THE SAMPLE AND RANDOMNESS COMPLEXITY OF POINTWISE REPLICABLE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Pointwise replicability requires independent training samples and shared randomness to produce agreeing predictions at each fixed test point with high probability. We close the additive sample-complexity gaps of Hopkins, Impagliazzo, and Ye (2026): binary VC classes admit agnostic learning with $\widetilde{O}(d/(\rho^2\alpha^2))$ samples on countable domains and realizable learning with $\widetilde{O}(d/(\rho^2\alpha))$ samples. At fixed small failure probability, these match known lower bounds up to logarithms, with the realizable comparison for $d \geq 2$. On arbitrary measurable domains, we improve the general agnostic finite-randomness sample bound from $\widetilde{O}(d/(\rho^4\alpha^7))$ to $\widetilde{O}(d/(\rho^2\alpha^4))$ using $\lceil \log_2(1/\rho) \rceil$ additional shared fair bits. With deterministic base learners these are total bits, at most one above the exact minimum $\lfloor \log_2(1/\rho) \rfloor$ when improper outputs are allowed. For three point functions, properness costs exactly one extra bit when $\rho \leq 1/8$ and $\beta < \rho/2$. One or two shared bits permit a $\Theta(\beta^2)$ improvement below disagreement $1/K$, where K is the seed count; for $K \geq 8$, the high-confidence infimum is exactly $1/K$. We determine matching proper sample complexity just above this boundary. We also prove a dimension-dependent sample lower bound near the bit boundary and matching joint tradeoffs, including confidence, for public products of countable VC-one classes, countable tree intervals, and real intervals, and for unions of intervals without a known partition. The agnostic screen estimates normalized label signals; interval scores combine mass and component costs to make confidence additive. The general finite-bit accuracy gap remains open; a statistical lower bound identifies a limitation of the weighted-CDF compiler.

1 INTRODUCTION

Retraining on independent samples can change the predictions of an accurate model. Pointwise replicability requires agreement at each test point fixed before the data and randomness are drawn. It is a per-query guarantee, rather than simultaneous agreement of the entire trained functions.

Hopkins et al. (2026) introduce this model and establish the countable-domain agnostic upper bound

$$\widetilde{O}\left(\frac{d}{\rho^2\alpha^2} + \frac{1}{\rho^2\alpha^4}\right),$$

against a lower bound of $\widetilde{\Omega}(d/(\rho^2\alpha^2))$ at fixed small failure probability. We resolve their question by removing the additive term, together with its counterpart in the realizable bound. Ignoring logarithms, the agnostic improvement factor is $1 + 1/(d\alpha^2)$; the gain is polynomial in accuracy for fixed d. The bounds coincide to leading order when $d \gtrsim \alpha^{-2}$.

Our second main result concerns finite randomness on arbitrary measurable domains. For general agnostic distributions, including model misspecification, we improve the bound $\widetilde{O}(d/(\rho^4\alpha^7))$ to $\widetilde{O}(d/(\rho^2\alpha^4))$ by risk-weighted selection of nested ensemble thresholds. The transformation uses $\lceil \log_2(1/\rho) \rceil$ additional shared bits. With a deterministic base learner and any finite sample size allowed, the exact minimum for improper learners is $\lfloor \log_2(1/\rho) \rfloor$. For three point functions, the proper minimum is exactly one bit larger when $\rho \leq 1/8$ and $\beta < \rho/2$. With one or two bits, the proper disagreement infimum improves below $1/K$ by $\Theta(\beta^2)$; with $K = 2^B \geq 8$ seeds, the high-confidence infimum is $1/K$ and no finite sample attains it. We determine the joint proper

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.

Accuracy setting and domain	Samples, up to logs	Shared fair bits
Agnostic, countable	$d/(\rho^2\alpha^2)$	Indexed randomness
Agnostic, finite N	$d/(\rho^2\alpha^2)$	$O(\log\log(4N) + L_\beta L_{\alpha\rho})$
Agnostic structured products	$d/(\rho^2\alpha^2)$	$O(L_\rho)$
Agnostic k-interval unions	$k/(\rho^2\alpha^2)$	$O(L_\rho)$
Agnostic Hamming ball, standard Borel	$d/(\rho^2\alpha^2)$	$O(L_\rho)$
Agnostic, arbitrary	$d/(\rho^2\alpha^4)$	$O(L_\rho)$
Bayes-risk matched, arbitrary	$d/(\rho^2\alpha^2)$	$O(L_\rho)$
Realizable, arbitrary	$d/(\rho^2\alpha)$	$O(L_\rho)$

Table 1: Upper bounds for $d \geq 1$, with $L_\beta = \log(1/\beta)$ and $L_\rho = \log(1/\rho)$, $L_{\alpha\rho} = \log(1/(\alpha\rho))$. Bit counts are total for deterministic base learners and otherwise count additional randomness. All outputs may be improper. Replication is required on all distributions, even in the last two rows where accuracy has a distribution restriction. Structured products have d public blocks of countable VC-one classes, tree intervals, or real intervals; their VC dimension is at most $2d$. Interval unions have VC dimension $2k$ on the real line. The full Hamming ball has radius d and VC dimension d when the domain contains at least d points.

sample complexity just above this boundary as a function of accuracy, confidence, seed count, and boundary distance. For improper learning, we give sample bounds for every feasible bit budget. Near the feasibility boundary, VC dimension multiplies the logarithmic cost of approaching the boundary in the sample lower bound. Matching joint tradeoffs, including confidence, hold for public products of countable VC-one classes, countable tree intervals, and real intervals. One seed serves all blocks, whose masses are unknown. The same joint law holds for interval unions without a public partition and full Hamming balls on standard Borel domains. Their uniform sample bound is $O(\alpha^{-2}[d\rho^{-2} + \log(2/\beta)])$, with d the number of blocks, intervals, or allowed point changes, respectively. Extending this joint rate to general VC classes remains open. Table 1 summarizes the sample and randomness bounds. For arbitrary domains, we use the base-learner measurability assumptions stated below.

Screening normalized label signals. Write $p_x = P(X = x)$ and $\mu_x = \mathbb{E}[Y \mid X = x]$. Changing a prediction changes risk by $p_x|\mu_x|$. Large atoms with nearly unbiased labels therefore contribute little to risk fluctuations. The earlier mass-based screen learns conditional labels at points of mass about α^2 to accuracy α, producing the α^{-4} cost. We instead estimate $\sqrt{p_x}\mu_x$ and screen large $p_x\mu_x^2$. Selected points receive Bayes signs; for the remaining points,

$$\sum_{x \notin B} (p_x|\mu_x|)^2 = \sum_{x \notin B} p_x(p_x\mu_x^2) \leq v \quad \text{if } p_x\mu_x^2 \leq v \text{ outside } B.$$

This bound controls the risk fluctuations required for high-probability accuracy. We retain the ensemble framework of Hopkins et al. (2026) and replace their Propositions 3.2–3.3 with signal screening and its risk analysis.

Coupling predictions rather than candidate indices. On arbitrary domains, we use all thresholds of one ensemble as candidates and weight them by the positive part of their improvement over the ensemble's average empirical risk plus a slack. The weights have a uniform lower bound on their total mass, and their perturbation is controlled by the *average* change in candidate risks. Their normalized cumulative sums define an inverse-CDF sampler. Since the candidates are nested, the prediction at a fixed point is determined by comparing one shared scalar to a cumulative sum. This identity couples predictions even when the selected indices differ. Discretizing the shared scalar on a dyadic grid gives the finite bit bound. On finite domains, almost independence preserves the optimal sample rate using a seed doubly logarithmic in the domain size.

Related work. Full-output replicability originates with Impagliazzo et al. (2022); its relations to privacy and generalization (Bun et al., 2023) and recent realizable PAC bounds (Larsen et al., 2026) concern a stronger agreement requirement. Dwork & Feldman (2018, Section 2.1) study private prediction using disjoint-sample ensembles and a related counting bound. Hopkins et al. (2026, Appendix A.1) use global shared ensemble thresholds. We restrict those thresholds to a central range and prove the resulting risk guarantee. Hopkins & Moran (2025, Theorem C.1) study randomness

complexity through full-output stability, using discrete thresholds and logarithmic additional randomness. Correlated sampling (Ghazi et al., 2021), inverse-CDF sampling, bounded-independence concentration (Schmidt et al., 1995), and small-bias generation (Alon et al., 1992) are established tools. Bun et al. (2023, Algorithm 10) combine random near-ERM thresholds with random orders. We use the VC-one tree structure characterized by Ben-David (2015, Theorem 4). Chase et al. (2023) prove properness separations for list replicability using topology and give a geometric characterization of that model. Our proper bounds instead quantify total fair bits, confidence, and pointwise disagreement. Signal screening removes the additive sample terms; nesting controls predictions through average risk perturbations; empirical-risk cuts and certified tree paths give short-seed learners for structured classes. The tree method controls aggregate risk through a cost per selected interval. Prefix-gain contraction and normalization by empirical block masses yield the joint product tradeoffs. For unknown interval partitions, local penalized values and normalized covering certificates give the short-seed guarantee.

2 Model and Main Result

Let $\mathcal{X}$ be a countable set and let $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$ be nonempty. All samples are drawn independently from an unknown distribution P on $\mathcal{X} \times \{-1,+1\}$. For a binary hypothesis h, define

$$\mathrm{R}_P(h) = \mathbb{P}_{(X,Y)\sim P}\{h(X) \neq Y\}, \qquad \mathrm{OPT}_{\mathcal{H}}(P) = \inf_{h\in\mathcal{H}} \mathrm{R}_P(h).$$

The infimum need not be attained. Put $\mu_x = 0$ when $p_x = 0$. For every binary h,

$$\mathrm{R}_P(h) = \frac{1}{2} - \frac{1}{2}\sum_{x\in\mathcal{X}} p_x \mu_x h(x). \tag{1}$$

The series is absolutely convergent.

Definition 2.1. A randomized algorithm is an agnostic (a,b)-learner for $\mathcal{H}$ if, for every P, its output is a binary function and

$$\mathbb{P}_{S,r}\{\mathrm{R}_P(A(S;r)) > \mathrm{OPT}_{\mathcal{H}}(P) + a\} \leq b.$$

It is ρ-pointwise replicable if, for every P and every fixed $x \in \mathcal{X}$,

$$\mathbb{P}_{S,S',r}\{A(S;r)(x) \neq A(S';r)(x)\} \leq \rho,$$

where S, S' are independent and r is shared.

The output need not belong to $\mathcal{H}$. The randomness model allows countably many independent uniform random variables indexed by $\mathcal{X}$, as well as independent tapes for finitely many calls to the base learner. Each component of this randomness is shared between corresponding executions. Sample blocks used by different subroutines within one execution are disjoint; samples are never shared between the two executions. All random outputs are understood with the coordinate sigma-algebra on the countable domain. This is the indexed shared-randomness model of Hopkins et al. (2026, Section 1.1, Shared Randomness). Section 4 gives the separate agnostic bound for finite shared randomness.

Theorem 2.2 (Generic transformation)**.** *Suppose A is a binary agnostic learner on a countable domain, using an integer number $m_A(a,b)$ of examples for parameters (a,b). For $0 < \alpha, \beta, \rho \leq 1/4$, set*

$$T = \left\lceil \frac{2}{\rho^2} \right\rceil, \qquad L = \log(3/\beta), \qquad \tau = \frac{\alpha}{\sqrt{18L}}.$$

There is an agnostic (α,β)-learner that is ρ-pointwise replicable and uses exactly

$$M = Tm_A\left(\frac{\alpha}{2}, \frac{\beta}{3T}\right) + M_0\left(\tau, \frac{\rho}{2}, \frac{\beta}{3}\right) + M_1\left(\tau, \frac{\rho}{2}, \frac{\beta}{3}\right) \tag{2}$$

examples, with M_0, M_1 as defined in Theorem A.1. In particular,

$$M \leq Tm_A\left(\frac{\alpha}{2}, \frac{\beta}{3T}\right) + O\left(\frac{L}{\rho^2\alpha^2} \log \frac{L}{\alpha\rho\beta}\right). \tag{3}$$

The transformation makes exactly T calls to A.

Corollary 2.3 (VC classes)**.** *Every nonempty binary hypothesis class of finite VC dimension $d \geq 1$ on a countable domain, for $0 < \alpha, \beta, \rho \leq 1/4$, has an improper agnostic (α, β)-learner that is ρ-pointwise replicable and uses*

$$\widetilde{O}\left(\frac{d}{\rho^2\alpha^2}\right)$$

examples. Here the suppressed factors are logarithmic in $1/\alpha, 1/\beta, 1/\rho$.

Proof. Apply Theorem 2.2 to an ordinary VC learner. The standard VC uniform-convergence bounds (Blumer et al., 1989; Haussler, 1992) give a choice with

$$m_A(a, b) = O\left(\frac{d\log(2/a) + \log(2/b)}{a^2}\right).$$

An empirical minimizer exists because empirical binary risks take finitely many values on every finite sample. For each sample fix one such minimizer. The space of finite samples of a given size is countable, so this deterministic selection is measurable in every output coordinate. Substitution in (3), using $T = O(\rho^{-2})$ and $d \geq 1$, proves the claim. □

Remark 2.4 (Matching lower bound). The corollary matches the agnostic pointwise lower bound of Hopkins et al. (2026, Theorem 1.5) up to logarithmic factors. Their statement uses excess error $0.01a$ and failure probability 10^{-4} and gives $\widetilde{\Omega}(d/(\rho^2 a^2))$ samples. Constant rescaling of accuracy gives matching rates at this fixed confidence, in the small-accuracy, small-ρ range of their lower bound. Confidence dependence and logarithmic factors remain unresolved in this comparison. A nonempty class of VC dimension zero contains just one distinct binary function; outputting it needs no samples.

Construction and analysis. Use the parameters and sample sizes of Theorem 2.2. Let C be the set of feature values observed in the discovery block. On the independent estimation block of size n, put $N_x = \sum_{i=1}^{n} \mathbf{1}\{X_i = x\}$ and $\widehat{t}_x = \sum_{i:X_i=x} Y_i/\sqrt{nN_x}$, with value zero when $N_x = 0$. Draw W uniformly from $[\tau, 2\tau]$ and set $B = \{x \in C : |\widehat{t}_x| \geq W\}$. On T disjoint remaining blocks, train $h_1, \ldots, h_T$ with parameters $(\alpha/2, \beta/(3T))$. Write $q_x = T^{-1}\sum_j \mathbf{1}\{h_j(x) = +1\}$ and return

$$\widehat{h}(x) = \begin{cases} \operatorname{sign}(\widehat{t}_x), & x \in B, \\ \mathbf{21}\{U_x \leq q_x\} - 1, & x \notin B. \end{cases}$$

The threshold W, all indexed uniforms $U_x \sim \mathrm{Unif}[0, 1]$, and all base tapes are mutually independent and independent of all samples; corresponding executions share each of these random components.

For $t_x = \sqrt{p_x}\mu_x$, estimation has the mass-independent tail bound

$$\mathbb{P}\{|\widehat{t}_x - t_x| \geq u\} \leq 4e^{-nu^2/8} \qquad (u > 0).$$

Independent discovery permits a union bound over the finite set C. With the prescribed screen sizes, except with probability $\beta/3$, all selected signs are Bayes labels and $|t_x| \leq 3\tau$ off B. The shared W also makes the screening states (selected sign or abstention) disagree at any fixed point with probability at most $\rho/2$; Appendix A proves these guarantees.

Condition on the samples, base tapes, and W on this good event. The risk-range calculation above bounds the sum of squared ranges by $9\tau^2$, and the conditional mean risk is at most the average base risk. Hence

$$\mathbb{P}_U\left\{\mathrm{R}_P(\widehat{h}) - \mathbb{E}_U\,\mathrm{R}_P(\widehat{h}) > \frac{\alpha}{2}\right\} \leq e^{-\alpha^2/(18\tau^2)} = \beta/3.$$

The base-accuracy event has failure probability at most $\beta/3$, giving (α, β) accuracy. Shared ensemble rounding adds at most $\mathbb{E}|q_x - q'_x| \leq 1/\sqrt{2T} \leq \rho/2$ to pointwise disagreement, completing replication. Appendices B–C give the full proofs.

3 FINITE RANDOMNESS FOR REALIZABLE AND BAYES-OPTIMAL LEARNING

This section applies to arbitrary domains. Assume that the base learner is defined on every sample, that its binary predictions are measurable jointly in its input and randomness, and that all risks and probability events below are measurable. A realizable (a, b)-learner has risk at most a with probability at least $1 - b$ whenever P is supported on $\{(x, f(x)) : x \in \mathcal{X}\}$ for some $f \in \mathcal{H}$. Pointwise replicability is required for *every* distribution, including those not realizable by $\mathcal{H}$.

Theorem 3.1 (Realizable transformation)**.** *Let A be a binary realizable learner satisfying the preceding measurability conditions, using an integer number $m_A(a, b)$ of examples. For $0 < \alpha, \beta, \rho \leq 1/4$, let*

$$k = 2^{\lceil \log_2(3/(2\rho^2)) \rceil}, \qquad T = 3k.$$

There is a realizable (α, β)-learner that is ρ-pointwise replicable on all distributions and uses exactly

$$T m_A(\alpha/3, \beta/T)$$

examples and T calls to A. The transformation uses exactly $\log_2 k$ additional shared fair random bits, beyond the independent base tapes shared between corresponding executions. If A is deterministic, these are all its random bits.

Proof sketch. Train $T = 3k$ independent base hypotheses and draw one shared K uniformly from $\{k + 1, \ldots, 2k\}$. Return $H_K(x) = 2\mathbf{1}\{N_x \geq K\} - 1$, where N_x is the positive vote count. A changed prediction requires K to lie between the two vote counts, so disagreement is at most $\sqrt{T/2}/k \leq \rho$. For every label y, a voting error requires at least k base errors. Thus, deterministically for every admissible K,

$$\mathrm{R}_P(H_K) \leq \frac{3}{T} \sum_{j=1}^{T} \mathrm{R}_P(h_j). \tag{4}$$

The base-accuracy union bound proves accuracy for every seed. Appendix D gives the full argument.

Corollary 3.2 (Realizable VC bound)**.** *For $0 < \alpha, \beta, \rho \leq 1/4$, every nonempty binary class of VC dimension $d \geq 1$ on a countable domain admits an improper realizable (α, β)-learner that is ρ-pointwise replicable on all distributions, uses*

$$\widetilde{O}\left(\frac{d}{\rho^2 \alpha}\right)$$

examples, and needs $O(\log(1/\rho))$ shared fair random bits in total. The suppressed factors are logarithmic in $1/\alpha, 1/\beta, 1/\rho$. The same statement holds on any domain for which a deterministic measurable base learner with the standard realizable VC sample bound exists.

Use a deterministic ERM with the standard realizable VC bound; Appendix D gives the substitution.

For $d \geq 2$, this matches the lower-bound construction of Hopkins et al. (2026, Theorem 3.10) up to logarithms at fixed small failure probability and sufficiently small α, ρ. Appendix D details the dimension qualification: a singleton domain has VC-one classes requiring no α^{-1} cost in the realizable case.

3.1 LABEL NOISE AND APPROXIMATION ERROR

The loss comparison also bounds excess risk above the Bayes optimum $\mathrm{R}^\star_P$ under label noise.

Proposition 3.3 (Finite randomness with approximation error)**.** *Apply the construction of Theorem 3.1 to an ordinary agnostic base learner with parameters $(\alpha/3, \beta/T)$. It is ρ-pointwise replicable on every distribution and, with probability at least $1 - \beta$, satisfies*

$$\mathrm{R}_P(H_K) \leq \mathrm{R}^\star_P + 3\big(\mathrm{OPT}_{\mathcal{H}}(P) - \mathrm{R}^\star_P\big) + \alpha. \tag{5}$$

When $\mathrm{OPT}_{\mathcal{H}}(P) = \mathrm{R}^\star_P$, this is additive agnostic accuracy under arbitrary label noise, even if the class-risk infimum is not attained.

The proof weights the error-counting inequality by label bias; Appendix D gives the details.

Deterministic VC learning gives $\widetilde{O}(d/(\rho^2\alpha^2))$ samples and $O(\log(1/\rho))$ total bits on Bayes-risk-matched distributions, with the preceding measurability assumptions.

Central voting under misspecification. On three uniformly distributed positive-labeled points, the three hypotheses positive at one point each all have optimal class risk $2/3$. An ensemble with k copies of each has $N_x = k$ everywhere, so every central threshold outputs constant -1, of risk one. Risk-weighted threshold selection in the next section resolves this failure of central voting.

3.2 A MATCHING BOUND ON SHARED FAIR BITS

A continuity argument gives a lower bound on total shared randomness (Impagliazzo et al., 2022; Hopkins et al., 2026). The bound counts all algorithmic randomness, including the bits used inside the base learner.

Proposition 3.4 (Total shared-bit lower bound)**.** *Suppose $\mathcal{H}$ has VC dimension at least one. Let a learner use a finite, fixed number of examples and at most B shared fair random bits, with no other randomness. Suppose it is ρ-pointwise replicable for every distribution and is an (α, β)-learner on realizable distributions, where $\alpha < 1$ and $\beta < 1/2$. Then*

$$\rho \geq \frac{1}{2} \quad \text{if } B = 0, \qquad \rho > 2^{-B-1} \quad \text{if } B \geq 1.$$

In particular, for $0 < \rho \leq 1/4$, $B \geq \lfloor \log_2(1/\rho) \rfloor$.

Proof sketch. At a shattered point, each deterministic-label distribution has more than half of the seeds predicting correctly. For $B \geq 1$, integrality gives at least two seeds correct at both endpoints. One seed has prediction probability $1/2$ at an intermediate label bias by continuity. The other has probability strictly between zero and one there, giving the strict inequality. The full argument is in Appendix F.1.

The intermediate distribution can have noisy labels, so the lower bound uses replication on non-realizable distributions. The next construction attains the exact integer lower bound when any finite sample size is allowed.

4 FINITE RANDOMNESS UNDER MODEL MISSPECIFICATION

Central voting can violate additive agnostic accuracy. We select thresholds by their validated risks instead. The selection law changes with the *average* perturbation of candidate risks, yielding finite-randomness learning on arbitrary domains under model misspecification.

Theorem 4.1 (Agnostic learning with a short shared seed)**.** *On any domain admitting a binary measurable agnostic learner A, there is an improper agnostic (α, β)-learner that is ρ-pointwise replicable and uses*

$$Tm_A(\alpha/8, \alpha/8) + O\left(T + \frac{1}{\alpha^2} \log \frac{T}{\beta}\right)$$

examples, where $T = \Theta(\alpha^{-2}(\rho^{-2} \log(1/\rho) + \log(1/\beta)))$. The transformation uses exactly

$$B_{\text{fast}} = \left\lceil \log_2 \frac{1}{\rho} \right\rceil$$

additional shared fair bits, independently of α, β and the domain. For VC dimension $d \geq 1$ and an ordinary learner with the standard agnostic VC bound, the sample count is $\widetilde{O}(d/(\rho^2\alpha^4))$. With a deterministic base learner, this is the total randomness. For $d \geq 1$, it uses at most one bit above the exact minimum.

Unless stated otherwise, $0 < \alpha, \beta, \rho \leq 1/4$. This improves the arbitrary-domain finite-randomness bound $\widetilde{O}(d/(\rho^4\alpha^7))$ of Hopkins et al. (2026, Theorem 1.13). An additional inverse-square accuracy factor separates this bound from the countable-domain rate. The exact construction and proof are in Appendix F.

Exact minimum and the sample tradeoff. For any prescribed integer B with $\Delta_B = \rho - 2^{-B-1} > 0$, the same construction uses exactly B additional bits and $\widetilde{O}(d\, 4^B \alpha^{-4} \log(1/\Delta_B))$ examples, with the explicit parameters in Theorem F.3. Together with Proposition 3.4, this proves

$$B_{\min} = \left\lfloor \log_2 \frac{1}{\rho} \right\rfloor$$

for total bits with improper outputs, for every nontrivial VC class admitting the stipulated deterministic ordinary learner, when the sample size may be any finite number. More precisely, for VC dimension at least d, any such agnostic learner using m examples and $B \geq 4$ total bits, with $K = 2^B$ and $0 < K\Delta_B < 1/2$, satisfies

$$m \geq \frac{3dK^2}{131072\alpha^2} \log \frac{1}{2K\Delta_B} \quad (0 < \alpha \leq 1/16,\ 0 < \beta \leq 1/4).$$

The dimension factor follows by selecting an edge of a shattered cube and applying a two-seed relative-entropy argument (Appendix F.2). Theorem 5.1 matches this bound jointly with accuracy and confidence for structured products, including the full binary class on d points. Matching this joint bound for general VC classes remains open. At $B_{\min}$, Δ_B can be arbitrarily small; choosing B_{fast} ensures $\Delta_B \geq \rho/2$ and gives the uniform sample bound in Theorem 4.1.

Properness couples randomness and confidence. Let $\mathcal{H}_3$ consist of exactly the three point functions on $\{1, 2, 3\}$, a VC-one class, and fix $0 < \alpha \leq 1/16$. For $B \in \{1, 2\}$ and $0 < \beta \leq 1/16$, allowing also $\rho > 1/4$, the optimal proper disagreement satisfies $2^{-B} - \rho^*_{\text{prop},B}(\mathcal{H}_3; \alpha, \beta) = \Theta(\beta^2)$. Rates $\rho = 2^{-B} - c\beta^2$ admit the optimal sample count $\Theta(\alpha^{-2} \log(1/\beta))$ (Theorem H.18); sample permutations implement the output laws without extra coins. For one bit, the same confidence and sample orders hold for all point functions on any standard Borel domain with at least three points, with or without the negative classifier (Theorem H.9).

For $\mathcal{H}_3$, $K = 2^B \geq 8$ and $\beta < 1/(2K)$ give $\rho^*_{\text{prop},B}(\mathcal{H}_3; \alpha, \beta) = 1/K$. Every proper learner using at most B total shared fair bits and no other randomness has $\rho > 1/K$ at any fixed finite sample size (Theorem H.21). The boundary follows from a parity obstruction for five continuous maps. For $\rho = 1/K + \Delta$ with $0 < K\Delta \leq 1/4$, the minimum sample size is

$$m^*_{\text{prop}} = \Theta\left(\frac{1}{\alpha^2}\left[K^2 \log \frac{1}{K\Delta} + \log \frac{1}{\beta}\right]\right) \quad \text{(Theorem H.23)}.$$

Five-seed subsets intersecting in at most one seed give $\Omega(K^2)$ witness distributions separated by statistical tests in a simplex of diameter $O(\alpha)$. Packing gives the lower bound; ordered selection matches it. Thus, when $0 < \rho \leq 1/8$ and $\beta < \rho/2$, the exact minimum is $\lfloor \log_2(1/\rho) \rfloor + 1$ total shared fair bits for proper learners, one more than for improper learners (Corollary H.22). These minima allow any fixed finite sample size and no other randomness. For N point functions, Theorem H.15 characterizes the disagreement infimum by continuous maps. For larger point classes, even with the negative classifier, a strict confidence-dependent lower bound has a logarithmic finite-sample penalty (Theorems H.25 and H.26).

