ON LEARNING WITH AMBIGUOUS LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study learning in the situation where each training example comes with a set of at most ℓ possible labels containing the correct one. The learner may also predict up to $L \geq \ell$ labels for each new input, and makes an "error" if its prediction omits the correct label. For countable input and label spaces and any fixed $\ell \geq 2$, we identify a combinatorial "dimension" whose finiteness guarantees that such learning is possible.

We then obtain upper and lower bounds on sample complexity that match up to logarithmic factors in the worst case for every fixed finite $L \geq \ell$. At a fixed sufficiently high success probability and for sufficiently small ϵ, the required number of examples grows, up to logarithmic factors, linearly with this "dimension" and with $1/\epsilon^{\ell}$, where ϵ is the desired error. We also obtain a separation between exact and ambiguous labels, by exhibiting a single class for which exact training labels allow learning from $O_{\ell}(1/\epsilon)$ examples, while ambiguous labels require $\Omega_{\ell,L}(1/\epsilon^{\ell})$ examples.

Finally, when training label sets omit the correct label with probability ρ, we achieve error at most $\rho + \epsilon$ with high probability, without knowing ρ, by allowing prediction width $O(\ell^2/\epsilon)$ and using a sample size nearly linear in $1/\epsilon$ at fixed dimension and confidence. Under the weaker assumption of finite ℓ-list DS dimension alone, the worst-case necessary and sufficient prediction width is $\Theta(\ell^2/\epsilon)$, with absolute constants, for sufficiently small error and sufficiently high fixed success probability.

1 INTRODUCTION

What does uncertainty in training labels cost when predictions may also express uncertainty? Suppose each observed input comes with a small set containing its correct label, and the learner may return a short list of labels at each new input. Unique identification can be impossible: at one input, the always-observed bag $\{0, 1\}$ has the same law whether the target label is zero or one. A list can preserve this ambiguity and still cover the target. We ask whether allowing lists also removes the statistical cost of ambiguous observations.

We study iid pairs of inputs and bags on countable input and label spaces, with a fixed target in a known class. Clean bags contain that target's label and have size at most $\ell \geq 2$, while predictions may contain up to $L \geq \ell$ labels. The learner returns one function, with all its randomness fixed after training; its error is the probability of omitting the target label. The learner receives no exact-label validation, and no bag is supplied at prediction time.

1.1 MAIN RESULTS

The statistical cost depends on both widths. To define our structural parameter, we fix $\ell - 1$ reference hypotheses, remove their labels, and retain a target value only where it and the reference values are distinct. The supremum of the resulting partial classes' DS dimensions is the reference-deletion dimension $\Delta_{\ell}(H)$, defined precisely in Section 2. Finite deletion dimension suffices for learning at every prediction width $L \geq \ell$.

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.

For every $L \geq \ell$, $1 \leq D = \Delta_\ell(H) < \infty$, and $0 < \epsilon, \delta < 1/4$, Theorem 3.1 gives a clean learner with error at most ϵ and failure probability at most δ, using

$$m \leq C_\ell \left[\left(\frac{D\epsilon^{-\ell}}{\log(1 + L/\ell)} + \frac{D}{\epsilon} \right) \log^2 \frac{eD}{\epsilon} + \frac{1}{\epsilon} \log \frac{2}{\delta} \right].$$

Here C_ℓ is independent of L. Increasing the prediction width reduces the leading ambiguity term, while the seed and confidence costs remain. At every fixed finite $L \geq \ell$, the accuracy exponent is ℓ up to logarithmic factors.

A second construction gives inverse-linear dependence on accuracy at a polynomial output width. Theorem 3.2 gives

$$L \geq \left\lceil \frac{384\ell^2}{\epsilon} \right\rceil, \qquad m \leq \frac{C\ell}{\epsilon} \left[D \log^2 \frac{e\ell D}{\epsilon} + \log \frac{6}{\delta} \right],$$

with an absolute constant C. If the bags omit the fixed target with probability ρ_h, the same learner has risk at most $\rho_h + \epsilon$, without knowing ρ_h. The clean guarantee is the case $\rho_h = 0$.

Our lower bound uses one fixed countable class of exact deletion dimension D, chosen before the output width, accuracy, and confidence. For every finite $L \geq \ell$, Theorem 5.1 requires

$$m \geq c_\ell \left[\frac{D}{L^\ell \epsilon^\ell} + \frac{D + \log(1/\delta)}{\epsilon} \right] \qquad \left(\epsilon L^2 \leq \frac{1}{102400}, \quad \delta \leq \frac{1}{100} \right).$$

The labels of this class form a countable alphabet. In particular, the exponent ℓ remains necessary for every fixed finite output width, while the displayed accuracy restriction must be retained when L grows.

Under the weaker assumption of finite ℓ-list DS dimension, the worst-case width threshold is exactly $\Theta(\ell^2/\epsilon)$, with absolute constants, for $0 < \epsilon < 1/64$ and any fixed failure probability below $1/6$ (Theorem 6.1). Here we allow any finite distribution-free sample bound. The lower example has infinite deletion dimension, so it does not establish the same width necessity for classes of finite deletion dimension.

For the general upper bound, we enumerate candidate labeled decoder inputs from bags and complete each narrow decoder separately, keeping their description costs additive. A larger update rate in the observable prediction game gives the logarithmic gain in L. The large-width construction instead accumulates fractional coverage from the decoded menus and rounds the resulting function. The statistical lower proof uses incidence systems to keep deletion dimension finite while leaving target labels uncertain at unobserved inputs; each prior is converted to one fixed clean target and observation law. We make no computational-efficiency claim for the general constructions.

1.2 Related work

Moses Charikar and Chirag Pabbaraju (Charikar & Pabbaraju, 2023) characterize exact-label list learnability; our weak learner uses a later quantitative bound (Pabbaraju, 2026). Partial-concept learning and multiclass disambiguation were studied by Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran (Alon et al., 2022), and by Jingyuan Xu, Xin Zou, Xiuwen Gong, and Weiwei Liu (Xu et al., 2026). Our deletion construction uses the partial-class viewpoint. Exactly labeled examples of a deleted target are unavailable to the bag learner, so sufficiency requires a construction using the observed bags.

Superset, fuzzy, and indirect supervision have established theories due to Liping Liu and Thomas Dietterich (Liu & Dietterich, 2014), Andrea Campagner (Campagner, 2024), and Kaifu Wang, Qiang Ning, and Dan Roth (Wang et al., 2020). Our list objective allows persistent distractors without requiring unique identification. Bags also encode demonstrated set-valued actions with target-containment reward in the framework of Nirmit Joshi, Gene Li, Siddharth Bhandari, Shiva Prasad Kasiviswanathan, Cong Ma, and Nathan Srebro (Joshi et al., 2026). The finite-class guarantees in that framework overlap with our setting.

Demonstrations and their predictions. There is an exact action encoding into learning from correct demonstrations (Joshi et al., 2026). For observation width a, take actions $\mathcal{A}_a = \{A \subseteq Y : 1 \leq |A| \leq a\}$ and rewards

$$r_h(x, A) = \mathbf{1}\{h(x) \in A\}.$$

An observed bag is a demonstrated action. The optimal reward at every input is one, so the demonstration's expected reward shortfall is exactly ρ_h.

Section 2 defines the model and dimension. Section 3 states the upper bounds and describes the construction from observed bags. Sections 4, 5, and 6 present the large-width argument, statistical separation, and prediction-width threshold. Appendices A and B contain the structural ingredients and deferred upper proofs. Appendices C and D give the shared lower-bound arguments and finite-list-dimension threshold proof.

2 THE MODEL AND DELETION DIMENSION

We work with nonempty countable sets X and Y, and a nonempty class $H \subseteq Y^X$ of total functions. The class H need not be countable. Fix an observation width $\ell \geq 2$ and a prediction width $L \geq 1$, both integers. A target $h \in H$ and a joint law Q of (Z, B) are chosen before sampling, where $Z \in X$ and $B \subseteq Y$ satisfies $1 \leq |B| \leq \ell$. The learner receives m independent observations (Z_i, B_i) from Q. We call B_i an observed bag; the learner does not know which of its labels, if any, equals $h(Z_i)$.

Write $\mu = Q_X$ for the input marginal. A prediction rule returns one function $M : X \to 2^Y$ with $|M(x)| \leq L$ at every input, and its risk and the omission probability of the observations are

$$R_h(M) = \Pr_{Z\sim\mu}\{h(Z) \notin M(Z)\}, \qquad \rho_h = \Pr_{(Z,B)\sim Q}\{h(Z) \notin B\}. \tag{1}$$

All randomness used by the learner is fixed with M. In particular, repeated queries at the same input return the same list, and no bag is supplied at prediction time. The learner may depend on H, the widths, and the accuracy and confidence parameters; it is not given h, Q, ρ_h, or correctness indicators for the observed labels. Distractors and omissions may depend arbitrarily on the input, subject to the fixed law Q.

A law is clean for h if $\rho_h = 0$. We say that H is (ℓ, L)-learnable if there is a family of learning rules and, for every $\epsilon, \delta \in (0, 1)$, a finite sample bound independent of h and Q such that $\Pr\{R_h(M) > \epsilon\} \leq \delta$ for every clean law. Let $m_H^{(\ell,L)}(\epsilon, \delta)$ be the least sufficient sample size. The two widths are independent parameters; a theorem may also prescribe L as a function of accuracy. We use (a, b) for the same convention when a local argument has observation width a and prediction width b. Bounds involving the actual ρ_h mean that a prescribed rule succeeds whenever its supplied sample budget exceeds the stated expression. Such a bound does not provide a stopping rule that estimates ρ_h. A known upper bound on ρ_h, including 1, gives a uniform sufficient budget.

The case $L < \ell$ is treated in Appendix A.3. Unless stated otherwise, the remaining results concern $L \geq \ell$.

The structural parameter in our upper bounds is determined by the observation width. Fix $\ell - 1$ reference functions $F = (f_1, \ldots, f_{\ell-1}) \in H^{\ell-1}$. On an input where these reference labels and $h(x)$ are distinct, the bag consisting of all ℓ labels is compatible with each reference and with h. The deletion parameter below supplies a sufficient condition and quantitative bounds for general L; its necessity is discussed in Remark 3.3. We define the partial functions using a symbol $\perp \notin Y$ for an undefined value:

$$h_F(x) = \begin{cases} h(x), & f_1(x), \ldots, f_{\ell-1}(x), h(x) \text{ are pairwise distinct,} \\ \perp, & \text{otherwise,} \end{cases} \qquad H_F = \{h_F : h \in H\}. \tag{2}$$

For example, when $\ell = 2$ and the reference value is 2, deletion preserves candidate values 0 and 1 but makes candidate value 2 undefined.

For a partial class $P \subseteq (Y \cup \{\perp\})^X$, a fully defined trace on distinct inputs $x_1, \ldots, x_n$ is a vector $(p(x_1), \ldots, p(x_n)) \in Y^n$ for some $p \in P$. The k-list DS dimension $d_{k\mathrm{DS}}(P)$ is the supremum of those n for which there is a finite nonempty family V of fully defined traces such that, for every $v \in V$ and every coordinate i, at least k other vectors in V differ from v only at i. Write $d_{\mathrm{DS}} = d_{1\mathrm{DS}}$. The dimension is zero if no positive n is possible, and infinity if such n are unbounded. Undefined values never serve as labels in these witnesses.

The reference-deletion dimension is

$$\Delta_\ell(H) = \sup_{F\in H^{\ell-1}} d_{\mathrm{DS}}(H_F). \tag{3}$$

The supremum requires a uniform bound over reference tuples. At its zero endpoint, each input has at most ℓ labels realized by H, so the fixed list $M(x) = \{h(x) : h \in H\}$ has zero risk without observations, even when bags omit the target. Indeed, $\ell + 1$ labels at one input would leave two distinct labels after deleting $\ell - 1$ references, giving a singleton DS witness. All logarithms below are natural.

3 Learning with general output width

Finite deletion dimension suffices for every prediction width $L \geq \ell$. Increasing L improves the sample bound, and a width of order ℓ^2/ϵ gives inverse-linear dependence on accuracy, up to logarithms.

Theorem 3.1 (Clean upper bound for general output width)**.** *Fix integers $2 \leq \ell \leq L$ and a class H with $1 \leq D = \Delta_\ell(H) < \infty$. For $0 < \epsilon, \delta < 1/4$,*

$$m_H^{(\ell,L)}(\epsilon, \delta) \leq C_\ell \left[\left(\frac{D\epsilon^{-\ell}}{\log(1 + L/\ell)} + \frac{D}{\epsilon} \right) \log^2 \frac{eD}{\epsilon} + \frac{1}{\epsilon} \log \frac{2}{\delta} \right]. \tag{4}$$

The constant C_ℓ depends only on ℓ, uniformly in L. If $D = 0$, no observations are needed.

Theorem 3.2 (The ℓ^2/ϵ sufficient width)**.** *Let $\ell \geq 2$, $0 < \epsilon, \delta < 1/4$, and $1 \leq s = d_{\ell\mathrm{DS}}(H) < \infty$. Whenever the integer L satisfies the width condition below, there is a learner using m observations within the displayed bound, for an absolute constant C:*

$$L \geq \left\lceil \frac{384\ell^2}{\epsilon} \right\rceil, \qquad m \leq \frac{C\ell}{\epsilon} \left[s \log^2 \frac{e\ell s}{\epsilon} + \log \frac{6}{\delta} \right] \tag{5}$$

The learner returns one width-L function with $\Pr\{R_h(M) > \rho_h + \epsilon\} \leq \delta$, for every fixed target and iid observation law. The learner is not given ρ_h. If $s = 0$, no observations are needed. In particular, since $s \leq \Delta_\ell(H)$, the theorem holds with s replaced by any finite bound $D \geq 1$ on the deletion dimension.

For clean bags, Theorem 3.2 gives $\widetilde{O}_\ell(D/\epsilon)$ samples at width $O_\ell(1/\epsilon)$. Theorem 3.1 is proved in Appendix B.7. Theorem 3.2 is proved in Section 4.

