CORRECT LABELS CAN DESTROY UNIFORM LEARNABILITY: A SHARP HONEST-INSERTION THRESHOLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Can adding correctly labeled examples make a learnable problem unlearnable? For each $\kappa \geq 3$, we construct one fixed class of total binary functions and one fixed known marginal. A learning rule achieves uniformly vanishing risk against all honest insertion channels with fewer than n additions to n IID observations, whereas n additions can force minimax risk toward $1/2$. The adversary retains every original example, uses only target-correct labels, and acts before the test query, but hides which observations were IID. The proof swaps the roles of clean and inserted samples between two target mixtures. Their observed data have a common probability mass approaching one, but their predictions disagree on most queries. An exact transport characterization gives the largest overlap that honest additions can create between two target mixtures' observed data at each budget. Our example is a countable class of binary-digit formulas under the uniform law on $[0, 1]$, not a geometrically regular low-dimensional class. The failure is uniform over targets; each fixed target remains consistently learnable. For this class, marking one original IID observation, chosen independently of its value, restores uniformly vanishing risk at every budget without adding data. Thus correct labels and reliable sampling information play different roles in uniform learning.

1 INTRODUCTION

Suppose a learner receives a random sample and then more correctly labeled examples. It seems that the extra examples should not hurt: none contradicts the target, and none replaces an original example. But the learner may not know which examples were sampled independently and which were selected after seeing the sample. We ask: *does adding only correct examples preserve uniform learnability?*

Monotone adversarial corruption makes this question precise (Larsen et al., 2026). A learner receives n independent examples from a realizable distribution. An adversary sees the clean sample, appends m examples labeled by the same target, and erases provenance by shuffling. The adversary may know the target and distribution, but it must act before an independent test query is drawn. No label is changed and no clean example is removed.

Worst-class guarantees are known: for finite-VC binary classes, empirical risk minimization remains learnable, while a sufficiently large insertion budget produces a logarithmic penalty in the VC rate (Larsen et al., 2026; Mehrotra, 2026). Destruction of learnability and competing-world constructions are also known for multiclass and partial binary PAC learning (Asilis et al., 2026). Our distinction is a sharp risk threshold for *total binary functions under one fixed known marginal*: every budget $m < n$ permits uniformly vanishing risk, whereas every $m \geq n$ forces risk toward $1/2$. This fixed-distribution guarantee complements, rather than dominates, distribution-free PAC results. Fixing the marginal matters: under every fixed purely atomic marginal, memorization remains uniformly consistent at every insertion budget (Proposition 1). Our nonatomic example has the opposite behavior. Knowing P does not reveal which observed examples were drawn from P.

Why exactly n additions? At high levels of our class, a target is almost constant. With high probability all n clean labels equal that constant. Fewer than n added labels cannot overturn this strict majority. A finite search handles the remaining, lower levels. At $m = n$, the two halves can instead exchange roles: what was an IID sample in one world is an inserted sample in the other.

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.

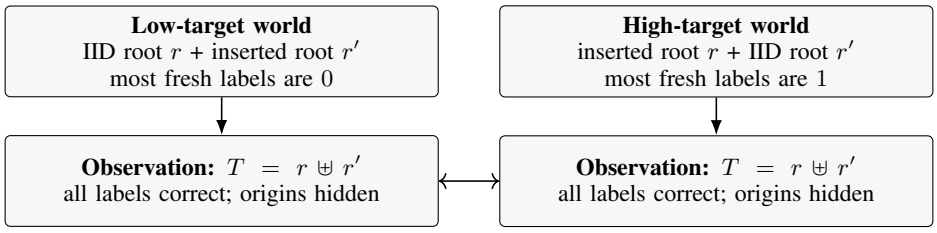


Figure 1: **Correct labels do not identify the IID sample.** Here a "root" means a clean sample of n feature points. In each world the two halves have opposite labels. The construction makes the shared observation have probability mass $1 - o(1)$ in both worlds, although the two target mixtures predict opposite labels on most fresh points. The learner sees neither the origins nor the target pair. For this fixed class, risk vanishes for every $m < n$, but tends to $1/2$ for $m \geq n$ (Theorem 2). These are theoretical limits, not measured risks.

The proof matches their routed observation sublaws while the two worlds predict opposite labels on most fresh points. This explains the threshold for our class; a tie in label counts alone would not prove the lower bound. For example, two clean zeros still outvote one inserted one. Two inserted ones can tie them, but only the equality of the full labeled-data laws prevents a learner from using features to identify the clean half. Figure 1 shows this sample swap. At $m = n$, half of the observed examples are inserted; the result shows that requiring correct labels does not remove this ambiguity.

Our class uses monotone formulas on independent digit blocks. Product panels provide an exact efficient conditional sampler (Proposition 9); they are not claimed necessary for every lower-bound construction.

Two fixed targets with disagreement d have maximum honest observation affinity $(1 - d)^n$ at every budget (Appendix H). Thus near-complementary fixed targets cannot give near-unit overlap; mixtures over compatible target pairs are essential.

Our contributions are:

- **Sharp loss of uniform learnability.** One fixed binary class and known marginal have uniformly vanishing risk for every $m < n$, but risk $1/2 - o(1)$ for every $m \geq n$ (Theorem 2). Atomic marginals remain robust (Proposition 1).
- **Exact budgeted transport.** For every m, maximum honest output affinity equals a clean-state transport value with a root-overlap constraint. It is monotone in m and constant for all $m \geq n$ (Theorem 5); a particular instance may become constant earlier.
- **Recovery from one provenance marker.** For this class, identifying one original IID observation restores uniformly vanishing risk at every budget, with no change to the observed examples (Proposition 3). The marker must be chosen independently of feature and label values.

This is an existence theorem for an infinite-precision class, not a claim about regular geometric classes or named learners. A fixed finite-VC class cannot exhibit this inconsistency: consistency with the enlarged sample implies consistency with its IID subset, so ordinary VC bounds apply (Larsen et al., 2026). Each fixed target remains learnable; the failure is uniform over targets (Appendix H).

2 Model and main result

Let P be a distribution on a standard Borel feature space $\mathcal{X}$ and let $\mathcal{H}$ be a class of measurable functions from $\mathcal{X}$ to $\{0, 1\}$. For target $h \in \mathcal{H}$, the clean sample is $S = ((X_i, h(X_i)))_{i=1}^n$ with $X_i \overset{\text{iid}}{\sim} P$. An exact-budget honest insertion channel K_h maps S to a multiset T that contains S, has $|T| = n + m$, and satisfies $y = h(x)$ for every $(x, y) \in T$. The learner knows $n, m, \mathcal{H}, P$ but observes only T; equivalently, the $n+m$ examples are uniformly shuffled. The channel may depend on h, P, $\mathcal{H}$, and the complete clean sample, but not on the independent test query $X \sim P$.

For a randomized learner A, define the distribution-specific clean and inserted minimax risks

$$\mathfrak{R}_n^{\mathrm{clean}}(\mathcal{H}, P) = \inf_A \sup_{h\in\mathcal{H}} \Pr[A(S)(X) \neq h(X)], \tag{1}$$

$$\mathfrak{R}_{n,m}^{\mathrm{ins}}(\mathcal{H}, P) = \inf_A \sup_{h\in\mathcal{H}} \sup_{K_h} \Pr[A(T)(X) \neq h(X)]. \tag{2}$$

The probabilities include all sample, channel, learner, and query randomness. For the fixed problem constructed below, the asymptotic conclusion is

$$\lim_{n\to\infty} \sup_{0\le m<n} \mathfrak{R}_{n,m}^{\mathrm{ins}} = 0, \qquad \lim_{n\to\infty} \inf_{m\ge n} \mathfrak{R}_{n,m}^{\mathrm{ins}} = \tfrac{1}{2}.$$

Thus even $n-1$ insertions and n insertions have different uniform-risk limits. This is neither an exact finite-n risk formula nor a universal threshold for all classes. Our lower bound uses randomized channels. On standard Borel spaces, a probability kernel has a representation $K_h(S) = \Phi_h(S, U)$ for one independent uniform seed U. Before fixing the seed, replace any invalid output of this representation by the clean sample plus m copies of its first occurrence. Validity is a Borel test (size, containment and target-correct labels), and the replacement has probability zero for every input under an honest kernel. The modified map is honest for every input and every seed. For each learner and finite-level prior, average over one seed for the entire target-indexed channel family. A seed with at least the prior-average risk exists; fixing it jointly for all target–root inputs gives a deterministic honest family with the same lower bound. The adversary may depend on the learner, but not its realized randomness or query.

Proposition 1 (Atomic-marginal converse)**.** *Suppose P is fixed and purely atomic on the standard Borel space $\mathcal{X}$, and let $\mathrm{At}(P) = \{x : P(\{x\}) > 0\}$. For every binary class $\mathcal{H}$, $n \ge 1$, and finite $m \ge 0$,*

$$\mathfrak{R}_{n,m}^{\mathrm{ins}}(\mathcal{H}, P) \le M_n(P) := \sum_{x\in\mathrm{At}(P)} P(\{x\})(1 - P(\{x\}))^n. \tag{3}$$

Hence $\mathfrak{R}_{n,m_n}^{\mathrm{ins}}(\mathcal{H}, P) \to 0$ for every finite budget sequence (m_n).

Proof. Given T, predict the unique observed label at each feature in T and zero elsewhere. Honesty makes the observed label unique, and $S \subseteq T$ implies $\mathrm{supp}(S) \subseteq \mathrm{supp}(T)$. Pointwise in the clean root and all channel randomness, the query error is therefore at most $P(X \notin \mathrm{supp}(S))$. Averaging the IID root yields Equation (3). Each summand tends to zero and is dominated by $P(\{x\})$, so dominated convergence completes the proof. □

Without insertions, this is the classical universal consistency mechanism of the Memorize rule on countable domains (Shalev-Shwartz & Ben-David, 2014); Proposition 1 records that target-correct additions cannot worsen its certificate.

2.1 Product-panel formulas

Fix integers s, L, w with $1 \le w \le L$ and set $D = sL$. Write a cube point as $x = (x_{j,k})_{j\in[s],k\in[L]}$ and let U_D be uniform on $\{0,1\}^D$. A target index $A = (A_1, \dots, A_s)$ chooses $A_j \subseteq [L]$ with $|A_j| = w$ in each panel. Define the low DNF and high CNF

$$f_A^-(x) = \bigvee_{j=1}^{s} \bigwedge_{k\in A_j} x_{j,k}, \qquad f_A^+(x) = \bigwedge_{j=1}^{s} \bigvee_{k\in A_j} x_{j,k}. \tag{4}$$

They are dual: $f_A^+(x) = 1 - f_A^-(\mathbf{1} - x)$. Let

$$\mathcal{H}_{s,L,w} = \{f_A^-, f_A^+ : |A_j| = w \text{ for every } j\}, \qquad M = \binom{L}{w}^s. \tag{5}$$

Every target is one monotone depth-two formula of size $O(sw)$ over the same D coordinates.