Risk-weighted threshold selection. Train T base hypotheses and let $g_j(x) = 2\mathbf{1}\{N_x \geq j\} - 1$ for $j = 1, \ldots, T$, where N_x is the positive vote count. On independent validation data, let $\widehat{R}_j$ denote the risk of g_j, and put $\overline{R} = T^{-1} \sum_j \widehat{R}_j$. This average is exactly the average empirical risk of the base hypotheses. For a dyadic $a = \Theta(\alpha)$ with $a \leq \alpha/4$, use the weights

$$w_j = \frac{1}{2}(\overline{R} + a - \widehat{R}_j)_+.$$

Their average is at least $a/2$ by convexity, on every training realization. Every positive-weight candidate has empirical risk below $\overline{R} + a$. Put $F(k) = \sum_{j \leq k} w_j / \sum_j w_j$, with $F(0) = 0$. For $K = 2^{B_{\text{fast}}}$, draw one shared uniform integer $D \in \{0, \ldots, K-1\}$, set $U = (D + 1/2)/K$, and select $J = \min\{j : F(j) \geq U\}$. Every selected index has positive weight, including when U equals a cumulative-sum boundary. Each base call has expected risk at most $\mathrm{OPT}_{\mathcal{H}}(P) + \alpha/4$. Concentration of their average and uniform validation therefore ensure risk at most $\mathrm{OPT}_{\mathcal{H}}(P) + \alpha$ for every seed on a common good event.

Coupling predictions through nested candidates. Nesting reduces the prediction to one scalar comparison:

$$g_J(x) = +1 \quad \Longleftrightarrow \quad U \leq F(N_x).$$

The average weight is at least $a/2$. Normalization and the positive-part map imply that replacing one base block changes $F(N_x)$ by at most $4/(aT)$; replacing one validation observation changes

it by at most $2/(an)$. Conditional on the shared base tapes, the two executions give independent, identically distributed scores. The bounded-differences inequality (McDiarmid, 1989) applies with squared sensitivity $S_* = 16/(a^2T) + 4/(a^2n)$. At most one grid midpoint is closer than $1/(2K)$ to the conditional mean. Its conditional disagreement is at most $1/2$; the remaining midpoints are controlled by the concentration tails. The resulting prediction disagreement is at most

$$\frac{1}{2K} + 2\exp\left(-\frac{1}{2K^2 S_*}\right).$$

Taking $T, n \geq 40K^2a^{-2}\log(4/\Delta_B)$ makes this below ρ. Uniform selection fails on the example in Appendix F.3. For one fixed countable VC-at-most-three class and deterministic ERM, the validated CDF compiler has disagreement at least $c_* \min\{1, 1/(\alpha\sqrt{T})\} - 1/K$ for every T and validation size $n \geq 1024/\alpha^2$ (Appendix N). Our ordered method learns the same class with $\widetilde{O}(1/(\rho^2\alpha^2))$ samples and $O(\log(1/\rho))$ total bits (Corollary N.8).

5 Optimal samples and short seeds for structured classes

For the structured classes below, a domain-independent short seed achieves the optimal agnostic sample rate under model misspecification.

Theorem 5.1 (Structured sample and randomness tradeoffs)**.** *For $0 < \alpha, \beta, \rho \leq 1/4$, there are improper agnostic (α, β)-learners that are ρ-pointwise replicable on every distribution and use $\lceil \log_2(2/\rho) \rceil$ total shared fair bits in each case:*

1. *For a product of $d \geq 1$ classes on known disjoint subdomains, each a nonempty countable-domain VC-at-most-one class, a tree-interval class on a countable rooted tree, or a real-interval class on a Borel copy of $\mathbb{R}$, the sample count is $O(\alpha^{-2}[d\rho^{-2} + \log(2/\beta)])$ (Theorem K.13).*
2. *Unions of at most k intervals on a countable linear order or the Borel real line admit $O(\alpha^{-2}[k\rho^{-2} + \log(2/\beta)])$ samples, with no public partition (Theorem L.8).*
3. *On a standard Borel domain, if every hypothesis differs from a known measurable classifier at at most $d \geq 1$ domain points, the sample count is $O(\alpha^{-2}[d\rho^{-2} + \log(2/\beta)])$ (Theorem M.3).*

The last rate uses the Hamming radius d, which equals the full ball's VC dimension when the domain has at least d points. The algorithms do not need the marginal atom masses or conditional label probabilities.

In the first case, if every factor is nontrivial, then for $0 < \alpha \leq 1/16$, $0 < \beta \leq 1/4$, integer $B \geq 4$, $K = 2^B$, and $0 < K\Delta_B \leq 1/4$, where $\Delta_B = \rho - 1/(2K)$, the minimum fixed sample count is

$$m^* = \Theta\left(\frac{1}{\alpha^2}\left[dK^2 \log\frac{1}{K\Delta_B} + \log\frac{1}{\beta}\right]\right)$$

among learners using at most B total shared fair bits and no other randomness, allowing improper outputs. Constants are absolute (Corollary K.14). The same formula holds for full Hamming balls on standard Borel domains with at least d points (Corollary M.5), and with d replaced by k for interval unions on a domain containing at least $2k$ points (Corollary L.10).

Certified tree paths and public products. On a VC-one tree, start where a backward path-gain score exceeds $W + \eta$ and follow a maximizing path while it exceeds $W - \eta$. The maximal selected prefixes are disjoint and pay 2η each. Capping the recursion at λ also creates a surplus: selected prefixes pay all of it, while any comparator interval gains at most λ plus the total surplus. Unions of m tree intervals have VC dimension at most $2m$; their deviation above the $2\eta m$ cost has tail $e^{-nt^2/2}$. The confidence term is therefore additive. For d public blocks use $\lambda_j = \alpha\sqrt{N_j/n}/(2\sqrt{d})$ and one shared threshold. Nested clipping contracts prefix-gain errors, so normalized scores concentrate on scale $1/\sqrt{n}$ despite unknown block masses. The same analysis covers real intervals with atoms (Appendix K).

Discovering intervals without a partition. With $\tau = \alpha/(4\sqrt{k})$ and $\lambda = \alpha/(4k)$, define the local penalized value

$$V(I) = \max_{A \subseteq I} \left\{ \widehat{a}(A) - \sum_{C \text{ a component of } A} \left(\tau\sqrt{\widehat{p}(C)} + \lambda\right) \right\},$$

$$F(x) = \max\left\{0, \sup_{I \ni x} \frac{\widehat{a}(I) - V(I) - \lambda}{\sqrt{\widehat{p}(I)}}\right\},$$

where zero-mass intervals contribute zero. An optimal certificate writes $\widehat{a} = q + e$, with $e \geq 0$ and $q(I) \leq \tau\sqrt{\widehat{p}(I)} + \lambda$. Thresholding at $w \in [4\tau/5, 9\tau/10]$ covers all surplus and leaves a gain margin of at least λ per component. A k-interval comparator pays only $\tau\sqrt{k} + \lambda k$. Without the component charge, this selector can overfit many noisy atoms (Proposition L.11). For unions of m intervals, let D_m be their largest signed empirical deviation. The residual $Z_\lambda = \sup_m (D_m - \lambda m)_+$ has expectation $O(1/(n\lambda))$ and sensitivity $2/n$, so confidence enters additively. For replication, an analysis-only population floor on interval masses bounds the score's replacement variance by $O(1/n)$. Relative-mass concentration on intervals anchored at the query transfers the resulting entropy bound to the actual score. One grid threshold contributes at most $1/(2K)$ to disagreement; the others have exponentially small tails (Appendix L). The score admits cubic preprocessing and logarithmic queries in the real-arithmetic model (Proposition L.9).

Open problems and computational scope. Crossing anchored rectangles rule out bounded hypothesis-cut orders (Proposition I.5). The general finite-bit agnostic sample gap and proper pointwise replication for general VC classes remain open. The general results concern statistical complexity and impose no running-time constraint. Full-output and approximate replicability are separate requirements.

Reproducibility Statement

The appendices provide complete proofs, explicit sample partitions, parameter choices, and shared-randomness constructions.

AI Use Statement

Generative AI assisted literature research, theorem development, proof writing and checking, manuscript preparation, and automated internal reviews. These reviews are not external peer review. The authors are responsible for all claims.

REFERENCES

Noga Alon, Oded Goldreich, Johan Håstad, and René Peralta. Simple constructions of almost k-wise independent random variables. *Random Structures & Algorithms*, 3(3):289–304, 1992. doi: 10.1002/rsa.3240030308.

Francis Bach. Learning with submodular functions: A convex optimization perspective, 2013. URL https://arxiv.org/abs/1111.6453.

Shai Ben-David. 2 notes on classes with Vapnik–Chervonenkis dimension 1. https://arxiv.org/abs/1507.05307, 2015.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik–Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989. doi: 10.1145/76359.76371.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. On concentration of self-bounding functions. *Electronic Journal of Probability*, 14(64):1884–1899, 2009. URL https://www.maths.tcd.ie/EMIS/journals/EJP-ECP/article/download/690/690-2335-1-PB.pdf.

Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization. https://arxiv.org/abs/2303.12921, 2023.

Zachary Chase, Shay Moran, and Amir Yehudayoff. Replicability and stability in learning. https://arxiv.org/abs/2304.03757, 2023.

Cynthia Dwork and Vitaly Feldman. Privacy-preserving prediction. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1693–1702. PMLR, 2018. URL https://proceedings.mlr.press/v75/dwork18a.html.

Peter Elias. The efficient construction of an unbiased random sequence. *The Annals of Mathematical Statistics*, 43(3):865–870, 1972. doi: 10.1214/aoms/1177692552.

Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. User-level differentially private learning via correlated sampling. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a89cf525e1d9f04d16ce31165e139a4b-Abstract.html.

Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002. URL https://pi.math.cornell.edu/~hatcher/AT/ATpage.html.

David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992. doi: 10.1016/0890-5401(92)90010-D.

David Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik–Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995. doi: 10.1016/0097-3165(95)90052-7.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Max Hopkins and Shay Moran. The role of randomness in stability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23805–23827. PMLR, 2025. URL https://proceedings.mlr.press/v267/hopkins25a.html.

Max Hopkins, Russell Impagliazzo, and Christopher Ye. Approximate replicability in learning. https://arxiv.org/abs/2510.20200v2, 2026. URL https://arxiv.org/abs/2510.20200v2. Version 2, April 7, 2026.

Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning. https://arxiv.org/abs/2201.08430, 2022.

Alexander S. Kechris. *Classical Descriptive Set Theory*, volume 156 of *Graduate Texts in Mathematics*. Springer, New York, 1995. doi: 10.1007/978-1-4612-4190-4.

Łukasz Kruk, John Lehoczky, Kavita Ramanan, and Steven Shreve. An explicit formula for the Skorokhod map on $[0, a]$. *The Annals of Probability*, 35(5):1740–1768, 2007. doi: 10.1214/009117906000000890.

Kasper Green Larsen, Markus Engelund Mathiasen, Chirag Pabbaraju, and Clement Svendsen. The sample complexity of replicable realizable PAC learning. https://arxiv.org/abs/2602.19552, 2026.

Colin McDiarmid. On the method of bounded differences. In J. Siemons (ed.), *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

Jeanette P. Schmidt, Alan Siegel, and Aravind Srinivasan. Chernoff–Hoeffding bounds for applications with limited independence. *SIAM Journal on Discrete Mathematics*, 8(2):223–250, 1995. doi: 10.1137/S089548019223872X.

A A JOINTLY ACCURATE, POINTWISE REPLICABLE SCREENING RULE

Let P be a distribution on $\mathcal{X} \times \{-1, +1\}$, where $\mathcal{X}$ is countable. Write $p_x = P(X = x)$, $\mu_x = \mathbb{E}[Y \mid X = x]$, and $t_x = \sqrt{p_x}\mu_x$. For $p_x = 0$, set $\mu_x = t_x = 0$. The abstention symbol $\perp$ is distinct from both labels. Throughout this section, logarithms are natural.

The screening rule. Take a discovery sample of size m_0 and let C be the set of its observed feature values. Independently, take an estimation sample $(X_i, Y_i)_{i=1}^n$. For every x, define

$$N_x = \sum_{i=1}^{n} \mathbf{1}\{X_i = x\}, \qquad \widehat{t}_x = \begin{cases} \dfrac{\sum_{i:X_i=x} Y_i}{\sqrt{nN_x}}, & N_x > 0, \\ 0, & N_x = 0. \end{cases}$$

Draw W uniformly from $[\tau, 2\tau]$, independently of both samples, and output the map

$$\Phi(x) = \begin{cases} \operatorname{sign}(\widehat{t}_x), & x \in C \text{ and } |\widehat{t}_x| \geq W, \\ \perp, & \text{otherwise.} \end{cases}$$

The sign at zero is immaterial because $W > 0$.

Theorem A.1 (Jointly accurate, pointwise replicable screening)**.** *Fix $0 < \tau \leq 1$, $0 < \rho \leq 1$, and $0 < \delta < 1$. Set*

$$\eta = \min\{\delta/2, \rho/4\}, \qquad \varepsilon = \frac{\rho\tau}{4},$$

$$m_0 = M_0(\tau, \rho, \delta) = \left\lceil \frac{4}{\tau^2} \log \frac{8}{\eta\tau^2} \right\rceil, \qquad n = M_1(\tau, \rho, \delta) = \left\lceil \frac{128}{\rho^2\tau^2} \log \frac{8m_0}{\eta} \right\rceil.$$

Run the rule twice on independent samples from P, using the same threshold W, and denote the outputs by Φ_1, Φ_2. Then, for every fixed $x \in \mathcal{X}$,

$$\Pr\{\Phi_1(x) \neq \Phi_2(x)\} \leq \rho. \tag{6}$$

Moreover, with probability at least $1 - \delta$, simultaneously for $j \in \{1, 2\}$ and all $x \in \mathcal{X}$,

$$\Phi_j(x) \neq \perp \implies \mu_x \neq 0 \text{ and } \Phi_j(x) = \operatorname{sign}(\mu_x), \tag{7}$$

$$\Phi_j(x) = \perp \implies |t_x| \leq 3\tau. \tag{8}$$

Each execution uses exactly $m_0 + n$ examples. Thus its sample requirement is $\widetilde{O}(1/(\rho^2\tau^2))$, with all logarithmic and confidence dependence given explicitly above. Equation (6) is a pointwise statement; it does not assert agreement of the entire maps with probability $1 - \rho$.

Proof. **A square-root binomial bound.** If $N \sim \mathrm{Bin}(n, p)$, then for every $s > 0$,

$$\Pr\left\{\left|\sqrt{N/n} - \sqrt{p}\right| \geq s\right\} \leq 2e^{-ns^2}. \tag{9}$$

Here is a derivation including the lower-tail endpoint. Put $v = np$. For $v > 0$, the moment-generating-function inequality

$$\mathbb{E}e^{\lambda N} = (1 + p(e^\lambda - 1))^n \leq \exp(v(e^\lambda - 1))$$

and exponential Markov's inequality imply $\Pr\{N \geq vr\} \leq e^{-vh(r)}$ for $r \geq 1$ and $\Pr\{N \leq vr\} \leq e^{-vh(r)}$ for $0 < r \leq 1$, where $h(r) = r \log r - r + 1$. At $r = 0$, the lower bound follows directly from $\Pr\{N = 0\} = (1 - p)^n \leq e^{-v}$, with $h(0) = 1$. For $r = z^2 \geq 0$,

$$h(r) - (\sqrt{r} - 1)^2 = 2z(z \log z - z + 1) \geq 0,$$

where the expression at zero is defined by continuity. Taking $r = (1 + a/\sqrt{v})^2$ bounds the upper square-root tail by e^{-a^2}. For $0 < a < \sqrt{v}$, taking $r = (1 - a/\sqrt{v})^2$ bounds the lower tail by e^{-a^2}; at $a = \sqrt{v}$, the lower-tail event is $\{N = 0\}$ and its probability is directly bounded by $e^{-v} = e^{-a^2}$ as above. For $a > \sqrt{v}$ that event is empty. If $v = 0$, $N = 0$ almost surely. The union bound and $a = s\sqrt{n}$ give (9).

Concentration of the normalized label sum. For every fixed x and every $u > 0$,

$$\Pr\{|\widehat{t}_x - t_x| \geq u\} \leq 4\exp(-nu^2/8). \tag{10}$$

Indeed, with a zero value assigned to the first term when $N_x = 0$, the exact decomposition is

$$\widehat{t}_x - t_x = \frac{\sum_{i:X_i=x}(Y_i - \mu_x)}{\sqrt{nN_x}} + \mu_x\left(\sqrt{N_x/n} - \sqrt{p_x}\right).$$

Conditionally on the indicators $(\mathbf{1}\{X_i = x\})_{i=1}^n$, the labels at indices having $X_i = x$ are independent with mean μ_x and range length two. Hoeffding's inequality therefore gives, for $N_x = k > 0$ and every $s > 0$, a conditional upper bound $2e^{-ns^2/2}$ for the absolute value of the first term exceeding or equaling s. The same bound holds unconditionally, since the term vanishes when $k = 0$. Because $|\mu_x| \leq 1$, (9) bounds the absolute second term by the tail $2e^{-ns^2}$. Apply these bounds at $s = u/2$ and take a union bound to obtain (10). When $p_x = 0$, the conclusion also holds directly.

A global event independent of the shared threshold. There are at most $4/\tau^2$ points with $p_x \geq \tau^2/4$. Consequently, the probability that discovery misses any such point is at most

$$\frac{4}{\tau^2}e^{-m_0\tau^2/4} \leq \frac{\eta}{2}.$$

Conditionally on the discovery sample, C is a fixed set of at most m_0 points and the estimation sample still has distribution P^n. Thus (10) and the union bound give

$$\Pr\{\exists x \in C : \ |\widehat{t}_x - t_x| > \varepsilon \mid C\} \leq 4m_0e^{-n\varepsilon^2/8} \leq \frac{\eta}{2}.$$

Let G be the event that every point of mass at least $\tau^2/4$ belongs to C, and that every point in C has estimation error at most ε. Then $\Pr(G) \geq 1 - \eta$, and G depends only on the samples, not on W.

On G, a selected point satisfies $|\widehat{t}_x| \geq \tau > \varepsilon$, so t_x is nonzero and has the same sign as $\widehat{t}_x$. An unselected point in C satisfies

$$|t_x| \leq |\widehat{t}_x| + \varepsilon < W + \varepsilon \leq 2\tau + \varepsilon \leq 9\tau/4.$$

An unselected point outside C satisfies $|t_x| \leq \sqrt{p_x} < \tau/2$. These conclusions hold for every $W \in [\tau, 2\tau]$. For each $j \in \{1, 2\}$, let G_j be the event G formed from the discovery and estimation samples in execution j. These are purely sample-dependent events, so their intersection is independent of the common threshold W. The union bound gives $\Pr(G_1 \cap G_2) \geq 1 - 2\eta \geq 1 - \delta$, proving (7) and (8).

Pointwise agreement. Fix x and condition on both data sets for which $G_1 \cap G_2$ holds. If x is absent from either candidate set, then $|t_x| < \tau/2$. In the other execution it is either absent as well, or its estimate has magnitude at most $\tau/2 + \varepsilon < \tau$. Both outputs are therefore $\perp$. Otherwise, $x \in C_1 \cap C_2$, and both estimated magnitudes lie in $[|t_x| - \varepsilon, |t_x| + \varepsilon]$. Whenever both executions select x, their labels agree by the sign conclusion above. Whenever their selection decisions differ, their common threshold must lie in that same interval, whose length is 2ε. Independence and uniformity of W therefore imply

$$\Pr\{\Phi_1(x) \neq \Phi_2(x), \ G_1 \cap G_2\} \leq \frac{2\varepsilon}{\tau}.$$

Adding the probability of failure of the two data events yields

$$\Pr\{\Phi_1(x) \neq \Phi_2(x)\} \leq 2\eta + \frac{2\varepsilon}{\tau} \leq \rho,$$

as required. No union bound over x is taken in this final argument. □

B Rounding averaged predictions

Two lemmas connect screening to the base learner: the first bounds disagreement between coupled training runs, and the second controls the risk of the rounded hypothesis.

Lemma B.1 (Shared-threshold coupling of an ensemble). *For $j = 1, \dots, T$, run a binary learner on independent sample blocks S_j, using independent random tapes R_j, and let h_j be its outputs. In a second execution use independent blocks S'_j and the same corresponding tapes R_j, obtaining h'_j. For a fixed x, define*

$$q_x = \frac{1}{T}\sum_{j=1}^{T} \mathbf{1}\{h_j(x) = +1\}, \qquad q'_x = \frac{1}{T}\sum_{j=1}^{T} \mathbf{1}\{h'_j(x) = +1\}.$$

If U_x is uniform on $[0,1]$ and independent of all these variables, then

$$\mathbb{P}\{\mathbf{1}\{U_x \le q_x\} \ne \mathbf{1}\{U_x \le q'_x\}\} = \mathbb{E}|q_x - q'_x| \le \frac{1}{\sqrt{2T}}.$$

Proof. Put $Z_j = \mathbf{1}\{h_j(x) = +1\}$ and $Z'_j = \mathbf{1}\{h'_j(x) = +1\}$. Given R_j, these are independent Bernoulli variables with a common parameter $\theta_j(R_j)$. Consequently,

$$\mathbb{E}[Z_j - Z'_j \mid R_j] = 0, \qquad \mathbb{E}[(Z_j - Z'_j)^2 \mid R_j] = 2\theta_j(R_j)(1 - \theta_j(R_j)) \le \tfrac{1}{2}.$$

The pairs indexed by different j are independent, because their sample blocks and tapes are independent. Hence

$$\mathbb{E}[(q_x - q'_x)^2] \le \frac{1}{2T}.$$

Cauchy–Schwarz proves the bound on the first absolute moment. Conditionally on q_x, q'_x, the shared uniform threshold causes a disagreement on an interval of length $|q_x - q'_x|$. The coupling argument applies even though q_x and q'_x can be dependent before conditioning on the shared tapes. □

Lemma B.2 (Risk concentration on points of small signal). *Fix binary functions $h_1, \dots, h_T$, a set $B \subseteq \mathcal{X}$, and labels $\ell_x = \operatorname{sign}(\mu_x)$ for all $x \in B$. Assume $\mu_x \ne 0$ on B and $p_x\mu_x^2 \le v$ on $\mathcal{X} \setminus B$. Let $q_x = T^{-1}\sum_j \mathbf{1}\{h_j(x) = +1\}$. Independently for every $x \notin B$, let $H(x) = 2\mathbf{1}\{U_x \le q_x\} - 1$, where U_x is uniform on $[0,1]$; set $H(x) = \ell_x$ on B. Then*

$$\mathbb{E}_U \mathrm{R}_P(H) \le \frac{1}{T}\sum_{j=1}^{T} \mathrm{R}_P(h_j). \tag{11}$$

For every $s > 0$ and $v > 0$,

$$\mathbb{P}_U\{\mathrm{R}_P(H) - \mathbb{E}_U \mathrm{R}_P(H) > s\} \le \exp(-2s^2/v). \tag{12}$$

When the sum of squared risk ranges is zero the risk is deterministic, and its upper-tail probability is zero.

Proof. At a point of B, the Bayes label minimizes conditional binary classification risk. Outside B, the expected conditional risk of $H(x)$ equals the average conditional risk of the $h_j(x)$. Multiplying by p_x and summing proves (11).

By (1),

$$\mathrm{R}_P(H) - \mathbb{E}_U \mathrm{R}_P(H) = -\frac{1}{2}\sum_{x \notin B} p_x\mu_x\big(H(x) - \mathbb{E}_U H(x)\big).$$

These summands are independent and centered. Their range lengths are $c_x = p_x|\mu_x|$, and

$$\sum_{x \notin B} c_x^2 = \sum_{x \notin B} p_x(p_x\mu_x^2) \le v. \tag{13}$$

For a finite domain, Hoeffding's inequality (Hoeffding, 1963) immediately yields the claim.

The bound extends to countable domains by dominated convergence. Enumerate its points and let V_k be the corresponding centered summands. They satisfy $|V_k| \le p_{x_k}$, so their series converges absolutely, with all partial sums bounded in absolute value by one. For every $\lambda > 0$, Hoeffding's moment bound for finite independent sums gives

$$\mathbb{E}\exp\left(\lambda\sum_{k=1}^{N} V_k\right) \le \exp\left(\frac{\lambda^2}{8}\sum_{k=1}^{N} c_{x_k}^2\right).$$

Dominated convergence passes this inequality to the full series. Exponential Markov's inequality and optimization over λ give $\exp(-2s^2/\sum c_x^2)$ when the denominator is positive. Using (13) proves (12). □

C THE LEARNER AND ITS GUARANTEE

Use the parameters of Theorem 2.2. Within each execution, partition the examples into disjoint blocks: the two samples for the screening rule, and T samples of size $m_A(\alpha/2, \beta/(3T))$.

1. Run the screening rule with parameters $(\tau, \rho/2, \beta/3)$ to obtain Φ. Put $B = \{x : \Phi(x) \neq \perp\}$.
2. On the remaining T blocks, run A with parameters $(\alpha/2, \beta/(3T))$ and independent tapes $R_1, \ldots, R_T$, obtaining $h_1, \ldots, h_T$.
3. For each point, define $q_x = T^{-1} \sum_j \mathbf{1}\{h_j(x) = +1\}$ and output

$$\widehat{h}(x) = \begin{cases} \Phi(x), & x \in B, \\ 2\mathbf{1}\{U_x \leq q_x\} - 1, & x \notin B. \end{cases} \tag{14}$$

The screening threshold W, the tapes $R_1, \ldots, R_T$, and the countable family $(U_x)_{x \in \mathcal{X}}$ are mutually independent and independent of all samples. In a second execution, the corresponding components of this randomness are shared. Screening is independent of the field (U_x).

Proof of Theorem 2.2. **Pointwise replicability.** Fix a distribution P and a point x. By Theorem A.1, the screening outputs disagree at x with probability at most $\rho/2$. If the screening states agree and are labels, the final predictions agree. If the screening states are both $\perp$, the predictions can differ only if the shared rounding of q_x, q'_x differs. It follows from Lemma B.1 that

$$\mathbb{P}\{\widehat{h}(x) \neq \widehat{h}'(x)\} \leq \frac{\rho}{2} + \frac{1}{\sqrt{2T}} \leq \rho.$$

The argument uses a union bound for the two possible causes of disagreement, without conditioning the ensemble guarantee on the screening states.

High-probability accuracy. Let F be the event that every base hypothesis has risk at most $\mathrm{OPT}_{\mathcal{H}}(P) + \alpha/2$. By the guarantee of A and a union bound, $\mathbb{P}(F^c) \leq \beta/3$. Let G be the global good event for a single screening execution, as constructed in the proof of Theorem A.1. Its failure probability is at most $\beta/3$. On G, simultaneously for every x, all selected labels are Bayes labels and all unselected points satisfy $p_x \mu_x^2 \leq 9\tau^2$. These assertions hold for every value of the screening threshold in its prescribed interval.