For every fixed $L \geq \ell$ and fixed sufficiently high confidence, Theorem 5.1 shows that the accuracy exponent ℓ remains necessary in the worst case. Its dependence on L is stated explicitly. We make no claim of computational efficiency for the general learners.

Remark 3.3. When $L = \ell$, clean learnability is equivalent to $\Delta_\ell(H) < \infty$: sufficiency follows from Theorem 3.1, and necessity from Proposition A.5. Necessity fails for larger output widths. On a countably infinite domain, all maps into a fixed set of $L > \ell$ labels admit a zero-data width-L predictor, but deleting $\ell - 1$ constant reference labels leaves binary cubes of arbitrary size, so their deletion dimension is infinite.

3.1 The construction from observed bags

For the clean upper bound, put $b = \max\{\ell, \lfloor L/2 \rfloor\}$. A fractional prediction assigns weights $p_y(x) \in [0, 1]$ with $\sum_y p_y(x) \leq b$. Its loss at target label $h(x)$ is $1 - p_{h(x)}(x)$. Appendix B.7 gives the proof, using the structural and fractional ingredients in Appendices A and B. We first reduce the labels that need to be considered, then treat large shattered sets without observations and learn on the remaining inputs. Each reduction keeps every candidate that may be needed for the analysis.

An exact-label learner can compress a finite sample into $O_\ell(D \log n)$ retained labeled examples. We cannot apply its encoder directly to bags, because the encoder uses the correct labels. We can, however, enumerate all ordered choices of observed examples and labels from their bags and apply the fixed decoder to each choice. This gives functions $\nu_1, \ldots, \nu_J$, each predicting at most ℓ labels, with $\log J = O_\ell(D \log^2 n)$. On a clean sample, one of these functions covers every hidden sample label. The compression generalization argument shows that, for a sufficiently large seed sample, one individual decoder also misses little target mass. Its identity is never supplied to the learner.

Fix the decoders and a tuple of remaining input locations. Retain a target value only where it belongs to the union of the decoded lists, marking other values undefined. This gives a finite partial trace

class on the tuple. On a set of K inputs, a Natarajan witness consists of two labels A_i at each input such that every choice from the K pairs is realized by a fully defined trace. We remove disjoint sets with such witnesses until the residual Natarajan dimension is less than K.

The deletion condition controls prediction on each removed block. For a distribution on its partial traces, take A_i and the $\ell - 2$ heaviest remaining labels as the prediction at input i, and let r_i be the probability of an actual label outside this list. Draw $\ell - 1$ reference traces independently. The probability that their labels at i are defined, distinct, and outside A_i is at least $r_i^{\ell-1}$. For any fixed reference tuple this can happen at at most D coordinates, since otherwise the binary block witness would give a deleted Natarajan witness of size greater than D. Hence

$$\sum_{i=1}^{K} r_i^{\ell-1} \le D, \qquad \frac{1}{K}\sum_{i=1}^{K} r_i \le (D/K)^{1/(\ell-1)}. \tag{6}$$

Finite minimax supplies a fractional prediction with this average defined-target loss for every trace. It uses only the partial class and the input locations, so no bag from the block is needed.

On the residual, we need finitely many total labelings of its indexed positions that extend its partial traces. A correction-transcript argument supplies such a family, but completing against the union of all decoded lists would make its complexity depend on the large union alphabet. Instead, for each j we restrict to labels in ν_j and complete this narrower partial class separately. After harmless padding, each coordinate has at most $\ell + 1$ available labels, and the resulting family $\mathcal{F}_j$ satisfies $\log |\mathcal{F}_j| = O_\ell(1 + K \log^2(2n))$. Only then do we take the union:

$$\log \left| \bigcup_{j=1}^{J} \mathcal{F}_j \right| \le \log J + \max_j \log |\mathcal{F}_j|. \tag{7}$$

This keeps the decoder cost additive. A completion associated with a good decoder agrees with the target wherever that decoder contains it, and all completions remain available to the learner.

These labelings serve as internal experts and may differ at positions carrying the same input. Their losses are computed from observed bags. Put $\tau = \ell/(b + 1)$. If f_i is an expert's label at position i, its gain there is

$$a_{f,i} = \mathbf{1}\{f_i \in B_i\}(1 - p_{f_i}(x_i)) - \tau \mathbf{1}\{f_i \notin B_i\}.$$

For any current expert weights, a finite minimax argument chooses p before observing B_i so that the weighted gain is nonpositive for every legal bag. We use the fixed update rate $\eta = (1 - \tau)/(2\tau)$ and multiply each weight by $1 + \eta a_{f,i}$. The logarithmic update bounds every expert's cumulative valid gain in terms of its bag omissions and the logarithm of the family size. The target and a good completion are compared only in the proof; their disagreement on missed decoder labels is charged explicitly.

To make a prediction at a queried input, we draw the decoder seed from the other observations. We form the block decomposition and completion families from the input tuple, including the query, without its bag. On the residual we run a random prefix of a random ordering, with the query placed next. Every prefix uses the same full expert family, initialization, and update rate. Averaging over these choices gives a symmetric fractional rule whose leave-one-out loss is bounded for every clean indexed state. The finite-population decoder estimate in Lemma B.3 remains valid even when bags are corrupted.

Choose K of order $D\epsilon^{-(\ell-1)}$ to control (6). The logarithmic gain per update is $\log(1 + \eta) = \Theta_\ell(\log(1 + L/\ell))$, yielding the improvement in the leading ambiguity term of (4). Reverse and forward concentration estimates give the additive $\epsilon^{-1} \log(e/\delta)$ confidence cost. Finally, we return the L largest positive fractional weights. An omitted label has weight at most $b/(L+1)$, so rounding multiplies fractional loss by at most $(L + 1)/(L + 1 - b) \le \ell + 1$. Appendix B.7 supplies the full calculation, including the separate decoder-seed cost.

4 Proof of the large-width upper bound

We prove Theorem 3.2. Throughout this section, ℓ is the maximum observed bag size and L is the returned list width. The clean proof in Appendix B.7 keeps constants depending on ℓ uniform

in L. The proof below keeps all numerical constants absolute, using the exact-label list dimension $s = d_{\ell\mathrm{DS}}(H)$.

4.1 Fractional coverage of a finite family

For the concentration bounds, put $u = \log(6/\delta)$.

We give the finite-family argument for an arbitrary iid law P on exact pairs (X, Y); no realizability is used here. Fix width-q menus $\nu_1, \dots, \nu_J$, an integer horizon T, and $0 < \gamma \le 1$. Initialize the menu weights uniformly and set $p_0(x, y) = 0$. At round t, select I_t according to the current normalized weights $w_{t,j}$, independently of the fresh pair conditional on the past. Use rewards

$$r_t(j) = \mathbf{1}\{Y_t \in \nu_j(X_t)\}(1 - p_{t-1}(X_t, Y_t)),$$

update each weight by $\exp(r_t(j)/2)$, and update the entire fractional function by

$$p_t(x, y) = p_{t-1}(x, y) + \gamma \mathbf{1}\{y \in \nu_{I_t}(x)\}(1 - p_{t-1}(x, y)). \tag{8}$$

This is a function update at every input, represented by the selected menu descriptions. All coordinates remain in $[0, 1]$, are nondecreasing in t, and satisfy $\sum_y p_T(x, y) \le \gamma q T$.

Put

$$a_t = \sum_j w_{t,j} r_t(j), \qquad z_t = r_t(I_t), \qquad d_t = \mathbb{E}_P[\mathbf{1}\{Y \in \nu_{I_t}(X)\}(1 - p_{t-1}(X, Y))].$$

For every realized path, population gains telescope:

$$\gamma \sum_{t=1}^{T} d_t = \mathbb{E}_P[p_T(X, Y)] \le 1. \tag{9}$$

The chord bound $e^{r/2} \le 1 + (e^{1/2} - 1)r \le 1 + r$ for $r \in [0, 1]$ gives the deterministic multiplicative-weights inequality

$$\sum_t r_t(j) \le 2 \log J + 2 \sum_t a_t \quad \text{for every } j. \tag{10}$$

For one comparator j fixed before the fresh stage, the concentration argument in Appendix B.10 gives

$$\mathbb{E}_P[\mathbf{1}\{Y \in \nu_j(X)\}(1 - p_T(X, Y))] \le \frac{4 \log J + 16/\gamma + 20u}{T} \tag{11}$$

with failure probability at most $3e^{-u}$. Direct summation gives $18u$, so the displayed constant is conservative. Only the one comparator fixed by the seed is used in the third event; there is no further union over menus or targets.

4.2 Bag transfer, rounding and the uniform sample bound

Proof of Theorem 3.2. Assume $s \ge 1$ and construct the seed family in Appendix B.9, with seed size n_1 from (46) and code length $r = r_{n_1}$ from Lemma B.7.

For a fixed target, define the population valid-miss mass

$$V_h(\nu) = Q\{h(X) \in B,\ h(X) \notin \nu(X)\}.$$

Thus, with probability at least $1 - \delta/2$, one seed-selected analytical comparator satisfies $V_h(\nu) \le \epsilon/6$. It is fixed before the independent second stage; its identity is not supplied to the learner.

On each fresh bag, choose one label uniformly and independently; these pairs have a fixed iid law P. Set $q = 2\ell$, retain $u = \log(6/\delta)$, put $\overline{J} = 1 + (\ell n_1)^r$, and take

$$K_0 = 4 \log \overline{J} + 16 + 20u, \qquad T = \lceil 6\ell K_0/\epsilon \rceil, \qquad \gamma = \min\{1, L/(2qT)\}. \tag{12}$$

The sample sizes of both stages are thus fixed before observing the seed, and $\log J \le \log \overline{J}$ in the coverage bound. The update (8) gives mass at most $L/2$. Return the list of up to L largest positive coordinates of p_T, with fixed ties. For any omitted label,

$$\mathbf{1}\{y \notin M(x)\} \le 2(1 - p_T(x, y)). \tag{13}$$

Indeed, an omitted positive entry is at most $L/[2(L+1)]$, and an omitted zero entry satisfies the inequality as well.

For every realized fixed comparator ν and output M,

$$R_h(M) \le \rho_h + V_h(\nu) + \ell P\{Y \in \nu(X),\ Y \notin M(X)\}. \tag{14}$$

Split a target omission into invalid bags, valid comparator misses, and the remaining valid omissions. On the last event the uniform label from the bag equals the omitted target with probability at least $1/\ell$, proving the inequality. It is a population comparison between fixed functions and applies after conditioning on their training data and tape.

On the good seed and coverage events, apply (13) only to the last term of (14). Since

$$\frac{1}{\gamma T} = \max\{1/T, 2q/L\} \le 1/T + 2q/L,$$

equation (11) yields

$$R_h(M) \le \rho_h + \frac{\epsilon}{6} + \frac{2\ell K_0}{T} + \frac{64\ell q}{L}. \tag{15}$$

For $L \ge \lceil 384\ell^2/\epsilon \rceil$ and the horizon in (12), the last two terms are each at most $\epsilon/3$. The excess over ρ_h is therefore at most $5\epsilon/6$. The combined failure probability is at most $\delta/2 + 3e^{-u} = \delta$. In particular, the learner need not know ρ_h, and its coefficient is one because only the observed-label residual was rounded.

The sample accounting is given in Appendix B.11.

At each input, at most qT coordinates of p_T are positive. All selected menus and sampled bag labels are part of the training tape, and their stored update sequence defines a single function on all inputs. Prediction uses no new random choices and no query bag. This also verifies the countable-label and repeated-input conventions. If $s = 0$, the pointwise target-label menu has size at most ℓ and yields zero risk without observations. □

5 THE STATISTICAL COST OF AMBIGUITY

For the lower bounds, we use classes fixed before the prediction width, accuracy, confidence, and sample size are chosen. The construction below separates exact labels from ambiguous observations at every fixed finite prediction width.

Theorem 5.1 (Sample lower bound at general prediction width)**.** *For all integers $\ell \ge 2$ and $D \ge 1$, there is a countable class $H_{\ell,D}$ on countable input and label spaces, independent of L, ϵ, δ, with $\Delta_\ell(H_{\ell,D}) = D$, such that*

$$m^{(\ell,L)}_{H_{\ell,D}}(\epsilon,\delta) \ge \frac{1}{6\,800^\ell}\left[\frac{D}{L^\ell \epsilon^\ell} + \frac{D + \log(1/\delta)}{\epsilon}\right] \tag{16}$$

for every integer $L \ge \ell$, $0 < \epsilon \le 1/(102400L^2)$, and $0 < \delta \le 1/100$. On the same class, for every finite $L \ge \ell$, $0 < \epsilon \le 1/8$, and $0 < \delta \le 1/100$,

$$m^{(\ell,L)}_{H_{\ell,D}}(\epsilon,\delta) > \frac{D}{32\epsilon}, \qquad m^{(\ell,L)}_{H_{\ell,D}}(\epsilon,\delta) \ge \frac{\log(1/\delta)}{8\epsilon}. \tag{17}$$

The accuracy lower bound uses bags of size two. For $D = 1$, this same class has infinite ordinary Natarajan and DS dimensions, but exact-label examples permit width-two predictions with risk at most ϵ and failure probability at most δ using $\lceil \epsilon^{-1} \log(1/\delta) \rceil$ examples.

5.1 THE SEPARATION FROM EXACT LABELS

For each fixed finite $L \ge \ell$, the ambiguity term has accuracy exponent ℓ, whereas the exact-label learner has exponent one. Both learners may return at most L labels, and the class is unchanged when L or ϵ varies. The theorem gives a worst-case lower bound at prescribed deletion dimension; it does not assert that every class of that dimension has this sample complexity.

The construction uses finitely many polynomial evaluations at each scale. A colour specifies one evaluation value, and its hypotheses may be nonzero on any subset of the corresponding fibre. Any

ℓ distinct colours have at most one common input, which controls the deletion dimension. At each input, we display an independently chosen incident colour together with zero. A hidden colour selects the target, so an unseen input retains uncertainty about its target label even after the entire training transcript is revealed. Section 5.2 gives the full class; Appendix C.1 gives the observation law and constant-confidence argument.