Panel independence gives the common minority mass

$$\epsilon_{s,w} = \Pr_{X\sim\mathsf{U}_D}[f_A^-(X) = 1] = \Pr_{X\sim\mathsf{U}_D}[f_A^+(X) = 0] = 1 - (1 - 2^{-w})^s \le s2^{-w}. \tag{6}$$

Thus the clean rule that predicts the first observed label everywhere has exact risk $2\epsilon_{s,w}(1-\epsilon_{s,w})$ for either target type. For $Z \sim \mathrm{Binomial}(L, 1/2)$, write

$$\beta_{L,w} = \Pr[Z < w] = 2^{-L} \sum_{k=0}^{w-1} \binom{L}{k}.$$

2.2 One fixed class and marginal

Fix any constant $\kappa \geq 3$. For every $s \geq 2$, let $w_s = \lceil \kappa \log_2 s \rceil$ and $L_s = 8w_s$. Form the countable coordinate set

$$\mathcal{C} = \bigsqcup_{s=2}^{\infty} \{s\} \times [s] \times [L_s], \qquad \mathcal{X}_\infty = \{0,1\}^{\mathcal{C}}, \qquad \mathsf{P}_\infty = \mathrm{Ber}(1/2)^{\mathcal{C}}.$$

Embed $\mathcal{H}_{s,L_s,w_s}$ on the level-s coordinates and let every target ignore all other coordinates. The fixed class is

$$\mathcal{H}_\infty = \bigcup_{s=2}^{\infty} \mathcal{H}_{s,L_s,w_s}. \tag{7}$$

This is one countable measurable class on one standard Borel cube; neither $\mathcal{H}_\infty$ nor P_∞ depends on n. Both are fixed after choosing κ; changing the requested rate parameter changes the constructed class. Write $\epsilon_s = \epsilon_{s,w_s}$ and $\beta_s = \beta_{L_s,w_s}$.

To realize the same construction on one real feature, enumerate $\mathcal{C}$ as $c_1, c_2, \ldots$ and let $d_k(x)$ be the kth canonical binary digit of $x \in [0,1]$ (with every digit of 1 set to one). Define $\Psi(x)_{c_k} = d_k(x)$, let $\mathsf{U}_{[0,1]}$ be Lebesgue measure, and set

$$\mathcal{G}_\infty = \{h \circ \Psi : h \in \mathcal{H}_\infty\}. \tag{8}$$

The digit map is Borel and $\Psi_{\#}\mathsf{U}_{[0,1]} = \mathsf{P}_\infty$; every target in $\mathcal{G}_\infty$ depends on finitely many digits. Appendix A handles the null binary-expansion ambiguity and transfers the honest channel on every input.

Theorem 2 (Sharp threshold for uniform learnability)**.** *The following bounds hold for both* $(\mathcal{G}_\infty, \mathsf{U}_{[0,1]})$ *and* $(\mathcal{H}_\infty, \mathsf{P}_\infty)$*. There is a constant* $C_\kappa < \infty$ *such that, for every* $n \geq 2$*,*

$$\sup_{0 \leq m < n} \mathfrak{R}^{\mathrm{ins}}_{n,m}(\mathcal{G}_\infty, \mathsf{U}_{[0,1]}) \leq C_\kappa n^{-1+2/\kappa} \log^{1-1/\kappa}(n+1), \tag{9}$$

$$\inf_{m \geq n} \mathfrak{R}^{\mathrm{ins}}_{n,m}(\mathcal{G}_\infty, \mathsf{U}_{[0,1]}) \geq \frac{1}{2}\left[1 - n\epsilon_n - n\beta_n\right]_+ (1 - 2\epsilon_n). \tag{10}$$

Consequently, choosing $\kappa = 2/\eta$ *gives risk* $O_\eta(n^{-1+\eta}\log^{1-\eta/2}(n+1))$ *for every fixed* $\eta \in (0, 2/3]$ *uniformly for* $m < n$ *(including clean learning at* $m = 0$*), while every* $m \geq n$ *forces minimax risk* $1/2 - o(1)$*. More explicitly, the latter lower bound is* $1/2 - O_\kappa(n^{2-\kappa})$*.*

The same threshold holds for a fixed problem in which every target has population label probability exactly $1/2$: Proposition 13 preserves both minimax risks by adding an observed fair feature bit and XOR-ing labels, not by hidden label noise.

For the upper bound, let $\mathcal{F}_t$ contain levels 2 through t. Return its first member consistent with T, if one exists; otherwise predict the majority label. Search must come first. Retained-root consistency bounds the search contribution by $(\log|\mathcal{F}_t| + 1)/n$. Fallback requires a target above level t; with probability at least $1 - n\epsilon_s$, all clean labels equal its population majority, which $m < n$ additions cannot overturn. Thus

$$\sup_{h, K_h} \mathbb{E}[\mathrm{risk}] \leq \frac{\log|\mathcal{F}_t| + 1}{n} + \frac{n+1}{(t+1)^{\kappa-1}}. \tag{11}$$

Taking $t = \max\{2, \lfloor n^{2/\kappa}/\log^{1/\kappa}(n+1)\rfloor\}$ gives Equation (9).

For the lower bound, choose the prior on level $s = n$. Apply the finite channel below to that level projection and fill every unused digit of every inserted row independently from the fair product law. Both latent worlds then have the same unused-coordinate law, while every target ignores those coordinates. The finite exchanged projections therefore retain exact honesty, query blindness, and output

overlap on the full cube, proving Equation (10). Appendix A supplies the complete uniform and measure-theoretic argument. For $m > n$, pad the common observation by repeating its canonically first labeled occurrence $m - n$ times. This common map preserves honesty and overlap; random guessing gives the matching upper bound $1/2$.

Proposition 3 (Recovery from one provenance marker)**.** *For either fixed problem in Theorem 2, reveal one original clean occurrence, selected by a fixed index or independently of all feature and label values. Using the same enlarged sample T, a learner achieves risk*

$$O_\kappa\Big(n^{-1+1/\kappa}\log^{1-1/\kappa}(n+1)\Big)$$

uniformly over targets and every finite insertion budget. The channel may know the marked occurrence.

Search a finite prefix first; if search fails, predict the marked label everywhere. Its unconditional error is twice the product of the two population label probabilities. At high levels that error is small. Appendix H proves the bound without conditioning on search failure. The data and budget are unchanged: only one provenance marker is added. This recovery is specific to our class, not a general one-example defense.

2.3 Finite level theorem

Theorem 4 (Finite-level insertion bound)**.** *For every integer $s \geq 2$, set*

$$n = m = s, \qquad w = \lceil \kappa \log_2 s \rceil, \qquad L = 8w, \qquad D = sL,$$

and let $\epsilon = \epsilon_{s,w}$ and $\beta = \beta_{L,w}$. Then

$$\mathfrak{R}_n^{\mathrm{clean}}(\mathcal{H}_{s,L,w}, \mathsf{U}_D) \leq 2\epsilon(1-\epsilon) \leq 2s^{1-\kappa}, \tag{12}$$

$$\mathfrak{R}_{n,n}^{\mathrm{ins}}(\mathcal{H}_{s,L,w}, \mathsf{U}_D) \geq \frac{1}{2}\left[1 - s\epsilon - s\beta\right]_+ (1-2\epsilon). \tag{13}$$

Consequently, $D = \Theta_\kappa(n \log n)$, the clean risk is $O(n^{1-\kappa})$, and the inserted risk is $1/2 - o(1)$ using exactly one inserted example per clean example. Moreover, the finite-level insertion channel has an exact expected-polynomial sampler with $O(nD)$ coordinate work.

The proof has four parts: formulas supply disagreement; compatible roots give legal insertions; balanced flow gives common observation mass; the prefix learner gives the upper bound. Sections 3–4 prove the finite lower bound.

3 When can correct additions hide the clean sample?

This section characterizes the largest possible observation overlap independently of the formula family. Let $\mathcal{H}^-$ and $\mathcal{H}^+$ be finite target sets with priors π^- and π^+, and let all targets share feature distribution P. A root $r = (x_1, \ldots, x_n) \in \mathcal{X}^n$ has law $p = P^n$. Targets $a \in \mathcal{H}^-$ and $b \in \mathcal{H}^+$ are *compatible at* r when

$$a(x_i) = b(x_i) \quad \text{for every } i \in [n]. \tag{14}$$

Write $Q^-(da, dr) = \pi^-(a)P^n(dr)$ and $Q^+(db, dr') = \pi^+(b)P^n(dr')$ for the two clean target–root laws. Let $N_r(x)$ be the multiplicity of x in r. For $m \in \mathbb{N}_0$, write $(a, r) \sim_m (b, r')$ when the targets are compatible at both roots and

$$\sum_x \min\{N_r(x), N_{r'}(x)\} \geq n - m. \tag{15}$$

Thus related roots share at least $n - m$ occurrences, counted with multiplicity. A budget-m subcoupling is a nonnegative measure γ on the product source space, supported on $\sim_m$, with marginals dominated by Q^- and Q^+. Define its supremum mass

$$\tau_m = \sup_\gamma |\gamma|. \tag{16}$$

For output laws W^- and W^+ on shuffled labeled multisets, let $\mathsf{Aff}(W^-, W^+) = |W^- \wedge W^+|$, the mass of their greatest common submeasure. On finite spaces this is $\sum_T \min\{W^-(T), W^+(T)\}$.

Theorem 5 (Budgeted honest-superposition transport)**.** *For standard Borel $\mathcal{X}$, finite target sets, $n \geq 1$, and every exact insertion budget $m \in \mathbb{N}_0$,*

$$\sup_{K^-,K^+ \text{ honest}} \mathsf{Aff}(W^-, W^+) = \tau_m, \tag{17}$$

where the channels may be randomized and depend arbitrarily on their source target and root, but not on the future query. The transport curve satisfies

$$\tau_0 \leq \tau_1 \leq \cdots \leq \tau_n \quad \textit{and} \quad \tau_m = \tau_n \textit{ for all } m \geq n. \tag{18}$$

Moreover, every feasible γ yields honest channels under whose equal type mixture every learner has Bayes risk at least

$$\frac{1}{2} \int P\{x : a(x) \neq b(x)\}\, \gamma(d(a,r), d(b,r')). \tag{19}$$

The equality isolates the role of budget. Any output common to two source states contains the multiset union of their roots. That union has size $2n - \sum_x \min\{N_r(x), N_{r'}(x)\}$, so it fits in $n + m$ positions exactly under Equation (15); target agreement on the union is also necessary. Conversely, duplicate any occurrence of a compatible union until it has exactly $n + m$ entries. The relations therefore expand with m and are unchanged for $m \geq n$, proving Equation (18). The optimal value may already be constant at a smaller budget. The proof routes any feasible subcoupling to common union outputs; conversely, it lifts common output mass back to paired source states. Appendix B makes both steps measurable and handles output collisions. On finite spaces the suprema are attained. Equation (17) characterizes observation affinity, not prediction risk. Equation (19) also weights each paired state by target disagreement. If the two priors concentrate on the same target, affinity is one but this risk certificate is zero. Our construction needs both large common mass and large disagreement.