Condition on all sample blocks, all base tapes, and W. The rounding variables U_x remain independent uniforms. For every such realization in $F \cap G$, Lemma B.2 gives

$$\mathbb{E}_U \, \mathrm{R}_P(\widehat{h}) \leq \frac{1}{T} \sum_j \mathrm{R}_P(h_j) \leq \mathrm{OPT}_{\mathcal{H}}(P) + \alpha/2.$$

The same lemma, with $v = 9\tau^2$ and $s = \alpha/2$, yields

$$\mathbb{P}_U\{\mathrm{R}_P(\widehat{h}) > \mathrm{OPT}_{\mathcal{H}}(P) + \alpha \mid \text{samples}, R_1, \ldots, R_T, W\} \leq \exp\left(-\frac{\alpha^2}{18\tau^2}\right) = \beta/3$$

on $F \cap G$. Averaging and adding the two failure probabilities proves the claimed (α, β) accuracy.

Sample bound. The algorithm uses precisely the samples in (2). To verify the displayed asymptotic bound with its logarithms, substitute $(\tau, \rho/2, \beta/3)$ into the screening theorem. Its auxiliary confidence parameter becomes

$$\eta = \min\{\beta/6, \rho/8\}.$$

Let $K = \log(L/(\alpha\rho\beta))$. Because $\alpha, \beta, \rho \leq 1/4$ and $L = \log(3/\beta)$,

$$\log \frac{8}{\eta\tau^2} = O(K), \qquad M_0 = O\left(\frac{L}{\alpha^2} K\right).$$

Consequently $\log(8M_0/\eta) = O(K)$ as well, including the logarithm of K, and

$$M_1 = O\left(\frac{L}{\rho^2\alpha^2} K\right).$$

Both ceiling terms are absorbed by these bounds. This proves (3) and completes the proof. □

D PROOF OF THE CENTRAL-VOTING TRANSFORMATION

Proof of Theorem 3.1. Run A on T independent sample blocks with parameters $(\alpha/3, \beta/T)$ and independent tapes, obtaining $h_1, \dots, h_T$. Draw one integer K uniformly from $\{k+1, \dots, 2k\}$, independently of the samples and base tapes, and output

$$H_K(x) = 2\mathbf{1}\{N_x \geq K\} - 1, \qquad N_x = \sum_{j=1}^{T} \mathbf{1}\{h_j(x) = +1\}.$$

The integer K and corresponding base tapes are shared in two executions. Since k is a power of two, a uniform integer $J \in \{0, \dots, k-1\}$ uses exactly $\log_2 k$ fair bits, and $K = k + 1 + J$. The resulting hypothesis uses no further randomness at query time.

Fix any distribution and any point x. For two independent training executions and fixed counts N_x, N'_x, at most $|N_x - N'_x|$ possible values of K produce distinct predictions. The variance calculation in Lemma B.1 therefore gives

$$\mathbb{P}\{H_K(x) \neq H'_K(x)\} \leq \frac{\mathbb{E}|N_x - N'_x|}{k} \leq \frac{\sqrt{T/2}}{k} = \sqrt{\frac{3}{2k}} \leq \rho.$$

The replication bound holds on arbitrary domains, independently of any accuracy assumption.

Accuracy follows from a deterministic loss comparison. For any x, any binary label y, and any admissible K, if $H_K(x) \neq y$, then at least k of the base hypotheses also predict a label distinct from y. Indeed, for $y = -1$ there are at least $K \geq k+1$ positive votes; for $y = +1$ there are at least $T - (K-1) \geq k+1$ negative votes. Thus

$$\mathbf{1}\{H_K(x) \neq y\} \leq \frac{1}{k} \sum_{j=1}^{T} \mathbf{1}\{h_j(x) \neq y\}.$$

Integrating this inequality yields, for every K and every distribution,

$$\mathrm{R}_P(H_K) \leq \frac{3}{T} \sum_{j=1}^{T} \mathrm{R}_P(h_j).$$

On a realizable distribution, a union bound ensures that all base risks are at most $\alpha/3$ with probability at least $1 - \beta$. On this event, (4) is at most α simultaneously for every K. This proves accuracy without any concentration over domain points. The sample and randomness bounds follow from the construction. □

Proof of Corollary 3.2. Use a deterministic empirical risk minimizer, with a fixed selection on every sample, including inconsistent samples. The standard realizable VC bound (Blumer et al., 1989) is

$$m_A(a, b) = O\left(\frac{d \log(2/a) + \log(2/b)}{a}\right).$$

On a countable domain the measurability of this selection follows as in Corollary 2.3. Apply Theorem 3.1 and use $T = O(\rho^{-2})$. □

The dimension range of the realizable lower-bound comparison. For $d \geq 2$, the realizable lower-bound construction of Hopkins, Impagliazzo, and Ye (Hopkins et al., 2026, Theorem 3.10) uses one heavy shattered point and the remaining light shattered points, giving $\widetilde{\Omega}(d/(\rho^2\alpha))$ samples at fixed small failure probability. For $d = 2, 3$, a shattered pair suffices: fix the heavy point's label and assign mass 100α to the light point, whose label must be correct on the accuracy event since $100\alpha > \alpha$. The same reduction to endpoint bias estimation (Hopkins et al., 2026, proof of Theorem 3.10, using Theorem 3.12) gives the stated rate up to constants and logarithms for sufficiently small α, ρ. Thus Corollary 3.2 matches this rate up to logarithms in that dimension and parameter range. The dimension restriction is necessary for a per-class lower bound: the two constant classifiers on a singleton domain have VC dimension one, yet a realizable base learner needs only one example. Theorem 3.1 then gives $O(\rho^{-2})$ examples independently of α. Thus a universal α^{-1} lower bound fails for VC-one classes.

Proof of Proposition 3.3. Write $\mu(x) = \mathbb{E}[Y \mid X = x]$, choose a measurable Bayes classifier $f^\star(x) = \operatorname{sign}(\mu(x))$ with an arbitrary sign at zero, and put $\mathrm{R}_P^\star = \mathrm{R}_P(f^\star)$. For binary labels, a measurable version of μ exists as the Radon–Nikodym derivative of the signed label measure with respect to P_X. Every measurable binary h satisfies

$$\mathrm{R}_P(h) - \mathrm{R}_P^\star = \int |\mu(x)| \mathbf{1}\{h(x) \neq f^\star(x)\}\, dP_X(x). \tag{15}$$

The error-counting argument in Theorem 3.1 applies with $y = f^\star(x)$ for every point. Multiply it by $|\mu(x)|$ and integrate. By (15), for every choice of the base hypotheses and every admissible K,

$$\mathrm{R}_P(H_K) - \mathrm{R}_P^\star \leq \frac{3}{T} \sum_{j=1}^{T} \big(\mathrm{R}_P(h_j) - \mathrm{R}_P^\star\big).$$

A union bound gives $\mathrm{R}_P(h_j) \leq \mathrm{OPT}_{\mathcal{H}}(P) + \alpha/3$ simultaneously for all j with probability at least $1 - \beta$. Substitution proves (5). The replication argument is unchanged. □

E FINITE SHARED RANDOMNESS ON FINITE DOMAINS

On a finite domain, a short shared seed replaces the indexed rounding variables without changing the VC sample rate. We choose a power-of-two ensemble size so that the rounding probabilities are represented exactly, use bounded independence across domain points, and discretize the single screening threshold. The constructions require a fixed public indexing of the domain. Exact bounded independence gives a seed logarithmic in the domain size; almost independence reduces this dependence to a double logarithm at the cost of additional accuracy dependence. Both field constructions retain domain-size dependence; extending them directly to arbitrary countable domains gives no finite seed. All logarithms without a subscript are natural.

The bounded-independence moment method (Schmidt et al., 1995) supplies the concentration bound used in the construction.

Lemma E.1 (A bounded-independence risk bound)**.** *Let $(Z_i)_{i \in I}$ be a finite family of centered, $2r$-wise independent real random variables, where $r \geq 1$ is an integer. Here $2r$-wise independence includes every subfamily of size at most $2r$. Suppose Z_i has range length at most c_i and $\sum_{i \in I} c_i^2 \leq v$. If $v > 0$, then for every $s > 0$,*

$$\mathbb{P}\left\{\sum_{i \in I} Z_i > s\right\} \leq \left(\frac{rv}{s^2}\right)^r. \tag{16}$$

If $v = 0$, the sum is zero almost surely.

Proof. Let $(\widetilde{Z}_i)_{i \in I}$ be fully independent variables with the same respective marginal distributions. Every monomial in the expansion of $(\sum_i Z_i)^{2r}$ involves at most $2r$ distinct indices. Bounded independence therefore gives

$$\mathbb{E}\left|\sum_i Z_i\right|^{2r} = \mathbb{E}\left|\sum_i \widetilde{Z}_i\right|^{2r}.$$

Hoeffding's inequality for the independent variables gives $\mathbb{P}\{|\sum_i \widetilde{Z}_i| > u\} \leq 2e^{-2u^2/v}$ for $u > 0$. Integrating this tail, we obtain

$$\begin{aligned}\mathbb{E}\left|\sum_i \widetilde{Z}_i\right|^{2r} &\leq 4r \int_0^\infty u^{2r-1} e^{-2u^2/v}\, du \\ &= 2\, r! \left(\frac{v}{2}\right)^r \leq (rv)^r.\end{aligned}$$

Markov's inequality proves (16). □

Theorem E.2 (Finite-domain transformation)**.** *Let $\mathcal{X}$ be a publicly indexed domain of size $N \geq 1$, and let A be a binary agnostic learner using an integer number $m_A(a, \delta)$ of examples for parameters (a, δ). For $0 < \alpha, \beta, \rho \leq 1/4$, set*

$$r = \lceil \log(3/\beta) \rceil, \qquad k = 2r, \qquad \tau = \frac{\alpha}{6\sqrt{er}},$$

$$T = 2^{\lceil \log_2(8/\rho^2) \rceil}, \qquad K = 2^{\lceil \log_2(4/\rho) \rceil}, \qquad b = \max\{1, \lceil \log_2 N \rceil, \log_2 T\}.$$

There is an agnostic (α, β)-learner that is ρ-pointwise replicable and uses exactly

$$M_{\text{fin}} = T m_A\left(\frac{\alpha}{2}, \frac{\beta}{3T}\right) + M_0\left(\tau, \frac{\rho}{2}, \frac{\beta}{3}\right) + M_1\left(\tau, \frac{\rho}{2}, \frac{\beta}{3}\right) \tag{17}$$

examples. Here M_0, M_1 are the functions in Theorem A.1. In particular,

$$M_{\text{fin}} \leq T m_A\left(\frac{\alpha}{2}, \frac{\beta}{3T}\right) + O\left(\frac{r}{\rho^2\alpha^2} \log \frac{r}{\alpha\rho\beta}\right). \tag{18}$$

The transformation makes exactly T calls to A and uses exactly

$$kb + \log_2 K = O\left(\log \frac{1}{\beta}\left(\log(2N) + \log \frac{1}{\rho}\right)\right) \tag{19}$$

shared fair bits in addition to the independent shared tapes for these calls. If each call uses at most B_A random bits, the total number of shared random bits is at most $TB_A + kb + \log_2 K$.

Proof. **The finite seeds and the learner.** Fix a public representation of $\mathbb{F}_{2^b}$ and an injection $\iota : \mathcal{X} \to \mathbb{F}_{2^b}$. Such an injection exists because $2^b \geq N$. Put $t = \log_2 T$, and identify field elements with their b coordinates in a fixed $\mathbb{F}_2$-basis. Let $\pi : \mathbb{F}_{2^b} \to \{0, \ldots, T-1\}$ read any fixed t coordinates as an integer. A uniform field element has a uniform image under π, since each fiber has size 2^{b-t}.

Use kb independent shared fair bits to sample independent uniform coefficients $a_0, \ldots, a_{k-1} \in \mathbb{F}_{2^b}$, and define

$$f(z) = \sum_{j=0}^{k-1} a_j z^j, \qquad J_x = \pi(f(\iota(x))).$$

For any $s \leq \min(k, N)$ distinct domain points, the first s columns of their polynomial evaluation matrix form an invertible Vandermonde matrix. The evaluation vector is therefore uniform on $(\mathbb{F}_{2^b})^s$. Applying π coordinatewise shows that $(J_x)_{x \in \mathcal{X}}$ is a k-wise independent family of uniform ranks in $\{0, \ldots, T-1\}$. This argument also applies when $k > N$.

Independently, use $\log_2 K$ shared fair bits to sample $D \in \{0, \ldots, K-1\}$ uniformly, and set

$$W = \tau\left(1 + \frac{D + 1/2}{K}\right). \tag{20}$$

The two seeds and all base-learner tapes are mutually independent and independent of the samples. Corresponding executions share these seeds and tapes.

Use the sample partition and learner of Section C, with the parameters in this theorem. Run the screening rule with $(\tau, \rho/2, \beta/3)$, using the threshold (20). Run the T base learners with $(\alpha/2, \beta/(3T))$. For their outputs define

$$m_x = \sum_{j=1}^{T} \mathbf{1}\{h_j(x) = +1\}, \qquad q_x = m_x/T.$$

On screened points use the screen's label; on every other point output

$$H(x) = 2\mathbf{1}\{J_x < m_x\} - 1. \tag{21}$$

The polynomial seed is independent of screening and base training.

Pointwise replicability. The sample-dependent good events in the proof of Theorem A.1 do not depend on the screening threshold. Their label and residual-signal conclusions hold for every

$W \in [\tau, 2\tau]$, so both remain valid for (20). For screen parameter $\rho_s = \rho/2$, the proof uses $\eta = \min\{\beta/6, \rho/8\}$ and $\varepsilon = \rho\tau/8$. On the intersection of the two good events, disagreement at a fixed point requires W to lie in an interval of length at most 2ε. A grid of spacing τ/K places at most $K\ell/\tau + 1$ points in an interval of length ℓ. The screen disagreement probability is thus at most

$$2\eta + \frac{2\varepsilon}{\tau} + \frac{1}{K} \le \frac{\rho}{2} + \frac{1}{K} \le \frac{3\rho}{4}.$$

For the two base ensembles at this fixed point, a shared uniform rank satisfies, conditionally on m_x, m'_x,

$$\mathbb{P}\{\mathbf{1}\{J_x < m_x\} \neq \mathbf{1}\{J_x < m'_x\}\} = \frac{|m_x - m'_x|}{T} = |q_x - q'_x|.$$

There is no discretization error because both probabilities are multiples of $1/T$. The second-moment argument of Lemma B.1 therefore bounds the unconditional rounding disagreement by $1/\sqrt{2T} \le \rho/4$. A union bound with screen disagreement gives the claimed ρ-pointwise replicability, without conditioning this ensemble bound on the screening states.

Accuracy. Let F be the event that every base hypothesis has risk at most $\mathrm{OPT}_{\mathcal{H}}(P) + \alpha/2$. A union bound gives $\mathbb{P}(F^c) \le \beta/3$. Let G be the single-execution good event from the screening proof. It has failure probability at most $\beta/3$, and on G the screened labels are Bayes labels while every unselected point satisfies $p_x\mu_x^2 \le 9\tau^2$.

Condition on all samples, all base tapes, and D. The ranks remain $2r$-wise independent, with their original uniform marginals. Fix any such realization in $F \cap G$, and write B for its screened set. At each $x \notin B$, $\mathbb{P}\{J_x < m_x\} = q_x$. Linearity of expectation and Bayes optimality of the labels on B therefore give

$$\mathbb{E}_J \mathrm{R}_P(H) \le \frac{1}{T}\sum_{j=1}^{T} \mathrm{R}_P(h_j) \le \mathrm{OPT}_{\mathcal{H}}(P) + \alpha/2.$$

The centered risk is a sum over $x \notin B$ of the variables

$$Z_x = -p_x\mu_x\big(\mathbf{1}\{J_x < m_x\} - q_x\big).$$

They are centered and $2r$-wise independent, with range lengths $c_x = p_x|\mu_x|$ satisfying

$$\sum_{x\notin B} c_x^2 = \sum_{x\notin B} p_x(p_x\mu_x^2) \le 9\tau^2.$$

Lemma E.1, with $s = \alpha/2$, yields

$$\mathbb{P}_J\{\mathrm{R}_P(H) > \mathrm{OPT}_{\mathcal{H}}(P) + \alpha \mid \text{samples, base tapes, } D\} \le \left(\frac{36r\tau^2}{\alpha^2}\right)^r = e^{-r} \le \beta/3.$$

Averaging over the conditioning and adding the probabilities of F^c and G^c proves (α, β) accuracy.

Samples and random bits. The sample partition gives (17). Since $T = \Theta(\rho^{-2})$, $\tau^{-2} = 36er/\alpha^2$, and $\eta = \min\{\beta/6, \rho/8\}$, substitution into the formulas for M_0, M_1 gives (18), just as in the proof of Theorem 2.2. The polynomial coefficients use kb fair bits and the independent grid index uses $\log_2 K$ more. Finally, $k = O(\log(1/\beta))$, $b = O(\log(2N) + \log(1/\rho))$, and $\log_2 K = O(\log(1/\rho))$, proving (19). □

Corollary E.3 (VC classes with a finite shared seed)**.** *Every nonempty binary class of VC dimension $d \ge 1$ on a publicly indexed finite domain of size N, for $0 < \alpha, \beta, \rho \le 1/4$, admits an improper agnostic (α, β)-learner that is ρ-pointwise replicable, uses $\widetilde{O}(d/(\rho^2\alpha^2))$ examples, and uses a total of*

$$O\left(\log\frac{1}{\beta}\left(\log(2N) + \log\frac{1}{\rho}\right)\right)$$

shared fair bits. The suppressed factors in the sample bound are logarithmic in $1/\alpha, 1/\beta, 1/\rho$ and do not depend on N.

Proof. Use the deterministic empirical risk minimizer and the ordinary VC sample bound from the proof of Corollary 2.3 as the base learner in Theorem E.2. Its calls require no random bits. Substitution gives the stated sample and total randomness bounds. □

Evaluation from a seed. For $\mathcal{X} \subseteq \{0,1\}^\ell$, no enumeration of the domain is needed: take $b = \max\{1, \ell, \log_2 T\}$ and embed each input directly in the public field representation. The seed then uses $O((\ell + \log(1/\rho)) \log(1/\beta))$ bits. Each evaluation of J_x uses $O(k)$ field operations by Horner's rule, followed by the coordinate projection. The final prediction also evaluates the T base hypotheses and consults the finite screening output. After training, predictions at any number of points are determined by the stored learner outputs and seed, without generating a table of random values over the domain. This evaluation statement assumes the public encoding and field representation are available; the cost of learning and evaluating the base hypotheses remains that of the chosen base learner.

E.1 Almost independence and shorter seeds

The exact independence used above can be weakened because the absolute coefficients in the risk sum have total mass at most one. We use the small-bias construction of Alon et al. (1992), followed by the same linear polynomial-evaluation map as before. The construction and moment comparison follow below. For finite probability spaces, write $\mathrm{TV}(P, Q) = \frac{1}{2} \sum_z |P(z) - Q(z)|$.

Lemma E.4 (Ranks from a small-bias coefficient seed)**.** *Let $N, k, t \geq 1$ be integers, let $T = 2^t$, and set*

$$b = \max\{1, \lceil \log_2 N \rceil, t\}, \qquad m = kb.$$

For $0 < \delta < 1$, put

$$s = \left\lceil \log_2 \frac{m}{\delta} + kt \right\rceil.$$

On a publicly indexed domain of size N, exactly $2s$ shared fair bits suffice to generate ranks (J_x) in $\{0, \dots, T-1\}$ such that every subfamily of at most k ranks has total variation distance at most δ from independent uniform ranks.

Proof. Fix public binary-basis representations of $\mathbb{F}_{2^s}$ and $\mathbb{F}_{2^b}$. Use $2s$ fair bits to choose independent uniform $X, Y \in \mathbb{F}_{2^s}$, and form the m coefficient bits

$$C_i = \langle Y, X^i \rangle \quad (0 \leq i < m),$$

where the bracket is the binary coordinate inner product and $X^0 = 1$, including when $X = 0$. For any nonzero vector $u \in \{0,1\}^m$, its parity is

$$\sum_{i=0}^{m-1} u_i C_i = \left\langle Y, \sum_{i=0}^{m-1} u_i X^i \right\rangle \pmod 2.$$

The polynomial $P_u(z) = \sum_i u_i z^i$ is nonzero and has degree at most $m-1$. Conditional on X, the displayed parity is uniform unless $P_u(X) = 0$, in which case it is zero. The polynomial root bound therefore gives

$$\left| \mathbb{E}(-1)^{\sum_i u_i C_i} \right| \leq \frac{m-1}{2^s} =: \varepsilon_{\mathrm{sb}}. \tag{22}$$

This is the inner-product small-bias construction of Alon et al. (1992).

Group the m bits into k blocks of b bits, interpreted as $a_0, \dots, a_{k-1} \in \mathbb{F}_{2^b}$. Fix the public injection ι and the t-coordinate projection π from the proof of Theorem E.2, and set

$$J_x = \pi \left(\sum_{j=0}^{k-1} a_j \iota(x)^j \right).$$

For any $j \leq \min(k, N)$ distinct points, the map from coefficient bits to their jt rank bits is a surjective $\mathbb{F}_2$-linear map. Indeed, the first j polynomial coefficients give an invertible Vandermonde evaluation map onto j field elements; projecting each field element onto t coordinates is also surjective. Consequently, every nonzero parity of the jt output bits pulls back to a nonzero parity of the coefficient bits and has bias at most $\varepsilon_{\mathrm{sb}}$.

Let ν be their joint distribution, expressed on $\{0,1\}^{jt}$. Fourier inversion on this binary cube yields, for each z,

$$|\nu(z) - 2^{-jt}| \leq 2^{-jt}(2^{jt} - 1)\varepsilon_{\mathrm{sb}} \leq \varepsilon_{\mathrm{sb}}.$$

Summing over z gives

$$\mathrm{TV}(\nu, \mathrm{Unif}(\{0,1\}^{jt})) \le 2^{jt-1}\varepsilon_{\mathrm{sb}} \le 2^{kt-1}\frac{m-1}{2^s} \le \delta.$$

This proves the claim, including when $k > N$. □

Lemma E.5 (Weighted moments under almost independence). *Let $r, T \ge 1$ be integers, $v \ge 0$, and $0 \le \delta \le 1$. Let $(J_i)_{i\in I}$ be a finite family of ranks in $\{0,\dots,T-1\}$ whose restriction to any at most $2r$ indices has total variation distance at most δ from independent uniform ranks. Fix integers $0 \le m_i \le T$, put $q_i = m_i/T$, and let real coefficients a_i satisfy $\sum_i |a_i| \le 1$ and $\sum_i a_i^2 \le v$. For*

$$Z = \sum_{i\in I} a_i\big(\mathbf{1}\{J_i < m_i\} - q_i\big),$$

one has, for every $u > 0$,

$$\mathbb{E}|Z|^{2r} \le (rv)^r + 2\delta, \qquad \mathbb{P}\{Z > u\} \le \left(\frac{rv}{u^2}\right)^r + \frac{2\delta}{u^{2r}}. \tag{23}$$

The centering uses the ideal uniform marginals; Z need not have mean zero under the actual rank distribution.

Proof. Let $\widetilde{Z}$ use independent uniform ranks with the same fixed a_i, m_i. The proof of Lemma E.1 gives $\mathbb{E}|\widetilde{Z}|^{2r} \le (rv)^r$. Expand Z^{2r} as a sum over ordered tuples $(i_1,\dots,i_{2r}) \in I^{2r}$. Each product involves at most $2r$ distinct ranks. After factoring out $a_{i_1}\cdots a_{i_{2r}}$, its remaining factor lies in $[-1,1]$. The total variation hypothesis bounds the difference of its expectations under the actual and ideal rank laws by 2δ. With repeated indices, the product remains a bounded function of the distinct ranks. Thus

$$\begin{aligned}\left|\mathbb{E}Z^{2r} - \mathbb{E}\widetilde{Z}^{2r}\right| &\le 2\delta \sum_{(i_1,\dots,i_{2r})\in I^{2r}} |a_{i_1}\cdots a_{i_{2r}}| \\ &= 2\delta\left(\sum_i |a_i|\right)^{2r} \le 2\delta.\end{aligned}$$

Evenness of the power and Markov's inequality prove (23). The comparison is independent of $|I|$. □

Theorem E.6 (Finite-domain transformation with almost independence). *Let $\mathcal{X}$ be a publicly indexed domain of size $N \ge 1$, and let A be a binary agnostic learner with integer sample count $m_A(a,b)$. For $0 < \alpha, \beta, \rho \le 1/4$, set*

$$r = \lceil \log(6/\beta) \rceil, \qquad k = 2r, \qquad \tau = \frac{\alpha}{6\sqrt{er}},$$

$$T = 2^{\lceil \log_2(32/\rho^2)\rceil}, \qquad t = \log_2 T, \qquad K = 2^{\lceil \log_2(4/\rho)\rceil},$$

$$b = \max\{1, \lceil \log_2 N \rceil, t\}, \qquad m = kb,$$

$$\delta = \min\left\{\frac{\rho}{8}, \frac{\beta}{12}\left(\frac{\alpha}{2}\right)^{2r}\right\}, \qquad s = \left\lceil \log_2 \frac{m}{\delta} + kt \right\rceil.$$

There is an agnostic (α,β)-learner that is ρ-pointwise replicable and uses exactly

$$Tm_A\left(\frac{\alpha}{2}, \frac{\beta}{3T}\right) + M_0\left(\tau, \frac{\rho}{2}, \frac{\beta}{3}\right) + M_1\left(\tau, \frac{\rho}{2}, \frac{\beta}{3}\right) \tag{24}$$

examples and T calls to A. Its number of additional shared fair bits is exactly

$$2s + \log_2 K = O\left(\log\log(4N) + \log\frac{1}{\beta}\log\frac{1}{\alpha\rho}\right). \tag{25}$$

These are all its random bits when A is deterministic. If each base call uses at most B_A shared fair bits, the total is at most $TB_A + 2s + \log_2 K$.

Proof. **Construction and replication.** Use the screening grid (20), sample partition, and prediction rule (21) from Theorem E.2, with the parameters above. Replace its independent coefficient bits by the $2s$-bit seed of Lemma E.4. This seed, the screening seed, and all base tapes are mutually independent and independent of the samples; corresponding runs share each of them.

The screen's pointwise disagreement remains at most $3\rho/4$, by the same grid argument. At a fixed point, the marginal rank law has total variation distance at most δ from uniform. For fixed paired ensembles, the set of ranks that yield different rounded predictions is an interval of $|m_x - m'_x|$ integers. Consequently its probability is at most $|q_x - q'_x| + \delta$. Averaging and applying Lemma B.1 gives

$$\mathbb{P}\{\text{rounded predictions differ at } x\} \le \frac{1}{\sqrt{2T}} + \delta \le \frac{\rho}{4}.$$

A union bound with the screen proves pointwise replication on every distribution.

