(\log(1/\delta)/\epsilon)$ even for randomized improper learners. The maximum of the two terms is within a factor two of their sum. □

F Additional model details

F.1 Minimax risk and randomization

The predictor may use fresh test-time randomization. Its risk is

$$\mathrm{err}_D(\widehat{h}, h_G) = \Pr_{A\sim D,\,\widehat{h}}[\widehat{h}(A) \neq h_G(A)].$$

Randomness used to construct the predictor remains in the outer failure probability. For causal selection, define

$$P_m(\epsilon) = \inf_{L_m} \sup_{D} \sup_{G\in\mathcal{G}} \sup_{\sigma\in\Sigma(G,D)} \Pr\{\mathrm{err}_D(L_m(\mathcal{T}_m), h_G) > \epsilon\}, \tag{55}$$

and

$$m^\star_{\mathrm{cau}}(\epsilon,\delta) = \min\{m : P_m(\epsilon) \le \delta\}. \tag{56}$$

The curve is nonincreasing: the learner can ignore its last record, and the prefix of a causal provider is again causal. Restricting the provider supremum to memoryless kernels defines P^{mem}_m and $m^\star_{\mathrm{mem}}$. Omitting witnesses defines the label-only complexity $m^\star_{\mathrm{lab}}$.

F.2 Bounded bag size

For $\beta \ge 1$, Sauer's lemma also gives $L \le \log_2|\mathcal{G}| \le \beta\log_2(e|R|/\beta)$. The countable extensions use finite bags on a countable universe; their observation and query domains are countable, so measurability is automatic.

On the restricted domain $\mathcal{A}_b = \{A \subseteq R : |A| \le b\}$, write $v_{\mathrm{bag}}(b) = \mathrm{VC}(\{h_G|_{\mathcal{A}_b} : G \in \mathcal{G}\})$. For $b \ge 1$, the classical MIL bound gives

$$v_{\mathrm{bag}}(b) \le \max\{16, 2\beta\log_2(2eb)\}. \tag{57}$$

On 2^R, one may take $b = |R|$ (Sabato & Tishby, 2012).

G Further related work

Explanation-based learning. Classical EBL generalizes an explanation of an example into sufficient conditions applicable to future instances (Mitchell et al., 1986; DeJong & Mooney, 1986). This reusability is a conceptual antecedent of our certificate semantics. In the EBG formulation, a domain theory and an operationality criterion guide proof construction and generalization. We assume an already-valid witness within a fixed certificate relation and ask how its selection affects PAC sample complexity. Our abstraction captures reusable sufficiency, without modeling EBL's proof-generalization procedure.

Learning from annotator explanations. Rationales supply training-time evidence (Zaidan et al., 2007), and modeling their selection can improve inference (Zaidan & Eisner, 2008). Explanation constraints restrict the predictor class (Pukdee et al., 2023). Zheng et al. (2024) analyze PAC learning through explanation-preserving graph perturbations, including benefits and risks of augmented data. We hold the bag distribution fixed and study selection among valid training certificates; arbitrary rationales need not satisfy this semantics.

Bags and privileged information. The prediction task is the OR model of multiple-instance learning (Dietterich et al., 1997; Blum & Kalai, 1998; Sabato & Tishby, 2012). Long & Tan (1998) give polynomial sample and time bounds for learning axis-aligned rectangles when bag members are drawn independently from a product distribution over instance coordinates. Our bags may have arbitrary internal dependence and range over an arbitrary bag distribution; the additional observation is a selected positive witness. Such witnesses are structured privileged information (Pechyony & Vapnik, 2010; Sharoni & Sabato, 2023), certifying positivity wherever the same instance occurs.

Teaching and cooperation. Teaching dimension measures the smallest labeled sets that identify a target (Goldman & Kearns, 1995); recursive teaching exploits a hierarchy of successively reduced concept classes (Zilles et al., 2011; Doliwa et al., 2014). Doliwa et al. (2014) show, in particular, that recursive teaching dimension is at most one for finite classes of VC dimension one. Our VC-one result concerns every legal witness choice, rather than the existence of a small cooperative teaching set. The provider selects only within an arriving i.i.d. bag, cannot choose the bags, and need not help; the objective is PAC prediction rather than exact identification. Collusion-freeness constrains cooperative teacher–learner pairs: Goldman & Mathias (1996) require successful identification to survive adding target-consistent examples. Our minimax formulation instead asks what is guaranteed without cooperation. These address complementary concerns about selection, but legality and worst-case selection do not impose the collusion-freeness criterion or establish a formal duality.

Robust and sequential learning. Robust-learning tools yield the memoryless benchmark (Montasser et al., 2019; Montasser, 2026); a direct argument based on Aden-Ali et al. (2024) already suffices at VC dimensions one and two. These i.i.d. guarantees do not directly apply to a causal transcript. Our VC-one learner uses classical tree representations (Ben-David, 2015; Chepoi et al., 2020), with witness canonization providing the new selection-independent constraint and PAC analysis. The general causal bound uses online mistake control (Littlestone, 1988; Cesa-Bianchi et al., 2004).