When the affinity does jump at n. Suppose every low target has minority mass at most ϵ, and so does every high target. For $m < n$, a related pair of roots must share an occurrence. If all low-root labels are zero and all high-root labels are one, compatibility at that occurrence is impossible. Every feasible subcoupling is therefore supported where at least one root contains a minority label. Marginal domination and a union bound give

$$\tau_m \leq 2n\epsilon \quad (m < n).$$

For our level-n priors, Proposition 9 instead gives $\tau_n \geq [1 - n\epsilon_n - n\beta_n]_+$. Thus this example has a jump from vanishing overlap below n to overlap approaching one at n. This observation concerns affinity; the prediction-risk upper bound still needs the prefix learner.

We next give a structured certificate for γ on a finite feature space, as used by each product-panel level.

Suppose a forward kernel $q_-(b \mid a, r)$ and reverse kernel $q_+(a \mid b, r)$ select only compatible partners; they may have total mass zero when no partner exists. Define the pair–root submeasures

$$u_{ab}(r) = \pi^-(a)p(r)q_-(b \mid a, r), \qquad v_{ab}(r) = \pi^+(b)p(r)q_+(a \mid b, r), \tag{20}$$

$$F_{ab} = \sum_r u_{ab}(r), \qquad G_{ab} = \sum_r v_{ab}(r). \tag{21}$$

Definition 6 (Balanced compatible-root flow)**.** *The two partner kernels are balanced when $F_{ab} = G_{ab}$ for every target pair (a, b).*

Balance is a statement about probability volume, not merely partner existence. It is the condition needed by the following construction. In plain terms, compatibility makes the added labels legal, balance makes the two observed-data laws agree, and disagreement makes the hidden choice of target matter at test time. If the common mass is ρ and each paired target disagrees on a fraction at least δ of queries, the error is at least $\rho\delta/2$.

Theorem 7 (Balanced-root exchange)**.** *Assume $\mathcal{X}$ is finite and the compatible-root flow is balanced. There are query-blind honest insertion channels with exact budget $m = n$ such that, under the equal mixture of the two target priors, every learner has Bayes risk at least*

$$\frac{1}{2} \sum_{a,b} F_{ab}\, P\{x : a(x) \neq b(x)\}. \tag{22}$$

The bound holds after aggregating all target-pair and output collisions; the learner does not observe a pair index.

Construction. For a positive-flow pair, normalize

$$\mu_{ab}(r) = u_{ab}(r)/F_{ab}, \qquad \nu_{ab}(r) = v_{ab}(r)/F_{ab}.$$

From the low world, route the observed root r to b with q_- and insert an independent root $r' \sim \nu_{ab}$. From the high world, route r' to a with q_+ and insert an independent root $r \sim \mu_{ab}$. The pair-output submeasures agree pointwise because

$$u_{ab}(r)\nu_{ab}(r') = \frac{u_{ab}(r)v_{ab}(r')}{F_{ab}} = v_{ab}(r')\mu_{ab}(r). \tag{23}$$

Both channels label insertions by the source target and output the shuffled multiset union, retaining every clean occurrence and its multiplicity. Compatibility makes every routed label correct under both paired targets; the inserted root is pair-conditioned, not an ordinary IID draw from P. On failed routing, append any n copies of clean occurrences (for $n \geq 1$); the implementation duplicates the clean root.

Equivalently, balanced exchange constructs a feasible budget-n subcoupling $\gamma((a,r),(b,r')) = u_{ab}(r)v_{ab}(r')/F_{ab}$ on positive-flow pairs. Its mass is $\sum_{a,b} F_{ab}$, so Theorem 5 shows exactly what portion of output-law overlap this certificate explains.

At a fixed output and query where a and b disagree, the paired worlds contribute equal mass to label zero and label one. Summing these nonnegative equal contributions over all colliding pairs proves Equation (22); Appendix C gives the aggregated-output proof. This collision argument is important: no injective decoding of the latent pair is assumed.

4 Product panels give an exact sample swap

We apply Theorem 7 with uniform priors over the M low and M high targets. The main construction uses the efficiently sampled barred kernels below. Appendix D gives the separate uniform-partner construction used in some exact checks.

4.1 One panel per row gives partners

Let $d_-(A,r)$ count the high targets compatible with low A at r, and define $d_+(B,r)$ symmetrically. Write

$$p^-_{\text{good}} = \Pr_{A,r}[d_-(A,r) > 0], \qquad p^+_{\text{good}} = \Pr_{B,r}[d_+(B,r) > 0].$$

Complement symmetry gives equality, while a direct construction lower-bounds each side.

Lemma 8 (Partner existence)**.** *If $n \leq s$, then*

$$p^-_{\text{good}} = p^+_{\text{good}} \geq [1 - n\epsilon_{s,w} - n\beta_{L,w}]_+ . \tag{24}$$

For a low root whose labels are all zero, assign clean row i to panel i. If that row contains at least w zeros in its assigned panel, choose the ith high-target block from those zero coordinates. The corresponding high clause is zero on row i, so the high target agrees there. Different rows modify disjoint panels and cannot conflict. The root has a nonzero low label with probability at most $n\epsilon_{s,w}$, and an assigned panel lacks w zeros with probability at most $n\beta_{L,w}$. Bit complement proves the high-root statement.

Choosing the assigned blocks independently gives the following efficient kernels.

Proposition 9 (Exact efficient panel exchange)**.** *Assume $n = s$, $L = 8w$, $w \geq 1$. On each channel invocation, first order the multiset occurrences uniformly at random, independently of other channel randomness, before selecting a partner. Write the resulting root as $r = (x_1, \ldots, x_s)$; row i is assigned to panel i. Let $Z_i(x)$ and $O_i(x)$ be the zero and one coordinates of panel i in x. The*

kernels

$$\bar{q}_-(B \mid A, r) = \mathbf{1}\{f_A^-(x_i) = 0 \; \forall i\} \prod_{i=1}^{s} \frac{\mathbf{1}\{B_i \subseteq Z_i(x_i)\}}{\binom{|Z_i(x_i)|}{w}}, \tag{25}$$

$$\bar{q}_+(A \mid B, r) = \mathbf{1}\{f_B^+(x_i) = 1 \; \forall i\} \prod_{i=1}^{s} \frac{\mathbf{1}\{A_i \subseteq O_i(x_i)\}}{\binom{|O_i(x_i)|}{w}} \tag{26}$$

are interpreted as zero when a denominator is unavailable. They form a balanced compatible-root flow, and each row sum is zero or one. Their total routed mass is at least the right side of Equation (24). Partner selection uses $O(D)$ coordinate work, and the pair-conditional s-root laws needed by balanced exchange can be sampled exactly with expected $O(nD)$ coordinate work.

For balance, π_{AB} swaps $A_i \setminus B_i$ with $B_i \setminus A_i$ within each panel. It preserves the uniform root law, every denominator, and the uniform target prior; it exchanges the two low-kernel indicators. With $c(r) = \mathbf{1} - r$, the pair–root measures therefore satisfy

$$(\pi_{AB})_{\#}\bar{u}_{AB} = \bar{u}_{BA}, \qquad c_{\#}\bar{u}_{BA} = \bar{v}_{AB}.$$

Taking masses gives $\bar{F}_{AB} = \bar{F}_{BA} = \bar{G}_{AB}$. Complement transposes the target indices: $\nu_{AB}(r) = \mu_{BA}(\mathbf{1} - r)$. For sampling, conditioning on (A, B) factors the root law over rows and panels. The only non-product-bit factor is a panel subset with weight $\mathbf{1}\{C \subseteq S,\ S \cap R \neq \varnothing\}/\binom{|S|}{w}$; an exact constant-acceptance rejection sampler is given in Appendix E.

4.2 Bayes ambiguity and proof of Theorem 4

For every pair A, B, a union bound under U_D gives

$$\Pr[f_A^-(X) \neq f_B^+(X)] \geq \Pr[f_A^-(X) = 0, f_B^+(X) = 1] \geq 1 - 2\epsilon_{s,w}. \tag{27}$$

Summing the efficient flows in Proposition 9 gives routed mass at least the right side of Equation (24). Theorem 7, Proposition 9, and Equation (27) prove Equation (13) and its sampling claim. The first-label rule and Equation (6) prove Equation (12). Finally, for $L = 8w$, a multiplicative Chernoff bound yields

$$\beta_{L,w} = \Pr[\mathrm{Binomial}(8w, 1/2) < w] \leq e^{-9w/8}.$$

With $w = \lceil \kappa \log_2 s \rceil$, both failure terms vanish and $D = 8sw = \Theta_\kappa(s \log s)$.

5 Checks and finite-scale illustration

Exact rational checks guard kernel orientation, normalization, collision aggregation, budgeted transport, and honest labeling. Appendix I separates the uniform-partner checks from those of the efficient kernels and implemented sampler. These finite checks do not replace the all-parameter proofs.

Finite-scale interpretation. Appendix J studies fixed trees, an MLP, and a panel GNN. Discarding insertions recovers the clean baseline by construction. Conditional label-count matching largely reproduces the degradation, so the experiments do not establish an extra paired-target feature-selection effect. The hard-example pool also supplies fewer minority-label candidates, confounding that comparison. These controls limit interpretation: near-chance errors for fitted learners do not prove the arbitrary-learner lower bound.

6 Related work

Monotone adversarial corruption. Larsen et al. (2026) introduced learning under adaptive insertion of correctly labeled examples and established the query-blind model studied here. Mehrotra (2026) determined the worst-budget VC rates and left the optimal minimax risk at intermediate budgets open. Our theorem fixes one class and one feature marginal and proves a prediction-risk threshold at $m = n$ for that problem, not optimal VC rates in general. Asilis et al. (2026) prove

destruction of learnability for multiclass and partial binary classes, including a linear-budget multiclass separation. Their hard marginals vary with the instance; our witness instead consists of total binary functions under a single known marginal. These are complementary guarantees, not a strict dominance claim. Johny (2026) studies insertion-stability as a structural route to positive guarantees beyond dimension alone.

Fixed-distribution and adaptive-addition learning. Clean learnability under one known distribution is classical; Benedek & Itai (1991) characterize it through finite covers. Our thresholded prefix learner is an elementary fixed-distribution argument, not a new clean-learning theory. Uniformity is the key distinction: nonuniform learning allows sample requirements to depend on the target (Benedek & Itai, 1994), while universal learning studies rates without uniformity over data distributions (Bousquet et al., 2021). Here P is fixed, the guarantee lost is uniform over targets, and each fixed target remains learnable after insertion (Appendix H). The universal consistency of memorization on countable domains is also classical (Shalev-Shwartz & Ben-David, 2014); our converse applies the same support argument to every purely atomic marginal and uses only the insertion-specific observation that honest additions cannot remove clean occurrences. For distribution estimation, Lechner et al. (2025) show that additive adaptive adversaries can make a learnable class of distributions unlearnable. Their adversary adds arbitrary samples while the unknown object is a distribution. Here the clean marginal is fixed, the unknown object is a binary target, and every appended label must remain correct under that target.