Accuracy. As before, the base accuracy event F and the single-execution screening good event G each fail with probability at most $\beta/3$. Condition on the samples, base tapes, and screening seed, at any realization in $F \cap G$. The ranks retain their almost-independence guarantee. Write B for the selected points and let R_0 be the mean risk obtained by replacing the unscreened ranks by independent uniform ranks. The selected labels are Bayes labels, so

$$R_0 \le \frac{1}{T}\sum_j \mathrm{R}_P(h_j) \le \mathrm{OPT}_{\mathcal{H}}(P) + \alpha/2.$$

The actual risk obeys the exact identity

$$\mathrm{R}_P(H) - R_0 = -\sum_{x \notin B} p_x \mu_x \big(\mathbf{1}\{J_x < m_x\} - q_x\big).$$

Its coefficients satisfy

$$\sum_{x \notin B} p_x |\mu_x| \le 1, \qquad \sum_{x \notin B} p_x^2 \mu_x^2 \le 9\tau^2.$$

Apply Lemma E.5 with $u = \alpha/2$:

$$\begin{aligned}&\mathbb{P}\{\mathrm{R}_P(H) > \mathrm{OPT}_{\mathcal{H}}(P) + \alpha \mid \text{samples, base tapes, screening seed}\}\\ &\le \left(\frac{36r\tau^2}{\alpha^2}\right)^r + \frac{2\delta}{(\alpha/2)^{2r}} \le \frac{\beta}{6} + \frac{\beta}{6} = \frac{\beta}{3}.\end{aligned}$$

Averaging and adding the probabilities of F^c, G^c proves accuracy. The use of R_0 avoids assuming exact centering under the compressed seed.

Resource bounds. The sample partition gives (24). Substitution in the screen sizes, with $r = \Theta(\log(1/\beta))$, again bounds the screening cost by

$$O\left(\frac{r}{\rho^2\alpha^2}\log\frac{r}{\alpha\rho\beta}\right).$$

For the seed, $t = O(\log(1/\rho))$ and

$$\log\frac{1}{\delta} = O\left(\log\frac{1}{\rho} + r\log\frac{1}{\alpha}\right), \qquad \log m = O(\log r + \log\log(4N) + \log(t+1)).$$

Together with $s = O(\log m + \log(1/\delta) + kt)$, these imply (25). The exact bit counts follow from the two uniform field elements and the independent screening grid index. □

Corollary E.7 (VC learning with two finite-seed alternatives)**.** *Every nonempty binary VC class of dimension $d \ge 1$ on a publicly indexed finite domain of size N, for $0 < \alpha, \beta, \rho \le 1/4$, has an improper agnostic (α, β)-learner that is ρ-pointwise replicable and uses $\widetilde{O}(d/(\rho^2\alpha^2))$ examples and*

$$O\left(\log\log(4N) + \log\frac{1}{\beta}\log\frac{1}{\alpha\rho}\right)$$

total shared fair bits. Both this bound and the bound of Corollary E.3 are available; one may choose the construction with the smaller seed, using only the public parameters. In particular, the achievable bit bound is the minimum of these two asymptotic bounds. The suppressed factors in the sample count are logarithmic in $1/\alpha, 1/\beta, 1/\rho$ and do not depend on N.

Proof. Apply Theorem E.6 to the deterministic empirical risk minimizer used in Corollary E.3. The ordinary VC sample bound and the displayed screening cost give the sample claim. A deterministic choice between the two constructions preserves both guarantees. □

Evaluation on encoded inputs. For $\mathcal{X} \subseteq \{0,1\}^\ell$, take $b = \max\{1, \ell, t\}$ and embed inputs directly in $\mathbb{F}_{2^b}$. The compressed seed bound becomes

$$O\left(\log(\ell+1) + \log\frac{1}{\beta}\log\frac{1}{\alpha\rho}\right).$$

Given the public field representations, all $m = kb$ coefficient bits can be generated by successive powers of the stored field element X and binary inner products with Y. This uses $O(m)$ field multiplications in $\mathbb{F}_{2^s}$ and $O(ms)$ binary operations. The resulting polynomial is evaluated by Horner's rule with $O(k)$ operations in $\mathbb{F}_{2^b}$ per input, or its coefficients can be regenerated from the seed when needed. These costs are polynomial in the input length and the displayed parameters, and require no listing of the input domain. As before, the final prediction also evaluates the base hypotheses and consults the screening output.

E.2 A LIMITATION OF OBLIVIOUS RANK FIELDS

For rank fields fixed independently of the samples and supplied ensemble, the next result gives matching domain-size dependence at fixed parameters. This is a lower bound for that rounding module. Risk-dependent threshold selection in Section 4 uses a different mechanism.

Proposition E.8 (Domain dependence for oblivious rounding)**.** *Let $T \geq 2$ be even. Suppose a publicly indexed domain of size N has a fixed rank field*

$$J : \mathcal{X} \times \{0,1\}^{B_0} \longrightarrow \{0, \ldots, T-1\}$$

with a uniform shared seed, satisfying for every x

$$\left|\mathbb{P}_r\{J_x(r) < T/2\} - \frac{1}{2}\right| \leq \varepsilon_0.$$

Fix $0 < v \leq 1$, $0 < u < 1/2$, $0 \leq \varepsilon_0 < 1/2$, and $0 \leq \beta_0 < 1/2 - \varepsilon_0$, and put $M = 2\lceil 1/(2v)\rceil$. Assume the unscreened rounding rule $H_r(x) = 2\mathbf{1}\{J_x(r) < T/2\} - 1$ satisfies

$$\mathbb{P}_r\{\mathrm{R}_P(H_r) > 1/2 + u\} \leq \beta_0$$

for every distribution with deterministic positive labels and atom masses at most v. Then

$$N \leq (M-1)2^{2^{B_0}}. \tag{26}$$

Thus, at fixed $v, u, \beta_0, \varepsilon_0$, such a universal rounding guarantee requires $B_0 = \Omega(\log\log N)$ as $N \to \infty$. On an infinite domain no finite B_0 suffices.

Proof. Associate to each point its binary signature

$$(\mathbf{1}\{J_x(r) < T/2\})_{r \in \{0,1\}^{B_0}}.$$

There are at most $2^{2^{B_0}}$ signatures. If (26) fails, the pigeonhole principle gives M points with the same signature. Put uniform feature mass on those points and label all of them $+1$. Since $M \geq 1/v$, each atom has mass at most v. For every seed, all rounded predictions on the support coincide, so the risk is either zero or one. Its probability of being one is at least $1/2 - \varepsilon_0 > \beta_0$, contradicting the assumed tail bound because $u < 1/2$. On an infinite domain, finitely many signatures likewise have a class containing at least M points. □

The vote fractions used in this obstruction can come from a supplied ensemble of exact class-risk minimizers. On the M points in the proof, let h predict $+1$ on exactly half and -1 on the other half, and take $\mathcal{H} = \{h, -h\}$, extending h arbitrarily elsewhere. This class has VC dimension one, and both functions have risk $1/2$. An ensemble with $T/2$ copies of each has $q_x = 1/2$ everywhere. With no selected points, it also satisfies the residual condition $p_x\mu_x^2 \leq v$. This example is about the supplied ensemble and the rounding module; it does not assert that a particular black-box base learner outputs that ensemble with appreciable probability.

F FINITE RANDOMNESS FOR AGNOSTIC LEARNING ON ARBITRARY DOMAINS

The agnostic construction in this section uses $\lceil \log_2(1/\rho) \rceil$ additional shared fair bits, independently of accuracy, confidence, and the domain, with sample bound $\widetilde{O}(d/(\rho^2\alpha^4))$, improving the $\widetilde{O}(d/(\rho^4\alpha^7))$ bound of Hopkins et al. (2026, Appendix A.1). It applies to general agnostic distributions, with no assumption on the approximation error above Bayes risk. For deterministic base learners, the exact minimum total bit count is $\lfloor \log_2(1/\rho) \rfloor$ when any finite sample size is allowed. At this minimum, the additional sample cost near a dyadic boundary is logarithmic in the inverse distance to that boundary. A matching lower bound establishes this logarithmic order when the other parameters and the class are fixed. The remaining sample-complexity gap is a factor α^{-2} relative to the countable-domain optimum.

We use all integer thresholds of one ensemble as candidates. Their average risk equals the average base risk. A candidate with empirical risk r_j receives weight proportional to $(c + a - r_j)_+$, where c is the average empirical risk and a is a small positive slack. These weights have sufficient total mass, have support only on candidates with empirical risk below $c + a$, and change by at most the average change in the risk curve, up to a constant. One shared grid point selects a candidate through the inverse CDF of the normalized weights. Nesting turns this selection rule into a pointwise coupling: each prediction is a comparison between that same grid point and one weighted prefix sum. This score has bounded differences in the independent sample blocks and validation observations. Concentration gives exponentially small disagreement at all but one grid midpoint. In contrast, the correlated-sampling selection argument of Hopkins et al. (2026, Appendix A.2), using Ghazi et al. (2021), controls agreement of the selected indices.

Throughout this section, the domain is a measurable space and the base learner has a jointly measurable evaluation map in its sample, random tape, and test point. Its output classifiers are measurable. These assumptions ensure measurability of the finite constructions and justify the integrals below. Corresponding base calls share their random tapes, and the tapes for different calls are independent.

Lemma F.1 (Weighted-CDF perturbation for nested predictions)**.** *Let $T \geq 1$ be an integer, and let $w, w' \in [0,1]^T$ have averages $z, z' \geq z_0 > 0$. For integers $0 \leq m \leq T$, define*

$$F_w(m) = \frac{\sum_{j=1}^{m} w_j}{Tz}, \qquad F_{w'}(m) = \frac{\sum_{j=1}^{m} w'_j}{Tz'}, \qquad \Delta_w = \frac{1}{T}\sum_{j=1}^{T} |w_j - w'_j|.$$

For every $m, m' \in \{0, \dots, T\}$,

$$|F_w(m) - F_{w'}(m')| \leq \frac{\Delta_w + |m - m'|/T}{z_0}. \tag{27}$$

For $u \in (0,1)$, set $J_w(u) = \min\{j : F_w(j) \geq u\}$. This index exists, has $w_{J_w(u)} > 0$, and satisfies

$$\mathbf{1}\{J_w(u) \leq m\} = \mathbf{1}\{u \leq F_w(m)\}. \tag{28}$$

Proof. Normalize the weights to probability vectors $p_j = w_j/(Tz)$ and $p'_j = w'_j/(Tz')$. Since $|z - z'| \leq \Delta_w$, the triangle inequality gives

$$\frac{1}{2}\sum_j |p_j - p'_j| \leq \frac{\Delta_w + |z - z'|}{2z} \leq \frac{\Delta_w}{z_0}.$$

The difference of probabilities of any prefix is at most this total variation distance. Moving the endpoint of a prefix from m to m' changes $F_{w'}$ by at most $|m - m'|/(Tz_0)$, because $w'_j \leq 1$. These bounds prove (27).

Since $F_w(T) = 1$, the index $J_w(u)$ exists. If its weight were zero and $J_w(u) > 1$, then $F_w(J_w(u) - 1) = F_w(J_w(u)) \geq u$, contradicting minimality. If $J_w(u) = 1$ and $w_1 = 0$, then $F_w(1) = 0 < u$, also a contradiction. Monotonicity of F_w gives (28), including at CDF boundaries and at $m = 0, T$. □

Lemma F.2 (Midpoint-grid coupling for concentrated independent scores)**.** *Let $K \geq 1$ be an integer. Let V, V' be independent and identically distributed random variables in $[0,1]$, with $\mu = \mathbb{E}V$. Suppose that for some $\sigma^2 > 0$ and every $t > 0$,*

$$\mathbb{P}\{V - \mu \geq t\} \leq e^{-2t^2/\sigma^2}, \qquad \mathbb{P}\{V - \mu \leq -t\} \leq e^{-2t^2/\sigma^2}.$$

If U is independent of V, V' and uniform on the K midpoints $\{(d+1/2)/K : 0 \le d < K\}$, then

$$\mathbb{P}\{\mathbf{1}\{U \le V\} \ne \mathbf{1}\{U \le V'\}\} \le \frac{1}{2K} + 2\exp\left(-\frac{1}{2K^2\sigma^2}\right). \tag{29}$$

The same assertion holds conditionally when its independence and tail assumptions hold for each conditional law.

Proof. Choose one grid midpoint u_* closest to μ, breaking ties arbitrarily. Every other midpoint u has $|u-\mu| \ge 1/(2K)$. Indeed, consecutive midpoints have spacing $1/K$; this remains true when μ lies outside their range. For a midpoint u, put $p_u = \mathbb{P}\{V \ge u\}$. Independence gives disagreement probability $2p_u(1-p_u) \le 1/2$ at u_*. At every other midpoint, the upper tail bounds p_u when $u > \mu$, and the lower tail bounds $1-p_u$ when $u < \mu$. Consequently,

$$2p_u(1-p_u) \le 2\min\{p_u, 1-p_u\} \le 2\exp\left(-\frac{1}{2K^2\sigma^2}\right).$$

Averaging over the K midpoints proves (29). For $K = 1$, there are no other midpoints and the bound follows directly from $2p_u(1-p_u) \le 1/2$. The conditional assertion follows by the same argument for each conditional law. □

Theorem F.3 (An agnostic transformation with a finite seed)**.** *Let A be an agnostic learner using an integer number $m_A(u,v)$ of examples for parameters (u,v) and satisfying the preceding measurability assumptions. For $0 < \alpha, \beta, \rho \le 1/4$, let $B \ge 0$ be an integer such that*

$$\Delta_B = \rho - 2^{-B-1} > 0.$$

Define

$$a = 2^{\lfloor \log_2(\alpha/4) \rfloor}, \qquad K = 2^B, \qquad L_B = \log\frac{4}{\Delta_B},$$

$$T = 2^{\left\lceil \log_2 \max\left\{ \frac{40K^2 L_B}{a^2}, \frac{8}{\alpha^2}\log\frac{2}{\beta} \right\} \right\rceil}, \qquad n = \left\lceil \max\left\{ \frac{40K^2 L_B}{a^2}, \frac{32}{\alpha^2}\log\frac{4T}{\beta} \right\} \right\rceil.$$

There is an agnostic (α,β)-learner that is ρ-pointwise replicable on every distribution and uses exactly

$$M_{\mathrm{arb}} = T m_A\left(\frac{\alpha}{8}, \frac{\alpha}{8}\right) + n \tag{30}$$

examples and exactly T calls to A. It uses exactly

$$B = \log_2 K \tag{31}$$

additional shared fair bits. If each base call uses at most B_A shared fair bits, the total number is at most $TB_A + B$. For a deterministic base learner, the additional B bits are the total randomness.

Proof. **The learner and the weights.** On T disjoint sample blocks, run A with parameters $(\alpha/8, \alpha/8)$ and independent tapes, obtaining $h_1, \dots, h_T$. Set

$$N_x = \sum_{t=1}^{T} \mathbf{1}\{h_t(x) = +1\}, \qquad q_x = N_x/T, \qquad g_j(x) = 2\mathbf{1}\{N_x \ge j\} - 1 \quad (1 \le j \le T).$$

Use an independent validation sample of size n to compute the empirical risks r_j of g_j, and put

$$c = \frac{1}{T}\sum_{j=1}^{T} r_j, \qquad w_j = \frac{1}{2}(c + a - r_j)_+, \qquad z = \frac{1}{T}\sum_{j=1}^{T} w_j. \tag{32}$$

Since $0 \le c, r_j \le 1$ and $a \le 1/16$, we have $0 \le w_j < 1$. The inequality $u_+ \ge u$ gives, for every possible sample,

$$z \ge \frac{1}{2T}\sum_{j}(c + a - r_j) = \frac{a}{2}. \tag{33}$$

Define the weighted CDF F_w as in Lemma F.1. Independently of all samples and base tapes, draw $D \in \{0, \dots, K-1\}$ uniformly using B shared fair bits, put $U = (D + 1/2)/K$, and return

$$H = g_{J_w(U)}, \qquad J_w(U) = \min\{j : F_w(j) \geq U\}.$$

Corresponding executions share D and their corresponding base tapes. For every possible sample and every seed, the selected index has positive weight by Lemma F.1. All comparisons are between rational numbers: a is dyadic, the empirical risks have denominator n, c has denominator dividing nT, and $z > 0$. Thus the CDF comparisons can be performed exactly. The finite sums, positive parts, and selection among finitely many indices preserve measurability. In particular, the final evaluation map is jointly measurable under the base learner assumptions. After training, the stored ensemble and selected index determine all predictions.

Accuracy. For every (x, y), threshold counting gives the exact identity

$$\frac{1}{T}\sum_{j=1}^{T} \mathbf{1}\{g_j(x) \neq y\} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{h_t(x) \neq y\}. \tag{34}$$

Write $R_j = \mathrm{R}_P(g_j)$ and $\bar{R} = T^{-1}\sum_j R_j$. For each base call, the risk lies in $[0, 1]$, so its accuracy and failure parameters imply

$$\mathbb{E}\,\mathrm{R}_P(h_t) \leq \mathrm{OPT}_{\mathcal{H}}(P) + \frac{\alpha}{8} + \frac{\alpha}{8} = \mathrm{OPT}_{\mathcal{H}}(P) + \frac{\alpha}{4}.$$

The base risks are independent over their sample blocks and tapes. By (34) and Hoeffding's inequality,

$$\mathbb{P}\left\{\bar{R} > \mathrm{OPT}_{\mathcal{H}}(P) + \frac{\alpha}{2}\right\} \leq \exp(-T\alpha^2/8) \leq \frac{\beta}{2}. \tag{35}$$

Thus the average risk has the required high-probability accuracy without requiring every base hypothesis to be accurate. Conditionally on the ensemble, Hoeffding's inequality and a union bound give

$$\mathbb{P}\left\{\max_j |r_j - R_j| > \frac{\alpha}{8} \,\middle|\, h_1, \dots, h_T\right\} \leq 2T\exp(-n\alpha^2/32) \leq \frac{\beta}{2}.$$

On the corresponding validation good event, also $|c - \bar{R}| \leq \alpha/8$. Every selected index $j = J_w(U)$ has $w_j > 0$, whence $r_j < c + a$. Consequently, on the two preceding good events,

$$\mathrm{R}_P(g_j) \leq c + a + \frac{\alpha}{8} \leq \bar{R} + a + \frac{\alpha}{4} \leq \mathrm{OPT}_{\mathcal{H}}(P) + \alpha.$$

This guarantee holds for every grid point U simultaneously. The two failure probabilities sum to β, proving accuracy.

Bounded differences of the prediction score. Fix a test point x. For any two empirical-risk curves, use primes for the second curve and its associated weights. Since the positive-part map is 1-Lipschitz and $|c - c'| \leq T^{-1}\sum_j |r_j - r'_j|$, the weights satisfy

$$\Delta_w := \frac{1}{T}\sum_j |w_j - w'_j| \leq \frac{1}{T}\sum_j |r_j - r'_j|. \tag{36}$$

For an ensemble and validation sample, put $V_x = F_w(N_x)$. By (33) and Lemma F.1, any two such realizations satisfy

$$|V_x - V'_x| \leq \frac{2}{a}\big(\Delta_w + |q_x - q'_x|\big). \tag{37}$$

Fix all base tapes and replace just one base sample block, holding the other blocks and the validation sample fixed. Only one base hypothesis can change, so $|N_y - N'_y| \leq 1$ at every point y. Threshold counting gives

$$\frac{1}{T}\sum_{j=1}^{T} \mathbf{1}\{g_j(y) \neq g'_j(y)\} = \frac{|N_y - N'_y|}{T} \leq \frac{1}{T}.$$

Average this inequality over the fixed validation sample and use the triangle inequality for its empirical risks. It follows that $T^{-1}\sum_j |r_j - r'_j| \le 1/T$, so (36) gives $\Delta_w \le 1/T$. Also $|q_x - q'_x| \le 1/T$. Thus this block replacement changes V_x by at most $4/(aT)$.

Instead replace just one validation observation, keeping the ensemble and all other observations fixed. Each r_j changes by at most $1/n$, whence $\Delta_w \le 1/n$, while N_x does not change. The score changes by at most $2/(an)$. These bounds hold for every input realization and fixed choice of base tapes, including realizations on which accuracy fails.

Pointwise replicability. Nesting gives the exact prediction identity for two executions:

$$H(x) = 2\mathbf{1}\{U \le V_x\} - 1, \qquad H'(x) = 2\mathbf{1}\{U \le V'_x\} - 1, \qquad V'_x = F_{w'}(N'_x).$$

Condition only on the shared base tapes $R = (R_1, \dots, R_T)$. The complete sample collections in the two executions, including their validation blocks, remain independent and identically distributed. Their scores V_x, V'_x are therefore conditionally independent and identically distributed. The shared grid point U is independent of these samples and tapes. Put

$$\mu_R = \mathbb{E}[V_x \mid R], \qquad \sigma^2 = \frac{16}{a^2 T} + \frac{4}{a^2 n}.$$

Conditionally on R, the score is a function of T independent base sample blocks and n independent validation observations. The preceding coordinate bounds have sum of squares σ^2. The bounded-differences inequality (McDiarmid, 1989) therefore gives, for every $t > 0$,

$$\mathbb{P}\{V_x - \mu_R \ge t \mid R\} \le e^{-2t^2/\sigma^2}, \qquad \mathbb{P}\{V_x - \mu_R \le -t \mid R\} \le e^{-2t^2/\sigma^2}.$$

The conditional version of Lemma F.2 implies

$$\mathbb{P}\{H(x) \ne H'(x) \mid R\} \le \frac{1}{2K} + 2\exp\left(-\frac{1}{2K^2\sigma^2}\right). \tag{38}$$

Our choices of T, n imply $\sigma^2 \le 1/(2K^2 L_B)$. Averaging over R, the pointwise disagreement probability is therefore at most

$$\frac{1}{2K} + 2e^{-L_B} = 2^{-B-1} + \frac{\Delta_B}{2} < \rho.$$

This proof holds on every distribution. It conditions only on the shared base tapes, and never on an accuracy event.

Finally, $\alpha/8 < a \le \alpha/4$ implies

$$\begin{aligned} T &= O\left(\frac{1}{\alpha^2}\left(K^2 L_B + \log\frac{1}{\beta}\right)\right), \\ n &= O\left(\frac{1}{\alpha^2}\left(K^2 L_B + \log\frac{T}{\beta}\right)\right). \end{aligned} \tag{39}$$

The exact sample count follows from the sample partition, and the only additional random choice is D, using exactly $\log_2 K = B$ shared fair bits. □

Corollary F.4 (VC classes on arbitrary measurable domains)**.** *Let $\mathcal{H}$ have VC dimension $d \ge 1$ and admit a jointly measurable agnostic learner with the standard VC sample bound*

$$m_A(u, v) = O\left(\frac{d\log(2/u) + \log(2/v)}{u^2}\right).$$

For $0 < \alpha, \beta, \rho \le 1/4$ and every integer $B \ge 0$ such that $\Delta_B = \rho - 2^{-B-1} > 0$, the class has an improper agnostic ρ-pointwise replicable (α, β)-learner using

$$O\left(\frac{d\log(2/\alpha)}{\alpha^4}\left(4^B \log\frac{4}{\Delta_B} + \log\frac{2}{\beta}\right)\right) \tag{40}$$

examples and exactly B additional shared fair bits. In particular, choosing

$$B = \left\lceil \log_2 \frac{1}{\rho} \right\rceil \tag{41}$$

gives the uniform sample bound $\widetilde{O}(d/(\rho^2\alpha^4))$, where the suppressed factors are logarithmic in $1/\alpha, 1/\beta, 1/\rho$.

If the ordinary base learner is deterministic, these are total bit counts, and the exact minimum total number of shared fair bits among learners using any finite, fixed sample size is

$$B_{\min} = \left\lfloor \log_2 \frac{1}{\rho} \right\rfloor . \tag{42}$$

Thus (41) uses at most one bit above the exact minimum. Every countable-domain VC class admits the required deterministic ordinary learner. The uniform finite-seed sample bound remains above the countable-domain sample optimum by a factor α^{-2}, up to logarithms.

Proof. Substitute the stated base sample bound into (30). With $G = K^2 L_B + \log(2/\beta)$, the base calls use at most $O(d\alpha^{-4}\log(2/\alpha)G)$ examples. Also $T = O(\alpha^{-2}G)$ and

$$\log(T/\beta) = O(\log(2/\alpha) + G) .$$

The validation count in (39) is therefore absorbed by the same bound, since $d \geq 1$, $G \geq 1$, and $\alpha \leq 1/4$. This proves (40). For countable domains, deterministic empirical risk minimization has the required measurability, as in Corollary 2.3. For the choice (41), $K < 2/\rho$ and $\Delta_B \geq \rho/2$, proving the uniform sample bound.

An agnostic learner is accurate on realizable distributions. Proposition 3.4 therefore rules out zero total bits and requires $\rho > 2^{-B-1}$ for any admissible positive total bit count B. For integers B, this condition is equivalent to $B \geq \lfloor \log_2(1/\rho) \rfloor$. Conversely, put $b = \lfloor \log_2(1/\rho) \rfloor$. Then

$$2^{-b-1} < \rho \leq 2^{-b}, \qquad \Delta_b = \rho - 2^{-b-1} > 0.$$

The transformation with $B = b$ and a deterministic ordinary learner has a finite sample size and exactly b total bits, proving (42). □

At $B = B_{\min}$, one has $4^B = \Theta(\rho^{-2})$, but $\log(4/\Delta_B)$ can grow without bound while ρ approaches the fixed dyadic endpoint 2^{-B-1} from above. The lower bound below shows that at the exact bit minimum no sample bound with only polynomial and logarithmic factors in $1/\rho, 1/\alpha, 1/\beta$ can hold uniformly across these dyadic intervals. For a prescribed budget larger than $\lceil \log_2(1/\rho) \rceil$, one may run the construction at that smaller budget and pad its seed with ignored bits; the factor 4^B in (40) need not be paid for an unnecessarily large budget.

F.1 Proof of the total shared-bit lower bound

Proof of Proposition 3.4. Let m be the learner's fixed, finite sample size. Pad any unused bits so that its seed is uniform on $\{0,1\}^B$, and put $K = 2^B$. There is no other internal randomness, so the learner is deterministic after fixing its seed. Choose a point x shattered by $\mathcal{H}$. Both singleton distributions supported on $(x, +1)$ and $(x, -1)$ are realizable, and their training samples are deterministic. Let C_+ and C_- be the sets of seeds producing the respective correct predictions at x on those samples. Since $\alpha < 1$, an incorrect prediction has risk exceeding the accuracy tolerance. Thus

$$|C_+|, |C_-| \geq (1-\beta)K > K/2.$$

In particular, $C_+ \cap C_-$ is nonempty. These accuracy requirements also rule out $m = 0$: the two samples would then be identical, so no seed could lie in the intersection.