The countable alphabet is substantive. Unlike a fixed finite alphabet, it cannot be covered by a sufficiently large constant list. No zero-data learner of any fixed finite width succeeds uniformly at the confidence in the theorem: the class contains an input at which arbitrarily many labels can occur. A finite polynomial block can have a wider zero-data list, but the theorem concerns one class containing all its scales and all finite prediction widths.

5.2 The fixed polynomial class

Fix powers of two k, Q with $k \geq 2\ell$ and $Q \geq k$. We take as inputs the polynomials over $\mathbb{F}_Q$ of degree less than ℓ, including zero, and choose distinct $a_1, \ldots, a_k \in \mathbb{F}_Q$. For each colour $(i, y) \in [k] \times \mathbb{F}_Q$, prescribe the support

$$S_{i,y} = \{f : f(a_i) = y\}. \tag{18}$$

There are $n = Q^\ell$ inputs and $K = kQ$ colours. Each input is incident to exactly k colours, and each colour has $Q^{\ell-1} = nk/K$ incident inputs.

Any ℓ distinct colours either specify conflicting values at one evaluation position, or prescribe ℓ distinct evaluations and determine a unique polynomial of degree less than ℓ. Thus their supports have intersection size at most one. Taking any input and ℓ of its incident colours gives equality. The finite-subset colour class of this block, defined in Appendix C.1, has deletion dimension exactly one by Lemma C.1.

Take the disjoint union of these blocks over all permitted k, Q, tagging their inputs and nonzero colours by the block. Add an input x_* with countably many additional colours, each supported only on $\{x_*\}$, and an anchor x_0 where every hypothesis is zero. Include all finite-subset colour hypotheses and zero. Call the resulting countable class $H_{\ell,1}$. Every ℓ-fold support intersection still has size at most one, and singleton intersections occur, so $\Delta_\ell(H_{\ell,1}) = 1$.

Let $H_{\ell,D}$ be the product of D independent copies of this class on disjoint input components. It is countable and is fixed before L, ϵ, δ are chosen. Any finite deleted DS witness projects to a deleted DS witness in each component: a neighbour changing a retained coordinate remains such a neighbour after projection. Each component therefore contributes at most one input, so $\Delta_\ell(H_{\ell,D}) \leq D$. Conversely, one x_* input in each component admits $\ell - 1$ distinct references and independent surviving binary choices of zero and one further colour. This gives a deleted binary cube on D inputs, proving equality. The same special inputs realize all label assignments required by Lemma C.3.

Proof of Theorem 5.1. We use the fixed class just constructed. Given $L \geq \ell$, let k be the least power of two at least $2L$, so $2L \leq k < 4L$. Given $\epsilon \leq 1/(102400L^2)$, let Q be the largest power of two at most $1/(100k\epsilon)$. Then

$$\frac{1}{200k\epsilon} < Q \leq \frac{1}{100k\epsilon}, \qquad Q \geq 64k. \tag{19}$$

For the second assertion, the accuracy bound implies $\epsilon \leq 1/(6400k^2)$, and $64k$ is itself a power of two. Hence this polynomial block already appears in every component of $H_{\ell,D}$.

Apply Lemma C.2 with $d = D$, $n = Q^\ell$, $K = kQ$, $q = r = 1$. Its degree condition holds with equality even without the factor two. Also $k \geq 2L$, $\epsilon \leq 1/(100K)$, and

$$\mu = \frac{DQ^{\ell-1}}{k} \geq \frac{Q}{k} \geq 64.$$

Thus every successful learner in the stated confidence range needs

$$m > \frac{D}{2}Q^\ell > \frac{D}{2(200k)^\ell\epsilon^\ell} \geq \frac{D}{2(800L)^\ell\epsilon^\ell}. \tag{20}$$

Lemma C.3 gives the two floors on this identical class, including for the larger accuracy range stated in the theorem. Taking the maximum of their three lower terms and using that a maximum is at least one third of their sum proves (16) with the displayed constant.

For $D = 1$, each target is zero or uses one nonzero colour on a finite support. Given exact labels, we retain the first nonzero observed colour j and return $\{0, j\}$ everywhere; before such an observation we return $\{0\}$. If the nonzero region has mass $p > \epsilon$, missing it in m examples has probability $(1-p)^m \leq e^{-m\epsilon}$. Otherwise the fallback already has risk at most ϵ. This proves the exact-label assertion on the same class. Meanwhile every evaluation fibre supports its full zero/colour binary cube, and its cardinality $Q^{\ell-1}$ is unbounded. The ordinary Natarajan and DS dimensions are therefore infinite.

Finally, the alphabet cannot be covered by a fixed finite list without data. Concentrate the input law at x_* and average over $K_* \geq 2L$ possible target labels. Any random width-L zero-data list omits some fixed target with probability at least one half. This verifies the stated distinction from the zero-data lists available on an individual finite block. □

6 THE WIDTH THRESHOLD AT FINITE LIST DIMENSION

Finite list dimension permits learning when the prediction width grows with the required accuracy. The necessary width is quadratic in the observation width, even for a class with one-example exact-label list compression.

Fix $\ell \geq 2$ and put $r = \ell - 1$. On $X = \mathbb{N}$ and $Y = \{0, 1, 2, \ldots\}$, partition the nonzero labels into the private groups

$$A_j = \{(j-1)r + 1, \ldots, jr\}, \qquad j \geq 1.$$

Each target uses labels from one private group and is zero outside a finite set. More precisely, let

$$\mathcal{G}_\ell = \bigcup_{j \geq 1} \{h : \mathbb{N} \to \{0\} \cup A_j : |\{x : h(x) \neq 0\}| < \infty\}. \tag{21}$$

This is one countable class for each ℓ. It is independent of the prediction width, accuracy, confidence, and sample size.

Theorem 6.1 (Prediction width at finite list dimension)**.** *For every integer $\ell \geq 2$, the class $\mathcal{G}_\ell$ has an exact-label width-ℓ sample compression scheme of size at most one and satisfies*

$$d_{\ell\mathrm{DS}}(\mathcal{G}_\ell) = 1, \qquad \Delta_\ell(\mathcal{G}_\ell) = \infty. \tag{22}$$

For every integer $L \geq \ell$, finite sample size $m \geq 0$, and randomized learner returning a width-L function, there are a fixed $h \in \mathcal{G}_\ell$ and a fixed clean iid law Q, with bags of size exactly ℓ, such that

$$\Pr\left\{R_h(M) > \frac{1}{64}\min\left\{1, \frac{\ell^2}{L}\right\}\right\} \geq \frac{1}{6}. \tag{23}$$

Consequently, for $0 < \epsilon < 1/64$ and $0 < \delta < 1/6$, any width admitting a finite distribution-free sample bound on this class must satisfy

$$L \geq \frac{\ell^2}{64\epsilon}. \tag{24}$$

Conversely, for every class with $d_{\ell\mathrm{DS}}(H) < \infty$, prediction width $L \geq \lceil 384\ell^2/\epsilon \rceil$ permits a finite distribution-free sample bound for risk at most ϵ with probability at least $1 - \delta$ on every clean law, by Theorem 3.2.

Thus the worst-case necessary and sufficient width under the finite-list-dimension assumption is $\Theta(\ell^2/\epsilon)$, with absolute constants, at fixed confidence below $1/6$. The lower bound allows any finite sample budget. Its target and law may depend on that budget and the learner, but are fixed before sampling. Since $\mathcal{G}_\ell$ has infinite deletion dimension, this statement does not give the same width necessity for a class with bounded Δ_ℓ.

The proof is in Appendix D.

AI-USE STATEMENT

The first completed version of this manuscript (with more results than covered here) was passed through an AI-based tool for recommendations on structure and presentation within the first 9 pages, parts to be moved to the appendix, and appropriate rewording of the prose in all sections to ensure smooth and coherent reading. Several of these recommendations were found reasonable, and then implemented with the help of the same tool. The formal mathematics is left untouched by the tool.

REPRODUCIBILITY STATEMENT

Complete proofs of all statements not proved in the main body are provided in the appendices.

REFERENCES

Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A theory of pac learnability of partial concept classes. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 658–671, 2022. doi: 10.1109/FOCS52979.2021.00070.

Andrea Campagner. Learning from fuzzy labels: Theoretical issues and algorithmic solutions. *International Journal of Approximate Reasoning*, 171:108969, 2024. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2023.108969. URL https://www.sciencedirect.com/science/article/pii/S0888613X23001007. Synergies between Machine Learning and Reasoning.

Moses Charikar and Chirag Pabbaraju. A characterization of list learnability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, pp. 1713–1726, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399135. doi: 10.1145/3564246.3585190. URL https://doi.org/10.1145/3564246.3585190.

Nirmit Joshi, Gene Li, Siddharth Bhandari, Shiva Kasiviswanathan, Cong Ma, and Nathan Srebro. Learning to answer from correct demonstrations. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=69fIHgLjyH.

Liping Liu and Thomas Dietterich. Learnability of the superset label learning problem. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1629–1637, Beijing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/liug14.html.

Chirag Pabbaraju. The Optimal Sample Complexity of Multiclass and List Learning. *arXiv:2604.24749v4*, 2026. https://arxiv.org/abs/2604.24749v4.

Kaifu Wang, Qiang Ning, and Dan Roth. Learnability with indirect supervision signals. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9145–9155. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/67ff32d40fb51f1a2fd2c4f1b1019785-Paper.pdf.

Jingyuan Xu, Xin Zou, Xiuwen Gong, and Weiwei Liu. On learnability and disambiguation of multiclass partial concept classes. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=LbYeOvDo3o.

A STRUCTURAL FACTS AND LEARNING INGREDIENTS

A.1 DIMENSIONS, RESTRICTIONS, AND ZERO CASES

The upper bound uses two parameters separately: one for prediction on shattered sets and one for learning from exact labels. For a partial class $P \subseteq (Y \cup \{\perp\})^X$, its partial Natarajan dimension $d_{\mathrm{N}}(P)$ is the supremum of the integers n for which there are distinct inputs $x_1, \dots, x_n$ and pairs of distinct labels $u_i, v_i \in Y$ such that every vector in $\prod_{i=1}^{n} \{u_i, v_i\}$ is a fully defined trace of P. As with DS dimension, the value is zero when no positive-size witness exists and infinity when witness sizes are unbounded. Put

$$d = \sup_{F \in H^{\ell-1}} d_{\mathrm{N}}(H_F), \qquad s = d_{\ell\mathrm{DS}}(H), \qquad D = \Delta_\ell(H). \tag{25}$$

Both parameters are bounded by the deletion dimension. Their zero cases give a pointwise menu of at most ℓ labels, requiring no observations.

Lemma A.1. *For every nonempty class $H \subseteq Y^X$ and integer $\ell \geq 2$, the parameters in* (25) *satisfy $d, s \leq D$. Moreover, the following conditions are equivalent:*

$$D = 0, \qquad d = 0, \qquad s = 0, \qquad |\{h(x) : h \in H\}| \leq \ell \quad \textit{for every } x \in X.$$

Under these conditions, the pointwise list $M(x) = \{h(x) : h \in H\}$ has zero risk without any observations, for every target and observation law.

Proof. A binary Natarajan witness is an ordinary DS witness. Applying this to each H_F and taking suprema gives $d \leq D$.

To prove $s \leq D$, let $V \subseteq Y^n$ be a finite ℓ-list DS witness on distinct inputs $x_1, \dots, x_n$. Each coordinate fibre of V contains at least $\ell + 1$ vectors, whose labels at that coordinate are distinct. We first select $\ell - 1$ vectors $v^{(1)}, \dots, v^{(\ell-1)} \in V$ whose labels are pairwise distinct at every coordinate.

Suppose the first $j - 1$ vectors have been selected. We start at any vector of V and process the coordinates in order. At coordinate i, we move within the current i-fibre to avoid the $j-1$ previously selected labels there. Such a choice exists because the fibre has at least $\ell+1$ labels. The move leaves every other coordinate unchanged, so choices made at earlier coordinates remain valid. The resulting vector is $v^{(j)}$.

Choose $f_j \in H$ with trace $v^{(j)}$ on these inputs, and put $F = (f_1, \dots, f_{\ell-1})$. Let V' contain the vectors of V that avoid all reference labels at every coordinate. The same coordinate-by-coordinate procedure, now avoiding $\ell - 1$ labels, shows that V' is nonempty. If $v \in V'$, its i-fibre in V already avoids the reference labels at all coordinates other than i. Removing the at most $\ell - 1$ forbidden labels at i therefore leaves at least two vectors in that fibre. Thus V' is an ordinary DS witness, and every vector in it is a fully defined trace of H_F. We obtain $n \leq D$ for every finite witness size n, proving $s \leq D$ also when the dimensions are unbounded.

If some input has at least $\ell + 1$ realized labels, choose hypotheses realizing $\ell + 1$ of them. Taking the first $\ell - 1$ as references leaves two distinct defined labels in H_F at that input, so $d, D \geq 1$. The $\ell + 1$ original singleton traces also give $s \geq 1$.

Conversely, suppose at most ℓ labels are realized at every input. For any reference tuple F, there is at most one actual label in H_F at each input: there are none if the reference labels repeat, and otherwise at most one remains after deleting them. No positive-size DS witness is possible, so $D = 0$, and the inequalities already proved give $d = s = 0$. The displayed pointwise list contains every target label and has size at most ℓ, which proves the last assertion. □

Later constructions retain target labels only inside a prescribed list. We describe this operation as erasure: a partial function may have some of its actual values replaced by $\perp$, with every remaining value unchanged. We also use injective changes of labels separately at each input.