Clean-label attacks. The clean-label attackable-rate model permits attacks tailored to the test instance while preserving correct labels (Blum et al., 2021). Our adversary never observes the fresh query. Polynomial-time p-tampering attacks can replace selected correctly labeled examples to bias a targeted learner (Mahloujifar et al., 2018); our channel instead appends examples and exactly balances provenance-free observation laws. The separation therefore comes from provenance erasure in the training experiment, not from placing a point around a known test input. Practical poisoning also includes backdoor objectives and algorithm-specific increases in test error (Wang et al., 2022). Our result concerns an average fresh query, permits knowledge of the full target, and proves a statement about all learners on an encoded class. It neither constructs a trigger nor establishes a vulnerability of a trained practical system.

Couplings and experiment comparison. Relation-supported couplings, probabilistic relational liftings, and comparison of statistical experiments are classical (Strassen, 1965; Barthe et al., 2015; Blackwell, 1953; Torgersen, 1991). Classical arbitrarily varying-channel theory likewise uses symmetrizability to characterize when adversarial state selection destroys deterministic capacity (Csiszár & Narayan, 1988). We do not claim a new coupling, total-variation, product-measure, or generic symmetrizability principle. The new statements are operational: Theorem 5 replaces the output space with an exact root-overlap transport problem for legal insertion channels and shows that its value is constant for $m \geq n$; balanced flow supplies an explicit certificate; and product panels give common mass approaching one with paired targets that disagree almost everywhere.

7 Limitations and conclusion

Our class uses an infinite digit dictionary, with $\Theta(n \log n)$ active digits at level n (Appendix G), not a regular Euclidean or finite-precision model. Atomic marginals and fixed finite-VC classes cannot exhibit this inconsistency; regular infinite-VC cases remain open. Each fixed target remains learnable. The clean learner may search a large prefix; the target-aware attack has an expected-polynomial finite projection but needs a lazy IID oracle for unused digits.

Population label imbalance is not necessary: an observed fair-bit transformation preserves both minimax risks and balances every target, but loses monotone syntax (Appendix H). An authenticated IID reserve recovers its clean-sample guarantee; for our class, one marked occurrence suffices. More generally, finite-cover learnability survives $m = o(n)$ insertions. Neither observation affinity alone nor these positive results characterize risk for arbitrary classes.

Correct labels need not identify representative data. For our fixed class, n insertions destroy uniform learning, while one provenance marker restores it. These are statistical statements about the specified model, not evidence of a deployed-system vulnerability.

USE OF GENERATIVE AI SYSTEMS

Generative AI assisted with literature discovery, formulation exploration of mathematical claims, checker design, code implementation, and manuscript drafting and editing.

ETHICS STATEMENT

This theoretical construction could inform data-poisoning analysis, but it is target-aware and demonstrated only on a synthetic countable formula class; it uses no human data or deployed system. Its intended use is defensive: the theorem motivates provenance preservation, limits on adaptive selection channels, and evaluation of learning pipelines under honest-but-endogenous data addition.

REPRODUCIBILITY STATEMENT

Appendices A–I contain complete proofs and checker specifications. The five exact rational enumerators and the prefix/majority check use only the Python standard library; Appendix J specifies the stochastic validation and its dependencies. The anonymous code supplement includes completed records, frozen generator sources, protocols, dependency versions, a SHA-256 inventory, and commands for exact checks and reconstruction outside Git. Its README distinguishes regeneration of diagnostics from new neural training. The imported-sampler regression uses NumPy and tests the implemented conditional laws separately from the standalone enumerators.

REFERENCES

Julian Asilis, Shaddin Dughmi, and Chirag Pabbaraju. When clean data hurts: Learning with monotone corruptions beyond binary classification, 2026. URL https://arxiv.org/abs/2608.20480.

Gilles Barthe, Thomas Espitau, Benjamin Grégoire, Justin Hsu, Léo Stefanesco, and Pierre-Yves Strub. Relational reasoning via probabilistic coupling. In *Logic for Programming, Artificial Intelligence, and Reasoning*, volume 9450 of *Lecture Notes in Computer Science*, pp. 387–401. Springer, 2015. doi: 10.1007/978-3-662-48899-7_27. URL https://doi.org/10.1007/978-3-662-48899-7_27.

Gyora M. Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, September 1991. ISSN 0304-3975. doi: 10.1016/0304-3975(91)90026-X. URL https://doi.org/10.1016/0304-3975(91)90026-X.

Gyora M. Benedek and Alon Itai. Nonuniform learnability. *Journal of Computer and System Sciences*, 48(2):311–323, April 1994. ISSN 0022-0000. doi: 10.1016/s0022-0000(05)80005-4. URL http://dx.doi.org/10.1016/s0022-0000(05)80005-4.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032. URL https://doi.org/10.1214/aoms/1177729032.

Avrim Blum, Steve Hanneke, Jian Qian, and Han Shao. Robust learning under clean-label attack. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 591–634. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/blum21a.html.

Olivier Bousquet, Steve Hanneke, Shay Moran, Ramon van Handel, and Amir Yehudayoff. A theory of universal learning. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC '21, pp. 532–541. ACM, June 2021. doi: 10.1145/3406325.3451087. URL http://dx.doi.org/10.1145/3406325.3451087.

Imre Csiszár and Prakash Narayan. The capacity of the arbitrarily varying channel revisited: Positivity, constraints. *IEEE Transactions on Information Theory*, 34(2):181–193, March 1988. doi: 10.1109/18.2627. URL https://doi.org/10.1109/18.2627.

Joseph Sankoorikal Johny. When does more correct data hurt? insertion-stability and the limits of dimension-based theory, 2026. URL https://arxiv.org/abs/2608.14020.

Kasper Green Larsen, Chirag Pabbaraju, and Abhishek Shetty. Learning with monotone adversarial corruptions, 2026. URL https://arxiv.org/abs/2601.02193.

Tosca Lechner, Alex Bie, and Gautam Kamath. On the learnability of distribution classes with adaptive adversaries. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32853–32877. PMLR, 13–19 July 2025. URL https://proceedings.mlr.press/v267/lechner25a.html.

Saeed Mahloujifar, Dimitrios I. Diochnos, and Mohammad Mahmoody. Learning under p-tampering attacks. In Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan (eds.), *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pp. 572–596. PMLR, 07–09 Apr 2018. URL https://proceedings.mlr.press/v83/mahloujifar18a.html.

Anay Mehrotra. Optimal rates for learning with monotone adversaries, 2026. URL https://arxiv.org/abs/2608.06337.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, May 2014. ISBN 9781107298019. doi: 10.1017/CBO9781107298019. URL https://doi.org/10.1017/CBO9781107298019.

V. Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965. doi: 10.1214/aoms/1177700153. URL https://doi.org/10.1214/aoms/1177700153.

Erik Torgersen. *Comparison of Statistical Experiments*. Cambridge University Press, 1991. ISBN 9780511666353. doi: 10.1017/CBO9780511666353. URL https://doi.org/10.1017/CBO9780511666353.

Zhibo Wang, Jingjing Ma, Xue Wang, Jiahui Hu, Zhan Qin, and Kui Ren. Threats to training: A survey of poisoning attacks and defenses on machine learning systems. *ACM Computing Surveys*, 55(7):1–36, 2022. doi: 10.1145/3538707. URL https://doi.org/10.1145/3538707.

A PROOF OF THE FIXED-CLASS THEOREM

The coordinate set $\mathcal{C}$ is countable, so $\mathcal{X}_\infty$ with its product sigma algebra is standard Borel. Every target is a measurable cylinder function and $\mathcal{H}_\infty$ is countable. Fix an enumeration within every level and then by increasing level; the learner described below is therefore measurable.

Uniform upper bound below the insertion threshold. Let $t = \max\{2, \lfloor n^{2/\kappa}/\log^{1/\kappa}(n+1)\rfloor\}$ and $\mathcal{F}_t = \bigcup_{s=2}^{t} \mathcal{H}_{s,L_s,w_s}$. The level sizes obey

$$|\mathcal{H}_{s,L_s,w_s}| = 2\binom{8w_s}{w_s}^s, \qquad \log|\mathcal{H}_{s,L_s,w_s}| \le \log 2 + sw_s\log(8e) \le c_\kappa s\log(s+1).$$

Thus, for $N_t = |\mathcal{F}_t|$,

$$\log N_t \le c'_\kappa t\log(t+1). \tag{28}$$

Search $\mathcal{F}_t$ for a hypothesis consistent with the entire multiset T. If one exists, return the first; otherwise return its majority label (break ties at zero). Let E denote search success and $L_P(g) = P(g(X) \ne h(X))$. For fixed g with $L_P(g) > q$, IID sampling gives $\Pr[g \text{ is consistent with } S] \le e^{-nq}$. Since $S \subseteq T$, for every honest adaptive channel,

$$\Pr(E \text{ and } L_P(\widehat{h}) > q) \le \min\{1, N_t e^{-nq}\}.$$

This does not require the target to belong to the prefix or condition the IID sample on E. Integrating over $q \in [0,1]$ yields $\mathbb{E}[\mathbf{1}_E L_P(\widehat{h})] \le (\log N_t + 1)/n$.

If the target belongs to a level $s \le t$, honesty makes it consistent with T, so E always holds. For $s > t$, its minority mass satisfies $\epsilon_s \le s^{1-\kappa} \le 1/4$. Let G be the event that all n clean labels equal its population-majority label. Then $\Pr(G^c) \le n\epsilon_s$. On G, the majority of T is still that label because $m < n$, regardless of the inserted features, labels or channel randomness. Thus the fallback risk contribution is at most

$$\mathbb{E}[\mathbf{1}_{E^c} L_P(\widehat{h})] \le \epsilon_s + \Pr(G^c) \le (n+1)/(t+1)^{\kappa-1}.$$

Combining the branches proves Equation (11), uniformly in the target, channel and $m < n$. Equation (28) and the stated choice of t give

$$\sup_{h\in\mathcal{H}_\infty} \mathbb{E}[\text{risk}] = O_\kappa\left(\frac{t\log(t+1)}{n} + \frac{n}{t^{\kappa-1}} + \frac{1}{n}\right) = O_\kappa\left(n^{-1+2/\kappa}\log^{1-1/\kappa}(n+1)\right).$$

Increasing C_κ handles the finitely many n for which $\log N_t > n$.

An explicit finite-sample certificate. The same argument holds for every integer $t \ge 2$, with the exact count $N_t = 2\sum_{s=2}^{t}\binom{8w_s}{w_s}^s$. Choosing the prefix from this bound, or using a random constant prediction, gives for $m < n$

$$\mathfrak{R}^{\text{ins}}_{n,m}(\mathcal{H}_\infty, \mathsf{P}_\infty) \le \min\left\{\frac{1}{2}, \inf_{t\ge 2}\left[\frac{\log N_t + 1}{n} + \frac{n+1}{(t+1)^{\kappa-1}}\right]\right\}.$$

This bound has no hidden constants. It need not improve on $1/2$ at small n, is not a matching finite-sample rate, and does not make prefix search efficient.