For $p \in [0,1]$, let P_p put all feature mass on x and give label $+1$ with probability p. For each fixed seed r, define

$$g_r(p) = \mathbb{P}_{S \sim P_p^m}\{A(S;r)(x) = +1\}.$$

This is a polynomial in p, obtained by summing the probabilities of the finitely many label strings that produce a positive prediction. For a seed $r_1 \in C_+ \cap C_-$, we have $g_{r_1}(0) = 0$ and $g_{r_1}(1) = 1$. Continuity gives $p_0 \in (0,1)$ with $g_{r_1}(p_0) = 1/2$. On two independent samples from P_{p_0}, conditional on r_1, the predictions disagree with probability $2g_{r_1}(p_0)(1-g_{r_1}(p_0)) = 1/2$. Averaging over all seeds and using replication on P_{p_0} therefore gives

$$\rho \geq \frac{1}{2K} = 2^{-B-1}.$$

If $B \geq 1$, then K is even and the integer sizes of C_+ and C_- are both at least $K/2+1$. Consequently, their intersection contains at least two seeds. Choose $r_2 \neq r_1$ in that intersection. Its predictions on the all-positive and all-negative label strings imply

$$0 < p_0^m \leq g_{r_2}(p_0) \leq 1 - (1-p_0)^m < 1.$$

Hence its conditional disagreement probability is strictly positive. Including both seeds in the average gives

$$\rho \geq \frac{1}{K}\left(\frac{1}{2} + 2g_{r_2}(p_0)(1 - g_{r_2}(p_0))\right) > 2^{-B-1}.$$

For $B = 0$, the earlier bound gives $\rho \geq 1/2$. In the parameter range $\rho \leq 1/4$, zero bits are impossible and the strict inequality for positive B implies $B \geq \lfloor \log_2(1/\rho) \rfloor$. The intermediate P_{p_0} need not be realizable; the proof uses the requirement of replication on every distribution. □

Proposition F.5 (A sample lower bound near the bit boundary)**.** *Let $\mathcal{H}$ have VC dimension at least one, and let $0 < \alpha, \beta \leq 1/4$. Suppose an agnostic (α, β)-learner uses a fixed, finite number m of examples and at most $B \geq 2$ shared fair bits, with no other randomness, and is ρ-pointwise replicable on every distribution. Put $K = 2^B$ and $\Delta_B = \rho - 1/(2K)$. If $0 < K\Delta_B < 1/2$, then*

$$\Delta_B \geq \frac{1}{2K} \exp\left(-\frac{2048 m \alpha^2}{3K^2}\right). \tag{43}$$

Consequently,

$$m \geq \frac{3K^2}{2048\alpha^2} \log \frac{1}{2K\Delta_B}. \tag{44}$$

This lower bound concerns all the learner's randomness, rather than the additional bits of a transformation. The interval $1/(2K) < \rho < 1/K$ lies within the range where B is the minimum feasible total bit budget.

Proof. Pad unused bits so that the seed is uniform on $\{0,1\}^B$. Choose a shattered point x and let P_p and $g_r(p)$ be as in the preceding proof. As there, accuracy on the two deterministic label distributions rules out $m = 0$, and each g_r is a continuous polynomial. Put $p_- = 1/2 - \alpha$ and $p_+ = 1/2 + \alpha$. On P_{p_-}, a positive prediction at x has excess risk $2\alpha > \alpha$; on P_{p_+}, a negative prediction has this excess risk. Both class labels at x are available because x is shattered. Agnostic accuracy therefore gives

$$\frac{1}{K}\sum_r g_r(p_-) \leq \beta, \qquad \frac{1}{K}\sum_r (1 - g_r(p_+)) \leq \beta.$$

Independence of the two samples, conditional on a seed, gives

$$\sum_r 2g_r(p)(1 - g_r(p)) \leq K\rho \quad \text{for every } p \in [0,1]. \tag{45}$$

Define

$$C_- = \{r : g_r(p_-) \leq 1/2\}, \qquad C_+ = \{r : g_r(p_+) \geq 1/2\}.$$

The inequality $\mathbf{1}\{u > 1/2\} \leq u + 2u(1-u)$ for $u \in [0,1]$, accuracy, and (45) imply $|C_-^c| \leq K(\beta + \rho)$. Applying the same inequality to $1 - g_r(p_+)$ gives $|C_+^c| \leq K(\beta + \rho)$. Hence, with $L = |C_- \cap C_+|$,

$$L \geq K(1 - 2\beta - 2\rho) > K/2 - 2,$$

where the strict inequality uses $K\rho = 1/2 + K\Delta_B < 1$. Since L is an integer, $L \geq K/2 - 1$. For $K \geq 8$ this gives $L - 1 \geq K/4$. For $K = 4$, accuracy alone gives $|C_-|, |C_+| \geq 3$: two seeds outside either set would contribute strictly more than 1 to its corresponding error sum, which is at most $\beta K \leq 1$. Thus $L \geq 2$ and $L - 1 \geq K/4$ also in this case.

For each $r \in C_- \cap C_+$, choose $p_r \in [p_-, p_+]$ with $g_r(p_r) = 1/2$, using continuity. Ordering these L locations in an interval of length 2α gives two distinct seeds r_1, r_2 with locations p_1, p_2 satisfying

$$|p_1 - p_2| \leq \frac{2\alpha}{L-1} \leq \frac{8\alpha}{K}.$$

Let E be the event that seed r_2 outputs $+1$, and put $q = g_{r_2}(p_1)$. Under $P_{p_2}^m$, the event has probability $1/2$. Grouping sample probabilities on E and E^c and applying the log-sum inequality gives

$$\begin{aligned} mD_{\mathrm{KL}}(\mathrm{Ber}(p_2)\|\,\mathrm{Ber}(p_1)) &\geq D_{\mathrm{KL}}(\mathrm{Ber}(1/2)\|\,\mathrm{Ber}(q)) \\ &= -\frac{1}{2}\log\big(4q(1-q)\big). \end{aligned}$$

All these sample laws have full support, so $0 < q < 1$. The inequality $\log u \leq u - 1$ implies, for $s, t \in (0, 1)$,

$$D_{\mathrm{KL}}(\mathrm{Ber}(s)\|\,\mathrm{Ber}(t)) \leq \frac{(s-t)^2}{t(1-t)}.$$

Since $p_1 \in [1/4, 3/4]$, the second seed's disagreement probability at p_1 is therefore at least

$$2q(1-q) \geq \frac{1}{2}\exp\left(-\frac{32m}{3}(p_1 - p_2)^2\right) \geq \frac{1}{2}\exp\left(-\frac{2048m\alpha^2}{3K^2}\right).$$

The first seed contributes disagreement $1/2$ at p_1. Averaging these two contributions over all K seeds proves (43); rearrangement gives (44). The argument covers arbitrary sample-dependent computations, whose output probabilities are encoded in $g_r(p)$. □

Corollary F.6 (Logarithmic sample divergence at fixed parameters)**.** *Fix $0 < \alpha, \beta \leq 1/4$, an integer $B \geq 2$, and a class $\mathcal{H}$ of VC dimension at least one admitting a deterministic agnostic learner with finite sample bounds and the measurability assumptions of Theorem F.3. Let $m^*_{\mathcal{H}}(\alpha, \beta, \rho; B)$ be the minimum fixed sample size of an agnostic (α, β)-learner that is ρ-pointwise replicable on every distribution and uses at most B shared fair bits with no other randomness. As $\Delta \downarrow 0$,*

$$m^*_{\mathcal{H}}(\alpha, \beta, 2^{-B-1} + \Delta; B) = \Theta_{\mathcal{H},\alpha,\beta,B}\left(\log\frac{1}{\Delta}\right). \tag{46}$$

For sufficiently small Δ, B is the exact minimum bit budget. The assertion concerns dependence on Δ with the other displayed parameters fixed; it does not characterize their joint optimal sample tradeoff.

Proof. The lower bound is Proposition F.5, since $K = 2^B$ and α are fixed, and $0 < K\Delta < 1/2$ holds for all sufficiently small Δ. For the upper bound, the deterministic base learner uses the fixed number $m_A(\alpha/8, \alpha/8)$ of examples per call. Theorem F.3 has $T = O(\log(1/\Delta))$ and $n = O(\log(1/\Delta) + \log\log(1/\Delta)) = O(\log(1/\Delta))$. Its exact total bit count is B, proving the upper bound. If $0 < \Delta \leq 2^{-B-1}$, then $2^{-B-1} < \rho \leq 2^{-B}$, so Corollary F.4's bit calculation, or directly Proposition 3.4 and this upper construction, gives $B_{\min} = B$. □

Proposition F.7 (A matching joint tradeoff on one point)**.** *Let $\mathcal{X} = \{x\}$ and let $\mathcal{H}$ contain both binary functions on $\mathcal{X}$. For $0 < \alpha, \beta \leq 1/4$, an integer $B \geq 2$, $K = 2^B$, and $0 < \Delta \leq 1/(2K)$, there is an agnostic (α, β)-learner that is $(1/(2K)+\Delta)$-pointwise replicable, uses exactly B shared fair bits with no other randomness, and uses*

$$m = \left\lceil \frac{8}{\alpha^2}\max\left\{K^2\log\frac{8}{K\Delta}, \log\frac{1}{\beta}\right\}\right\rceil \tag{47}$$

examples. Consequently, uniformly for $0 < K\Delta \leq 1/4$,

$$m^*_{\mathcal{H}}(\alpha, \beta, 1/(2K) + \Delta; B) = \Theta\left(\frac{1}{\alpha^2}\left[K^2\log\frac{1}{K\Delta} + \log\frac{1}{\beta}\right]\right), \tag{48}$$

with absolute implicit constants. This joint characterization is for the singleton class; it does not give a matching joint sample bound for general VC classes.

Proof. Let $\widehat{p}$ be the empirical fraction of positive labels. Use the B shared bits to choose J uniformly from $\{0, \ldots, K-1\}$, and output $+1$ exactly when $\widehat{p} \geq u_J$, where

$$u_j = \frac{1}{2} - \frac{\alpha}{4} + \left(j + \frac{1}{2}\right)\frac{\alpha}{2K}.$$

Under P_p, either prediction has excess risk at most α if $|p-1/2| \le \alpha/2$. Otherwise every threshold is at distance at least $\alpha/4$ from p on its correct side. Hoeffding's inequality thus bounds the accuracy failure probability by $e^{-m\alpha^2/8} \le \beta$.

For replicability, choose one grid point nearest to p, breaking ties arbitrarily. Its conditional disagreement probability is at most $1/2$. Put $h = \alpha/(2K)$ and $t = m\alpha^2/(8K^2)$. Each other grid point at index distance $\ell \ge 1$ from the nearest one is at distance at least $(\ell - 1/2)h$ from p, and there are at most two such points for each ℓ. This also holds when p is outside the grid's range. Writing $q_j = \mathbb{P}\{\widehat{p} \ge u_j\}$, Hoeffding's inequality gives

$$2q_j(1-q_j) \le 2\min\{q_j, 1-q_j\} \le 2e^{-2m(u_j-p)^2}$$

at these other grid points. Since the samples are independent conditional on J, averaging yields

$$\begin{aligned}\mathbb{P}\{H(x) \ne H'(x)\} &\le \frac{1}{2K} + \frac{4}{K}\sum_{\ell \ge 1} e^{-t(2\ell-1)^2} \\ &\le \frac{1}{2K} + \frac{4}{K}\frac{e^{-t}}{1-e^{-8t}}. \end{aligned} \tag{49}$$

The second inequality uses $(2\ell-1)^2 \ge 1 + 8(\ell-1)$. Our sample choice gives $t \ge \log(8/(K\Delta)) \ge \log 16$, so $1 - e^{-8t} \ge 1/2$. The last term in (49) is at most $8e^{-t}/K \le \Delta$, proving the upper bound.

For the confidence contribution to the lower bound, consider $p_- = 1/2 - \alpha$ and $p_+ = 1/2 + \alpha$ again. Include the independent uniform seed in the laws of the complete inputs to an arbitrary learner. Their relative entropy remains $mD_{\mathrm{KL}}(\mathrm{Ber}(p_+) \| \mathrm{Ber}(p_-))$. The event of outputting $+1$ has probability at least $1-\beta$ under p_+ and at most β under p_-. Grouping on this event, as in the preceding lower bound, and using monotonicity of binary relative entropy on these probability ranges gives

$$\begin{aligned}\frac{64m\alpha^2}{3} &\ge mD_{\mathrm{KL}}(\mathrm{Ber}(p_+) \| \mathrm{Ber}(p_-)) \\ &\ge (1-2\beta)\log\frac{1-\beta}{\beta} \ge \frac{1}{4}\log\frac{1}{\beta}.\end{aligned}$$

The first inequality uses $p_-(1-p_-) \ge 3/16$; the last uses $1 - 2\beta \ge 1/2$ and $1 - \beta \ge \sqrt{\beta}$ for $\beta \le 1/4$. Thus $m \ge 3\log(1/\beta)/(256\alpha^2)$. Combining this with Proposition F.5 proves the lower half of (48): when $K\Delta \le 1/4$, the logarithms $\log(1/(2K\Delta))$, $\log(1/(K\Delta))$, and $\log(8/(K\Delta))$ differ by at most absolute constant factors. Formula (47) proves the upper half. □

F.2 Dimension and the joint tradeoff at the bit boundary

The boundary cost also grows linearly with VC dimension. The lower bound below averages accuracy over the edges of a shattered cube. For the full binary class on a finite domain, one shared threshold and the normalized-signal estimate give a matching joint upper bound.

Proposition F.8 (A dimension-dependent bit-boundary lower bound)**.** *Let $\mathcal{H}$ have VC dimension at least an integer $d \ge 1$, and let $0 < \alpha \le 1/16$ and $0 < \beta \le 1/4$. Suppose an agnostic (α, β)-learner uses a fixed, finite number m of examples and at most an integer $B \ge 4$ shared fair bits, with no other randomness, and is ρ-pointwise replicable on every distribution. Put $K = 2^B$ and $\Delta = \rho - 1/(2K)$. If $0 < K\Delta < 1/2$, then*

$$m \ge \frac{3dK^2}{131072\alpha^2}\log\frac{1}{2K\Delta}. \tag{50}$$

Proof. Fix shattered points $x_1, \dots, x_d$. For each $\sigma \in \{-1,+1\}^d$, let P_σ give each point mass $1/d$ and let

$$P_\sigma(Y = +1 \mid X = x_i) = \frac{1}{2} + 4\alpha\sigma_i.$$

The Bayes labels σ are realized by the class on this support. Thus any binary output h, including an improper one, has excess risk

$$\mathrm{R}_{P_\sigma}(h) - \mathrm{OPT}_{\mathcal{H}}(P_\sigma) = \frac{8\alpha}{d}\sum_{i=1}^{d} \mathbf{1}\{h(x_i) \ne \sigma_i\}.$$

On the accuracy event the fraction of wrong coordinates is at most $1/8$, and on its complement it is at most one. Writing

$$e_i(\sigma) = \mathbb{P}_{S,r}\{A(S;r)(x_i) \neq \sigma_i\}, \qquad S \sim P_\sigma^m,$$

we consequently have $d^{-1}\sum_i e_i(\sigma) \leq 1/8 + \beta \leq 3/8$ for every σ.

Average over σ and i, and group the terms by the undirected edges of the binary cube. The average sum of the two endpoint errors on an edge is at most $3/4$. Therefore there are a coordinate i and a fixed choice of the other $d-1$ signs such that its two endpoint errors, denoted by e_- and e_+, satisfy

$$e_- + e_+ \leq \frac{3}{4}.$$

Keep these other signs fixed and vary only $p = P(Y = +1 \mid X = x_i)$ over $[p_-, p_+] = [1/2 - 4\alpha, 1/2 + 4\alpha]$. Write P_p for the resulting distribution. Pad the seed to be uniform on $[K]$ and define the continuous polynomials

$$g_r(p) = \mathbb{P}_{S\sim P_p^m}\{A(S;r)(x_i) = +1\}.$$

The endpoint errors give $K^{-1}\sum_r g_r(p_-) = e_-$ and $K^{-1}\sum_r (1 - g_r(p_+)) = e_+$. Replication gives $\sum_r 2g_r(p)(1-g_r(p)) \leq K\rho$ for every p.

As in Proposition F.5, the inequality $\mathbf{1}\{u > 1/2\} \leq u + 2u(1-u)$ shows that at least

$$L \geq K(1 - e_- - e_+ - 2\rho) \geq K/4 - 2K\rho > K/4 - 2$$

seeds have $g_r(p_-) \leq 1/2$ and $g_r(p_+) \geq 1/2$. Here $K\rho = 1/2 + K\Delta < 1$. Since L is an integer, $L \geq K/4 - 1$; as $K \geq 16$, we have $L - 1 \geq K/4 - 2 \geq K/8$. Choose one location p_r with $g_r(p_r) = 1/2$ for each of these seeds. Two adjacent locations, belonging to distinct seeds r_1, r_2, satisfy

$$|p_1 - p_2| \leq \frac{8\alpha}{L-1} \leq \frac{64\alpha}{K}.$$

Only the conditional label law at x_i changes between P_{p_2} and P_{p_1}, and that point has mass $1/d$. Thus the single-observation relative entropy is exactly

$$D_{\mathrm{KL}}(P_{p_2}\|P_{p_1}) = \frac{1}{d} D_{\mathrm{KL}}(\mathrm{Ber}(p_2)\|\,\mathrm{Ber}(p_1)) \leq \frac{16(p_2-p_1)^2}{3d},$$

using $p_1 \in [1/4, 3/4]$ and the elementary Bernoulli bound in Proposition F.5. Put $q = g_{r_2}(p_1)$. These sample laws have the same finite support, so $0 < q < 1$. The same event-level log-sum argument used there, with $g_{r_2}(p_2) = 1/2$, gives

$$2q(1-q) \geq \frac{1}{2} e^{-2mD_{\mathrm{KL}}(P_{p_2}\|P_{p_1})}.$$

At p_1, seed r_1 contributes disagreement $1/2$. Averaging these two seeds' contributions yields

$$\Delta \geq \frac{1}{2K}\exp\left(-\frac{131072m\alpha^2}{3dK^2}\right). \tag{51}$$

Taking logarithms proves (50). □

Theorem F.9 (The joint tradeoff for the full binary cube)**.** *Let $d \geq 1$ be an integer and let $\mathcal{H}_d = \{-1,+1\}^{[d]}$. For $0 < \alpha, \beta \leq 1/4$, an integer $B \geq 2$, $K = 2^B$, and $0 < \Delta \leq 1/(2K)$, there is a proper agnostic (α,β)-learner for $\mathcal{H}_d$ that is $(1/(2K)+\Delta)$-pointwise replicable on every distribution. It uses exactly B shared fair bits, with no other randomness, and exactly*

$$n = \left\lceil \max\left\{ \frac{128dK^2}{\alpha^2}\log\frac{32}{K\Delta}, \frac{8}{\alpha^2}\left((d+1)\log 2 + \log\frac{1}{\beta}\right)\right\}\right\rceil \tag{52}$$

examples. In particular, uniformly for $0 < \alpha \leq 1/16$, $0 < \beta \leq 1/4$, $B \geq 4$, and $0 < K\Delta \leq 1/4$,

$$m^*_{\mathcal{H}_d}(\alpha,\beta,1/(2K)+\Delta;B) = \Theta\left(\frac{dK^2\log(1/(K\Delta)) + \log(1/\beta)}{\alpha^2}\right), \tag{53}$$

with absolute implicit constants. The matching upper bound in this theorem is for the full binary cube; the preceding lower bound applies to every class of VC dimension at least d.

Proof. **The learner and its accuracy.** For each $i \in [d]$, let N_i be its observed count and put

$$Z_i = \sum_{\ell: X_\ell = i} Y_\ell, \qquad \widehat{t}_i = \begin{cases} Z_i/\sqrt{nN_i}, & N_i > 0, \\ 0, & N_i = 0. \end{cases}$$

Set $G_i = 2\sqrt{d}\,(\widehat{t}_i)_+/\alpha$ and $V_i = \min\{1, G_i\}$. Draw one shared U uniformly from the K midpoints of $[0,1]$ and output

$$H(i) = 2\mathbf{1}\{U < V_i\} - 1 \qquad (i \in [d]).$$

In particular, an unobserved point receives label -1. The rule uses neither the point masses nor the conditional label means. It is the singleton-block, one-cut instance of the normalized product selector in Theorem I.9.

Its empirical excess risk admits the following direct bound. At coordinate i, the loss count of label -1 minus that of label $+1$ equals Z_i. If $Z_i \le 0$, the output is an empirical minimizer there. If $Z_i > 0$ and the output is -1, then $G_i \le U$: clipping preserves the strict comparison because $U \in (0,1)$. Hence the coordinate's excess loss count is at most $\alpha U\sqrt{nN_i}/(2\sqrt{d})$, including equality at a grid point. Therefore, for every sample and every seed,

$$\widehat{\mathrm{R}}_S(H) - \min_{h \in \mathcal{H}_d} \widehat{\mathrm{R}}_S(h) \le \frac{\alpha U}{2\sqrt{d}} \sum_{i=1}^{d} \sqrt{N_i/n} \le \frac{\alpha}{2}.$$

The last inequality uses $\sum_i N_i = n$ and Cauchy–Schwarz. There are exactly 2^d hypotheses. Hoeffding's inequality and a union bound consequently give

$$\mathbb{P}\left\{ \max_{h \in \mathcal{H}_d} |\widehat{\mathrm{R}}_S(h) - \mathrm{R}_P(h)| > \alpha/4 \right\} \le 2^{d+1} e^{-n\alpha^2/8} \le \beta.$$

On the complementary event, the output has risk at most $\mathrm{OPT}_{\mathcal{H}_d}(P) + \alpha$, simultaneously for all seeds. This proves accuracy without any additional logarithmic dependence on d or α.

Replication at a fixed coordinate. Write $p_i = P(X = i)$ and $\mu_i = \mathbb{E}[Y \mid X = i]$, setting $\mu_i = 0$ when $p_i = 0$, and put $t_i = \sqrt{p_i}\mu_i$. The normalized-signal bound (10) applies to $\widehat{t}_i$ on this sample, including $N_i = 0$:

$$\mathbb{P}\{|\widehat{t}_i - t_i| \ge u\} \le 4e^{-nu^2/8} \qquad (u > 0).$$

Taking the positive part and clipping are 1-Lipschitz. Thus, for the deterministic center $v_i = \min\{1, 2\sqrt{d}\,(t_i)_+/\alpha\}$,

$$\mathbb{P}\{|V_i - v_i| \ge s\} \le 4\exp\left(-\frac{n\alpha^2 s^2}{32d}\right) \qquad (s > 0). \tag{54}$$

The center need not equal the score's expectation.

Choose a midpoint nearest to v_i. Its conditional disagreement probability is at most $1/2$, since the two samples give independent, identically distributed scores. For any other midpoint u, put $q_u = \mathbb{P}\{V_i > u\}$. If $u > v_i$, the upper tail in (54) bounds q_u; if $u < v_i$, the lower tail bounds $1 - q_u = \mathbb{P}\{V_i \le u\}$. Consequently

$$2q_u(1 - q_u) \le 8\exp\left(-\frac{n\alpha^2(u - v_i)^2}{32d}\right).$$

This also verifies the strict-comparison convention at atoms of the score law. At index distance $\ell \ge 1$ from the nearest midpoint there are at most two grid points, each at distance at least $(\ell - 1/2)/K$ from v_i. This statement includes centers outside the grid's range. With $t = n\alpha^2/(128dK^2)$, averaging over the shared midpoint therefore yields

$$\begin{aligned} \mathbb{P}\{H(i) \ne H'(i)\} &\le \frac{1}{2K} + \frac{16}{K} \sum_{\ell \ge 1} e^{-t(2\ell-1)^2} \\ &\le \frac{1}{2K} + \frac{16}{K} \frac{e^{-t}}{1 - e^{-8t}} \\ &\le \frac{1}{2K} + \frac{32}{K} e^{-t} \le \frac{1}{2K} + \Delta. \end{aligned} \tag{55}$$

As in (49), the geometric sum uses $(2\ell-1)^2 \geq 1+8(\ell-1)$. Our sample choice gives $t \geq \log(32/(K\Delta)) \geq \log 64$, so the penultimate bound on the denominator holds. This argument applies to every distribution and every fixed coordinate, without conditioning on accuracy.

The matching lower bound. Under the additional parameters in (53), Proposition F.8 supplies the $dK^2\alpha^{-2}\log(1/(K\Delta))$ term. To obtain the confidence term, concentrate the feature distribution on one coordinate. Both possible labels on that coordinate belong to $\mathcal{H}_d$, so the testing argument in Proposition F.7 gives

$$m \geq \frac{3}{256\alpha^2}\log\frac{1}{\beta}.$$

Allowing predictions on the other, zero-mass coordinates does not affect that argument. The maximum of these two lower bounds is at least half their sum. For $K\Delta \leq 1/4$, replacing $\log(1/(2K\Delta))$ or $\log(32/(K\Delta))$ by $\log(1/(K\Delta))$ changes only an absolute constant. Also the d/α^2 term in (52) is absorbed by its first term. This proves the joint characterization. □

F.3 Few accurate thresholds despite optimal base risks

The selection step is needed even when all base hypotheses are exact class-risk minimizers. Here is a finite example of constant VC dimension. Choose $0 < \alpha \leq 1/16$ and an integer T such that $T/4$ and $4\alpha T$ are positive integers. Let the domain consist of two disjoint copies of the cycle $\mathbb{Z}_T$, each of marginal mass $1/2$, uniform within a copy. Labels are -1 on the first copy and $+1$ on the second. Let $\mathcal{H}$ contain all classifiers whose positive set is one cyclic interval of length $T/4$ in the first copy and one cyclic interval of length $T/4+4\alpha T$ in the second. Each classifier has risk

$$\frac{1}{2}\left(\frac{1}{4}+1-\left(\frac{1}{4}+4\alpha\right)\right) = \frac{1}{2}-2\alpha = \mathrm{OPT}_{\mathcal{H}}(P).$$

The VC dimension is at most six: cyclic intervals cannot shatter four cyclically ordered points, since alternating labels cannot be realized by one interval; shattering seven points across two copies would require shattering four in one copy.

Take an ensemble of T hypotheses using every cyclic shift, with the same shift in the two copies. The positive vote fractions are identically $1/4$ and $1/4+4\alpha$ on the respective copies. Among the T integer-threshold classifiers, exactly $4\alpha T$ predict both copies correctly and have risk zero; every other threshold has risk $1/2$. Thus a uniform threshold is accurate to $\mathrm{OPT}_{\mathcal{H}}(P)+\alpha$ with probability exactly 4α, even though every base hypothesis is optimal within the class. The example rules out uniform threshold selection as a general agnostic transformation. Optimal sample complexity for other transformations requires a separate lower bound.

Proposition F.10 (Sensitivity of the weighted-CDF operator)**.** *Let T be divisible by four, let $1 \leq r \leq T/4$ be an integer, put $\alpha = r/(4T)$, and choose $0 < a < 2\alpha$. There are a class of VC dimension at most seven, a distribution P, and a point x_* of zero P_X-mass with the following property. For any integers $N, N' \in \{T/4, \ldots, T/4+r\}$, there are two ensembles of T exact class-risk minimizers that agree term by term on the support of P_X, have positive vote counts N, N' at x_*, and whose weighted CDFs, formed from true threshold risks using (32), satisfy*

$$|F_w(N) - F_{w'}(N')| = \frac{|N/T - N'/T|}{4\alpha}. \tag{56}$$

Thus the $1/\alpha$ sensitivity of this CDF operator is attained up to constants. This is a deterministic statement about the operator, with no lower-bound claim for a learning algorithm's sample complexity.