Lemma A.2. *Restricting the input set, taking a subclass, or erasing values cannot increase partial Natarajan dimension or any k-list DS dimension. Applying an injective map to the actual labels separately at each input, while keeping $\perp$ undefined, preserves these dimensions.*

Let P be a class of partial functions, each obtained by erasing values of some member of H. For a tuple $G = (g_1, \dots, g_{\ell-1}) \in P^{\ell-1}$, define $p_G(x) = p(x)$ when all of $g_1(x), \dots, g_{\ell-1}(x), p(x)$ are actual, pairwise distinct labels, and set $p_G(x) = \perp$ otherwise. Then

$$d_{\mathrm{N}}(P_G) \le d, \qquad d_{\mathrm{DS}}(P_G) \le D, \qquad P_G = \{p_G : p \in P\}. \tag{26}$$

Proof. Every fully defined trace remaining after a restriction, passage to a subclass, or erasure was a fully defined trace before that operation. The same witness therefore proves each asserted monotonicity. An injective relabeling at each input preserves equality and inequality of labels there, so it maps coordinate fibres and binary witness pairs bijectively to their counterparts in the relabeled class. Applying the inverse on the image gives equality of dimensions.

For the final assertion, choose $f_j \in H$ from which g_j was obtained, and set $F = (f_1, \dots, f_{\ell-1})$. For each $p \in P$, choose an original hypothesis $h \in H$. Wherever p_G is defined, all the chosen original functions have the same values as their partial versions, so h_F is defined there and agrees with p_G. Consequently every fully defined trace of P_G is a fully defined trace of H_F. The two bounds follow from the definitions of d and D. □

These observations also apply when a finite sample is represented by indexed coordinates with repeated inputs. A fully defined DS witness cannot use two coordinates corresponding to the same input: a neighbour changing one would have to change the other. A Natarajan witness cannot use both either, since it requires all four independent choices at those coordinates. Erasure and injective coordinatewise recoding do not change this conclusion. After discarding this impossible repetition in a witness, its coordinates correspond to distinct original inputs, and the preceding bounds apply.

Enlarging the alphabet by symbols that no target uses also leaves every fully defined witness unchanged. Adding or removing such a symbol from a prediction list does not change target inclusion, provided the required width bound is respected. Likewise, filling an empty observed bag with a new non-target symbol preserves its omission probability.

A.2 Necessity from common observations

We first give the finite-witness obstruction for ordinary prediction from exact labels. Its sample bound will be uniform over the reference tuples used in the reduction.

Lemma A.3. *Suppose a partial class P has a finite DS witness V on $n \ge 1$ distinct inputs. For any randomized learner receiving m exact-label examples and returning a single label at each input, there is a partial target in P, defined on those inputs, for which its expected risk under the uniform input law is at least*

$$\frac{1}{2}\left(1 - \frac{m}{n}\right). \tag{27}$$

The expectation includes the training sample and the learner's randomness.

Proof. Choose a trace v uniformly from the finite family V and choose one representing partial target for each trace. Draw the m training indices independently and uniformly from $[n]$, independently of v. Fix an index i that was not sampled, and reveal to the learner all labels v_j for $j \ne i$ in addition to its actual data.

Conditional on these revealed labels and the sampled indices, the trace is uniform in its i-fibre. That fibre contains at least two vectors, with distinct labels at i. Hence no single prediction, even randomized and with the additional information, can have conditional success probability greater than $1/2$. The original learner has no greater success probability. At least $n - m$ indices are unseen, so averaging their contributions to the uniform-input risk gives (27). Finally, averaging over the finite prior on V selects one fixed partial target with at least this expected risk. □

The same observed bags can be used under several targets. This turns a bag learner into a single-label learner for every deleted partial class.

Lemma A.4. *Suppose a width-ℓ bag learner, using m samples, has risk at most α with probability at least $1 - \beta$ on every clean law for H. For each $F \in H^{\ell-1}$, there is a single-label learner from m exact-label examples such that, for every partial target h_F and every input law supported on its*

defined region, its risk is at most $\ell\alpha$ with probability at least $1 - \ell\beta$. The same sample size works for all F.

Proof. Fix $F = (f_1, \ldots, f_{\ell-1})$, a partial target h_F, and an input law μ supported on its defined region. Replace each exact-label observation (x, y) by the bag $\{f_1(x), \ldots, f_{\ell-1}(x), y\}$, and apply the assumed learner once to these bags. This transformation uses the reference functions and the observed exact label, and does not require the unknown original hypothesis h.

The resulting law is clean both for h and for each f_j. We couple these ℓ target worlds using the same bag sample and the same learner randomness, so they return the same function M. For each target the learner's guarantee applies under this common law. A union bound shows that, with probability at least $1 - \ell\beta$, all ℓ population omission risks are at most α.

Fix a label $y_0 \in Y$. At an input x, remove the reference labels from $M(x)$; output its sole remaining label if there is exactly one, and output y_0 otherwise. On the defined region of h_F, whenever $M(x)$ contains $h(x)$ and all reference labels, these are ℓ distinct labels, so the remaining singleton is $h(x)$. On the simultaneous good event, the error region of this single-label rule therefore has mass at most

$$R_h(M) + \sum_{j=1}^{\ell-1} R_{f_j}(M) \leq \ell\alpha.$$

All choices in the constructed learner depend only on F and its exact-label sample. The assumed bag guarantee is uniform, so its sample size is unchanged as the tuple and partial target vary. □

Proposition A.5. *If H is (ℓ, ℓ)-learnable, then $\Delta_\ell(H) < \infty$.*

Proof. Choose a bag learner at accuracy and failure probability both $1/(16\ell)$, and let m be its finite uniform sample size. Lemma A.4 gives, for every reference tuple, a single-label partial learner with risk at most $1/16$ except on an event of probability at most $1/16$. Its expected risk is therefore at most $1/8$.

If $\Delta_\ell(H) = \infty$, there are a tuple F and a finite DS witness for H_F of size $n \geq 2m+1$. Lemma A.3 gives a defined partial target and a uniform input law on which the expected risk of that same learner is greater than $1/4$, a contradiction. The tuple is chosen after the uniform budget m has been fixed; no single tuple attaining infinite DS dimension is required. □

A.3 Prediction widths below the observation width

When $L < \ell$, learnability holds if and only if $|\{h(x) : h \in H\}| \leq L$ for every x. Sufficiency follows by returning this known pointwise set. For necessity, if $L + 1$ target labels occur at one input, concentrate the input law there and always display their common bag, of size $L + 1 \leq \ell$. The observed law is identical for all these targets. Any returned list omits at least one, so some fixed target has failure probability at least $1/(L+1)$ for every accuracy less than one, regardless of sample size.

A.4 Exact-label learning and a fixed decoder

We use an exact-label learner only to construct a decoder. Later, the bag learner will enumerate its possible labeled inputs; it will not be asked to identify a correct labeling. The imported deterministic guarantee below includes the extension to infinite label spaces.

Theorem A.6 (Pabbaraju, 2026, Corollary A.1 and Remark 4)**.** *Let $s = d_{\ell\mathrm{DS}}(H) < \infty$. There is a deterministic exact-label learner returning width-ℓ predictions such that, for every realizable distribution and $\alpha, \beta \in (0, 1)$, its risk is at most α with probability at least $1 - \beta$, whenever*

$$m \geq 4.82(\ell + 1)\frac{s + \log(2/\beta)}{\alpha}.$$

The result applies to the countable input and label spaces used here.

Assume $s \geq 1$ and set $\lambda_\ell = \ell + 1$. Apply Theorem A.6 with $\alpha = \beta = 1/(8\lambda_\ell)$ and fix the resulting deterministic rule A at sample size

$$m_0 = \left\lceil 38.56\lambda_\ell^2\big(s + \log(16\lambda_\ell)\big)\right\rceil.$$

Its expected omission risk is at most $\alpha+\beta = 1/(4\lambda_\ell)$. Here the expectation is only over the training sample. Define A on every ordered labeled sample of length m_0, assigning a fixed width-ℓ output on any sample outside its promise if needed. Fix an enumeration of Y for all subsequent ties.

The following compression argument is an instance of the minimax and voting construction (Charikar & Pabbaraju, 2023, Lemma 7.5). We state a common sample-length cap because the later enumeration must also include decoders for an unknown clean subset.

Lemma A.7. *For every integer $N \geq 1$, put*

$$T_N = \lceil 4\lambda_\ell \log(2N)\rceil, \qquad r_N = m_0 T_N = O_\ell(s\log(2N)).$$

There is a fixed decoder $\mathcal{D}_N$ mapping an ordered sequence of r_N labeled examples to a width-ℓ function such that every nonempty realizable labeled sample of at most N indexed examples has a code, with repetitions allowed, whose decoded function contains every sample label. The decoder uses no additional labels, weights, or data-dependent block boundaries. An empty code has a separately fixed output.

Proof. Let $S = ((x_i, y_i))_{i=1}^n$ be realizable, with $1 \leq n \leq N$. Consider the finite game whose rows are indexed examples $i \in [n]$ and whose columns are ordered sequences $I \in [n]^{m_0}$. The loss in entry (i, I) is $\mathbf{1}\{y_i \notin A(S_I)(x_i)\}$. For every row distribution w, the exact-label distribution assigning mass w_i to indexed example i is realizable. Drawing I independently from w at each position therefore gives average loss at most $1/(4\lambda_\ell)$. Finite minimax gives a distribution π on columns such that, for every row i,

$$\Pr_{I\sim\pi}\{y_i \notin A(S_I)(x_i)\} \leq \frac{1}{4\lambda_\ell}.$$

Draw T_N independent columns from π, and let V_i count the resulting lists that omit y_i at x_i. For each summand U and $t = \log 4$, the bound $\Pr(U = 1) \leq 1/(4\lambda_\ell)$ gives $\mathbb{E}e^{tU} \leq \exp(3/(4\lambda_\ell))$. Exponential Markov's inequality and independence give

$$\Pr\{V_i \geq T_N/\lambda_\ell\} \leq \exp\left(-\frac{(4\log 4 - 3)T_N}{4\lambda_\ell}\right).$$

Since $4\log 4 - 3 > 1$ and $T_N \geq 4\lambda_\ell \log(2N)$, the probability that this event occurs for any of the n rows is at most

$$n(2N)^{-(4\log 4-3)} \leq N(2N)^{-(4\log 4-3)} < 1.$$

Choose columns for which every $V_i < T_N/\lambda_\ell$.

At an arbitrary input x, let $q_y(x)$ be the fraction of the chosen weak lists containing y. These fractions sum to at most ℓ. Return the ℓ labels of largest positive frequency, or all such labels if there are fewer than ℓ. If a label y with positive frequency is omitted, there are ℓ other labels of at least that frequency, so $(\ell + 1)q_y(x) \leq \ell$. At every sample input, however, the chosen columns give $q_{y_i}(x_i) > 1 - 1/\lambda_\ell = \ell/(\ell + 1)$. Thus every sample label is returned.

We concatenate the T_N selected ordered subsamples. This stores exactly r_N labeled examples from S, with repetition. The decoder splits the code into the predetermined blocks of length m_0, applies A to each, and takes the vote just defined. It does not receive π, its weights, or the row distribution. On arbitrary codes the same procedure remains a legal width-ℓ function; no correctness guarantee for an unrealizable code is used. For the empty sample, we use the empty code and the empty prediction at every input. This proves all the stated conventions. □

The encoder in this proof uses exact labels and is only an existence argument. Given bags, we retain all codes obtained by choosing observed indices and labels from their bags. The existence of one good decoded function is used in the analysis; the learner does not select that function by its unobserved target risk.

B THE UPPER BOUND

This appendix proves the structural and fractional ingredients used by the general-width learners. The complete clean proof is in Appendix B.7; the deferred large-width verifications begin in Appendix B.8. For the fractional ingredients, $d, s \geq 1$ are the finite parameters in (25); a zero parameter is covered by Lemma A.1.

Parameter	Role
d, s	Deleted Natarajan and list DS dimensions in (25)
D	Reference-deletion dimension, with $d, s \leq D$
t	Full indexed population size in a leave-one-out experiment
m_{seed}	Number of observed bags used to enumerate decoders
r_N	Fixed decoder code length for samples of size at most N
K	Size of each Natarajan-shattered block
J	Number of retained decoders
D_t	Upper bound on the logarithm of the completion-family size
b, η	Prediction mass and update rate in the clean proof
η_{p}	Internal target for valid decoder omissions
α, δ	Fractional error and failure probability
n	Final iid sample budget, chosen even

B.1 COMPLETING PARTIAL TRACES

The expert construction needs total predictions that agree with each partial target wherever that target is defined. A completion family for a partial class P is a family of total functions with this property. We first bound its size when each coordinate has a finite label set.

Lemma B.1. *Let $P \subseteq ([q] \cup \{\perp\})^{[v]}$ be nonempty, where $q \geq 1$ and $v \geq 1$, and put $p = d_{\mathrm{N}}(P)$. There is a completion family $\mathcal{C} \subseteq [q]^{[v]}$ satisfying*

$$\log(2|\mathcal{C}|) \leq C\big[1 + p \log(2vq) \log(2v)\big] \tag{28}$$

for an absolute constant C. If $p = 0$ or $q = 1$, one completion suffices. If $v = 0$, the family consists of the unique empty function.

Proof. For a partial class A, let $\Sigma(A)$ count its Natarajan witnesses, including both their coordinate subsets and their unordered pairs of actual labels. Count the empty witness once if $A \neq \varnothing$, and set $\Sigma(\varnothing) = 0$. Then

$$1 \leq \Sigma(P) \leq \sum_{j=0}^{p} \binom{v}{j} \binom{q}{2}^{j}. \tag{29}$$

Here $p \leq v$, and the terms with $j > 0$ vanish when $q = 1$.

For $A_{i=y} = \{a \in A : a(i) = y\}$, we claim that

$$\sum_{y=1}^{q} \Sigma(A_{i=y}) \leq \Sigma(A). \tag{30}$$

No witness in a fixed-label branch uses coordinate i. Fix a witness away from i that occurs in $r \geq 1$ actual-label branches. In A, it gives its original witness and one larger witness for each pair of these branches, by adjoining their two labels at i. These are $1 + \binom{r}{2} \geq r$ distinct witnesses. The witnesses charged for different original witnesses cannot coincide, since deleting coordinate i recovers the original witness. Undefined values belong to no branch, so summing these charges proves (30).