Inserted lower bound on the full cube. Fix $n \ge 2$ and place the equal-type prior on the embedded level-n low and high subclasses. Let $B_n \subset \mathcal{C}$ be the finite level-n coordinate bank. Projecting an IID clean root from P_∞ onto B_n gives exactly the finite uniform root used by Theorem 4. Run its balanced-root-exchange channel on that projection. For every inserted row, independently sample all coordinates in $\mathcal{C} \setminus B_n$ from their fair product law, independently of the finite channel and the clean root.

This construction is a probability kernel on a standard Borel space. All $2n$ unused-coordinate row vectors are IID with the same product law in both latent worlds and are independent of the routed finite projections. The equality of the shuffled finite-projection sublaws therefore lifts to equality of the routed full-vector sublaws. Formally, attach an IID unused-coordinate mark to each occurrence

of a finite labeled multiset and then forget its ordering. This defines one common Markov kernel K_{mark} on multisets: permuting even repeated occurrences leaves its law unchanged. For each paired finite common submeasure c_{ab}, both lifted worlds therefore dominate $c_{ab}K_{\mathrm{mark}}$, of the same mass. The query label depends only on the finite projection. Integrating the paired error against $c_{ab}K_{\mathrm{mark}}$ gives the same lower bound as against c_{ab}, even for a learner that uses every unused coordinate. Every level-n target ignores $\mathcal{C} \setminus B_n$, so finite compatibility makes each completed inserted row correctly labeled. The channel contains the complete source root, inserts exactly n rows, and never observes the future query. Applying the finite Bayes bound to this prior gives

$$\mathfrak{R}^{\mathrm{ins}}_{n,n}(\mathcal{H}_\infty, \mathsf{P}_\infty) \geq \frac{1}{2}\,[1 - n\epsilon_n - n\beta_n]_+\,(1 - 2\epsilon_n).$$

Since $n\epsilon_n \to 0$ and $n\beta_n \to 0$, the lower bound tends to $1/2$. The prior is supported inside the one fixed class, so the conclusion is a minimax lower bound for $\mathcal{H}_\infty$. For any finite $m > n$, apply a common deterministic map to the observation: repeat its first labeled occurrence in a fixed Borel ordering $m - n$ times. Every repeated label is honest, and the original sample is retained. Pushing forward the common routed sublaw preserves its mass and its paired target–query disagreement. The same lower bound therefore holds for every $m \geq n$; random constant prediction has risk $1/2$.

Proposition 9 samples the active level-n projection with expected $O(nD_n)$ coordinate work. Unused coordinates can be generated independently when queried, so the same construction has a polynomial random-access coordinate implementation on $\mathcal{X}_\infty$. This statement is an oracle representation of the product measure, not a finite-string encoding of an infinite feature vector.

Unit-interval transfer. For $x \in [0, 1)$, the canonical digit map in Equation (8) is

$$d_k(x) = \lfloor 2^k x \rfloor - 2\lfloor 2^{k-1} x \rfloor;$$

set $d_k(1) = 1$. Let $B \subseteq \{0,1\}^{\mathcal{C}}$ contain the sequences that are not eventually all one together with the all-one sequence. Then B is Borel and has P_∞-measure one, and Ψ is a Borel bijection from $[0, 1]$ to B with inverse $z \mapsto \sum_{k\geq 1} z_{c_k} 2^{-k}$. Every length-$K$ digit cylinder has a Lebesgue preimage of length 2^{-K}, so $\Psi_{\#}\mathsf{U}_{[0,1]} = \mathsf{P}_\infty$ and the clean proof transfers exactly.

For an arbitrary real root $(x_1, \ldots, x_n)$, encode every row by Ψ and run the cube channel above. Preserve each clean row using $\Psi^{-1}(\Psi(x_i)) = x_i$. Map a proposed inserted row $z \in B$ to $\Psi^{-1}(z)$; if $z \notin B$, instead duplicate x_1 and give it its target-correct label. This defines a Borel exact-budget channel that contains the root and is honest for every input and every realization. Conditional on every active finite projection, each inserted cube row retains infinitely many independent fair digits, so $z \in B$ almost surely. The fallback branch consequently has zero probability in both routed worlds. Rowwise conjugation therefore preserves their common shuffled sublaw and the query disagreement probability, proving Equations (9) and (10) for $(\mathcal{G}_\infty, \mathsf{U}_{[0,1]})$. The transfer is a classical representation of a nonatomic standard probability space; it does not provide Euclidean regularity or finite-precision features.

B Proof of the budgeted honest-superposition characterization

Let $u = (a, r)$ and $v = (b, r')$ denote low and high source states. Both source spaces are standard Borel. Represent shuffled size-$(n + m)$ labeled multisets by sorting tuples under a fixed Borel order on $\mathcal{X} \times \{0, 1\}$; this makes the output space standard Borel. First, $u \sim_m v$ if and only if one size-$(n + m)$ labeled multiset is an honest extension of both roots. For necessity, a common extension contains the multiset union of r and r', whose size is $2n - \sum_x \min\{N_r(x), N_{r'}(x)\} \leq n + m$, and honesty forces target agreement at every feature in that union. For sufficiency, label the multiset union under either target and duplicate any one of its occurrences until the size is $n + m$; the union is nonempty because $n \geq 1$. Call the sorted result $T_m(u, v)$. The relation $\sim_m$ is Borel: overlap at least k is a finite union over size-k position matchings of finite intersections of coordinate equalities. The map T_m is Borel as well. Indeed, scan r' and retain an occurrence exactly when its prefix rank among equal features exceeds its multiplicity in r, then pad and sort; every step uses finitely many Borel equality and target-evaluation tests.

For the lower bound, fix a budget-m subcoupling γ and let γ^- and γ^+ be its marginals. Disintegrate $\gamma(du, dv) = \gamma^-(du)\kappa^-(dv \mid u) = \gamma^+(dv)\kappa^+(du \mid v)$ and choose Radon–Nikodym derivatives

$\rho^- = d\gamma^-/dQ^-$ and $\rho^+ = d\gamma^+/dQ^+$ in $[0,1]$. From u, route with probability $\rho^-(u)$, draw $v \sim \kappa^-(\cdot \mid u)$, and insert the multiset difference $T_m(u,v) \setminus r$. Define the high channel symmetrically; residual source mass duplicates its first root occurrence m times. Disintegrations are specified only almost everywhere. To obtain channels honest on every input, check the Borel relation $u \sim_m v$ before using a proposed pair; if it fails, duplicate the source's first labeled occurrence instead. This modifies no γ-mass and ensures exact budget and honesty even on exceptional source states. Each routed output is $T_m(u,v)$ in both worlds and every inserted label is correct. Pushing γ through this common-extension map therefore gives a common output submeasure of mass $|\gamma|$, so $\mathsf{Aff}(W^-, W^+) \geq |\gamma|$. Taking the supremum proves the lower bound in Equation (17).

For the converse, fix arbitrary honest channels and write

$$J^-(du, dT) = Q^-(du)K^-(dT \mid u), \qquad J^+(dv, dT) = Q^+(dv)K^+(dT \mid v).$$

Let $c = W^- \wedge W^+$ and restrict J^- and J^+ by the output densities dc/dW^- and dc/dW^+. The resulting measures L^- and L^+ both have output marginal c. Disintegrate them as $L^-(du, dT) = \alpha_T(du)c(dT)$ and $L^+(dv, dT) = \beta_T(dv)c(dT)$, then set

$$\gamma(du, dv) = \int \alpha_T(du)\beta_T(dv)c(dT). \tag{29}$$

Its source marginals are dominated by Q^- and Q^+, and its total mass is $|c| = \mathsf{Aff}(W^-, W^+)$.

For c-almost every T, the posteriors α_T and β_T are supported on sources for which that same labeled multiset is an honest extension. Each such T contains both source roots. Honesty under both targets therefore forces a and b to agree on every occurrence in r and r'. Because $|T| = n + m$, the contained root union also has size at most $n + m$, which is equivalent to Equation (15). Thus γ is a budget-m subcoupling, proving $\mathsf{Aff}(W^-, W^+) \leq \tau_m$.

Finally, $\sim_m$ is nested in $\sim_{m+1}$, proving monotonicity. For $m \geq n$, the overlap threshold is nonpositive and $\sim_m = \sim_n$, proving saturation. On finite spaces the feasible subcoupling polytope and channel polytopes are compact, so both suprema are attained.

Finally, under the equal type mixture, the two routed worlds have the same latent measure $\gamma(du, dv)/2$. Whenever $a(x) \neq b(x)$, they contribute that measure to opposite query-label bins. Pair and output collisions preserve the lower bound after aggregation, while fallback adds nonnegative mass. Integrating the pointwise Bayes error over outputs and queries proves Equation (19).

C Proof of the balanced-root exchange theorem

We prove Theorem 7 on finite spaces; the product-panel application is finite. For each pair with $F_{ab} = G_{ab} > 0$, define $\mu_{ab} = u_{ab}/F_{ab}$ and $\nu_{ab} = v_{ab}/F_{ab}$.

In the low world, the subprobability of choosing pair (a, b), observing root r, and inserting root r' is

$$W^-_{ab}(r, r') = \frac{1}{2}u_{ab}(r)\nu_{ab}(r') = \frac{u_{ab}(r)v_{ab}(r')}{2F_{ab}}. \tag{30}$$

The factor $1/2$ is the prior probability of the low type. In the high world, consistently writing the low root first, the corresponding subprobability is

$$W^+_{ab}(r, r') = \frac{1}{2}v_{ab}(r')\mu_{ab}(r) = \frac{v_{ab}(r')u_{ab}(r)}{2F_{ab}} = W^-_{ab}(r, r'). \tag{31}$$

Every r in the support of u_{ab} satisfies $a(x_i) = b(x_i)$ for all its coordinates, and the same is true of every r' in the support of v_{ab}. Consequently the labeled multiset

$$T_{ab}(r, r') = \{(x, a(x)) : x \in r\} \uplus \{(x, a(x)) : x \in r'\}$$

equals the multiset obtained by labeling both roots with b. It contains the source clean root, adds exactly n examples, and is honest in both worlds. The conditional laws depend on (a, b, P, r) but not on the independent query.

It remains to prove the risk statement without exposing (a, b). Let z range over observed labeled multisets and let $w_{ab}(z)$ be the common mass obtained by pushing $2W^-_{ab} = 2W^+_{ab}$ through

$(r, r') \mapsto T_{ab}(r, r')$. At query x with $a(x) \neq b(x)$, the two worlds contribute $w_{ab}(z)/2$ to each of the two query-label bins. All pair collisions and fallback outputs add nonnegative mass. Therefore, pointwise in (z, x),

$$\min\{\Pr[Y = 0, z, x], \Pr[Y = 1, z, x]\} \geq \frac{1}{2} \sum_{a,b:a(x)\neq b(x)} w_{ab}(z)P(x).$$

Summing over z, x and using $\sum_z w_{ab}(z) = F_{ab}$ gives Equation (22). The left side is the Bayes error under the constructed prior and channels, so it lower-bounds the minimax value by Yao's principle.

If randomized adversaries are disallowed, sample in advance every partner and conditional-root choice for every finite target–root input. This produces a distribution over deterministic honest exact-budget channel families that is independent of learner randomness. For each learner, at least one family in the support attains the average loss above; the adversarial supremum follows the learner, as required by Equation (2).