Proof. Use the preceding two-cycle construction with $4\alpha T = r$, add the zero-mass point x_*, and give every class function both possible extensions to x_*. Adding one point increases the VC dimension by at most one. For both ensembles, use the same T cyclic shifts on the support and choose their extensions independently to realize counts N, N' at x_*. Every base risk remains $c = 1/2-2\alpha$. In either ensemble, threshold risk is zero exactly for $j \in \{T/4+1, \ldots, T/4+r\}$ and is $1/2$ otherwise. Because $a < 2\alpha$, the weights are a common positive constant on those r indices and zero elsewhere. Hence $F_w(N) = (N - T/4)/r$ and $F_{w'}(N') = (N' - T/4)/r$, which gives (56) since $r = 4\alpha T$. □

G FINITE RANDOMNESS FOR AGNOSTIC POINT FUNCTIONS

For point functions, the construction below removes the finite-randomness accuracy gap under general agnostic distributions. The construction estimates all empirical gains from one sample. A shared seed sets the tolerance below the largest gain; a fixed enumeration then selects the first eligible classifier. At a fixed test point, selection is an interval event with two concentrated endpoints.

Lemma G.1 (Midpoint coupling from mean absolute deviation)**.** *Let V, V' be independent and identically distributed random variables in $[0,1]$. Let U be independent of them and uniform on the K midpoints of $[0,1]$. Then*

$$\mathbb{P}\{\mathbf{1}\{U \le V\} \ne \mathbf{1}\{U \le V'\}\} \le \frac{1}{2K} + \mathbb{E}|V - V'| \le \frac{1}{2K} + \sqrt{2\,\mathrm{Var}(V)}.$$

The same bound holds for comparisons $U \ge V$ and their complements.

Proof. Put $p(u) = \mathbb{P}\{V \ge u\}$ and $f(u) = 2p(u)(1 - p(u))$. Independence makes $f(u)$ the disagreement probability at a fixed threshold u. Since p is nonincreasing, each strict superlevel set of f is an interval, possibly empty, and $0 \le f \le 1/2$. Every interval $I \subseteq [0,1]$, with any endpoint convention, contains at most $K|I| + 1$ grid points. Layer-cake integration therefore gives

$$\frac{1}{K} \sum_{u \text{ on the grid}} f(u) \le \int_0^1 f(u)\,du + \frac{1}{2K}.$$

For fixed values of V, V', the set of thresholds at which their comparisons disagree has Lebesgue measure $|V - V'|$. Thus the integral equals $\mathbb{E}|V - V'|$. Finally, $\mathbb{E}|V - V'| \le \sqrt{\mathbb{E}(V - V')^2} = \sqrt{2\,\mathrm{Var}(V)}$. Reflecting both the grid and the scores by $z \mapsto 1 - z$ covers $U \ge V$, and complementation preserves disagreement. □

Theorem G.2 (Agnostic point functions with a finite seed)**.** *Let $\mathcal{X}$ be a nonempty countable domain with a fixed enumeration, and let*

$$\mathcal{H}_{\mathrm{pt}} = \{h_0\} \cup \{h_x : x \in \mathcal{X}\}, \qquad h_0 \equiv -1, \qquad h_x(z) = 2\mathbf{1}\{z = x\} - 1.$$

For $0 < \alpha, \beta, \rho \le 1/4$, put

$$B = \left\lceil \log_2 \frac{2}{\rho} \right\rceil, \qquad K = 2^B, \qquad m = \left\lceil \frac{1}{\alpha^2} \max\left\{ \frac{128}{\rho^2}, 256 \log \frac{1}{\beta} \right\} \right\rceil.$$

There is a proper agnostic (α, β)-learner for $\mathcal{H}_{\mathrm{pt}}$ that is ρ-pointwise replicable on every distribution, uses m examples, and uses exactly B shared fair bits with no other randomness. In particular,

$$m = O\left(\frac{1}{\alpha^2} \left[\frac{1}{\rho^2} + \log \frac{1}{\beta} \right] \right).$$

No knowledge of the marginal atom masses or conditional label means is required.

Proof. **The learner.** Order the candidates as $h_0, h_{x_1}, h_{x_2}, \ldots$, placing h_0 first. From one sample $S = ((X_i, Y_i))_{i=1}^m$, compute

$$\widehat{s}_x = \frac{1}{m} \sum_{i=1}^m Y_i \mathbf{1}\{X_i = x\}, \qquad \widehat{s}_0 = 0, \qquad c = \max\{0, \sup_x \widehat{s}_x\}.$$

Only finitely many point scores are nonzero, so this maximum is attained. Independently of the sample, use the B shared bits to draw U uniformly from the K midpoints of $[0,1]$. Put $\lambda = c - (\alpha/2)U$ and output the first candidate whose empirical score is at least λ. Such a candidate exists. If $\lambda \le 0$, the output is h_0; otherwise every eligible point occurs in the sample. Thus the selection uses only a finite set of observed points and the fixed enumeration. All its operations are measurable on the countable sample space.

Accuracy. Write

$$s_x = \mathbb{E}[Y \mathbf{1}\{X = x\}], \qquad s_0 = 0, \qquad s^* = \max\{0, \sup_x s_x\}.$$

The risk identity

$$\mathrm{R}_P(h_x) = \mathrm{R}_P(h_0) - s_x$$

gives $\mathrm{OPT}_{\mathcal{H}_{\mathrm{pt}}}(P) = \mathrm{R}_P(h_0) - s^*$. Let $D = \sup_x |\widehat{s}_x - s_x|$. For completeness, symmetrization with independent Rademacher signs $\varepsilon_1, \ldots, \varepsilon_m$ gives

$$\begin{aligned}\mathbb{E}D &\le \frac{2}{m}\mathbb{E}\sup_x \left|\sum_{i=1}^m \varepsilon_i Y_i \mathbf{1}\{X_i = x\}\right| \\ &\le \frac{2}{m}\mathbb{E}\left[\sum_x \left(\sum_{i=1}^m \varepsilon_i Y_i \mathbf{1}\{X_i = x\}\right)^2\right]^{1/2} \le \frac{2}{\sqrt{m}}.\end{aligned}$$

The last inequality follows from Jensen's inequality: conditional on the sample, the expectation over the signs of the sum inside the square root is m. Replacing one observation changes D by at most $2/m$, so bounded differences gives

$$\mathbb{P}\{D > \mathbb{E}D + t\} \le \exp(-mt^2/2).$$

Our choice of m, together with $\beta \le 1/4$, implies $\mathbb{E}D \le \alpha/8$ and

$$\mathbb{P}\{D > \alpha/4\} \le \exp(-m\alpha^2/128) \le \beta.$$

On the complementary event, the selected candidate, say h_j, satisfies, for every possible seed,

$$s_j \ge \widehat{s}_j - \frac{\alpha}{4} \ge c - \frac{\alpha}{2} - \frac{\alpha}{4} \ge s^* - \alpha.$$

The risk identity proves (α, β) accuracy.

The two prediction boundaries. Fix $x = x_j$ and let

$$b_x = \max\{0, \widehat{s}_{x_1}, \ldots, \widehat{s}_{x_{j-1}}\}, \qquad A_x = \frac{2}{\alpha}(c - \widehat{s}_x), \qquad C_x = \frac{2}{\alpha}(c - b_x).$$

Only h_x predicts positively at x. By the rule selecting the first eligible candidate, its selection is equivalent to $\widehat{s}_x \ge \lambda$ and $b_x < \lambda$. Consequently,

$$\mathbf{1}\{H(x) = +1\} = \mathbf{1}\{U \ge A_x\}\mathbf{1}\{U < C_x\}.$$

Both boundaries are nonnegative. Clipping them to $[0, 1]$ preserves this identity because every grid midpoint lies strictly between zero and one.

We record a joint bounded-differences observation to avoid paying separately for two maxima. Include coordinate 0 in the empirical score vector, and let δ_i be the score increments when one observation is replaced. Then

$$\max_i \delta_i - \min_i \delta_i \le \frac{2}{m}.$$

Indeed, if the two feature values differ there are at most two nonzero increments, each of magnitude $1/m$; if they agree, there is at most one nonzero increment of magnitude $2/m$. The increment of the maximum over any fixed nonempty set of coordinates lies between $\min_i \delta_i$ and $\max_i \delta_i$. The same is true of an individual coordinate. It follows that each of $c - \widehat{s}_x$ and $c - b_x$ changes by at most $2/m$. Each clipped boundary V therefore has bounded differences $4/(m\alpha)$ per observation. The Efron–Stein inequality, followed by the bound of one quarter of the squared range for each conditional variance, gives

$$\mathrm{Var}(V) \le \sum_{i=1}^m \mathbb{E}\,\mathrm{Var}(V \mid (X_\ell, Y_\ell)_{\ell \ne i}) \le \frac{1}{4} m \left(\frac{4}{m\alpha}\right)^2 = \frac{4}{m\alpha^2}.$$

Replication. For two independent training samples, each pair of corresponding clipped boundaries is independent and identically distributed. Lemma G.1 applies also to comparisons of the form $U \ge V$: reflect both the grid and the score by the map $z \mapsto 1 - z$. Complementing such a

comparison also covers the strict comparison $U < V$, with all equality conventions preserved. The probability that either of the two boundary comparisons changes is consequently at most

$$\frac{1}{K} + \frac{4\sqrt{2}}{\alpha\sqrt{m}} \le \frac{\rho}{2} + \frac{\rho}{2} = \rho.$$

If neither comparison changes, the displayed prediction identity gives identical predictions at x. This proves pointwise replicability. The midpoint choice is the learner's only randomness, proving the total-bit assertion. □

Remark G.3 (Misspecification and fixed centers). This theorem is not a Bayes-risk-matched special case. For example, under uniform mass on $N \ge 2$ points and deterministically positive labels, its class optimum is $1 - 1/N$, whereas Bayes risk is zero. More generally, the same proof applies to a fixed classifier h_0 and the class obtained by flipping its prediction at at most one point: replace the signed gain by $-\mathbb{E}[Y h_0(X)\mathbf{1}\{X = x\}]$ and its empirical counterpart.

Proposition G.4 (Point functions at a prescribed bit budget)**.** *Let $\mathcal{H}_{\mathrm{pt}}$ be the point-function class on a nonempty countable domain as in Theorem G.2. For $0 < \alpha, \beta \le 1/4$, an integer $B \ge 2$, $K = 2^B$, and $0 < \Delta \le 1/(2K)$, there is an improper agnostic (α, β)-learner that is $(1/(2K)+\Delta)$-pointwise replicable on every distribution. It uses exactly B shared fair bits with no other randomness and exactly*

$$m = \left\lceil \frac{1}{\alpha^2} \max\left\{ 128K^2 \log\frac{16}{K\Delta}, 128 \log\frac{1}{\beta} \right\} \right\rceil \tag{57}$$

examples. Consequently, uniformly for $0 < K\Delta \le 1/4$,

$$m^*_{\mathcal{H}_{\mathrm{pt}}}(\alpha, \beta, 1/(2K) + \Delta; B) = \Theta\left(\frac{1}{\alpha^2} \left[K^2 \log\frac{1}{K\Delta} + \log\frac{1}{\beta} \right] \right), \tag{58}$$

with absolute implicit constants. The minimum is over fixed sample sizes of all learners with the displayed accuracy, replication, and bit constraints, allowing improper outputs. This conclusion includes the infinite point-function class.

Proof. Use one sample to compute

$$\widehat{s}_x = \frac{1}{m}\sum_{i=1}^{m} Y_i \mathbf{1}\{X_i = x\}, \qquad s_x = \mathbb{E}[Y\mathbf{1}\{X = x\}], \qquad s^* = \max\{0, \sup_x s_x\}.$$

Use the B shared bits to draw U uniformly from the K midpoints of $[0, 1]$, independently of the sample, and put

$$W = \frac{\alpha}{2} + \frac{\alpha}{4}U.$$

Output $+1$ at every point with $\widehat{s}_x \ge W$, and output -1 at every other point. The positive set is a subset of the finitely many observed points. Several points can be positive, so the output is generally improper. The rule is measurable and has no randomness other than U.

Accuracy. Put $D = \sup_x |\widehat{s}_x - s_x|$. The symmetrization and bounded-differences calculation in the proof of Theorem G.2 gives

$$\mathbb{E}D \le \frac{2}{\sqrt{m}}, \qquad \mathbb{P}\{D > \mathbb{E}D + t\} \le e^{-mt^2/2} \quad (t > 0).$$

The sample choice has $m \ge 256/\alpha^2$, so

$$\mathbb{P}\{D > \alpha/4\} \le e^{-m\alpha^2/128} \le \beta.$$

On the complementary event, every selected point has $s_x \ge W - \alpha/4 > \alpha/4 > 0$. If $s^* \le \alpha$, the all-negative classifier already has risk at most $\mathrm{OPT}_{\mathcal{H}_{\mathrm{pt}}}(P) + \alpha$, and every selected point improves its risk. If $s^* > \alpha$, every point with $s_x > \alpha$ is selected, since

$$\widehat{s}_x \ge s_x - \alpha/4 > 3\alpha/4 > W.$$

The gains satisfy $\sum_x |s_x| \le \sum_x P(X = x) = 1$. The sum of the selected positive gains is then at least s^*, by taking points with gains approaching the supremum. The risk identity for this finite positive set gives

$$\mathrm{R}_P(H) = \mathrm{R}_P(h_0) - \sum_{x:H(x)=+1} s_x \le \mathrm{R}_P(h_0) - s^* = \mathrm{OPT}_{\mathcal{H}_{\mathrm{pt}}}(P).$$

Both cases prove accuracy simultaneously for every seed on the same event.

One prediction boundary. Write $\mathrm{clip}(v) = \min\{1, \max\{0, v\}\}$ and, for a fixed test point x, put

$$V_x = \mathrm{clip}(4\widehat{s}_x/\alpha - 2), \qquad v_x = \mathrm{clip}(4s_x/\alpha - 2).$$

The prediction is positive exactly when $U \le V_x$, including equality at the threshold. Each summand defining $\widehat{s}_x$ lies in $[-1, 1]$. Hoeffding's inequality and clipping therefore give

$$\mathbb{P}\{|V_x - v_x| \ge t\} \le 2e^{-m\alpha^2 t^2/32} \qquad (t > 0).$$

The center v_x is deterministic. Let V'_x come from an independent sample and set $q(u) = \mathbb{P}\{V_x \ge u\}$. At the midpoint closest to v_x, disagreement is at most $1/2$. At every other midpoint it is bounded by

$$2q(u)(1 - q(u)) \le 4e^{-m\alpha^2|u-v_x|^2/32}.$$

There are at most two grid points at each index distance $j \ge 1$ from the nearest midpoint, and their distances from v_x are at least $(j - 1/2)/K$. Hence, with $t_0 = m\alpha^2/(32K^2) \ge 1$,

$$\begin{aligned}\mathbb{P}\{H(x) \ne H'(x)\} &\le \frac{1}{2K} + \frac{8}{K}\sum_{j\ge1} e^{-t_0(j-1/2)^2} \\ &\le \frac{1}{2K} + \frac{8}{K}\frac{e^{-t_0/4}}{1 - e^{-t_0}} \\ &\le \frac{1}{2K} + \frac{16}{K}e^{-m\alpha^2/(128K^2)} \le \frac{1}{2K} + \Delta.\end{aligned}$$

Here $(j-1/2)^2 \ge 1/4 + (j-1)$ gives the geometric bound. The argument averages over the shared seed and uses complete independence of the two samples. Thus it proves unconditional pointwise replicability with exactly B shared fair bits.

Matching lower bound. Concentrate the feature distribution on any one fixed point of the domain. The restrictions of h_0 and its point function give both possible labels there. Any learner for $\mathcal{H}_{\mathrm{pt}}$, including an improper learner, therefore induces a learner for the singleton class with the same sample size, bits, accuracy, and replication guarantee. Proposition F.7 supplies its joint lower bound. For $K\Delta \le 1/4$, $\log(16/(K\Delta))$ and $\log(1/(K\Delta))$ differ by at most an absolute constant factor, and a maximum of the two sample terms in (57) is within a factor of two of their sum. This proves (58). □

H PROPERNESS AND TOTAL SHARED RANDOMNESS

Chase et al. (2023, Theorem 6 and Section 6.3) use topology to prove properness separations for list replicability of singleton classes, with a global-stability interpretation. We study total shared fair bits under a per-query agreement requirement. For exactly three point functions, Theorems H.21 and H.23 determine the high-confidence feasibility boundary and the joint sample cost just above it for $K = 2^B \ge 8$. Theorem H.18 handles the different one- and two-bit regimes. The preliminary probability-column bounds also extend to hollow stars and a fixed infinite point class.

Lemma H.1 (An interior target for a simplex map)**.** *For an integer $N \ge 2$, let*

$$\Delta_N = \{x \in [0,1]^N : \textstyle\sum_i x_i = 1\},$$

and let $f : \Delta_N \to \Delta_N$ be continuous. Suppose $0 \le \eta < 1/N$ and $f_i(x) \le \eta$ whenever $x_i = 0$. Then every $t \in \Delta_N$ with $t_i > \eta$ for all i belongs to the image of f.

Proof. Define

$$g_i(x) = \frac{(f_i(x) - \eta)_+}{\sum_{\ell=1}^N (f_\ell(x) - \eta)_+}, \qquad v_i = \frac{t_i - \eta}{1 - N\eta}.$$

The denominator for g is at least $1 - N\eta > 0$. Thus g is continuous and satisfies $g_i(x) = 0$ on $x_i = 0$, while v is in the interior of Δ_N. Consider the continuous self-map

$$T_i(x) = \frac{(x_i + v_i - g_i(x))_+}{\sum_{\ell=1}^N (x_\ell + v_\ell - g_\ell(x))_+}.$$

Its denominator is at least one, since the unclipped numerators sum to one. Brouwer's fixed-point theorem gives a fixed point x. This point is interior: if $x_i = 0$, then $g_i(x) = 0$ and $T_i(x) > 0$, a contradiction. At an interior fixed point all numerators are positive, so their normalizing sum is one and $g(x) = v$. All coordinates of $g(x)$ are positive. Hence all $f_i(x) > \eta$, the normalizer in g is $1 - N\eta$, and $f(x) = t$. □

Lemma H.2 (Packing small probabilities)**.** *Let $u_1, \dots, u_\ell \in [0, 1/2]$ and $\sum_s u_s \geq r$, where $1/2 \leq r \leq 1$. Then*

$$\sum_s 2u_s(1 - u_s) \geq 1 - 2(1 - r)^2.$$

Proof. Put $U = \sum_s u_s$. If $U \geq 1$, then $\sum_s 2u_s(1 - u_s) \geq U \geq 1$. Otherwise $1/2 \leq U \leq 1$. Moving mass from a smaller positive coordinate to a larger one, up to $1/2$, increases the sum of squares. Thus $\sum_s u_s^2 \leq 1/4 + (U - 1/2)^2$, and

$$\sum_s 2u_s(1 - u_s) \geq 4U - 2U^2 - 1 = 1 - 2(1 - U)^2 \geq 1 - 2(1 - r)^2.$$

□

Theorem H.3 (A proper lower bound on three point functions)**.** *Let $\mathcal{X} = \{1, 2, 3\}$ and*

$$\mathcal{H}_3 = \{h_1, h_2, h_3\}, \qquad h_i(x) = 2\mathbf{1}\{x = i\} - 1.$$

This class has VC dimension one. Fix $0 < \alpha \leq 1/16$, $0 < \beta \leq 1/4$, an integer $B \geq 3$, and $K = 2^B$. Every proper agnostic (α, β)-learner for $\mathcal{H}_3$ that uses a fixed finite sample size, at most B total shared fair bits, and no other randomness, and is ρ-pointwise replicable on every distribution, must satisfy

$$\rho \geq \frac{7}{9K}. \tag{59}$$

Proof. The class shatters any one point, but cannot label two distinct points positively, so its VC dimension is one. Pad unused bits to view the learner as K equally likely deterministic seed rules. For $\theta \in \Delta_3$, let $P_\theta(X = i, Y = +1) = \theta_i$, and let

$$q_{s,i}(\theta) = \mathbb{P}_{S \sim P_\theta^n}\{A(S; s) = h_i\}, \qquad \bar{q}_i(\theta) = \frac{1}{K} \sum_{s=1}^K q_{s,i}(\theta).$$

Each $q_{s,i}$ is a polynomial in θ, since the sample size and sample space are finite. Properness gives $\sum_i q_{s,i} = 1$, so $\bar{q}$ is a continuous map from Δ_3 to itself.

The risk of h_i under P_θ is $1 - \theta_i$. If $\theta_i = 0$, some other coordinate is at least $1/2$, so h_i has excess risk at least $1/2 > \alpha$. Accuracy therefore gives $\bar{q}_i(\theta) \leq \beta$ on that face. Choose nonnegative integers m_1, m_2, m_3 that sum to $K - 2$ and differ by at most one, and put

$$t_i = \frac{m_i + 2/3}{K}.$$

These coordinates sum to one. For $K = 8$ they all equal $1/3$; for powers of two $K \geq 16$, their minimum is at least $1/3 - 2/(3K) \geq 7/24$. In particular $t_i > \beta$. Lemma H.1 supplies a θ_* with $\bar{q}(\theta_*) = t$. In the rest of the proof write $q_{s,i} = q_{s,i}(\theta_*)$.

Let c_i be the number of seeds with $q_{s,i} > 1/2$. Each seed contributes to at most one c_i, so $\sum_i c_i \le K$. Since $\sum_i (m_i + 1) = K + 1$, some coordinate i satisfies $c_i \le m_i$. For that coordinate set

$$U = \sum_{s:q_{s,i}\le 1/2} q_{s,i}.$$

The remaining c_i probabilities sum to at most c_i. Consequently $U \ge Kt_i - c_i \ge 2/3$. Lemma H.2, with $r = 2/3$, therefore gives

$$\sum_s 2q_{s,i}(1-q_{s,i}) \ge \sum_{s:q_{s,i}\le 1/2} 2q_{s,i}(1-q_{s,i}) \ge \frac{7}{9}.$$

Finally, at query i the proper learner predicts positively exactly when it outputs h_i. For two independent samples and the same seed, its pointwise disagreement probability is

$$\frac{1}{K}\sum_{s=1}^{K} 2q_{s,i}(1-q_{s,i}) \ge \frac{7}{9K}.$$

This proves the stated pointwise bound directly. □

Corollary H.4 (An exact one-bit properness cost)**.** *For $\mathcal{H}_3$ and the accuracy parameters in Theorem H.3, let $B \ge 3$, $K = 2^B$, and*

$$\frac{1}{2K} < \rho < \frac{7}{9K}. \tag{60}$$

Among learners using any fixed finite sample size and no randomness other than shared fair bits, the minimum total number of bits is B when improper outputs are allowed, and is exactly $B + 1$ when outputs must belong to $\mathcal{H}_3$.

Proof. The class $\mathcal{H}_3$ is contained in the radius-one Hamming ball around the constant negative classifier. Theorem M.3 therefore gives an improper learner using B bits. Proposition 3.4, or (42), rules out fewer: on (60), $\lfloor \log_2(1/\rho) \rfloor = B$. Theorem H.3 rules out every proper learner using at most B bits.

The ordered selector gives a proper learner with exactly $B + 1$ bits. Put $K' = 2^{B+1} = 2K$ and $\delta = \rho - 1/K' > 0$. Order $h_1 < h_2 < h_3$, draw U uniformly from the K' midpoints of $[0, 1]$, and output the first hypothesis with

$$\widehat{\mathrm{R}}_S(h_i) \le \min_j \widehat{\mathrm{R}}_S(h_j) + (\alpha/2)U.$$

This is the proper finite-order case of Theorem I.1 and Remark I.2; its analysis applies to this prescribed grid. For either nonempty proper initial segment C, write

$$V_C = \mathrm{clip}_{[0,1]}\left(\frac{2}{\alpha}\left[\min_{h\in C}\widehat{\mathrm{R}}_S(h) - \min_{h\in\mathcal{H}_3}\widehat{\mathrm{R}}_S(h)\right]\right).$$

Replacing an observation changes V_C by at most $4/(n\alpha)$, so

$$\mathbb{P}\{|V_C - \mathbb{E}V_C| \ge u\} \le 2e^{-n\alpha^2u^2/8}.$$

The nearest-grid-point argument, including strict comparisons, therefore bounds the disagreement of one cut by

$$\frac{1}{2K'} + \frac{16}{K'}e^{-n\alpha^2/(32K'^2)}$$

whenever $n\alpha^2/K'^2 \ge 8$. Each query uses at most two cuts. Hence the learner's pointwise disagreement is at most

$$\frac{1}{K'} + \frac{32}{K'}e^{-n\alpha^2/(32K'^2)}.$$

It suffices to take

$$n = \left\lceil \max\left\{\frac{32K'^2}{\alpha^2}\log\left(2 + \frac{32}{K'\delta}\right), \frac{8}{\alpha^2}\log\frac{6}{\beta}\right\}\right\rceil.$$

The first term makes the last disagreement bound at most $1/K' + \delta = \rho$. The second gives uniform empirical-risk deviation at most $\alpha/4$ over the three hypotheses with probability at least $1 - \beta$, by Hoeffding's inequality. On that event every eligible hypothesis has excess risk at most α, for every seed. This proves the proper upper bound with a fixed finite sample size and completes the exact bit separation. □

Remark H.5 (The preliminary separation). The constant $7/9$ identifies a band of exact bit separations for a VC-one class; determining the full proper feasibility threshold requires the sharper arguments below. The lower-bound distributions are agnostic: $\mathrm{OPT}_{\mathcal{H}_3}(P_\theta) = 1 - \max_i \theta_i$. The argument bounds disagreement at an individual query.

H.1 A CONFIDENCE-DEPENDENT OBSTRUCTION FOR POINT FUNCTIONS

Lemma H.6 (Two simplex maps have a common near-half coordinate)**.** *Let f, g be continuous maps from Δ_3 to $\{u \in [0,1]^3 : \sum_i u_i \le 1\}$, with $f(e_i) = g(e_i) = e_i$ at each vertex. Suppose that $0 \le \eta < 1/2$ and, on each edge joining e_i and e_j,*

$$(1 - f_i(x) - f_j(x)) + (1 - g_i(x) - g_j(x)) \le \eta.$$

Then some $x \in \Delta_3$ and $i \in \{1,2,3\}$ satisfy

$$2f_i(x)(1 - f_i(x)) + 2g_i(x)(1 - g_i(x)) \ge 1 - \frac{\eta^2}{2}.$$

In particular, two face-preserving simplex maps have a coordinate equal to $1/2$ simultaneously at some point.