Process coordinates in the order $1, \ldots, v$. In a current nonempty class A, predict a label maximizing $\Sigma(A_{i=y})$, with fixed ties; if all branch counts vanish, use label 1. For a fixed partial target, make a correction only when its value is defined and differs from this prediction. At a correction, output the target value and restrict A to that branch. At every other coordinate, leave A unchanged, including when the target is undefined. The target remains in the current class throughout.

A correction branch is nonempty and differs from a maximizing branch. By (30), its count is at most half the current count. Thus every target causes at most

$$R = \min\{v, \lfloor \log_2 \Sigma(P) \rfloor\}$$

corrections.

Fix a set of correction positions $I = \{i_1 < \cdots < i_r\}$, where $r \le R$. Enumerate correction-label sequences that choose a nonempty branch different from the prediction at each prescribed position. The prediction and state between these positions are determined by the previous corrections. If the successive states are $A_0 = P, A_1, \ldots, A_r$, assign that branch transition the weight $\Sigma(A_j)/\Sigma(A_{j-1})$. At each state, the outgoing weights sum to at most one by (30). Therefore the sum of the products of transition weights over all such sequences is at most one. Each product is

$$\frac{\Sigma(A_r)}{\Sigma(P)} \ge \frac{1}{\Sigma(P)}.$$

There are consequently at most $\Sigma(P)$ valid label sequences for this fixed position set. This also holds for the empty position set, which has one transcript.

Each transcript reconstructs a total function by the stated prediction and correction procedure. Retain all these functions for all position sets of size at most R. A target's own correction transcript is included, and its reconstruction agrees with that target at every defined coordinate. Hence

$$|\mathcal{C}| \le \Sigma(P) \sum_{r=0}^{R} \binom{v}{r}. \tag{31}$$

No correction label or other information has been omitted from this count.

If $\Sigma(P) = 1$, then $R = 0$ and one completion suffices. Otherwise the subsets of $[v]$ of size at most R can be encoded by their increasing lists, padded with zeros to length R, so their number is at most $(v+1)^R$. Thus

$$\log |\mathcal{C}| \le \log \Sigma(P) \left(1 + \frac{\log(v+1)}{\log 2}\right).$$

For $p \ge 1$, (29) gives $\Sigma(P) \le (1 + v\binom{q}{2})^p$, and substituting this proves (28). For $p = 0$ or $q = 1$, (29) gives $\Sigma(P) = 1$. The empty-domain assertion is immediate. □

B.2 Prediction on shattered blocks

On a large binary-shattered set, reference deletion limits the mass outside a small prediction list. This gives a prediction rule that uses the class and inputs without observing their bags.

Lemma B.2. *Let P be a finite partial trace class obtained by erasing values of H on an indexed input tuple. Suppose a coordinate set C of size $K \ge 1$ is Natarajan-shattered by P. There are fractional predictions $p_{i,y} \in [0,1]$, supported on the labels realized by P at coordinate i, with $\sum_y p_{i,y} \le \ell$, such that every $u \in P$ satisfies*

$$\sum_{i \in C : u(i) \ne \perp} (1 - p_{i,u(i)}) \le \theta |C|, \qquad \theta = (d/K)^{1/(\ell-1)}. \tag{32}$$

These predictions depend only on P and the chosen block.

Proof. Let A_i be the pair of actual labels witnessing shattering at coordinate $i \in C$. For any $\ell - 1$ reference traces from P, consider the coordinates where the reference values are defined, pairwise distinct, and outside A_i. If there are r such coordinates, the binary witness restricted to them remains fully defined after deleting these references. Lemma A.2 therefore gives $r \le d$.

Fix any probability distribution on P. At coordinate i, take A_i together with the $\ell - 2$ most probable actual labels outside A_i, or all such labels if there are fewer. Let r_i be the remaining probability of a defined value outside this list; undefined mass is not included. Draw $\ell - 1$ reference traces independently from the chosen distribution. At each draw, conditional on the previous reference values being defined, distinct, and outside A_i, at most $\ell - 2$ previous labels have to be excluded.

Their total mass is no greater than that of the $\ell - 2$ heaviest labels outside A_i. The conditional chance of another permissible value is therefore at least r_i.

The probability that all references are permissible at i is at least $r_i^{\ell-1}$. Every realized reference tuple is permissible at at most d coordinates, as shown above. Taking expectations gives

$$\sum_{i \in C} r_i^{\ell-1} \le d, \qquad \frac{1}{K} \sum_{i \in C} r_i \le \left(\frac{1}{K} \sum_{i \in C} r_i^{\ell-1} \right)^{1/(\ell-1)} \le \theta.$$

The middle inequality is the power-mean inequality, with equality of the exponents when $\ell = 2$.

Consider the finite game between a target trace and a choice of at most ℓ realized labels at each block coordinate, with loss equal to average defined-target omission. For every distribution on target traces, the lists just constructed give loss at most θ. Finite minimax supplies a distribution on list choices with this bound for every target. Taking its inclusion marginals gives the required fractional predictions. This construction uses no observed bags on the block. □

We use disjoint blocks of size K. Repeatedly choose a K-coordinate Natarajan witness on the remaining coordinates and remove that block, until none remains. The residual class has partial Natarajan dimension less than K. Fixing orders on coordinates and labels makes these choices depend only on the finite partial class. Summing (32) over the blocks gives at most θ times their total size for each target. Repeated copies of one input cannot occur within a witness, by the observation following Lemma A.2.

B.3 Observed decoder families and a finite-population bound

The decoder of Lemma A.7 uses exact labels, but the bag learner can enumerate every possible supplied labeling of its code. Let $m_{\mathrm{seed}} \ge 1$ be the number of seed observations, and use the decoder $\mathcal{D}_{m_{\mathrm{seed}}}$ with fixed code length $r_{m_{\mathrm{seed}}}$. Enumerate every sequence of $r_{m_{\mathrm{seed}}}$ seed positions and every choice of a label from the bag at each occurrence. Include the fixed fallback decoder, and add a fixed label $y_0 \in Y$ to every decoded menu. This gives nonempty menus $\nu_1, \dots, \nu_J$ of width at most $q_0 = \ell + 1$, with

$$J \le 1 + (\ell m_{\mathrm{seed}})^{r_{m_{\mathrm{seed}}}}, \qquad \log(2J) = O_\ell\big(s \log^2(2m_{\mathrm{seed}})\big) \quad (s \ge 1). \tag{33}$$

The extra label is internal to the construction; the final prediction satisfies the prescribed width bound L.

Lemma B.3. *Assume $1 \le s < \infty$. Fix $h \in H$ and a full indexed state $((x_i, B_i))_{i=1}^t$, with $1 \le |B_i| \le \ell$, allowing arbitrary erroneous bags and repeated observations. Select an ordered seed of m_{seed} distinct indices uniformly without replacement, where $m_{\mathrm{seed}} \le t/2$, and construct the menus above. Let V be the complement of the seed, $v = t - m_{\mathrm{seed}}$, and put*

$$U_j = |\{i \in V : h(x_i) \in B_i,\ h(x_i) \notin \nu_j(x_i)\}|.$$

For every $\eta_p > 0$,

$$\Pr\left\{ \min_j U_j / v > \eta_p \right\} \le (m_{\mathrm{seed}}^{r_{m_{\mathrm{seed}}}} + 1) \exp\big[-\eta_p (m_{\mathrm{seed}} - r_{m_{\mathrm{seed}}})/2\big]. \tag{34}$$

For $s \ge 1$ and $0 < \eta_p < 1/4$, an integer seed size

$$m_{\mathrm{seed}} = O_\ell\left(s \eta_p^{-1} \log^2 \frac{es}{\eta_p} \right) \tag{35}$$

can be chosen so that $m_{\mathrm{seed}} \ge 2 r_{m_{\mathrm{seed}}}$ and, whenever $t \ge 2m_{\mathrm{seed}}$,

$$\mathbb{E} \min_j U_j / v \le 2\eta_p, \qquad \log(2J) = O_\ell\left(s \log^2 \frac{es}{\eta_p} \right). \tag{36}$$

The probability and expectation here are only over the seed selection.

Proof. In the analysis, keep the selected observations whose bags contain $h(x_i)$. If this valid subset is nonempty, Lemma A.7, with cap m_{seed}, supplies a code of length $r_{m_{\mathrm{seed}}}$ using its true labels and covering all its labels. This code appears in the observable enumeration. If the valid subset is empty, use the fallback. The learner does not identify the subset or this code.

Fix one sequence of code positions among the m_{seed} ordered seed positions. Let $r' \le \min\{r_{m_{\mathrm{seed}}}, m_{\mathrm{seed}}\}$ be the number of distinct positions it uses, and condition on the full-state indices occupying those positions. The true-label decoder and its padded menu are now fixed, as is the set T of all full-state indices where the bag is valid and that menu omits $h(x_i)$. The remaining $m_{\mathrm{seed}} - r'$ seed indices are still uniformly ordered without replacement from the other $t - r'$ indices.

For this decoder to cover all valid seed labels, every selected index must avoid T. If a conditioned code index is in T, or there are too few indices outside T for the selection, this event is impossible. In all other cases with $|T| > \eta_p(t - m_{\mathrm{seed}})$, its conditional probability is at most

$$\prod_{a=0}^{m_{\mathrm{seed}}-r'-1} \left(1 - \frac{|T|}{t - r' - a}\right) \le \exp\left[-\frac{|T|(m_{\mathrm{seed}} - r')}{t - r'}\right] \le \exp[-\eta_p(m_{\mathrm{seed}} - r_{m_{\mathrm{seed}}})/2].$$

The last inequality uses $t - m_{\mathrm{seed}} \ge t/2$ and $r' \le r_{m_{\mathrm{seed}}}$. When the decoder covers all valid seed labels, T lies entirely outside the seed, so $U_j = |T|$ for this decoder.

There are $m_{\mathrm{seed}}^{r_{m_{\mathrm{seed}}}}$ code-position sequences. Their true labels are fixed by h, so no additional union over guessed labels or targets is needed. The fallback has the same estimate with $r' = 0$: consistency again requires the seed to avoid its valid-miss set. If every observable menu has valid-miss fraction greater than η_p, the consistent code or fallback supplied in the first paragraph is one of these bad consistent candidates. Taking the union bound proves (34). Equal observations at different indices remain separate throughout the conditioning argument.

For completeness, write $L_p = \log(es/\eta_p) \ge 1$ and use a width-only constant c_ℓ such that $r_{m_{\mathrm{seed}}} \le c_\ell s \log(2m_{\mathrm{seed}})$. Choose

$$m_{\mathrm{seed}} = \left\lceil A_{\mathrm{seed}} s \eta_p^{-1} L_p^2 \right\rceil.$$

For $A_{\mathrm{seed}} \ge 1$, we have $\log(2m_{\mathrm{seed}}) \le \log(4A_{\mathrm{seed}}) + 3L_p$. Taking A_{seed} large enough ensures both $m_{\mathrm{seed}} \ge 2r_{m_{\mathrm{seed}}}$ and

$$\frac{\eta_p(m_{\mathrm{seed}} - r_{m_{\mathrm{seed}}})}{2} \ge \log(m_{\mathrm{seed}}^{r_{m_{\mathrm{seed}}}} + 1) + \log(1/\eta_p).$$

Indeed, the left side is at least $A_{\mathrm{seed}} s L_p^2/4$, while the right side is at most $[c_\ell(3 + \log(4A_{\mathrm{seed}}))^2 + 2]sL_p^2$. A fixed sufficiently large A_{seed} dominates this squared logarithm; enlarging it also gives $m_{\mathrm{seed}} \ge 2r_{m_{\mathrm{seed}}}$. The tail probability is then at most η_p. Since $0 \le \min_j U_j/v \le 1$, its expectation is at most $2\eta_p$. Substituting the same logarithmic bound into (33) gives the remaining assertion. □

B.4 Completing each decoder separately

The block decomposition may use the union of all decoded menus. The completion count must instead be applied within each individual menu, before the resulting expert families are united.

Lemma B.4. *Fix the observed seed, its menus $\nu_1, \dots, \nu_J$, and the full unlabeled tuple. On the complementary index set V, let P be the class obtained from H by retaining a value at i only when it belongs to $\nu(x_i) = \bigcup_{j=1}^{J} \nu_j(x_i)$. Form the disjoint K-block decomposition above, and let R be its residual, of size w, where $K \ge 1$ is an integer. There is a finite nonempty family $\mathcal{F}$ of total label assignments on R such that*

$$\log(2|\mathcal{F}|) \le \log J + C_\ell\big[1 + K \log^2(2t)\big], \tag{37}$$

and for every $h \in H$ and every decoder index j, some $g \in \mathcal{F}$ agrees with $h(x_i)$ at all $i \in R$ where $h(x_i) \in \nu_j(x_i)$. Every expert value belongs to the union menu at its coordinate. The family depends on the observed seed and unlabeled inputs, and not on any residual bag or the hidden target.

Proof. The residual class $P|_R$ has partial Natarajan dimension less than K. For each j, retain $h(x_i)$ on R only when it belongs to $\nu_j(x_i)$, obtaining a partial class P_j. Each of its fully defined traces is a fully defined trace of $P|_R$, since ν_j is contained in ν. Thus $d_{\mathrm{N}}(P_j) < K$.

Injectively recode the nonempty menu $\nu_j(x_i)$ into $[q_0]$ at each residual coordinate, where $q_0 = \ell+1$. Lemmas A.2 and B.1 give a completion family $\mathcal{F}_j$ with

$$\log(2|\mathcal{F}_j|) \le C_\ell\big[1 + K\log^2(2w)\big] \quad (w \ge 1).$$

Decode each active symbol by its inverse map. An inactive symbol is decoded to a fixed member of $\nu_j(x_i)$; a defined target never uses an inactive symbol, so this does not affect agreement on defined coordinates. If $w = 0$, take the single empty assignment instead, without evaluating the logarithm at zero.