D Proof of product-orbit balance

Here the partner is chosen uniformly from all compatible targets, with probability zero if none exists:

$$q_-(B \mid A, r) = \frac{\mathbf{1}\{A, B \text{ compatible at } r\}}{d_-(A, r)}, \qquad q_+(A \mid B, r) = \frac{\mathbf{1}\{A, B \text{ compatible at } r\}}{d_+(B, r)}.$$

These kernels require enumeration and differ from the barred kernels used by the efficient sampler.

Lemma 10 (Product-orbit balance)**.** *For every s, L, w, n and every target pair A, B, the uniform partner kernels satisfy $F_{AB} = G_{AB}$.*

Fix target indices A, B. For panel j, the sets $A_j \setminus B_j$ and $B_j \setminus A_j$ have equal cardinality. Choose any bijection between them, swap each matched coordinate, and fix all other panel coordinates. The product over panels is an involutive coordinate permutation π_{AB} that preserves U_D and satisfies

$$f_A^-(\pi_{AB}x) = f_B^-(x), \qquad f_B^-(\pi_{AB}x) = f_A^-(x),$$
$$f_A^+(\pi_{AB}x) = f_B^+(x), \qquad f_B^+(\pi_{AB}x) = f_A^+(x).$$

Apply the permutation rowwise to a root r. High targets compatible with low A at r are carried bijectively to high targets compatible with low B at $\pi_{AB}r$; in particular,

$$d_-(A, r) = d_-(B, \pi_{AB}r).$$

Moreover, A is compatible with B at r exactly when B is compatible with A at $\pi_{AB}r$. Changing variables in the uniform root sum therefore gives

$$F_{AB} = F_{BA}. \tag{32}$$

Let $c(r)$ complement every bit of every root row. The duality $f_C^+(x) = 1 - f_C^-(\mathbf{1} - x)$ implies that low A and high B agree on r exactly when low B and high A agree on $c(r)$. Complement also bijects the low partners of high B at r with the high partners of low B at $c(r)$, so $d_+(B, r) = d_-(B, c(r))$. Changing variables gives

$$G_{AB} = F_{BA}. \tag{33}$$

Combining Equations (32) and (33) proves $F_{AB} = G_{AB}$.

E Proof of exact efficient panel exchange

Fix $n = s$ and the ordered root $r = (x_1, \ldots, x_s)$. If the clean input is represented as a multiset, first apply an independent uniform ordering to its occurrences. For the low kernel in Equation (25), summing over all high targets B gives one exactly when every low label is zero and each assigned panel i of row x_i contains at least w zeros; otherwise it gives zero. Indeed, on this event each factor is the uniform distribution over the w-subsets of $Z_i(x_i)$. Whenever the kernel is positive, the ith high clause is zero on row x_i, so $f_B^+(x_i) = 0 = f_A^-(x_i)$ for every i. The reverse statement follows identically from one-coordinates. Thus both kernels are subprobability kernels supported

on compatible partners, and the forward routed event has probability at least $1 - s\epsilon_{s,w} - s\beta_{L,w}$ by the union bound used in Lemma 8.

We next prove exact balance. The coordinate involution π_{AB} from Appendix D preserves the number of zeros in every panel and satisfies

$$B_i \subseteq Z_i(x_i) \quad \Longleftrightarrow \quad A_i \subseteq Z_i(\pi_{AB} x_i).$$

It also maps $f_A^-(x_i) = 0$ to $f_B^-(\pi_{AB} x_i) = 0$. Consequently,

$$\bar{q}_-(B \mid A, r) = \bar{q}_-(A \mid B, \pi_{AB} r) \tag{34}$$

pointwise. Uniform-cube invariance and uniform target priors give $\bar{F}_{AB} = \bar{F}_{BA}$. Bit complement gives $\bar{q}_+(A \mid B, r) = \bar{q}_-(A \mid B, \mathbf{1} - r)$ and hence $\bar{G}_{AB} = \bar{F}_{BA} = \bar{F}_{AB}$.

It remains to sample the conditional roots used by Theorem 7. For fixed (A, B), the unnormalized density $p(r)\bar{q}_-(B \mid A, r)$ factors over rows. Within low row i, it also factors over panels. Every off-diagonal panel $j \neq i$ is a fair bit panel conditioned on the zero set intersecting A_j; rejection from fair bits accepts with probability $1 - 2^{-w} \geq 1/2$. In the diagonal panel, the zero set $S \subseteq [L]$ has probability proportional to

$$\frac{\mathbf{1}\{B_i \subseteq S,\ S \cap A_i \neq \varnothing\}}{\binom{|S|}{w}}. \tag{35}$$

For arbitrary fixed $C, R \subseteq [L]$ of size w, sample the template in Equation (35) as follows: draw $K \sim \mathrm{Binomial}(L, 1/2)$ conditioned on $K \geq w$; draw a uniform K-subset S containing C; and reject if $S \cap R = \varnothing$. For every admissible S of size K the proposal mass is, up to a constant,

$$\frac{\binom{L}{K}}{\binom{L-w}{K-w}} = \frac{\binom{L}{w}}{\binom{K}{w}},$$

which proves the exact target law after rejection. If $C \cap R \neq \varnothing$, the final rejection never occurs. If they are disjoint and $L = 8w$, conditional probability at least $1/2$ lies on $K \geq 4w$; there,

$$\Pr[S \cap R = \varnothing \mid K] = \frac{\binom{6w}{K-w}}{\binom{7w}{K-w}} \leq (6/7)^{K-w} \leq (6/7)^{3w}.$$

The acceptance probability is therefore at least $\frac{1}{2}(1 - (6/7)^3) = 127/686$. Sampling the truncated binomial itself takes fewer than two proposals in expectation.

Each row contains $sL = D$ bits and is sampled in expected $O(D)$ coordinate work; all $s = n$ rows cost expected $O(nD)$. The reverse conditional root sampler follows by complement. Partner selection reads only the assigned panels and costs $O(sL) = O(D)$. This proves Proposition 9.

F Proofs of abundance, disagreement, and clean risk

F.1 Partner abundance

Consider a uniformly random low target A and IID root $r = (X_1, \ldots, X_n)$. The probability that some low root label is one is at most $n\epsilon_{s,w}$. Condition only on the complementary event that all labels are zero. Assign row i to panel i; if the clean sample is presented as an unordered multiset, the channel first applies an independent random ordering.

For every row i, let E_i be the event that assigned panel i contains fewer than w zero coordinates. Unconditionally $\Pr(E_i) = \beta_{L,w}$, so the union of the label-failure event and the panel-failure events has probability at most $n\epsilon_{s,w} + n\beta_{L,w}$. Outside this union, choose B_i to be any w zero coordinates of row i in panel i. Choose the unused blocks B_j, $j > n$, arbitrarily. For each row i, the ith disjunction in f_B^+ is zero, hence $f_B^+(X_i) = 0 = f_A^-(X_i)$. Thus B is a compatible high target and $d_-(A, r) > 0$.

Bit complement maps a high root with all labels one to the same construction for a low partner. This proves the lower bound for both p^-_{good} and p^+_{good}. Finally,

$$\sum_{A,B} F_{AB} = p^-_{\mathrm{good}}, \qquad \sum_{A,B} G_{AB} = p^+_{\mathrm{good}}.$$

Here F, G refer to the uniform partner kernels of Appendix D. Their balance gives the same equality as bit complement, completing this alternative proof of Lemma 8.

F.2 Pairwise query disagreement

For any A, B,

$$\begin{aligned}\Pr[f_A^-(X) = 0, f_B^+(X) = 1] &\geq 1 - \Pr[f_A^-(X) = 1] - \Pr[f_B^+(X) = 0] \\ &= 1 - 2\epsilon_{s,w}.\end{aligned}$$

This event is contained in the target disagreement set, proving Equation (27).

F.3 Clean risk and asymptotics

Predict the first clean label Y_1 at every query. For a low target, Y_1 and the independent test label Y are IID Bernoulli$(\epsilon_{s,w})$; for a high target their complements are IID Bernoulli$(\epsilon_{s,w})$. In either case,

$$\Pr[Y_1 \neq Y] = 2\epsilon_{s,w}(1 - \epsilon_{s,w}).$$

When $w = \lceil \kappa \log_2 s \rceil$, $\epsilon_{s,w} \leq s2^{-w} \leq s^{1-\kappa}$.

For $Z \sim \mathrm{Binomial}(8w, 1/2)$, the mean is $4w$. The multiplicative Chernoff inequality with relative deviation $3/4$ gives

$$\Pr[Z < w] \leq \exp\left(-\frac{(3/4)^2(4w)}{2}\right) = e^{-9w/8}.$$

Therefore $s\epsilon_{s,w} \leq s^{2-\kappa}$ and $s\beta_{8w,w} \leq s\exp(-9\lceil \kappa \log_2 s \rceil/8) = o(1)$. Theorem 4 follows.

G Proof of the panel lower law

Proposition 11 (Panel lower law)**.** *Consider a sequence using a distinct panel for each of n root obligations, with $s \geq n$, $L \geq w$ and $\epsilon = 1 - (1 - 2^{-w})^s$. If $n\epsilon \to 0$, then*

$$w \geq \log_2(ns) + \omega(1), \qquad D = sL = \Omega(n \log n).$$

Set $q = 2^{-w}$ and $z = sq$. Because $(1 - q)^s \leq e^{-sq}$,

$$\epsilon = 1 - (1 - q)^s \geq 1 - e^{-z}.$$

For $0 \leq z \leq 1$, $1 - e^{-z} \geq z/2$; for $z > 1$, it is at least $1 - e^{-1}$. The condition $n\epsilon \to 0$ first forces $z \to 0$ and then forces $nz \to 0$. Hence

$$ns2^{-w} \to 0,$$

which is equivalent to $w - \log_2(ns) \to \infty$. Since $s \geq n$ and $L \geq w$,

$$D = sL \geq sw \geq n(2\log_2 n + \omega(1)) = \Omega(n \log n).$$

H Fixed-target consistency and authenticated provenance

Lemma 12 (Two fixed targets cannot supply near-unit camouflage)**.** *Fix total binary targets a, b, a common marginal P, $n \geq 1$, and finite $m \geq 0$. Put $d = P(a(X) \neq b(X))$. For their full observation laws,*

$$\sup_{K_a, K_b} \mathsf{Aff}(W_a, W_b) = (1 - d)^n.$$

Proof. An observation in the common sublaw must be honest under both targets, hence every feature it contains lies in $E = \{x : a(x) = b(x)\}$. Since the clean root is retained, its probability in either world is at most $P(E)^n$. This bounds the common mass. For attainment, use in both worlds the channel that appends m copies of the canonically first labeled clean occurrence. The two clean-root sublaws restricted to E^n coincide and have mass $P(E)^n$; applying the same padding map preserves that common mass. This argument includes duplicate occurrences and does not condition the adversary on the test query. □

This elementary identity explains why balancing only label histograms is not our arbitrary-learner proof. The feature configuration may distinguish fixed targets even with identical counts. Balanced-root exchange instead routes across target mixtures and aggregates collisions without revealing the pair index.