Proof. Suppose the displayed conclusion fails everywhere, and set

$$H_i(x) = (f_i(x) - 1/2, g_i(x) - 1/2) \in \mathbb{R}^2.$$

The sum of the two Bernoulli disagreement probabilities is $1 - 2\|H_i(x)\|_2^2$. Hence $\|H_i(x)\|_2 > \eta/2$ for every i, x. Since the triangle is simply connected, each continuous nonzero vector field H_i has a continuous real argument ϕ_i on the entire triangle. On the opposite face $x_i = 0$, the edge assumption and the subprobability row sums give $f_i + g_i \le \eta$. Both coordinates of H_i are therefore strictly negative. Normalize ϕ_i to take values in $(-\pi, -\pi/2)$ on this face. At its two endpoints it equals $-3\pi/4$, whereas

$$\phi_i(e_i) = \pi/4 + 2\pi k_i \quad \text{for an integer } k_i.$$

On the edge joining e_i and e_j, write $a = 1 - f_i - f_j$ and $b = 1 - g_i - g_j$. There

$$H_i + H_j = -(a, b), \qquad \|H_i + H_j\|_2 \le a + b \le \eta.$$

The vectors H_i, H_j cannot point in the same direction: their norms would then add to more than η. Consequently $\phi_j - \phi_i$ avoids $2\pi\mathbb{Z}$ along this connected edge. Its endpoint values are $-\pi - 2\pi k_i$ and $\pi + 2\pi k_j$. These odd multiples of π must belong to the same component of $\mathbb{R} \setminus 2\pi\mathbb{Z}$, so they are equal and $k_i + k_j = -1$. Applying this to all three edges gives $2(k_1 + k_2 + k_3) = -3$, a contradiction. When $\eta = 0$, equality in the upper bound 1 for the sum of two Bernoulli disagreement probabilities requires $f_i = g_i = 1/2$, proving the last assertion. □

Theorem H.7 (Proper point functions at high confidence)**.** *Let $\mathcal{X}$ be a standard Borel space with at least three points, and let $\mathcal{H}$ consist of all point functions on $\mathcal{X}$, with or without the constant negative classifier. Let $K = 2^B \ge 2$, $0 < \alpha < 1/2$, and $0 < \beta < 1/(2K)$. Every proper agnostic (α, β)-learner using a fixed finite sample size and at most B shared fair bits, with no other randomness, that is pointwise ρ-replicable on every distribution satisfies*

$$\rho \ge \frac{1 - (K\beta)^2/2}{K}. \tag{61}$$

Proof. Pad unused bits to obtain K equally likely deterministic seed rules. Choose three distinct domain points, indexed by $1, 2, 3$, and put $P_\theta(X = i, Y = +1) = \theta_i$ for $\theta \in \Delta_3$. Write $q_{s,i} = \mathbb{P}\{A(S; s) = h_i\}$ for these three point functions. Each row is a continuous subprobability vector: its coordinates are polynomials determined by the finitely many possible sample strings. Set $\eta = K\beta$. At vertex e_i the sample is deterministic. Each seed therefore either always outputs h_i or always fails accuracy, since every other proper output has excess risk one. As $\beta < 1/K$, every seed outputs h_i there. Choose any two distinct seeds s, t, and apply Lemma H.6 to $f = (q_{s,1}, q_{s,2}, q_{s,3})$ and

$g = (q_{t,1}, q_{t,2}, q_{t,3})$. On the edge joining e_i and e_j, every output other than h_i, h_j is negative on the support. Such an output has excess risk at least $1/2 > \alpha$. Therefore

$$\sum_{s=1}^{K} (1 - q_{s,i} - q_{s,j}) \le K\beta = \eta.$$

In particular the two chosen rows satisfy the lemma's edge condition, including all outputs outside the three chosen point functions and the negative classifier when present. The point and coordinate supplied by the lemma contribute at least $1 - \eta^2/2$ to the sum of the K seedwise pointwise disagreement probabilities. The other seeds contribute nonnegative terms, giving (61). □

Corollary H.8 (A confidence floor below $1/K$)**.** *For either point class in Theorem H.7, fix $K = 2^B \ge 2$, $0 < \alpha < 1/2$, and $\rho < 1/K$. Every proper agnostic (α, β)-learner, with any $0 < \beta < 1$, a fixed finite sample size, at most B total shared fair bits and no other randomness, that is pointwise ρ-replicable on every distribution must have*

$$\beta \ge \frac{1}{K} \min\left\{\frac{1}{2}, \sqrt{2(1 - K\rho)}\right\}.$$

Proof. If β were strictly below the displayed quantity, then $\beta < 1/(2K)$ and Theorem H.7 would give $K\rho \ge 1 - (K\beta)^2/2 > K\rho$. □

This bound approaches $1/K$ as β tends to zero; $\rho \ge 1/K$ is therefore a limiting conclusion, rather than a bound at fixed positive failure probability.

Theorem H.9 (The one-bit confidence law for point functions)**.** *Let $\mathcal{X}$ be a standard Borel space with at least three points, and let $\mathcal{H}$ consist of all point functions on $\mathcal{X}$, with or without the constant negative classifier. For $0 < \alpha < 1/2$ and $0 < \beta \le 1/8$, let $\rho^*_{\mathrm{prop},1}(\mathcal{H}; \alpha, \beta)$ denote the infimum of achievable pointwise replication parameters over proper agnostic (α, β)-learners for $\mathcal{H}$ using any fixed finite sample size, at most one total shared fair bit, no other randomness, and the accuracy and replication guarantees on every distribution. Then*

$$\frac{1}{2} - \beta^2 \le \rho^*_{\mathrm{prop},1}(\mathcal{H}; \alpha, \beta) \le \frac{1}{2} - \frac{\beta^2}{4608}. \tag{62}$$

The upper bound uses exactly one shared bit and $O(\alpha^{-2} \log(1/\beta))$ examples, independently of the domain size. This sample order is optimal even without imposing replication. Thus $1/2 - \rho^*_{\mathrm{prop},1}(\mathcal{H}; \alpha, \beta) = \Theta(\beta^2)$. This statement concerns replication parameters near $1/2$, beyond the range $\rho \le 1/4$ used in most of the sample–randomness tradeoffs above.

Proof. The lower bound follows from Theorem H.7 with $K = 2$. For the upper bound put

$$\varepsilon = \frac{\beta}{2}, \qquad \kappa = \frac{\alpha}{8}, \qquad a = \frac{\alpha}{32}, \qquad L = \left\lceil \log_2 \frac{12}{\beta} \right\rceil,$$

$$r = \left\lceil \frac{8L}{a} \right\rceil, \qquad m = \left\lceil \frac{512}{\alpha^2} \log \frac{8}{\beta} \right\rceil.$$

Index point functions by $\mathcal{X}$ and, when present, index h_0 by a separate isolated symbol. This index space is standard Borel; fix a Borel injection into $\mathbb{R}$ (Kechris, 1995, Section 15) to order finite candidate sets. Also fix such an encoding of the raw observation space $\mathcal{Z} = \mathcal{X} \times \{-1, +1\}$ to order finite strings of observations. Fix three distinct points a_1, a_2, a_3; when the domain is infinite, extend them to a fixed sequence of distinct points $a_1, a_2, \ldots$. Define fixed penalties by

$$\pi(h_0) = 0 \quad \text{when present}, \qquad \pi(h_{a_1}) = \kappa,$$
$$\pi(h_{a_2}) = 2\kappa, \qquad \pi(h_x) = 3\kappa \quad (x \notin \{a_1, a_2\}).$$

On a block of m examples, the base rule maximizes $\widehat{s}(h) - \pi(h)$, where $\widehat{s}(h_x) = m^{-1} \sum_j Y_j \mathbf{1}\{X_j = x\}$ and $\widehat{s}(h_0) = 0$. This biased empirical maximizer is a finite computation: consider the observed points and a_1, a_2, and include h_0 when it is present; when h_0 is absent, also include the first unused point in the fixed sequence on an infinite domain, or the first unobserved point of the finite domain in the fixed order if one exists. In the infinite case the first $m + 1$ fixed

points contain an unused point. Every unobserved point function has empirical gain zero; the explicitly included anchors cover the smaller penalties, and all other unobserved points have the same penalty. If the finite domain has been fully observed and h_0 is absent, simply maximize over its observed points. Use the fixed encoding to resolve ties. Equality on a standard Borel space is Borel, so these finite operations give a measurable output index.

Write $s_x = \mathbb{E}[Y\mathbf{1}\{X = x\}]$ and let $\widehat{s}_x$ be its empirical counterpart. For $D = \sup_x |\widehat{s}_x - s_x|$, Lemma M.1 with $d = 1$ and negative baseline gives measurability; its moment and replacement bounds, together with McDiarmid's inequality, give

$$\mathbb{E}D \le \frac{1}{\sqrt{m}}, \qquad \mathbb{P}\{D > \mathbb{E}D + t\} \le e^{-mt^2/2}.$$

Uniform gain error D bounds the excess risk of an exact biased empirical maximizer by $2D + 3\kappa$, including when the true gain supremum is not attained. Since $m \ge 64/\alpha^2$, the event $D \le \alpha/4$ fails with probability at most $e^{-m\alpha^2/128} \le (\beta/8)^4 \le \beta/2$. On that event the excess risk is at most $7\alpha/8$. Write C for the output of this single base block.

The penalties also make the base rule constant near any raw point mass. If a sample has a fraction at least $1 - d$ of one observation $z = (x, y)$, with $d \le \kappa/2$, its unique maximizing hypothesis is h_x when $y = +1$; when $y = -1$, it is h_0 if present, and otherwise the first of h_{a_1}, h_{a_2} whose point differs from x. For the negative case, two point functions supported away from x have empirical gain difference at most d, whereas the chosen anchor's penalty advantage is at least κ. With h_0 present its penalty advantage gives the same conclusion. Comparisons involving h_x in either label case have gain gap at least $1 - 2d$ in the required direction; subtracting any penalty difference leaves at least $1 - 4\kappa > 0$. Thus the stated choices are unique. Consequently, if P has a raw atom z of mass at least $1 - a$, Hoeffding's inequality gives

$$\mathbb{P}\{C \ne h(z)\} \le \mathbb{P}\{\text{empirical minority at } z > 2a\} \le e^{-2ma^2} \le \frac{\beta}{8},$$

where $h(z)$ is that fixed choice and $2a = \kappa/2$.

Use $2r$ additional independent raw observations $Z_1, \dots, Z_{2r} \in \mathcal{Z}$ as the auxiliary block, independent of C.

We use the classical equal-probability type-class principle of Elias (1972) to extract a uniform index on a rare event, without an additional random seed. For an auxiliary sequence, let n_z count its occurrences of $z \in \mathcal{Z}$, and set

$$M = \frac{(2r)!}{\prod_{z:n_z>0} n_z!}, \qquad A = 3\left\lfloor \frac{\varepsilon M}{3} \right\rfloor.$$

Only finitely many counts are nonzero, regardless of the domain's cardinality. Order the M sequences with that full multiset lexicographically using the fixed Borel encoding, and let $R \in \{0, \dots, M-1\}$ be the observed sequence's rank. Declare acceptance when $R < A$ and, on acceptance, put $J = 1 + (R \bmod 3)$. The shared bit chooses one of two rules: rule zero outputs C; rule one outputs h_{a_J} on acceptance and C otherwise. All the block rules, the ranking, and acceptance are deterministic functions of the sample. The shared bit is the only external randomness.

For a fixed distribution P and query x, let v_x be the probability that one base rule outputs h_x. Let T be the sorted auxiliary sequence, representing its full multiset. Sorting, equality counts, and finite ranks are Borel under the fixed encoding. Write $Z_\sigma = (Z_{\sigma(1)}, \dots, Z_{\sigma(2r)})$. For every bounded Borel function φ, exchangeability under the permutation group $\mathfrak{S}_{2r}$ gives

$$\begin{aligned}\mathbb{E}\varphi(T, R) &= \mathbb{E}\left[\frac{1}{(2r)!}\sum_{\sigma\in\mathfrak{S}_{2r}} \varphi(T, R(Z_\sigma))\right] \\ &= \mathbb{E}\left[\frac{1}{M}\sum_{j=0}^{M-1} \varphi(T, j)\right].\end{aligned}$$

Indeed, every distinct ordering of the multiset occurs $\prod_z n_z!$ times among its permutations. This proves the uniform conditional rank law also for nonatomic raw distributions; no positive-probability

multiset assumption is needed. Thus conditional on T, acceptance has probability $A/M \le \varepsilon$ and, conditionally on acceptance, J is exactly uniform on the three indices. Let h denote the unconditional acceptance probability. Since C is independent of the auxiliary sequence, the two seedwise positive-prediction probabilities at x are

$$q_{0,x} = v_x, \qquad q_{1,x} = (1-h)v_x + hw_x, \qquad 0 \le h \le \varepsilon.$$

Here $w_x = 1/3$ for $x \in \{a_1, a_2, a_3\}$ and $w_x = 0$ otherwise. The negative classifier, when present, is an available output but contributes no positive prediction. Each seed has accuracy failure probability at most $\beta/2 + h \le \beta$.

If some $v_x \in [1/4, 3/4]$, the near-atom property implies that every raw atom has mass less than $1 - a$: otherwise one hypothesis would be selected with probability at least $1 - \beta/8 > 3/4$, and each other one with probability at most $\beta/8 < 1/4$. Thus the probability that two independent raw observations are unequal satisfies

$$\theta = \mathbb{P}\{Z_1 \ne Z_2\} = 1 - \sum_{z:\, P\{z\}>0} P\{z\}^2 \ge a.$$

The sum is over the countably many atoms of P; the formula also includes its nonatomic part. Let J_0 count the unequal pairs $(Z_1, Z_2), \dots, (Z_{2r-1}, Z_{2r})$. Then $J_0 \sim \mathrm{Bin}(r, \theta)$ has mean $\mu \ge 8L$, so the multiplicative Chernoff bound gives

$$\mathbb{P}\{J_0 < L\} \le e^{-\mu/8} \le e^{-L} < \frac{1}{2}.$$

Independently swapping each unequal pair produces 2^{J_0} distinct orderings of the same raw multiset. Hence $M \ge 2^{J_0}$, and

$$\mathbb{P}\{M \ge 6/\varepsilon\} \ge \mathbb{P}\{J_0 \ge L\} > \frac{1}{2}.$$

We condition the uniform rank law on the full multiset, not on J_0. On every multiset with $M \ge 6/\varepsilon$, $A/M \ge \varepsilon - 3/M \ge \varepsilon/2$. It follows that $h \ge \varepsilon/4 = \beta/8$.

Fix a query x with $v_x \in [1/4, 3/4]$, and put $z = v_x - 1/2$ and $d_x = 1/2 - w_x \ge 1/6$. The sum of its two seedwise disagreement probabilities is

$$\begin{aligned} 2q_{0,x}(1-q_{0,x}) + 2q_{1,x}(1-q_{1,x}) &= 1 - 2\left[z^2 + ((1-h)z - hd_x)^2\right] \\ &\le 1 - \frac{h^2}{36} \le 1 - \frac{\beta^2}{2304}. \end{aligned}$$

Here completing the square bounds the bracket from below by $h^2 d_x^2/[1 + (1-h)^2] \ge h^2/72$. If v_x is outside that interval, its first disagreement term is at most $3/8$ and the second is at most $1/2$; their sum is at most $7/8 \le 1 - \beta^2/2304$. The uniform shared bit divides these sums by two. This proves the upper bound in (62) on every P, with the fixed sample size $m + 2r = O(\alpha^{-2}\log(1/\beta))$.

For the matching sample lower bound, fix one point x and let $X = x$ almost surely under two distributions with

$$P_{\pm}\{Y = +1\} = \frac{1}{2} \pm \frac{3\alpha}{4}.$$

A wrong prediction at x has excess risk $3\alpha/2 > \alpha$, since the point-function class contains hypotheses of both signs there. Thus any agnostic learner, even an improper one and without a replication requirement, gives a test of P_+ against P_- with both errors at most β. An independent random seed does not increase the relative entropy of the two sample laws. Data processing therefore gives

$$n \,\mathrm{kl}\left(\frac{1}{2} + \frac{3\alpha}{4}, \frac{1}{2} - \frac{3\alpha}{4}\right) \ge \frac{1}{2}\log\frac{1}{\beta}.$$

For $0 < \alpha < 1/2$ the displayed one-observation relative entropy is at most $(72/7)\alpha^2 < 16\alpha^2$. Consequently $n \ge (32\alpha^2)^{-1}\log(1/\beta)$, which proves optimality of the sample order throughout the stated range. □

H.2 HOLLOW-STAR WITNESSES

Say that distinct points $x_1, \dots, x_N$ and labels $y_1, \dots, y_N$ form a hollow-star witness for $\mathcal{H}$ if, writing

$$S(h) = \{i : h(x_i) \neq y_i\},$$

every $S(h)$ is nonempty and each singleton $\{i\}$ occurs as $S(h)$ for some $h \in \mathcal{H}$. This is the missing center and its one-coordinate neighbors used in the proper list-replicability lower bound of Chase et al. (2023). Here the witness yields a lower bound directly on pointwise disagreement with a fixed uniform seed.

Theorem H.10 (A hollow-star proper lower bound)**.** *Suppose $\mathcal{H}$ has a hollow-star witness of size $N \geq 2$ and $|S(h)| \leq r$ for every $h \in \mathcal{H}$, where r is an integer with $1 \leq r \leq N$. Let $K = 2^B \geq N - 1$, and suppose*

$$0 < \alpha < \frac{1}{(N-1)(2N-1)}, \qquad 0 < \beta < \frac{\lfloor (K - N + 1)/N \rfloor + (N-1)/N}{K}.$$

Every proper agnostic (α, β)-learner with a fixed finite sample size and at most B total shared fair bits, no other randomness, and pointwise replication parameter ρ on every distribution satisfies

$$\rho \geq \frac{1}{K}\left[1 - \frac{2}{N^2} - 2\left(1 - \frac{1}{r}\right)^2 (K\beta)^2\right]. \tag{63}$$

Proof. Pad unused bits to obtain K deterministic seed rules with the uniform seed distribution. For $\theta \in \Delta_N$, put deterministic label y_i at x_i and assign it mass

$$p_i(\theta) = \frac{2 - \theta_i}{2N - 1}.$$

These masses sum to one. Since every trace is nonempty and each singleton trace is present, the optimal risk is $\min_i p_i \leq 1/N$. Every trace with at least two flips has risk at least $3/(2N-1)$ and hence excess risk at least $(N+1)/(N(2N-1)) > \alpha$. On the face $\theta_i = 0$, a trace containing i has risk at least $2/(2N-1)$, whereas some other singleton has risk at most $(2 - 1/(N-1))/(2N-1)$. Its excess risk is therefore at least $1/((N-1)(2N-1)) > \alpha$.

For each deterministic seed, let S_s denote the random flip trace of its output on a fresh sample, and define

$$q_{s,i} = \mathbb{E}\frac{\mathbf{1}\{i \in S_s\}}{|S_s|}, \qquad a_{s,i} = \mathbb{P}\{i \in S_s\}, \qquad d_{s,i} = a_{s,i} - q_{s,i}.$$

This uniform projection analyzes output probabilities; the learner itself uses no projection randomness. Each q_s is a probability vector. All these functions are continuous in θ, since the sample has finitely many possible strings. Moreover,

$$0 \leq d_{s,i} \leq \left(1 - \frac{1}{r}\right)\mathbb{P}\{|S_s| \geq 2\}, \qquad \sum_s d_{s,i} \leq \left(1 - \frac{1}{r}\right) K\beta.$$

The last inequality follows because every trace with multiple flips fails accuracy. The opposite-face accuracy bound also gives $\bar{q}_i \leq \bar{a}_i \leq \beta$ whenever $\theta_i = 0$.

Choose balanced nonnegative integers m_i with sum $K - N + 1$ and set

$$t_i = \frac{m_i + (N-1)/N}{K}.$$

They sum to one and satisfy $t_i > \beta$ by hypothesis. In particular $\beta < 1/N$, so Lemma H.1 applies to $\bar{q}$ and supplies a parameter with $\bar{q} = t$. Let $c_i = |\{s : q_{s,i} > 1/2\}|$ there. Since $\sum_i c_i \leq K$ but $\sum_i (m_i + 1) = K + 1$, some i has $c_i \leq m_i$. Thus

$$\sum_{s : q_{s,i} \leq 1/2} q_{s,i} \geq \frac{N-1}{N}.$$

Lemma H.2 bounds the sum of the projected disagreement terms on these seeds from below by $1 - 2/N^2$. For $q \leq 1/2$ and $d \geq 0$ with $q + d \leq 1$,

$$2(q+d)(1-q-d) - 2q(1-q) = 2d(1-2q) - 2d^2 \geq -2d^2.$$

It follows that the actual disagreement sum at query x_i is at least

$$\sum_s 2a_{s,i}(1-a_{s,i}) \geq 1 - \frac{2}{N^2} - 2 \sum_{s:q_{s,i}\leq 1/2} d_{s,i}^2$$
$$\geq 1 - \frac{2}{N^2} - 2\left(1-\frac{1}{r}\right)^2 (K\beta)^2.$$

Dividing by K proves the pointwise lower bound. □

Corollary H.11 (A three-point hollow star)**.** *If $\mathcal{H}$ has a three-point hollow-star witness, then for $K = 2^B \geq 2$, $0 < \alpha \leq 1/16$, and $0 < \beta \leq 1/(2K)$, every learner with the other properties in Theorem H.10 satisfies $\rho \geq 5/(9K)$.*

Proof. Use $N = r = 3$ in Theorem H.10. Its target bound is at least $2/(3K) > \beta$, and its accuracy bound is $1/10 > \alpha$. The bracket in (63) is at least $7/9 - 2(2/3)^2(1/2)^2 = 5/9$. □

This corollary is a proper lower bound for the witnessed class. It does not supply a matching proper upper bound or an improper bit optimum for an arbitrary such class.

H.3 A FIXED INFINITE POINT CLASS

The next bound allows $\beta < 1/(K+1)$, subject to $\beta \leq 1/4$, extending beyond the range $\beta < 1/(2K)$ of the circle argument, but requires the stronger accuracy condition $\alpha < 1/(K+1)$.

Theorem H.12 (Proper point functions with escape outputs)**.** *On $\mathcal{X} = \mathbb{N}$, let $h_i(x) = 2\mathbf{1}\{x = i\} - 1$ and let $\mathcal{H}$ be either $\{h_i : i \geq 1\}$ or $\{h_0\} \cup \{h_i : i \geq 1\}$, where $h_0 \equiv -1$. Fix $K = 2^B \geq 2$ and*

$$0 < \alpha < \frac{1}{K+1}, \qquad 0 < \beta < \frac{1}{K+1}, \qquad \beta \leq \frac{1}{4}.$$

Every proper agnostic (α, β)-learner using a fixed finite sample size and at most B total shared fair bits, with no other randomness, that is pointwise ρ-replicable on every distribution satisfies

$$\rho \geq \frac{c_K(\beta)}{K}, \qquad c_K(\beta) = 1 - 2\left(\frac{1+K\beta}{K+1}\right)^2. \tag{64}$$

Proof. Put $N = K + 1$ and use the distributions $P_\theta(X = i, Y = +1) = \theta_i$, $\theta \in \Delta_N$, supported on the first N points. For each padded deterministic seed s define

$$q_{s,i}(\theta) = \mathbb{P}\{A(S;s) = h_i\}, \qquad e_s(\theta) = \mathbb{P}\{A(S;s) \notin \{h_1, \ldots, h_N\}\}, \qquad e = \frac{1}{K}\sum_s e_s.$$

These functions are polynomials: only the N^n possible sample strings are involved, even though the output class is infinite. Every escape output, including h_0 when present, is negative on the support and has risk one. Its excess risk is $\max_i \theta_i \geq 1/N > \alpha$. Consequently $e \leq \beta$. On the face $\theta_i = 0$, output h_i is also bad, so

$$\bar{q}_i + e \leq \beta, \qquad \bar{q}_i = \frac{1}{K}\sum_s q_{s,i}.$$

Complete each probability row only for the analysis by $\widetilde{q}_{s,i} = q_{s,i} + e_s/N$. The mean $f_i = \bar{q}_i + e/N$ defines a continuous map $f : \Delta_N \to \Delta_N$ with $f_i \leq \beta$ on its opposite face. Since $\beta < 1/N$, Lemma H.1 gives a parameter at which $f_i = 1/N$ for every i. At this parameter, each original column has the same sum

$$u = \sum_s q_{s,i} = \frac{K(1-e)}{N}.$$

Each row has sum at most one and hence at most one entry strictly greater than $1/2$. Since there are $K + 1$ columns but only K rows, some column has all its entries at most $1/2$. Also

$$\frac{1}{2} \leq \frac{K(1-\beta)}{K+1} \leq u < 1,$$

where the first inequality uses $K \geq 2$ and $\beta \leq 1/4$. Lemma H.2 yields

$$\sum_s 2q_{s,i}(1-q_{s,i}) \geq 1 - 2(1-u)^2 \geq 1 - 2\left(\frac{1+K\beta}{K+1}\right)^2.$$

At query i, the original learner is positive exactly when it outputs h_i. Thus this sum divided by K is its actual pointwise disagreement probability. The probability completion has added no learner randomness and has not changed this final calculation. □

Corollary H.13 (An exact extra bit for the infinite point class)**.** *For either class in Theorem H.12, let $K = 2^B \geq 4$, take the accuracy parameters in that theorem, and suppose*

$$\frac{1}{2K} < \rho < \frac{c_K(\beta)}{K}.$$

Allowing any fixed finite sample size, the minimum total number of shared fair bits is B for improper learning and exactly $B+1$ for proper learning.

Proof. The band is nonempty: $K\beta < 1$ and $K \geq 4$ imply $c_K(\beta) > 1 - 8/(K+1)^2 > 1/2$. The improper upper bound follows from Theorem M.3, with radius one and negative baseline; the baseline is not required to belong to the comparator class. Proposition 3.4 excludes fewer than B bits. Theorem H.12 excludes proper learners with at most B bits.

For the remaining upper bound, set $K' = 2K$ and order the point functions by their indices, placing h_0 first when it is present. With U uniform on the K' midpoints of $[0,1]$, output the first hypothesis whose empirical risk is within $(\alpha/2)U$ of the empirical infimum. This is a finite, measurable computation: all unobserved point functions have the same empirical risk as the negative baseline, so the observed indices, the first unobserved index, and h_0 when present suffice. The empirical infimum is therefore attained.

The score calculation in Theorem G.2 applies also when h_0 is omitted. In that case the supremum of the true point gains is still nonnegative, since the gains are absolutely summable on an infinite domain. Uniform point-gain deviation at most $\alpha/4$ gives accuracy for every seed; its probability is at least $1-\beta$ for a sufficiently large fixed sample size n, by that theorem's bound $\mathbb{E}D \leq 2/\sqrt{n}$ and bounded differences. At each query, selection is determined by at most two initial-segment cuts. The two clipped risk-gap scores have replacement sensitivity $4/(n\alpha)$ and variance at most $4/(n\alpha^2)$, exactly as in the proof of Theorem G.2; an empty initial segment contributes a constant cut. Hence Lemma G.1 gives pointwise disagreement at most

$$\frac{1}{K'} + \frac{4\sqrt{2}}{\alpha\sqrt{n}}.$$

Because $\rho > 1/K'$, some finite n meets both this replication requirement and the preceding accuracy requirement. The midpoint choice uses exactly $B+1$ bits and is the only randomness. □

The class in Theorem H.12 is fixed as K varies, and the witness distributions have finite support. The accuracy parameters satisfy the displayed bounds depending on K. The limit as K grows with fixed α and β is outside this parameter regime.

H.4 An exact variational characterization

Geometric characterizations of list replicability were given by Chase et al. (2023, Theorem 7). Here we characterize the pointwise disagreement infimum at fixed seed and confidence budgets.