Set $\mathcal{F} = \bigcup_{j=1}^{J} \mathcal{F}_j$. Then

$$\log(2|\mathcal{F}|) \le \log J + \max_j \log(2|\mathcal{F}_j|) \le \log J + C_\ell\big[1 + K\log^2(2t)\big].$$

For each h, j, the completion from $\mathcal{F}_j$ has the required agreement. All families are retained, so the construction does not need to know which decoder best covers the target. Every operation uses only the seed menus and the unlabeled tuple. □

These lemmas supply the following inputs to the prediction game. With a fixed state of size $t \ge 2m_{\mathrm{seed}}$, choose a uniformly ordered seed and keep all its decoded menus. The blocks have total defined-target omission at most θ times their size. The remaining experts have logarithmic size at most

$$D_t = C_\ell\big[1 + \log(2J) + K\log^2(2t)\big], \tag{38}$$

and contain a completion for every target and every individual decoder. Replacing J in this expression by $1 + (\ell m_{\mathrm{seed}})^{r_{m_{\mathrm{seed}}}}$ gives an envelope independent of the seed contents. The expected smallest valid-miss fraction on the seed complement is at most $2\eta_p$.

None of these constructions uses the query bag. A legal prediction first chooses its seed among the other indices; equivalently, in the analysis we may choose the seed first and then a uniform query in its complement. These procedures give the same joint law of seed and query, since each ordered seed and disjoint query has the same probability under either order. The blocks and expert family may use all unlabeled inputs but only the seed bags. Choosing the best decoder or an agreeing expert remains an analytical comparison, and is never an operation performed by the learner.

B.5 An observable finite-expert game

The completion family supplies a finite set of candidate labels at each input, but the learner cannot tell which candidate agrees with the target. We therefore compare all experts through gains that can be evaluated from the observed bag.

Write

$$\tau = \frac{\ell}{\ell+1}.$$

A fractional prediction is a vector $p = (p_y)_{y\in Y}$ with $0 \le p_y \le 1$ and $\sum_y p_y \le \ell$. For an expert g and an observed pair (x, B), define its gain by

$$a_g(p; x, B) = \mathbf{1}\{g(x) \in B\}(1 - p_{g(x)}) - \tau\mathbf{1}\{g(x) \notin B\}. \tag{39}$$

The gain is observable once B is given. The prediction p must be chosen before seeing B. Every gain lies in $[-\tau, 1]$.

The following action controls the weighted mean gain for every possible bag at once.

Lemma B.5. *Let $\mathcal{F}$ be a finite nonempty family of total experts, let x be an input, and let $(u_g)_{g\in\mathcal{F}}$ be nonnegative weights summing to one. There is a fractional prediction supported on $A = \{g(x) : g \in \mathcal{F}\}$ such that*

$$\sum_{g\in\mathcal{F}} u_g a_g(p; x, B) \le 0$$

for every $B \subseteq Y$ of size at most ℓ, including $B = \varnothing$. The prediction depends only on the weights and the expert labels at x.

Proof. Only $B \cap A$ affects the gain, so the opponent has finitely many distinct actions. The feasible prediction vectors on A form the polytope

$$\mathcal{P} = \{p \in [0,1]^A : \sum_{y \in A} p_y \le \ell\}.$$

Its vertices are indicators of subsets of A of size at most ℓ. Indeed, two fractional coordinates can be perturbed in opposite directions while preserving their sum, contradicting extremality. If exactly one coordinate is fractional and the sum constraint is strict, that coordinate can be perturbed in both directions. If exactly one coordinate is fractional and the sum constraint is tight, the integral value ℓ would equal an integer plus a proper fraction, which is impossible. Thus finite minimax applies, with mixed prediction actions identified with their inclusion vectors.

Fix a mixed bag action, and let r_y be the probability that its bag contains y. Then $0 \le r_y \le 1$ and $\sum_{y \in A} r_y \le \ell$. Define

$$f(r) = \begin{cases} \max\{0, 1 + \tau - \tau/r\}, & r > 0, \\ 0, & r = 0. \end{cases}$$

For each label, the choice $p_y = f(r_y)$ makes

$$r_y(1 - p_y) - \tau(1 - r_y) \le 0.$$

When $r_y = 0$, the expression equals $-\tau$. When $r_y > 0$ and $f(r_y) = 0$, it equals $(1+\tau)r_y - \tau \le 0$. When $f(r_y) > 0$, substitution gives zero.

It remains to check the mass constraint. We have $0 \le f(r) \le 1$ for $0 \le r \le 1$. If at most ℓ terms are positive, their sum is at most ℓ. Otherwise let $k \ge \ell + 1$ be their number. Cauchy-Schwarz on the positive coordinates gives

$$\begin{aligned} \sum_{y \in A} f(r_y) &= k(1+\tau) - \tau \sum_{y : f(r_y) > 0} \frac{1}{r_y} \\ &\le k(1+\tau) - \frac{\tau k^2}{\sum_{y : f(r_y) > 0} r_y} \\ &\le k(1+\tau) - \frac{\tau k^2}{\ell} = \frac{k(2\ell + 1 - k)}{\ell + 1} \le \ell. \end{aligned}$$

The second inequality uses the bag-size bound. The last inequality is equivalent to $(k - \ell)(k - \ell - 1) \ge 0$, which holds for the integer $k \ge \ell + 1$. Thus every mixed bag action has a feasible response making each label's expected gain nonpositive. Its weighted expected gain is also nonpositive. Finite minimax gives one prediction whose weighted gain is nonpositive against every bag. □

B.6 A BOUNDED-MARTINGALE INEQUALITY

We will apply the following inequality under reverse exposure of a random ordering and under the forward iid filtration.

Lemma B.6. *Let $Y_1, \dots, Y_T$ be martingale differences with $|Y_j| \le 2$. If $0 < \lambda \le 1/4$ and $\Lambda > 0$, then, with probability at least $1 - e^{-\Lambda}$,*

$$\sum_{j=1}^{T} Y_j \le \lambda \sum_{j=1}^{T} \mathbb{E}[Y_j^2 \mid \mathcal{F}_{j-1}] + \frac{\Lambda}{\lambda}.$$

Proof. For $|u| \le 1/2$, we have $e^u \le 1 + u + u^2$. Conditional expectation and $\mathbb{E}[Y_j \mid \mathcal{F}_{j-1}] = 0$ give

$$\mathbb{E}[e^{\lambda Y_j} \mid \mathcal{F}_{j-1}] \le 1 + \lambda^2 \mathbb{E}[Y_j^2 \mid \mathcal{F}_{j-1}] \le \exp\big(\lambda^2 \mathbb{E}[Y_j^2 \mid \mathcal{F}_{j-1}]\big).$$

Successive conditioning shows that the expectation of the exponential of $\lambda \sum_j Y_j - \lambda^2 \sum_j \mathbb{E}[Y_j^2 \mid \mathcal{F}_{j-1}]$ is at most one. Markov's inequality proves the claim. □

B.7 THE CLEAN BOUND FOR EVERY OUTPUT WIDTH

Proof of Theorem 3.1. If $D = 0$, Lemma A.1 supplies a pointwise menu of at most ℓ labels containing every target. Assume $D \geq 1$. The following construction applies to every $L \geq \ell$.

Put

$$\begin{aligned} b &= \max\{\ell, \lfloor L/2 \rfloor\}, \qquad \tau = \frac{\ell}{b+1}, \qquad \eta = \frac{1-\tau}{2\tau}, \\ A &= \log(1+\eta), \qquad B_0 = -\log(1-\eta\tau), \quad \beta = \frac{B_0}{A}. \end{aligned} \tag{40}$$

Here $b \geq \ell$ and $0 < \tau < 1$. We allow fractional prediction mass b while retaining the original seed, block and completion constructions, whose widths and entropy depend only on ℓ.

The larger-mass prediction game. For a scalar expert g, use the observable gain

$$a_g(p; x, B) = \mathbf{1}\{g(x) \in B\}(1 - p_{g(x)}) - \tau \mathbf{1}\{g(x) \notin B\}.$$

There is an action of mass at most b, chosen before the bag, making the weighted mean gain nonpositive for every bag of size at most ℓ. To verify this, fix a mixed bag action with inclusion marginals r_y, so $\sum_y r_y \leq \ell$, and set

$$p_y = \begin{cases} \max\{0, 1 + \tau - \tau/r_y\}, & r_y > 0, \\ 0, & r_y = 0. \end{cases}$$

Each label's expected gain is nonpositive. If $k \geq b+1$ of these coordinates are positive, Cauchy–Schwarz gives

$$\sum_y p_y \leq k(1+\tau) - \frac{\tau k^2}{\ell} = \frac{k(\ell + b + 1 - k)}{b+1} \leq \ell \leq b,$$

because the numerator minus $\ell(b+1)$ equals $-(k-\ell)(k-b-1) \leq 0$. With at most b positive coordinates, feasibility is immediate. The finite minimax argument of Lemma B.5, on the labels of the finite expert family, now gives the required action against every bag, including bags disjoint from that label set.

For a family $\mathcal{F}$ of N experts, initialize weights uniformly and update by $W_g \leftarrow W_g(1 + \eta a_g)$. The factors are positive, since $\eta\tau = (1-\tau)/2 < 1/2$, and total weight does not increase. Write

$$u_{i,g} = \mathbf{1}\{g(x_i) \in B_i\}(1 - p_{g(x_i)}), \qquad c_{i,g} = \mathbf{1}\{g(x_i) \notin B_i\}.$$

When $c_{i,g} = 0$, concavity gives $\log(1 + \eta u_{i,g}) \geq A u_{i,g}$; when $c_{i,g} = 1$, the logarithmic update is $-B_0$. Consequently, simultaneously for every expert and every prefix of the same initialized run,

$$\sum_i (u_{i,g} - \beta c_{i,g}) \leq \frac{\log N}{A}. \tag{41}$$

Moreover $0 < \beta < 1$, because

$$(1+\eta)(1-\eta\tau) = \frac{(1+\tau)^2}{4\tau} > 1.$$

Thus increasing the update rate introduces no additive error per round.

A legal statewise rule. Fix a seed size m_{seed}, a block size K, and a tolerance η_{p}. By Lemmas B.3, B.2 and B.4, the unchanged constructions give a deterministic decoder-count bound J, block loss

$$\theta = (D/K)^{1/(\ell-1)},$$

and a nondecreasing entropy envelope

$$E_t = C_\ell[1 + \log(2J) + K \log^2(2t)]. \tag{42}$$

The block actions have mass at most $\ell \leq b$. The residual completion family has logarithmic size at most E_t, irrespective of L.

On a clean state, let g be the completion agreeing with h wherever one particular decoder ν_j contains h. At an agreement position, its transformed gain $u_{i,g} - \beta c_{i,g}$ equals the target fractional loss. At a decoder miss, their difference is at most $1 + \beta < 2$. Hence on the residual R,

$$\sum_{i \in R} (1 - p_{i,h(x_i)}) \le \frac{\log |\mathcal{F}|}{A} + 2U_R, \tag{43}$$

where U_R counts that decoder's target misses. This is a comparison with one individual decoder; the learner retains the whole family.

The comparator is fixed before the random residual order. Since (41) holds simultaneously for every expert, no part of the algorithm selects j or g.

Write $S = ((x_i, B_i))_{i=1}^t$ for the indexed state, and let S^{-i} denote the state with record i deleted.

On an indexed state of size $t \ge 2m_{\mathrm{seed}}$, choose the ordered seed uniformly from the other observations. Construct blocks and completions using its bags and all complementary inputs, including the query, but no other complementary bag. Average the block-order choices to make the rule symmetric. On the residual, insert the query at a uniform position in a uniform residual order and run only the preceding bags. The full family, initialization and rate are fixed across all these internal prefixes. Experts are assignments to indexed residual positions; repeated input values need not have equal expert values.

For a query at index $i \in R$, write $w = |R|$. Choose a position q uniformly from $[w]$, order $R \setminus \{i\}$ uniformly, and insert i at position q. Initialize the full family $\mathcal{F}$ uniformly and run the update leading to (41) on the first $q - 1$ pairs in this order. At the query input use its next action, without revealing the query bag. The family, rate η, horizon w, and initialization are those of the full residual problem; none is recomputed from the selected prefix. Thus every prefix used here is a prefix of the same full-horizon sequential procedure for its chosen order. When $w = 1$, the prefix is empty. When R is empty, this part is never used.

Choosing a uniform query and then a uniform seed excluding it has exactly the same distribution as choosing a uniform seed first and a uniform query from its complement. For an unordered seed I and $i \notin I$, the two probabilities agree because

$$\frac{1}{t\binom{t-1}{m_{\mathrm{seed}}}} = \frac{1}{\binom{t}{m_{\mathrm{seed}}}(t - m_{\mathrm{seed}})}.$$

Both experiments also order the seed uniformly. Conditional on the seed and the block construction, a uniform query in R, a uniform insertion position, and a uniform ordering of the other residual indices are equivalent to a uniform full residual permutation with a uniformly marked position. Consequently, the average query loss on R is exactly $1/w$ times the expected sum of losses in the common full sequential run.

In the analysis, a uniform query followed by a seed excluding that query is equivalent to a uniform seed followed by a uniform query in its complement V. Conditional on the seed, a uniform residual query with a uniform insertion position is a uniform marked position in one full residual run. Thus (43) can be summed over that run before dividing by $v = |V| \ge t/2$. The block contribution is at most $\theta |V \setminus R| + U_{V \setminus R}$. Choosing the best decoder only in the analysis, Lemma B.3 gives $\mathbb{E}(U_V / v) \le 2\eta_{\mathrm{p}}$. Averaging all finite internal choices produces a symmetric fractional rule of mass at most b satisfying, on every clean indexed state,

$$\frac{1}{t} \sum_{i=1}^{t} \big(1 - p_{t,h(x_i)}(S^{-i}; x_i)\big) \le \theta + 4\eta_{\mathrm{p}} + \frac{2E_t}{At}. \tag{44}$$

No component of this rule observes its query bag.

Finally average all internal actions into p_t. Feasibility is preserved by averaging, and the loss is affine in p, so the bound remains unchanged. For each fixed indexed state there are finitely many seed, ordering, and prefix choices, with fixed choices of the finite-game actions. Countability of the input and label spaces makes the resulting rule measurable. It is invariant under permutation of the observed indices, and every component withheld the query bag. This completes the construction.