Proposition 13 (Balanced-label reduction)**.** *For any binary problem $(\mathcal{H}, P)$ on a standard Borel space, define $\widetilde{P} = P \otimes \mathrm{Ber}(1/2)$ and $\widetilde{\mathcal{H}} = \{\widetilde{h} : \widetilde{h}(x, b) = h(x) \oplus b,\ h \in \mathcal{H}\}$. Every lifted target satisfies $\widetilde{P}(\widetilde{h} = 1) = 1/2$, and for randomized learners, $n \geq 1$ and every finite $m \geq 0$,*

$$\mathfrak{R}^{\mathrm{ins}}_{n,m}(\widetilde{\mathcal{H}}, \widetilde{P}) = \mathfrak{R}^{\mathrm{ins}}_{n,m}(\mathcal{H}, P).$$

The same equality holds for clean minimax risk.

Proof. For the upper bound, decode each labeled occurrence $((x, b), y)$ to $(x, y \oplus b)$, apply a base learner, and XOR its prediction with the query bit. A lifted honest channel induces a base honest channel by drawing the clean bits independently and integrating out all bits after its action. The decoded root is IID from P, retained with multiplicity, and every decoded insertion is correct. The query bit is independent, so loss equals base loss pointwise. Taking the channel and target suprema and learner infimum gives the upper bound.

Conversely, fix a lifted learner. Construct a randomized base learner by giving each observed occurrence an independent fair bit, replacing its label by its XOR with that bit, and drawing an independent fair query bit before invoking the lifted learner and decoding its answer. For any base channel, a legal lifted channel ignores the clean bits, applies that base channel to the decoded root, and independently marks each inserted occurrence. Conditional on the base output, all occurrence marks are IID fair, including at repeated features: forgetting the clean/inserted assignment gives exactly the base learner's marked multiset law. The losses agree. Hence the worst lifted risk of this learner is at least the worst base risk of the constructed learner, which is at least base minimax risk. Taking the infimum proves the reverse bound. With $m = 0$, the same constructions prove the clean equality. □

Applying the reduction to $(\mathcal{H}_\infty, \mathsf{P}_\infty)$ transfers the sharp threshold to one fixed known product marginal with balanced labels for every target. The extra fair bit is observed as part of the feature, not a hidden label flip. This is an equivalent statistical problem, not a claim that marginal balance alone makes data hard. The lifted targets need not be monotone formulas, and the reduction supplies neither geometric regularity nor an empirical selector-specific effect.

Proposition 14 (Targetwise consistency of countable classes)**.** *Let $\mathcal{H} = \{h_1, h_2, \ldots\}$ be a countable class of measurable binary functions. The learner returning the first hypothesis consistent with T satisfies, for every fixed target h_j, distribution P, $n \geq 1$ and finite m,*

$$\sup_{K_{h_j}} \mathbb{E}[L_P(\widehat{h})] \leq \min\{1, (\log j + 1)/n\}.$$

Proof. Honesty makes h_j consistent, so the returned index is at most j. Every returned hypothesis is consistent with the retained IID root. For $q > 0$, a union bound over the first j hypotheses gives $\Pr(L_P(\widehat{h}) > q) \leq \min\{1, je^{-nq}\}$, uniformly over channels. Integrating proves the bound. Countable search and measurable evaluations make the rule measurable; no computational efficiency is asserted. □

The index j is fixed in this statement; the bound is not uniform over the entire class. Thus Theorem 2 does not contradict targetwise consistency. More strongly, for any fixed probability prior $(p_j)_{j \geq 1}$ on this countable class and any finite budget sequence m_n, the prior-average worst-channel risk is at most $\sum_j p_j \min\{1, (\log j + 1)/n\} \to 0$ by dominated convergence. Thus our lower bound does not describe persistent failure under a fixed prior: its hard target mixture changes with n.

Knowing the feature marginal does not prevent the lower bound: the learner is already given P exactly. Estimating P cannot supply more information than that. This is not a robustness guarantee under a changed or misspecified marginal.

Corollary 15 (Authenticated IID reserve)**.** *Suppose the learner additionally receives an authenticated subset of k clean observations, selected by fixed indices or independently of all feature and label values. Its minimax risk is at most $\mathfrak{R}_k^{\mathrm{clean}}(\mathcal{H}, P)$, regardless of any finite insertion budget. In particular, for our fixed class it tends uniformly to zero whenever $k \to \infty$.*

Proof. Ignore all unauthenticated observations and apply a clean learner to the reserve. The reserve is IID from the target's labeled marginal and independent of the fresh query. Taking arbitrarily near-optimal clean learners gives the minimax inequality even if its infimum is not attained. □

Authentication without the stated sampling property is insufficient for this reduction: selecting the reserve based on observed values can bias its law.

Proof of Proposition 3. Let I be the fixed or value-independent index of the marked clean occurrence. Search $\mathcal{F}_t$ for a hypothesis consistent with T; on failure predict $h(X_I)$. The search-success contribution remains at most $(\log N_t + 1)/n$: every returned prefix hypothesis is consistent with the IID root, regardless of the channel's use of the marker. If the target belongs to the prefix, search cannot fail. Otherwise its minority mass is at most $(t+1)^{1-\kappa}$. Since X_I and the fresh query are independent with marginal P, the unconditional error of the constant $h(X_I)$ is $2p(1-p) \le 2(t+1)^{1-\kappa}$, where $p = P(h(X) = 1)$. Its contribution on search failure is no larger; independence from that event is not needed. Thus

$$\sup_{h, K_h} \mathbb{E}[\mathrm{risk}] \le \frac{\log N_t + 1}{n} + \frac{2}{(t+1)^{\kappa-1}}.$$

Take t of order $(n/\log(n+1))^{1/\kappa}$, using Equation (28). The proof also applies to one additional independent authenticated IID example. Neither version identifies all clean occurrences; it resolves the nearly constant target's fallback label. Value-dependent selection of the marker need not have the required marginal law.

Proposition 16 (A finite-cover bound for general budgets)**.** *Let $\mathcal{C}_\epsilon$ be a nonempty finite set of measurable binary predictors such that every $h \in \mathcal{H}$ is within disagreement probability ϵ of one member under P. For $n \ge 2$ and every $m \ge 0$,*

$$\mathfrak{R}_{n,m}^{\mathrm{ins}}(\mathcal{H}, P) \le \epsilon + \frac{m}{n} + \sqrt{\frac{2\log(2n|\mathcal{C}_\epsilon|)}{n}} + \frac{1}{n}.$$

If such covers exist for every $\epsilon > 0$, every $m = o(n)$ budget preserves uniformly vanishing risk.

Proof. Choose the first cover member $\widehat{g}$ minimizing disagreements with T, counting multiplicities. A cover member $g^\star$ within ϵ of the target has at most m more disagreements on T than on S. Thus $\widehat{g}$ has at most m more clean disagreements than $g^\star$. By Hoeffding's inequality and a union bound, with probability at least $1 - 1/n$, all cover members' clean empirical errors differ from their population errors by at most $\sqrt{\log(2n|\mathcal{C}_\epsilon|)/(2n)}$. This event concerns only the IID root, so it holds against every adaptive channel. Applying the bound twice and adding $1/n$ for its complement proves the inequality. For the limit, fix ϵ, let $n \to \infty$, then let $\epsilon \downarrow 0$. □

In particular, at $m/n \to c < 1/2$ such a problem cannot have minimax risk tending to $1/2$: its limit superior is at most c. This does not identify a sharp threshold for arbitrary classes.

I EXACT-CHECK SPECIFICATION

Table 1: Exact uniform-partner checks. Output laws are aggregated without revealing target-pair indices. The first cell tests normalization, not a nontrivial separation.

(s, L, w, n)	targets/type	outputs	paired mass	max. gap
$(1, 2, 1, 1)$	2	17	1	0
$(2, 2, 1, 1)$	4	135	$3/4$	0
$(2, 2, 1, 2)$	4	2053	$61/128$	0

These complete cells check the *uniform* partner kernels. The gap is between fully aggregated routed output laws, with target-pair indices hidden. The first cell is a normalization check, not a nontrivial separation. The efficient barred kernels are checked separately below.

The first checker enumerates every target and ordered root for each registered cell. For every root it computes all compatible opposite-type targets, uses exact fractions for the uniform partner kernel, and asserts

$$\max_{A,B} |F_{AB} - G_{AB}| = 0.$$

It also constructs one-panel-per-row partners and checks Equations (24) and (27) exactly.

The second checker constructs u_{AB} and v_{AB}, enumerates every exchanged root pair, converts the pair to a sorted labeled multiset, and sums exact fractions under that output key. The key intentionally omits (A, B). It then asserts equality of the complete routed low and high output dictionaries. The largest registered cell contains 2,053 distinct colliding output keys and has zero gap.

The third checker directly computes the maximum budget-m subcoupling mass by bipartite max flow, independently computes the maximum common mass mediated by enumerated legal outputs, and compares both with the affinity of constructed full channels including fallback. It checks every $m \in \{0, 1, 2, 3\}$ at $n = 2$, asserts that the transport edges equal the common-output edges, and verifies monotonicity and saturation. The registered cells are deliberately asymmetric in priors and feature masses and contain output collisions.

The fourth checker independently enumerates the constructive kernels underlying Proposition 9 with exact fractions. It verifies every row sum, all pair-flow equalities, full-root factorization into conditional row laws, and Equation (35) against direct panel enumeration. Its three complete cells are $(s, L, w) = (1, 3, 1), (1, 4, 2), (2, 2, 1)$; together they include overlapping and disjoint target blocks and a two-row exchange. These small cells test algebraic identities for the kernels, not the runtime bound, whose proof assumes $L = 8w$.

The fifth checker enumerates every binary target, clean root, and exact insertion budget $m \in \{0, 1, 2\}$ for three nonuniform finite distributions. It verifies the pointwise memorizer bound, averages with exact root probabilities, and checks the identity in Equation (3).

These checks are executable falsifiers for normalization, orientation, honesty, and collision handling. The NumPy-based `check_runner_laws.py` also imports the production sampler: 45 diagonal conditional laws are compared with exact rational proposal weights, 192 off-diagonal proposals check acceptance, and 24 routing traces check the low/high helper arguments. In particular, complement transposes target indices: $\nu_{AB}(x) = \mu_{BA}(\mathbf{1} - x)$, not generally $\mu_{AB}(\mathbf{1} - x)$. These are bounded author-side checks, not an independent audit of the full theorem. They are not used in place of the all-parameter proofs. An additional exhaustive three-point check tests the prefix-first learner, its retained-root consistency, the strict-majority argument for $m < n$, and common multiset padding. It does not certify the infinite-class theorem. It also checks 512 fixed-pair padding affinities, fair-bit population balance, loss preservation, and mark exchangeability with duplicate occurrences. These finite identities do not replace the standard-Borel reduction proof. The stochastic regression `test_high_source_off_diagonal_is_not_forced_all_zero` separately guards the corrected high-source sampler.