Clean concentration. For a fractional function p, write

$$R_h^{\mathrm{frac}}(p) := \mathbb{E}_{Z \sim \mu}[1 - p_{h(Z)}(Z)].$$

Take even $n \geq 4m_{\text{seed}}$ and average the rules trained on the first $j-1$ observations for $n/2 < j \leq n$, with all numerical parameters fixed.

Denote the resulting fractional rule trained on the first $j-1$ observations by $p^{(j)}$. Put

$$\overline{p} = \frac{2}{n} \sum_{j=n/2+1}^{n} p^{(j)}.$$

Their common statewise bound is

$$e_* = \theta + 4\eta_{\text{p}} + \frac{4E_n}{An}.$$

Reverse exposure. Condition on the unordered indexed sample and expose its order from the end. At the step removing position j, that observation is uniform among the remaining j observations. Symmetry and the statewise inequality bound its conditional mean loss by e_*.

Let $G_j \in [0,1]$ be the queried fractional loss. Under reverse exposure of a uniform ordering of the indexed sample, its conditional mean is at most e_* by (44), and its conditional variance is at most that mean.

Write

$$\overline{G} = \frac{2}{n} \sum_{j=n/2+1}^{n} G_j.$$

Lemma B.6, with $\lambda = 1/4$ and $\Lambda = \log(2/\delta)$, gives

$$\overline{G} \leq \tfrac{5}{4} e_* + \frac{8\Lambda}{n}$$

except with probability $e^{-\Lambda}$. Under forward iid exposure let v_j be the population fractional risk of the prefix rule.

Write

$$\overline{v} = \frac{2}{n} \sum_{j=n/2+1}^{n} v_j.$$

Applying the same lemma to $v_j - G_j$, whose conditional variance is at most v_j, gives

$$\tfrac{3}{4} \overline{v} \leq \overline{G} + \frac{8\Lambda}{n}$$

outside a second such event. For the averaged predictor $\overline{p}$, linearity of fractional loss therefore yields

$$R_h^{\text{frac}}(\overline{p}) \leq 2e_* + \frac{22\Lambda}{n} \tag{45}$$

with probability at least $1-\delta$. The reverse exposure is justified by symmetry of the statewise rule; cleanliness permits the absolute choice $\lambda = 1/4$ without any coefficient margin.

Accounting uniformly in L. For desired fractional error α, choose

$$\eta_{\text{p}} = \alpha/64, \qquad K = \left\lceil D(8/\alpha)^{\ell-1} \right\rceil, \qquad H_\alpha = \log \frac{eD}{\alpha}.$$

The seed lemma permits $m_{\text{seed}} = O_\ell(D\alpha^{-1}H_\alpha^2)$ and $\log(2J) = O_\ell(DH_\alpha^2)$. Set

$$T_0 = C_0(D + K/A)H_\alpha^2$$

and take the least even integer $n \geq C'\alpha^{-1}(T_0 + \Lambda)$. The constants C_0, C' depend only on ℓ. In particular, the seed cost is not divided by A.

Since

$$A = \log \frac{b+1+\ell}{2\ell} \geq \log\left(1 + \frac{1}{2\ell}\right),$$

we have $T_0 \le C_\ell(D+K)H_\alpha^2$ and $\log(T_0/\alpha) = O_\ell(H_\alpha)$ uniformly in L. Split $\log(2n)$ into $O_\ell(H_\alpha) + \log C'$ and $\log(1+\Lambda/T_0)$. Using $\log^2(1+x) \le Cx$ and $T_0 \ge C_0(K/A)H_\alpha^2$ in (42) gives

$$E_n/A \le C_\ell(1+\log^2 C')(T_0+\Lambda).$$

Choose C' sufficiently large to ensure $n \ge 4m_{\mathrm{seed}}$, $4E_n/(An) \le \alpha/8$, and $22\Lambda/n \le \alpha/4$. Then $\theta \le \alpha/8$ and $4\eta_{\mathrm{p}} = \alpha/16$, so (45) is at most α. This proves the sample bound

$$C_\ell\left[\left(\frac{D\alpha^{-\ell}}{A} + \frac{D}{\alpha}\right)H_\alpha^2 + \frac{\Lambda}{\alpha}\right].$$

Return up to L largest positive coordinates of $\overline{p}$.

Use the fixed order to break ties. This selection is well defined for a countable summable family: whenever at least one positive weight remains, its supremum is attained, since only finitely many weights can exceed any fixed positive threshold.

Since $\ell \le b \le L$, an omitted label has weight at most $b/(L+1)$, so

$$\mathbf{1}\{h(x) \notin M(x)\} \le \frac{L+1}{L+1-b}(1-\overline{p}_{h(x)}(x)), \qquad \frac{L+1}{L+1-b} \le \ell+1.$$

The last inequality follows from $b = \max\{\ell, \lfloor L/2 \rfloor\}$ and $L \ge \ell$. Set $\alpha = \epsilon/(\ell+1)$. To compare the gain with the stated width dependence, note that $L \le 2b+1$ and $b \ge \ell$ give $b+1-\ell \ge L/(2\ell+1)$. Concavity of $\log(1+x)$ yields

$$\frac{1}{4\ell+2}\log(1+L/\ell) \le \log\left(1+\frac{L/\ell}{4\ell+2}\right) \le A \le \log(1+L/\ell).$$

Thus (4) holds for every $L \ge \ell$, including all integer ceilings after enlarging C_ℓ. For a larger supplied budget, we discard surplus observations. The averaged fractional rule and its rounded output are fixed functions of the training data. □

B.8 A DECODER WITH AN ABSOLUTE COMPRESSION CONSTANT

The larger-width theorem needs a compression constant independent of ℓ. We obtain it by allowing the internal decoder width to be 2ℓ.

Lemma B.7. *Suppose $1 \le s = d_{\ell\mathrm{DS}}(H) < \infty$. For every integer cap $N \ge 1$, there is a fixed decoder taking an ordered code of*

$$r_N = 4s\lceil 16\log(2N)\rceil \le 72s\log(2N)$$

exact-labeled examples and returning a function of width at most 2ℓ. Every nonempty realizable indexed sample of at most N examples has such a code, using its own records with repetition, whose decoded function contains every sample label. The empty sample uses one fixed fallback.

Proof of Lemma B.7. The deterministic one-inclusion predictor in Pabbaraju (2026, Appendix A.1.1, equation (12), and Remark 4) has at most s leave-one-out omissions on each realizable exact-label state. Its output width is ℓ. We fix one admissible orientation on every finite unlabeled input set, together with a fixed rule for ties. At a query input already present in a consistent sample, return its observed label. Repeated inputs then cause no leave-one-out omission when another copy remains; on unique query inputs the pointwise orientation bound applies. Exchangeability gives expected risk at most $s/(k+1)$ from k iid exact-label examples, as explicitly used at $k = 3s$ in that source's Appendix A.2.1. At $k = 4s$ the expected risk is less than $1/4$. Fix arbitrary width-ℓ outputs on unrealizable codes as well.

On a realizable indexed sample of $n \le N$ records, every distribution on rows has an ordered $4s$-subsample with average omission at most $1/4$. Finite minimax supplies a distribution over such subsamples for which each row is omitted with probability at most $1/4$. Draw $T_N = \lceil 16\log(2N)\rceil$ independent subsamples from this distribution. If V_i counts their omissions at row i, then

$$\Pr\{V_i \ge T_N/2\} \le \exp\left[-\left(\frac{\log 2}{2} - \frac{1}{4}\right)T_N\right],$$

since an omission indicator U satisfies $\mathbb{E}2^U \le 5/4 \le e^{1/4}$. The union over $n \le N$ rows has probability less than one, because $16((\log 2)/2 - 1/4) > 1$. Choose subsamples for which every sample label has inclusion frequency greater than $1/2$.

At each input the inclusion frequencies sum to at most ℓ. Returning up to 2ℓ largest positive frequencies includes every label of frequency greater than $1/2$: an omitted positive label has frequency at most $\ell/(2\ell+1)$. Concatenate the fixed-length $4s$ blocks. Their number and boundaries are prescribed by N, so no weights, correction bits or variable boundaries enter the code. Finally, $4s\lceil 16\log(2N)\rceil \le 72s\log(2N)$. Countability permits all finite-set orientations, fallback outputs and tie conventions to be fixed before sampling. □

B.9 An observable seed with additive confidence

Fix $0 < \epsilon, \delta < 1/4$ and write

$$\begin{aligned} H_0 &= \log\frac{e\ell s}{\epsilon}, \qquad u = \log\frac{6}{\delta}, \\ B_1 &= sH_0^2 + u, \quad n_1 = \lceil C_0 B_1/\epsilon \rceil, \end{aligned} \tag{46}$$

where C_0 is a sufficiently large absolute constant. Apply Lemma B.7 with cap n_1, and let $r = r_{n_1}$. Enumerate all ordered r-tuples of seed positions and all choices of one observed bag label at each occurrence, together with the fallback. The resulting family $\nu_1, \dots, \nu_J$ has width $q = 2\ell$ and satisfies

$$J \le 1 + (\ell n_1)^r. \tag{47}$$

The valid-miss mass V_h is defined in Section 4.2. The decoder applied analytically to the valid seed subset supplies an enumerated candidate with zero valid empirical miss. If that subset is empty, use the fallback. For each fixed sequence of code positions, condition on its distinct selected records and decode their true labels. This candidate is now fixed, while the remaining records are independent. If its valid-miss mass exceeds α, the probability of zero valid miss on those records is at most $e^{-\alpha(n_1-r)}$. There are n_1^r position sequences; their true labels are already determined by h, so no factor for guessed labels is needed in this union. Including the fallback gives

$$\Pr\{\min_j V_h(\nu_j) > \alpha\} \le (n_1^r + 1)e^{-\alpha(n_1-r)}. \tag{48}$$

This argument conditions neither on the encoder's chosen code nor on the entire hidden valid subset. Repeated code positions and repeated input values are allowed.

With $\alpha = \epsilon/6$, the choice in (46) ensures $n_1 \ge 2r$ and makes (48) at most $\delta/2$. To see that the constants are absolute, put $x = u/(sH_0^2)$. Then

$$\log(2n_1) \le C_1 H_0 + \log(1+x),$$

where $C_1 = O(1 + \log C_0)$. Since $H_0 \ge \log \ell$, the same form bounds $\log(\ell n_1)$. The inequality $\log^2(1+x) \le Cx$ therefore implies

$$r\log n_1 \le C_2 B_1, \qquad \log J \le C_2 B_1, \tag{49}$$

where $C_2 = O((1+\log C_0)^2)$. Choosing a fixed C_0 large enough makes $\epsilon n_1/12$ dominate $r\log n_1 + \log 2 + \log(2/\delta)$ and also gives $n_1 \ge 2r$.

B.10 The coverage concentration events

We use the notation of Section 4.1.

We use three concentration events. First we expose the current example before the current menu. Then $\mathbb{E}[z_t \mid \text{past, example}] = a_t$, and

$$\mathbb{E}[e^{a_t/2 - (\log 2) z_t} \mid \text{past, example}] \le e^{a_t/2}(1 - a_t/2) \le 1.$$

Iteration and exponential Markov give $\sum_t a_t \le 2\sum_t z_t + 2u$ outside an event of probability e^{-u}. Second we expose the menu before the example. Its conditional mean reward is d_t; since $e^z \le 1 + 2z$ on $[0,1]$,

$$\mathbb{E}[e^{z_t - 2d_t} \mid \text{past, menu}] \le 1.$$

Together with (9), this yields $\sum_t z_t \le 2/\gamma + u$ outside a second such event. These are two valid exposure orders of the same conditionally independent choice and example, and the rewards always use p_{t-1}. Thus (10) implies

$$\sum_t r_t(j) \le 2\log J + 8/\gamma + 8u.$$

Finally we fix one comparator j before the fresh stage and put

$$b_t(j) = \mathbb{E}_P[\mathbf{1}\{Y \in \nu_j(X)\}(1 - p_{t-1}(X, Y))].$$

The first exponential bound applied to $r_t(j)$ conditional on the past gives $\sum_t b_t(j) \le 2\sum_t r_t(j) + 2u$ outside a third event of probability e^{-u}. Monotonicity of p_t bounds its final comparator residual by every $b_t(j)$. We obtain (11).

B.11 The final sample count

Use the horizon T from (12) in the proof of Theorem 3.2.

By (46) and (49), the total $n_1 + T$ is bounded by

$$\frac{C\ell}{\epsilon}\left[s\log^2\frac{e\ell s}{\epsilon} + \log\frac{6}{\delta}\right]$$

for an absolute constant C, proving (5). Ceilings are absorbed in this constant. For every larger integer budget the learner discards surplus observations.

C Lower-bound ingredients

C.1 Supporting arguments for the general-width lower bounds

We first record two dimension facts and one finite statistical experiment. They separate the geometric restriction on the class from the posterior and confidence calculations. The polynomial construction uses this experiment.

C.1.1 Support intersections and a finite observation experiment

Let $\mathcal{J}$ be a countable set of nonzero colours, and let $S_j \subseteq X$ be prescribed supports. The corresponding finite-subset colour class consists of zero and all $j\mathbf{1}_T$ with $T \subseteq S_j$ finite. Write

$$I_b = \sup_{\substack{A\subseteq\mathcal{J}\\|A|=b}} \left|\bigcap_{j\in A} S_j\right|,$$

with supremum zero if fewer than b colours exist.

Lemma C.1. *For every integer $b \ge 2$, the finite-subset colour class has deletion dimension $\Delta_b = I_b$.*

Proof. Choose a finite set E in the intersection of b colour supports. Use $b-1$ of the functions $j\mathbf{1}_E$ as references. The remaining colour and zero realize a full binary cube on E, and every value in that cube survives deletion. This proves $\Delta_b \ge I_b$, including when the supremum is infinite.

For the converse, we fix $b-1$ references and a fully defined deleted DS witness on E, where $|E| \ge 2$. Each trace has at most one nonzero colour. We start with a trace having a nonzero value j, which exists because every coordinate has at least two possible values. While retaining that value at one coordinate, we change every zero coordinate through a DS neighbour. Each changed value must be j. This produces the all-j trace. A neighbour changing any one coordinate of that trace, with another j retained, must put zero at the changed coordinate. Thus zero survives deletion at every input of E.

All references are consequently nonzero everywhere on E. They use $b-1$ distinct fixed colours, each different from j. Their supports and the support of the all-j trace contain E, so $|E| \le I_b$. A singleton witness requires two surviving labels besides the $b-1$ distinct reference values. At least b nonzero colours therefore occur there, which gives $I_b \ge 1$ and proves the remaining case. □

We will also take unions of classes on disjoint input blocks, with a common zero function and every hypothesis extended by zero outside its own block. The deletion dimension of such a union is the supremum of the block dimensions. To see the upper bound, a deleted DS witness cannot meet two blocks: any trace with a nonzero value in one block has no neighbour changing a coordinate in another block while retaining that value. Within one block, references from other blocks restrict to the available zero function. The lower bound follows by retaining a witness from one block. We use this observation only for classes containing zero.

Lemma C.2 (Finite row experiment)**.** *Consider d components, each with n indexed inputs and K private label groups of size q. At each input exactly k groups are incident, and every group is incident to at most $2nk/K$ inputs in its component. Let $1 \le r \le k \le K$, let $r+1 \le \ell \le L$, and suppose*

$$kq \ge 2L, \qquad \mu := \frac{dnr}{K} \ge 60.$$

Suppose the class contains every target obtained by selecting one group in each component, choosing an arbitrary subset of its incident inputs, and assigning any of its private labels on that subset, with zero elsewhere. Then a uniform clean PAC guarantee at

$$\epsilon \le \frac{r}{100K}, \qquad \delta \le \frac{1}{100}$$

requires more than $dn/2$ observations. The experiment uses bags of size $r+1$ and an input law uniform on the dn indexed inputs.

Proof. Write $E_{c,x}$ for the set of groups incident to input x of component c, and $A_{c,j}$ for group j's private labels. Independently at every input, draw an r-subset $C_{c,x}$ uniformly from $E_{c,x}$, and independently choose a uniform label $Y_{c,x,j} \in A_{c,j}$ for each selected group. Display

$$B(c,x) = \{0\} \cup \{Y_{c,x,j} : j \in C_{c,x}\}.$$

Independently select J_c uniformly from $[K]$ in each component. Set $h(c,x) = Y_{c,x,J_c}$ when $J_c \in C_{c,x}$, and set $h(c,x) = 0$ otherwise. This target is in the assumed class, and every displayed bag contains its target value.

The finite row array and hidden groups are drawn before sampling. Each atom of this finite prior specifies a fixed target, a fixed deterministic bag map, and a fixed uniform input law. The training transcript depends on the row array and input draws, and is independent of the hidden groups. Repeated visits reveal the same bag. Conditional on the whole transcript, an unobserved input still has its original independent row, and conditioning further on the learner's independent tape does not change this fact.

Put $p = r/K$. At such an unobserved input, the posterior mass of zero is $1-p$, and each of the kq incident private labels has mass $p/(kq)$. A width-L list covers at most L private labels, so its omission mass is at least $p(1 - L/(kq)) \ge p/2$. At least $dn - m$ input indices remain unobserved after m observations. Every learner therefore has prior-averaged risk

$$\mathbb{E}R_h(M) \ge \frac{p}{2}\left(1 - \frac{m}{dn}\right) \ge \frac{p}{4} \qquad (m \le dn/2). \tag{50}$$

This includes any postprocessing of a learner and remains true if the learner is given the finite incidence design and uniform input law.

Let W be the target's nonzero mass. Conditional on the hidden groups, $Z = dnW$ is binomial: its number of trials is the sum of their incident-input counts, and its success probability is r/k. The degree assumption gives $\theta := \mathbb{E}Z \le 2dnr/K = 2\mu$. For a returned function, let $z(M)$ be the uniform mass of inputs at which it omits zero. Replace the whole output by $\{0\}$ if $z(M) > 8p$, and otherwise keep it; denote the result by $\widehat{M}$. This is observable from the returned function and the known input law. Always

$$R_h(\widehat{M}) \le W + 8p.$$

On $G = \{W \le 4p\}$ its risk is at most $12p$. It preserves every original success at accuracy $\epsilon \le p/100$, because a replaced output had risk at least $z(M) - W > 4p$.

The binomial exponential bound with upper mean 2μ gives

$$\mathbb{P}(G^c) \le e^{-2\mu/3}.$$

For the first-moment tail, if there are N_0 trials, binomial size bias gives

$$\mathbb{E}[Z\mathbf{1}\{Z > 4\mu\}] = \theta\,\mathbb{P}\{1 + \mathrm{Bin}(N_0 - 1, r/k) > 4\mu\}.$$

If $N_0 = 0$, the left side is zero. Otherwise $4\mu - 1 \geq 3\mu$, and a binomial variable with mean at most 2μ exceeds 3μ with probability at most $e^{-\mu/6}$. Both bounds follow from $\mathbb{E}e^{sZ} \leq \exp(2\mu(e^s - 1))$, using $s = \log 2$ and $s = \log(3/2)$, respectively. Dividing the size-biased bound by dn yields

$$\mathbb{P}(G^c) \leq e^{-2\mu/3}, \qquad \mathbb{E}[W\mathbf{1}_{G^c}] \leq 2pe^{-\mu/6}. \tag{51}$$

The estimates are uniform in the hidden groups, so averaging over them preserves both inequalities.

If the original learner were uniformly PAC at the stated accuracy and confidence, its failure probability averaged over the finite prior would be at most δ. On G, its successes are preserved by the cap; every other capped output has risk at most $12p$. On G^c, we use the bound $W + 8p$. Consequently

$$\mathbb{E}R_h(\widehat{M}) \leq \epsilon + 12p\delta + 8p\mathbb{P}(G^c) + \mathbb{E}[W\mathbf{1}_{G^c}] \leq p\left[\frac{13}{100} + 8e^{-2\mu/3} + 2e^{-\mu/6}\right] < \frac{p}{4}.$$

The last inequality holds for $\mu \geq 60$; the two exponential terms are each less than $1/100$. This contradicts (50) when $m \leq dn/2$. Finite-prior averaging therefore selects a fixed target and a fixed clean iid law on which the proposed guarantee fails. □

Lemma C.3 (Dimension and confidence floors)**.** *Suppose a class realizes every assignment from a countably infinite label set on D specified inputs, with zero at a common anchor. Then both bounds in* (17) *hold for every finite L, $0 < \epsilon \leq 1/8$, and $0 < \delta \leq 1/100$, using singleton bags.*

Proof. Give the D inputs equal total mass 4ϵ, and put the remaining mass at the anchor. Choose each target label independently and uniformly from $K_* \geq 4L$ nonzero labels, and display the true singleton label. Replace the returned list at the anchor by $\{0\}$, preserving successes and bounding risk by 4ϵ. If $m \leq D/(32\epsilon)$, the expected number of distinct observed nonanchor inputs is at most $4\epsilon m \leq D/8$. Their expected unseen mass is at least $7\epsilon/2$. Every unseen label is still uniform, so the expected omission risk is at least $(7\epsilon/2)(1 - L/K_*) \geq 21\epsilon/8$. A PAC guarantee would instead give at most $\epsilon + 4\epsilon\delta$, a contradiction.

For the confidence bound, we give one specified input mass 2ϵ and the anchor mass $1 - 2\epsilon$. Choose its target uniformly among $K_* \geq 2L$ nonzero labels. Conditional on no visit to that input, a width-L list misses with probability at least one half, and that omission has risk 2ϵ. Thus

$$\delta \geq \tfrac{1}{2}(1 - 2\epsilon)^m \geq \tfrac{1}{2}e^{-4\epsilon m} \qquad (\epsilon \leq 1/4).$$

It follows that $m \geq \log(1/(2\delta))/(4\epsilon) \geq \log(1/\delta)/(8\epsilon)$ in the stated confidence range. Both experiments have finite priors on fixed targets and fixed singleton observation laws. □

D Proof of the finite-list-dimension threshold

We use the class and private groups from Section 6, with $r = \ell - 1$.

Proof of Theorem 6.1. We first verify the class properties. In any realizable exact-label sample, a nonzero label identifies the target's private group. We retain one such example, if it occurs, and decode it to the constant menu $\{0\} \cup A_j$. If every observed label is zero, we retain nothing and decode to $\{0\}$. These menus have width at most ℓ and cover every original example. No side information is needed.

At a single input, any $\ell + 1$ distinct labels form an ℓ-list DS witness, so $d_{\ell\mathrm{DS}}(\mathcal{G}_\ell) \geq 1$. Suppose that a witness uses at least two inputs and contains a trace with a nonzero value in A_j at one coordinate. Every trace that preserves that value comes from the same group. At any other coordinate, such traces have at most $|\{0\} \cup A_j| = \ell$ possible values, so the chosen trace has at most $\ell - 1$ neighbours changing only that coordinate. This contradicts the witness requirement of ℓ neighbours. A trace family containing no nonzero value consists only of the zero vector and has no neighbours. Hence the dimension is exactly one.

To prove infinite deletion dimension, fix any finite set $E \subset \mathbb{N}$ and one label $c \in A_1$. Use the r distinct labels in A_2 as reference values on E, taking each reference to be zero outside E. All these

functions belong to $\mathcal{G}_\ell$. The targets taking arbitrary values in $\{0, c\}$ on E and zero elsewhere also belong to the class. Their traces survive deletion by these references and form a full binary cube on E. Thus $d_{\mathrm{DS}}((\mathcal{G}_\ell)_F) \geq |E|$, and the arbitrariness of E proves $\Delta_\ell(\mathcal{G}_\ell) = \infty$.

For the statistical lower bound, we construct a finite prior over fixed clean laws and targets. Balance will make every target's nonzero mass deterministic. Observing only finitely many input rows then leaves enough uncertainty about the private labels at unseen inputs to force the claimed risk.

Fix L, m, and the learner. Set

$$K = 2\max\{r, \lceil L/r \rceil\}, \qquad T = \binom{K}{r} r^r.$$

A row selects r distinct groups from $[K]$ and one label from each selected group. There are exactly T row types. Choose a positive multiple N of T large enough that

$$N \geq 4K \max\{m, 1\}. \tag{52}$$

Repeat every row type N/T times and assign this multiset uniformly at random to the inputs $[N]$. Write C_x for the row assigned to input x, and reveal the bag $B(x) = \{0\} \cup C_x$. Every bag has size exactly ℓ. Each group occurs in exactly Nr/K rows, and each of its r labels occurs in exactly N/K rows: a uniform row selects a given group with probability r/K and then a given label with probability $1/r$.

Independently choose a hidden group J uniformly from $[K]$. At input $x \in [N]$, let $h_J(x)$ be the unique label in $C_x \cap A_J$ if that intersection is nonempty, and let it be zero otherwise. Set $h_J = 0$ outside $[N]$. This target belongs to $\mathcal{G}_\ell$. With the uniform input law on $[N]$, the displayed bags give a clean law for every choice of J. After fixing the row assignment and J, the target and observation law are fixed, and the training observations are iid. The entire observed bag law is independent of J.

Each possible target has nonzero support mass exactly $w = r/K$. For the analysis, we augment the learner's output to $M^+(x) = M(x) \cup \{0\}$ and let $U = R_{h_J}(M^+)$. This augmentation can only decrease risk and leaves at most L nonzero predictions at every input, even though its total width may be $L + 1$. Since it is always correct where the target is zero,

$$0 \leq U \leq w. \tag{53}$$

Condition on the complete training transcript and the learner's independent random tape, which fixes its returned function. Let $v \leq m$ be the number of distinct observed inputs, and let v_y count those distinct inputs whose rows contain a given nonzero label y. The hidden group remains uniform and independent. For each unseen input, the unobserved rows are exchangeable. Thus, for every $y \in A_j$ with $j \leq K$,

$$\Pr\{h_J(x) = y \mid \text{transcript, tape}\} = \frac{1}{K} \frac{N/K - v_y}{N - v}. \tag{54}$$

At each unseen input, M^+ omits at least $Kr - L$ of the private labels in the first K groups. Use $v_y \leq m$, sum their posterior masses, and discard the nonnegative risk on seen inputs. Since the unseen inputs have total mass $(N - v)/N$, we obtain

$$\begin{aligned} \mathbb{E}[U \mid \text{transcript, tape}] &\geq \frac{Kr - L}{K} \left(\frac{1}{K} - \frac{m}{N} \right) \\ &\geq \frac{Kr/2}{K} \frac{3}{4K} = \frac{3r}{8K} = \frac{3w}{8}. \end{aligned} \tag{55}$$

Here $Kr \geq 2L$ follows from the choice of K, and $m/N \leq 1/(4K)$ follows from (52). Repeated observations of the same input reveal no new row and are accounted for by v.

Taking expectations in (55) and using (53) yields

$$\frac{3w}{8} \leq \mathbb{E}U \leq \frac{w}{4} + \frac{3w}{4} \Pr\{U > w/4\}.$$

Consequently $\Pr\{U > w/4\} \geq 1/6$. The prior has finite support, so averaging fixes one row assignment and one target for which this probability, now only over the iid sample and learner randomness, is at least $1/6$. The actual risk is at least U.

Finally, $K \le 4\max\{r, L/r\}$ and $r^2 \ge \ell^2/4$, so

$$\frac{w}{4} = \frac{r}{4K} \ge \frac{1}{16}\min\left\{1, \frac{r^2}{L}\right\} \ge \frac{1}{64}\min\left\{1, \frac{\ell^2}{L}\right\}.$$

This proves (23). If $\epsilon < 1/64$ and $L < \ell^2/(64\epsilon)$, its risk threshold exceeds ϵ, contradicting a uniform failure probability below $1/6$. This gives (24); the converse is the stated specialization of Theorem 3.2. □