J FINITE-SCALE EMPIRICAL VALIDATION

The experiments illustrate the finite product-panel construction and its effect on fixed learners. Every run uses $m = n$; these runs do not test the threshold across budgets. We retain the controls because they limit what can be inferred from learner errors: label-count matching largely reproduces the harm. The theorem, not these fitted models, establishes the arbitrary-learner result. All completed scaling, tree, and neural records are reported; an incomplete natural-family study is excluded in full.

Design and pairing. Set $n = m = s$, $w = \lceil 3 \log_2 s \rceil$, $L = 8w$, and draw 16,384 fresh uniform-cube test points per scale and seed. The scaling grid uses $s \in \{2, 4, 8, 16, 32\}$ and seeds $0, \ldots, 29$ (150 records). The decision-tree grid uses $s \in \{16, 32\}$, the same 30 seeds, and six paired conditions (360 records). Targets alternate low and high types by seed. The clean condition uses only the original root; IID-hidden adds n independent target-correct examples. Hidden camouflage uses

balanced-root exchange. Visible camouflage contains the identical examples and labels, but filters inserted occurrences using their origin bits before fitting. This contrast measures the benefit of access to provenance, not a learned provenance detector.

The label-count-matched control matches the camouflage insertion-label histogram using target-correct examples; the hard-example control selects from an IID candidate pool of $\max(4096, 64m)$ examples using clean-tree error/confidence and breaks ties by pool order. There is one protocol deviation: the implemented label-matched control draws exactly from the uniform cube conditional on each requested label, instead of selecting from the specified finite IID pool and declaring exhaustion invalid. This change preceded the completed runs, but was not accompanied by a protocol amendment. We therefore treat its contrasts as sensitivity analyses, not confirmatory tests of the original finite-pool control. Scaling and provenance comparisons are unaffected. The transparent learner is an untuned `DecisionTreeClassifier(random_state=seed)`. All intervals below are percentile 95% paired-seed bootstrap intervals with 10,000 resamples. They quantify seed variability on this synthetic family, not uncertainty over deployment domains; intervals are not multiplicity-adjusted.

Finite scaling. The analytical first-label clean-risk log–log slope on $s \in \{8, 16, 32\}$ is -1.9848. The scaling records also compare routing and target disagreement with the finite-level certificate; they do not estimate the fixed-class minimax risk. Specifically, seed r records $\widehat{C}_r = \frac{1}{2} I_r \widehat{d}_r$, where I_r indicates successful routing and $\widehat{d}_r$ is the paired targets' disagreement fraction on the 16,384 fresh queries. The empirical number is the mean over 30 seeds. The analytical number is the conservative product $\frac{1}{2}[1 - s\epsilon - s\beta]_+[1 - 2\epsilon]_+$, not the expectation of an identical estimator. Their proximity is descriptive; sampling error can place the empirical mean below this analytical bound.

Provenance and selector controls. Table 2 reports all decision-tree contrasts. Filtering to clean observations reduces risk at both scales. Conditional label-count matching largely reproduces the observed harm, so these results do not isolate pair adaptivity as an additional cause beyond adversarial label composition. The original finite-pool comparison remains unevaluated because of the deviation above; the supplement preserves the prespecified decision thresholds.

Table 2: Decision-tree risk differences, 30 paired seeds per scale. Label-count matching is a protocol-deviating sensitivity analysis; other rows are confirmatory. Each entry is a mean difference with its paired 95% interval. Positive values mean greater risk for hidden camouflage.

Comparator	$s = 16$	$s = 32$
Visible camouflage	$0.4485\ [0.3844, 0.5024]$	$0.5340\ [0.5116, 0.5558]$
IID hidden	$0.4319\ [0.3687, 0.4861]$	$0.5299\ [0.5032, 0.5547]$
Label-count matched	$0.0375\ [0.0005, 0.0731]$	$0.0141\ [-0.0201, 0.0486]$
Hard example	$0.0353\ [-0.0063, 0.0824]$	$0.4051\ [0.3803, 0.4291]$

Fixed neural learners. The MLP has widths $D \to 64 \to 32 \to 1$ with ReLU hidden activations. The panel GNN uses width-16 Boolean embeddings, two feature-to-panel mean/max message-passing layers, panel-to-global mean/max aggregation, and a linear readout. Its graph uses only the public panel partition, not target blocks, target kind, labels, or provenance. Both use full-batch Adam, learning rate 10^{-3}, binary cross-entropy with logits, and float32 parameters. Clean-only calibration at scales 8 and 32, seeds 0–2, compared 100, 300 and 600 steps; 100 steps met the locked loss-within-0.01 rule for both models. No insertion-conditioned tuning was used. Evaluation fixes 100 steps, scales 8, 16 and 32, six conditions, and seeds 0–19: all 720 fits completed, with finite recorded losses and matching hidden/visible dataset hashes. Finite loss is not a guarantee of global optimization.

All six hidden-minus-visible intervals exclude zero; all six hidden-minus-label-matched intervals contain zero. Since visible training discards the insertions, the former is also a hidden-minus-clean contrast, not a separate learned provenance defense. Absence of significance is not equivalence: the latter does not establish an extra effect of the exchange selector. The GNN's mean final training loss is about 0.61–0.69 on hidden and label-matched data, versus 0.01–0.08 on clean data. Clean-only calibration therefore did not establish convergence on insertion conditions. These fixed-budget fits

Table 3: Secondary neural contrasts, 20 paired seeds per model and scale. Insertion harm is hidden minus clean risk (visible filtering gives the same result); selector is hidden camouflage minus label-count matching. Intervals are paired 95% bootstrap intervals.

Model	s	Insertion harm	Selector
MLP	8	$0.3763\ [0.2841, 0.4645]$	$0.0011\ [-0.0378, 0.0447]$
MLP	16	$0.4819\ [0.4435, 0.5243]$	$0.0102\ [-0.0223, 0.0414]$
MLP	32	$0.4976\ [0.4327, 0.5479]$	$-0.0015\ [-0.0294, 0.0252]$
Panel GNN	8	$0.3879\ [0.2715, 0.5022]$	$-0.0056\ [-0.0670, 0.0606]$
Panel GNN	16	$0.5055\ [0.4459, 0.5671]$	$-0.0268\ [-0.1447, 0.0764]$
Panel GNN	32	$0.4372\ [0.3248, 0.5545]$	$-0.0565\ [-0.1989, 0.0559]$

cannot support an architecture ranking or an optimized-robustness claim; no additional tuning or seed exclusion is used here.

Table 4: Absolute mean prediction risks from all completed learner grids. This post-hoc descriptive summary uses the unchanged records, with 30 seeds for trees and 20 for each neural model. Clean and visible risks coincide in these records. Label-matched results retain their sensitivity-only status. Paired uncertainty for the principal contrasts appears in the preceding tables; these means are not additional significance tests.

Model	s	Clean/visible	Hidden	IID	Label-matched	Hard
Tree	16	.0284	.4769	.0450	.4394	.4416
Tree	32	.0010	.5350	.0051	.5209	.1299
MLP	8	.0150	.3913	.0150	.3902	.4018
MLP	16	.0039	.4858	.0039	.4756	.3202
MLP	32	.0011	.4987	.0011	.5002	.0011
Panel GNN	8	.0150	.4029	.0150	.4086	.3985
Panel GNN	16	.0039	.5094	.0039	.5362	.3175
Panel GNN	32	.0011	.4383	.0011	.4948	.0011

An independent fair random constant classifier has expected risk exactly $1/2$ for any target and dataset; this is an analytical reference, not a trained baseline. The theorem's arbitrary-learner lower bound should not be inferred from any one fitted learner approaching $1/2$.

Post-hoc majority and label-supply diagnostics. We reconstructed all completed datasets using their original seed schedules and unchanged generator source. Every one of the 720 recorded neural training-dataset hashes and all 360 refitted tree risks matched exactly. Query sets were reconstructed deterministically; original query hashes were not stored, so direct archival query identity cannot be checked. No neural model was refitted. For N fitted labels, k positives and test-positive fraction $\widehat{q}$, majority risk is $\widehat{q}$ when $2k < N$, $1 - \widehat{q}$ when $2k > N$, and exactly $1/2$ on a fair randomized tie. Visible conditions use only authenticated clean occurrences, as do the fitted learners.

Table 5: Post-hoc majority risks and hard-pool minority supply, all completed seeds. C/V/I denotes clean, visible and IID conditions, whose majority risks coincide in these grids; H/M denotes hidden exchange and label matching, equal on every paired dataset. Pool columns give mean minority counts separately for low/high targets in the 4,096-point hard-example pool. These are descriptive diagnostics, not new confirmatory tests.

Dataset grid	s	C/V/I	H/M	Hard	Pool low	Pool high
Tree	16	.0038	.4504	.2353	16.07	17.13
Tree	32	.0010	.5000	.0010	3.60	4.00
Neural	8	.0150	.4031	.4031	62.60	63.10
Neural	16	.0039	.5000	.2767	15.80	16.00
Neural	32	.0011	.4751	.0011	4.30	4.00

Routed exchange has exactly n labels of each type, so its fair-tie majority risk is $1/2$; fall-back explains smaller means. Low/high hidden-majority means are $.4339/.4669$ and $.5000/.5000$ for the two tree scales, and $.4031/.4031$, $.5000/.5000$, $.5000/.4501$ for the three neural scales.

The corresponding routed counts are $13/15, 14/15$; $15/15, 15/15$; $8/10, 8/10$; $10/10, 10/10$; and $10/10, 9/10$. Both target types and every seed remain included.

At $s = 32$, the hard pool supplies about four minority labels, whereas a routed exchange inserts 32. Thus the hard-control contrast confounds feature selection with available label supply; it does not isolate pair adaptivity. The conditional label-matched sensitivity control avoids finite-pool exhaustion and largely reproduces the fitted-learner harm. These diagnostics explain much of the observed effect through label composition without equating the full observation laws. Per-seed label counts, routing, pool counts, majority risks and reconstructed tree-positive prediction fractions are provided by `review_diagnostics.py`; neural prediction fractions were not retained and are not inferred from error rates.

Reproduction and exclusions. The tree intervals use the seeds in the committed runner; neural intervals are recomputed from the unchanged records using the committed figure generator's seeds $300 + s + \text{len}(\texttt{model})$ for provenance and $400 + s + \text{len}(\texttt{model})$ for label matching, with model identifiers `mlp` and `panel_gnn`. Small differences from the earlier prose report reflect bootstrap resampling, not new fits or changed outcomes. The supplemental experiment directory contains the locked protocol and neural addendum, raw `scaling.json`, `transparent.json`, and `neural.json`, phase-specific analyses, and `test_validation.py`. The runner's `--phase scaling`, `--phase transparent`, and `--phase neural` reproduce these grids using NumPy/SciPy, scikit-learn and PyTorch as appropriate; neural calibration is recorded separately. CPU scaling/transparent execution took under one minute after a shared sampler defect was corrected; the complete neural grid took 526 seconds on local CUDA. Earlier failed stochastic attempts produced no result file and were invalidated before the corrected run; no completed seed was removed. The exact natural-family grid has a separate completeness gate and is excluded in full, not selected cell by cell. No conclusion here depends on its eventual outcome.