Let $\mathcal{H}_N$ be the N point functions on $\{1,\dots,N\}$, where $N \geq 2$, and let $K = 2^B$. For a nonempty $I \subseteq \{1,\dots,N\}$, write $\Delta_I = \{v \in \Delta_N : v_j = 0 \text{ for } j \notin I\}$. For $b \geq 0$, let $\mathcal{Q}_{N,K}(b)$ consist of the continuous maps $Q_s : \Delta_N \to \Delta_N$, $1 \leq s \leq K$, that satisfy

$$Q_s(e_i) = e_i \quad \text{and} \quad \frac{1}{K}\sum_{s=1}^{K}\sum_{j\notin I} Q_{s,j}(v) \leq b \quad (v \in \Delta_I)$$

for every nonempty proper I. The face condition bounds the total probability assigned outside the face, rather than its coordinates separately. Define

$$\mathcal{E}(Q) = \sup_{v\in\Delta_N} \max_{1\le i\le N} \frac{1}{K}\sum_{s=1}^{K} 2Q_{s,i}(v)(1-Q_{s,i}(v)), \qquad \Gamma_{N,K}(b) = \inf_{Q\in\mathcal{Q}_{N,K}(b)} \mathcal{E}(Q).$$

This family is nonempty, since all Q_s can equal the identity. The following realization lemma uses the same classical equal-probability type-class principle as above (Elias, 1972).

Lemma H.14 (Finite realization of continuous output maps)**.** *Suppose $Q \in \mathcal{Q}_{N,K}(b)$, $0 \le b < \beta < 1$, $\alpha > 0$, and $\rho > \mathcal{E}(Q)$. There is a proper agnostic (α,β)-learner for $\mathcal{H}_N$ with a fixed finite sample size, exactly B shared fair bits and no other randomness, that is ρ-pointwise replicable on every distribution.*

Proof. We first realize the maps approximately from an i.i.d. categorical sequence $Z_1,\dots,Z_m$ of arbitrary law $v \in \Delta_N$. Let $\widehat{v}$ be its empirical type, with counts $n_1,\dots,n_N$, and put

$$M = \frac{m!}{\prod_i n_i!}.$$

For each type and seed s, round the numbers $MQ_{s,i}(\widehat{v})$ to nonnegative integers $A_{s,i}$ with sum M and coordinate errors at most one. For example, take their floors and give the remaining units to the largest fractional parts, with a fixed tie rule. Order the M strings of this type lexicographically and partition their ranks into consecutive blocks of sizes $A_{s,1},\dots,A_{s,N}$. Output the index of the block containing the observed string's rank. All allocations form a finite integer lookup table.

Conditionally on a type of positive probability, its strings are equiprobable. The conditional output probabilities are therefore $A_{s,i}/M$. On a pure type the rounding is exact, by $Q_s(e_i) = e_i$. On any nonpure type, some $1 \le n_i \le m-1$, so $M \ge \binom{m}{n_i} \ge m$. Let $\widetilde{Q}_{s,i}(v)$ denote the resulting unconditional output probabilities. Define the common coordinate modulus of continuity

$$\omega(a) = \max_{s,i} \sup_{\|v-w\|_1\le a} |Q_{s,i}(v) - Q_{s,i}(w)|.$$

Compactness and continuity give $\omega(a) \to 0$ as $a \downarrow 0$. Coordinatewise Hoeffding bounds yield, for $m \ge 2$ and $a > 0$,

$$\sup_{v,s,i} |\widetilde{Q}_{s,i}(v) - Q_{s,i}(v)| \le \frac{1}{m} + \omega(a) + 2Ne^{-2ma^2/N^2}.$$

Thus the realized maps approximate Q uniformly to any prescribed coordinate error $\varepsilon > 0$ using one fixed finite m.

Choose ε with $N\varepsilon < (\beta-b)/2$ and $2\varepsilon < \rho - \mathcal{E}(Q)$. Choose $0 < \delta < 1$ so small that $N\omega(2\delta) < (\beta-b)/2$. Run a deterministic empirical risk minimizer over $\mathcal{H}_N$ on m independent blocks, each large enough to have excess risk at most α except with probability δ. For example, $\lceil 2\alpha^{-2}\log(2N/\delta)\rceil$ observations per block suffice by Hoeffding's inequality. These block outputs supply the categorical sequence above; the uniformly chosen shared seed selects its rank table.

Fix any distribution P, and let

$$I = \{i : \mathrm{R}_P(h_i) \le \mathrm{OPT}_{\mathcal{H}_N}(P) + \alpha\}.$$

The set I is nonempty. If v is the law of one block output, then $v(I^c) \le \delta$. When I is proper, normalize v on I to obtain $v' \in \Delta_I$; then $\|v - v'\|_1 = 2v(I^c) \le 2\delta$. The mean probability of an inaccurate final output is at most

$$\frac{1}{K}\sum_s \sum_{i\notin I} \widetilde{Q}_{s,i}(v) \le b + N\omega(2\delta) + N\varepsilon < \beta.$$

When I is the full set, accuracy is automatic. At query i, two independent executions with the same seed disagree with probability

$$\frac{1}{K}\sum_s 2\widetilde{Q}_{s,i}(v)(1-\widetilde{Q}_{s,i}(v)) \le \mathcal{E}(Q) + 2\varepsilon < \rho,$$

since $u \mapsto 2u(1-u)$ is 2-Lipschitz on $[0,1]$. The block rules, type ranks and lookup tables are all deterministic functions of the sample. Their fixed total sample size and the B shared bits therefore have the asserted properties. □

Theorem H.15 (The proper replication profile). *Let $N \geq 2$, $K = 2^B \geq 1$, and*

$$0 < \alpha < \frac{1}{N-1}, \qquad 0 < \beta < \frac{1}{K}.$$

Let $\rho^*_{\mathrm{prop}}(N, K; \alpha, \beta)$ be the infimum of achievable pointwise replication parameters over proper agnostic (α, β)-learners for $\mathcal{H}_N$, using any fixed finite sample size, at most B total shared fair bits and no other randomness, with guarantees on every distribution. Then

$$\rho^*_{\mathrm{prop}}(N, K; \alpha, \beta) = \Gamma_{N,K}(\beta). \tag{65}$$

In particular this infimum is independent of α throughout the displayed range.

Proof. For necessity, pad the learner's bits and apply it to $P_v(X = i, Y = +1) = v_i$, $v \in \Delta_N$. Its seedwise output probabilities define continuous maps $Q_s : \Delta_N \to \Delta_N$, since the sample space is finite and the sample size is fixed. At e_i the sample is deterministic and every output other than h_i has excess risk one. Since $\beta < 1/K$, each seed must output h_i there. On a nonempty proper face Δ_I, every h_j with $j \notin I$ has excess risk

$$\max_{i \in I} v_i \geq \frac{1}{|I|} \geq \frac{1}{N-1} > \alpha.$$

Accuracy therefore supplies the full face condition with budget β. Pointwise replication bounds $\mathcal{E}(Q)$ by the learner's replication parameter, proving the lower bound in (65).

For the reverse bound, take any $Q \in \mathcal{Q}_{N,K}(\beta)$ and $0 < t < 1$. The maps

$$Q^t_s(v) = (1-t)Q_s(v) + tv$$

belong to $\mathcal{Q}_{N,K}((1-t)\beta)$ and satisfy $\mathcal{E}(Q^t) \leq \mathcal{E}(Q) + 2t$. Lemma H.14 realizes every replication parameter strictly above this value with accuracy β and the prescribed bits. Take $\mathcal{E}(Q)$ arbitrarily close to $\Gamma_{N,K}(\beta)$ and then let $t \downarrow 0$. This proves the reverse inequality. □

This is an equality of infima. Existence of a minimizing map, finite-sample attainment, and optimal sample size require separate arguments. Evaluating $\Gamma_{N,K}$ is a geometric optimization problem.

Identity mixing also gives, for $0 < b_1 < b_2$,

$$0 \leq \Gamma_{N,K}(b_1) - \Gamma_{N,K}(b_2) \leq 2(1 - b_1/b_2).$$

Thus the profile is locally Lipschitz on positive confidence budgets.

H.5 Four seed maps below the 1/4 threshold

Lemma H.16 (Four continuous seed maps). *For every $0 < \varepsilon \leq 1/6$, there are maps $Q \in \mathcal{Q}_{3,4}(\varepsilon)$ such that*

$$\mathcal{E}(Q) \leq \frac{1}{4} - \frac{\varepsilon^2}{400}. \tag{66}$$

More precisely, put $\delta = 1/100$ and let the rows of the threshold matrix be

$$(t_{s,i})_{s,i} = \begin{pmatrix} 1/5 & 17/20 & 19/20 \\ 19/20 & 1/5 & 17/20 \\ 17/20 & 19/20 & 1/5 \\ 11/20 & 11/20 & 11/20 \end{pmatrix}. \tag{67}$$

The maps can be chosen to have, for every s, i, v,

$$Q_{s,i}(v) \leq \tfrac{1}{2} \iff v_i \leq t_{s,i}, \tag{68}$$

$$|Q_{s,i}(v) - \tfrac{1}{2}| \geq \varepsilon |v_i - t_{s,i}|, \tag{69}$$

$$v_i \geq t_{s,i} + \delta \implies Q_s(v) = e_i. \tag{70}$$

At each v, at most two rows $Q_s(v)$ are not vertices. Each row assigns mass at most ε to the missing coordinate on every edge of Δ_3.

Proof. Fix a row s and temporarily write $t_i = t_{s,i}$. Every pair of its thresholds has sum greater than one. Consequently

$$C_s = \{v \in \Delta_3 : v_i \le t_i \text{ for all } i\}$$

is a hexagon with vertices $v^{ij} = t_i e_i + (1 - t_i)e_j$, $i \neq j$. Its three inner edges lie on $v_i = t_i$, and its three remaining edges lie on the edges of Δ_3. The point

$$c_s = \frac{t_s}{\sum_i t_{s,i}}$$

lies strictly inside C_s. Define $B_s : C_s \to \Delta_3$ at its six vertices by

$$B_s(v^{ij}) = \tfrac{1}{2}e_i + (\tfrac{1}{2} - \varepsilon)e_j + \varepsilon e_k, \qquad \{i, j, k\} = \{1, 2, 3\},$$

and set $B_s(c_s) = (1/3, 1/3, 1/3)$. Extend affinely on the six triangles joining c_s to the hexagon's edges. All coordinates of B_s are at most $1/2$. On the inner edge $v_i = t_i$, its ith coordinate is $1/2$; on the outer edge $v_k = 0$, its kth coordinate is ε. On any other inner edge $v_j = t_j$, its ith coordinate is at most $1/2 - \varepsilon$.

Outside C_s there is a unique i with $v_i > t_i$. For $v \neq e_i$, define its projection onto the inner edge and its interpolation weight by

$$\pi_i(v) = t_i e_i + \frac{1 - t_i}{1 - v_i} \sum_{j \neq i} v_j e_j, \qquad p_i(v) = \min\{(v_i - t_i)/\delta, 1\}.$$

The projection belongs to C_s, since each of its other coordinates is at most $1 - t_i < t_j$. Set

$$Q_s(v) = \begin{cases} B_s(v), & v \in C_s, \\ (1 - p_i(v))B_s(\pi_i(v)) + p_i(v)e_i, & v_i > t_i, \end{cases} \qquad Q_s(e_i) = e_i.$$

The definitions agree on every inner edge. Since $t_i + \delta < 1$, the map is already constant e_i near that vertex, so Q_s is continuous everywhere. The asserted sign rule and purity follow directly. On an outer edge with missing coordinate k, the projection has that same missing coordinate; hence $Q_{s,k}$ equals ε inside C_s and $(1 - p_i)\varepsilon$ in a corner. Thus Q has the required face budget and preserves each vertex.

To prove (69) on C_s, it suffices to check the affine inequality

$$\tfrac{1}{2} - B_{s,i}(v) - \varepsilon(t_i - v_i) \ge 0$$

at the vertices of each fan triangle. At c_s the left side is at least $1/6 - \varepsilon$. At a hexagon vertex on $v_i = t_i$ it is zero. At every other hexagon vertex, $B_{s,i}$ is either $1/2 - \varepsilon$ or ε; the left side is then at least zero or $1/2 - 2\varepsilon$, respectively. In the ith corner, $Q_{s,i} - 1/2 = p_i/2 \ge (v_i - t_i)/2$. In a different corner j, one has $v_i < 1 - t_j < t_i$ and $Q_{s,i} \le 1/2 - \varepsilon$. These facts prove the separation inequality everywhere.

A nonvertex row must satisfy $v_i < t_{s,i} + \delta$ for every i. For any three rows, the sum of their coordinatewise minimum thresholds is $3/5$ if they are rows $1, 2, 3$, and $19/20$ otherwise. Even after adding 3δ, this sum is less than one. Thus no point belongs to three enlarged cores, and at most two rows are nonvertices.

Fix v and choose two distinct rows s, t containing all its nonvertex rows, adding arbitrary rows if necessary. Every other row has zero Bernoulli disagreement at every coordinate. The separation inequality and $g(q) = 2q(1 - q) = 1/2 - 2(q - 1/2)^2$ give

$$\begin{aligned} \sum_{r=1}^{4} g(Q_{r,i}(v)) &\le 1 - 2\varepsilon^2\big((v_i - t_{s,i})^2 + (v_i - t_{t,i})^2\big) \\ &\le 1 - \varepsilon^2(t_{s,i} - t_{t,i})^2 \le 1 - \varepsilon^2/100. \end{aligned}$$

The last step uses the minimum separation $1/10$ between distinct entries in each column of (67). Dividing by four proves (66). □

Theorem H.17 (The two-bit confidence law)**.** *For $0 < \alpha < 1/2$ and $0 < \beta < 1/8$, the proper replication infimum for $\mathcal{H}_3$ with at most two shared fair bits satisfies*

$$\frac{1}{4} - 2\beta^2 \le \rho^*_{\text{prop}}(3, 4; \alpha, \beta) = \Gamma_{3,4}(\beta) \le \frac{1}{4} - \frac{\beta^2}{1600}. \tag{71}$$

In particular a proper agnostic (α, β)-learner with a fixed finite sample size and exactly two shared fair bits achieves $\rho = 1/4 - \beta^2/3200$ on every distribution, with no other randomness.

Proof. The lower bound is Theorem H.7 with $K = 4$. The equality is Theorem H.15. For the upper bound, apply Lemma H.16 with $\varepsilon = \beta/2$. Its face budget is strictly below β and its energy is at most $1/4 - \beta^2/1600$. Lemma H.14 realizes every strictly larger replication parameter with a fixed finite sample size, proving both assertions. □

H.6 Optimal sample order with one or two shared bits

Theorem H.18 (Optimal samples for small proper seed budgets)**.** *There is a universal constant C such that, for $\mathcal{H}_3$, $B \in \{1, 2\}$ and $0 < \alpha, \beta \le 1/16$, a proper agnostic (α, β)-learner using exactly B shared fair bits and no other randomness achieves*

$$\rho = 2^{-B} - \frac{\beta^2}{40000} \quad \textit{with} \quad n = \left\lceil C\alpha^{-2}\log(1/\beta)\right\rceil \tag{72}$$

on every distribution. Conversely, every agnostic (α, β)-learner for this class requires $\Omega(\alpha^{-2}\log(1/\beta))$ examples, even without a replication requirement or a bound on independent randomness. Thus the displayed sample order is optimal.

Proof. Put $K = 2^B$ and $\varepsilon = \beta/2$. Use all four maps from Lemma H.16 when $K = 4$, and its first two maps when $K = 2$. Their rowwise face budgets are at most ε. We realize these maps directly from the type of the original sample, rather than from independent blocks of empirical risk minimizers.

For a risk vector $r = (r_1, r_2, r_3)$, define

$$o_i = \frac{(i-1)\alpha}{8}, \qquad \tau = \frac{\alpha}{16}, \qquad c_i = r_i + o_i, \qquad b = \min_i c_i,$$
$$w_i = (b + \tau - c_i)_+, \qquad \Psi_i(r) = \frac{w_i}{\sum_j w_j}.$$

The denominator is at least τ. Write $\widehat{R}_i$ for the empirical risk of h_i, and set $\widehat{v} = \Psi(\widehat{R})$. Every index in the support of $\widehat{v}$ satisfies

$$\widehat{R}_i \le \min_j \widehat{R}_j + \frac{5\alpha}{16}. \tag{73}$$

Indeed, the offsets range from zero to $\alpha/4$, and $w_i > 0$ implies $c_i < b + \tau$.

The sample alphabet consists of the six pairs $(x, y) \in \{1, 2, 3\} \times \{\pm 1\}$. For its empirical counts $(n_z)_z$, let $M = n!/\prod_z n_z!$. For each type and shared seed s, allocate its M lexicographically ordered strings to the three outputs by largest-remainder rounding of $MQ_{s,i}(\widehat{v})$, as in Lemma H.14. The rule outputs the class assigned to the observed string's rank. Its conditional output probabilities differ from $Q_{s,i}(\widehat{v})$ by at most $1/M$, and are exact whenever $Q_s(\widehat{v})$ is a vertex. This is a deterministic finite table for each seed.

Here the rounding error is exponentially small uniformly over types. At each of the six point-mass distributions, the smallest adjusted risk $R_i + o_i$ is unique and is separated from the others by at least $\alpha/8$. For a positive label this follows from the risk gap one; for a negative label the two zero-risk point functions are separated by their distinct offsets. If a type puts mass at least $1 - \alpha/32$ on one symbol, every risk changes from that point-mass risk by at most $\alpha/32$. The adjusted minimum therefore retains a gap at least $\alpha/16 = \tau$, making $\widehat{v}$ a vertex. All seed maps and their rounded versions are then pure. Otherwise, if k is the largest raw count,

$$\frac{n}{6} \le k < n(1 - \alpha/32), \qquad M \ge \binom{n}{k} \ge 2^{\min\{k, n-k\}} \ge 2^{n\alpha/32}.$$

The middle inequality follows by using the smaller of $k, n-k$ in the product formula for the binomial coefficient. Consequently every conditional coordinate rounding error is at most

$$e_n = 2^{-n\alpha/32}. \tag{74}$$

Fix a distribution P, put $R_i = \mathrm{R}_P(h_i)$ and $v = \Psi(R)$, and let $q_{s,i}$ denote the actual, unconditional output probabilities of the rank rule. If $\|r - r'\|_\infty \le u$, the corresponding minima differ by at most u, each weight by at most $2u$, and the weight sums by at most $6u$. Since both sums are at least τ,

$$\|\Psi(r) - \Psi(r')\|_\infty \le \frac{8u}{\tau} = \frac{128u}{\alpha}.$$

Take $\gamma = 1/1000$ and define

$$G = \left\{ \max_i |\widehat{R}_i - R_i| \le \frac{\alpha\gamma}{128} \right\}, \qquad p = \mathbb{P}(G^c) \le 6 \exp\left(-\frac{n\alpha^2\gamma^2}{8192} \right).$$

On G, $\|\widehat{v} - v\|_\infty \le \gamma$. Increase the universal constant C in the sample size so that

$$z := p + e_n \le \frac{\varepsilon^2}{10000}. \tag{75}$$

Such a constant exists uniformly for the stated parameters: both error bounds decay as $\exp(-cn\alpha^2)$ for a universal $c > 0$, using $\alpha \le 1$.

Let $\delta = 1/100$ and retain the thresholds from (67). The set of rows that can be nonpure on G is contained in

$$J(v) = \{s \le K : v_i \le t_{s,i} + \delta + \gamma \text{ for all } i\}.$$

It has at most two members: for any three of the four rows, the sum of the coordinatewise minimum thresholds, after this enlargement, is at most $19/20 + 3(\delta+\gamma) = 983/1000 < 1$. For $s \notin J(v)$, some fixed coordinate j satisfies $v_j > t_{s,j} + \delta + \gamma$. On every type in G, (70) therefore gives $Q_s(\widehat{v}) = e_j$. Rounding is exact on such a row, so its actual prediction at any query has disagreement probability at most $2p$.

Suppose first that $J(v)$ has two members and fix a query i. Their ith thresholds are separated by at least $1/10$, so one of these rows, say s, obeys $|v_i - t_{s,i}| \ge 1/20$. On G its half-probability deviation has a fixed sign, by (68), and (69) gives

$$|Q_{s,i}(\widehat{v}) - 1/2| \ge \varepsilon(1/20 - \gamma) \ge \varepsilon/25.$$

The bad types contribute at most p to the loss of this signed margin, and rounding contributes at most e_n. Thus (75) implies

$$|q_{s,i} - 1/2| \ge \varepsilon/25 - z \ge \varepsilon/50.$$

This row contributes at most $1/2 - \varepsilon^2/1250$ to the disagreement sum; the other row contributes at most $1/2$. There are at most two remaining rows, each contributing at most $2p$. Hence

$$\sum_{s=1}^{K} 2q_{s,i}(1 - q_{s,i}) \le 1 - \frac{\varepsilon^2}{1250} + 4p \le 1 - \frac{\varepsilon^2}{2500}.$$

If $|J(v)| \le 1$, the same bound follows from the smaller upper bound $1/2 + 6p$. After averaging over the K shared seeds, the pointwise disagreement is at most

$$\frac{1}{K} - \frac{\varepsilon^2}{2500K} \le \frac{1}{K} - \frac{\beta^2}{40000}.$$

This argument applies to every P and every query, without adding an accuracy-failure term to replication.

For accuracy, G implies uniform empirical risk error at most $\alpha/4$. By (73), every index in the support I of $\widehat{v}$ then has true excess risk at most $13\alpha/16 < \alpha$. If $|I| = 2$, each seed map gives its missing index mass at most ε; if $|I| = 1$, vertex preservation makes that mass zero; and if $|I| = 3$, no output is inaccurate on G. The conditional rank allocation increases the total probability outside I by at most $3e_n$. The accuracy failure probability is consequently at most $\varepsilon + 3e_n + p \le \beta$. The entire learner uses the original n examples and only its uniformly chosen B-bit seed.

Finally, consider the two distributions with $X = 1$ almost surely and positive-label probabilities $1/2 + \alpha$ and $1/2 - \alpha$. An incorrect prediction at point 1 has excess risk $2\alpha > \alpha$ under the corresponding distribution. Thus the event that the learned classifier predicts positive at 1 is a test with both error probabilities at most β, even if improper outputs are allowed. Data processing, including any independent random seed, gives

$$n \,\mathrm{kl}(1/2 + \alpha, 1/2 - \alpha) \ge \tfrac{1}{2} \log(1/\beta).$$

For $\alpha \le 1/16$, the Bernoulli relative entropy on the left is at most $16\alpha^2$. This proves the required sample lower bound. □

H.7 AN EXACT HIGH-CONFIDENCE BOUNDARY FOR THREE POINT FUNCTIONS

The two-bit construction cannot be extended to arbitrarily many seed rows while keeping every column's disagreement sum below one. A finite cellular-cohomology argument gives the five-row obstruction. All cochains in that argument have coefficients in $\mathbb{F}_2$, so orientations do not affect their values. A cellular k-cochain assigns a value to each k-cell; its coboundary on a $(k+1)$-cell is the sum of its values on the boundary k-cells. A cocycle has zero coboundary. We use the standard facts that a cocycle on a contractible cell complex is a coboundary in positive degree, and that a cocycle evaluates to zero on the boundary of a cellular chain (Hatcher, 2002, Chapter 3).

Let $Y = \{te_i : 0 \le t \le 1,\ i \in \{1,2,3\}\}$ be a three-arm star, with center $o = 0$ and leaves e_1, e_2, e_3. Define a cubical complex $L \subseteq Y^5$ as follows: at most two coordinates may be nonleaves, and two coordinates on open arms must have different arm labels. Thus its two-dimensional cells are the squares with two different arm labels and three fixed leaf coordinates. For distinct labels a, b, let

$$Y_{ab} = [o, e_a] \cup [o, e_b], \qquad M_{ab} = L \cap Y^5_{ab}, \qquad p_i = (e_i, e_i, e_i, e_i, e_i).$$

Let γ_{ab} be the edge path from p_a to p_b that changes coordinates $1, \dots, 5$ in that order, moving each coordinate from e_a through o to e_b.

Lemma H.19 (A relative obstruction in the five-fold star complex)**.** *There is a cellular one-cocycle φ on L whose restriction to each M_{ab} is a coboundary and such that*

$$\langle \varphi, \gamma_{12} + \gamma_{23} + \gamma_{31} \rangle = 1.$$

Consequently there is no continuous map $F : \Delta_3 \to L$ that sends e_i to p_i and sends each edge $[e_a, e_b]$ into M_{ab}.

Proof. Use the product cell structure on $P = (\Delta_3)^5$, with the three leaves as the vertices of each simplex. This auxiliary complex records the leaf labels; the cochain on $L \subseteq Y^5$ will be constructed separately below. Its two-cells are single-coordinate triangles and products of two coordinate edges. For a triple σ of pure labels, put $R(\sigma) = 1$ if its labels are $1, 2, 3$ in some order, and put $R(\sigma) = 0$ otherwise. Define a two-cochain ω on P to be zero on every single-coordinate triangle. On a square whose two coordinate edges have endpoint-label sets $A, B \subseteq \{1, 2, 3\}$, define

$$\omega(\text{square}) = |A \cap B|\, R(\sigma) \pmod 2,$$

where σ records the other three leaf coordinates.

We verify that $d\omega = 0$ on every three-cell. For a product of a triangle and an edge, each color occurs twice among the triangle's three edges, so the square-face values sum to zero; its two triangle faces also have value zero. For a product of three edges, fix the other two leaf colors. If they coincide, every square-face value is zero. Otherwise let ℓ be their missing color, and let $u, v, w \in \mathbb{F}_2^3$ be the indicator vectors of the endpoint-label sets of the three varying edges. The sum of the six square-face values is

$$\langle u, v \rangle w_\ell + \langle u, w \rangle v_\ell + \langle v, w \rangle u_\ell = 0.$$

Indeed, u, v, w lie in the two-dimensional even-parity subspace of $\mathbb{F}_2^3$. The dot product restricts to an alternating bilinear form there, and $\langle u, v \rangle w + \langle u, w \rangle v + \langle v, w \rangle u = 0$ in that two-dimensional space. These are all types of three-cells of P.

Since P is contractible, there is a cellular one-cochain h with $dh = \omega$. On a single-coordinate triangle $dh = 0$. Consequently, for each coordinate r and fixed pure background, there are values $\lambda_r(i)$, $i = 1, 2, 3$, such that

$$h(\text{coordinate } r \text{ edge from } i \text{ to } j) = \lambda_r(i) + \lambda_r(j).$$

The values depend on the fixed background; this dependence is left implicit. Assign the value $\lambda_r(i)$ to the corresponding edge of L from a one-center vertex to the pure vertex with coordinate r equal to e_i.

It remains to assign values to edges incident to two-center vertices. Fix coordinates $r < t$ and their three pure background labels σ, and write $R = R(\sigma)$. Define

$$D_{ij} = \lambda_r(i;\, t = j) + \lambda_t(j;\, r = i).$$

The notation specifies the additional pure coordinate in each background. Applying $dh = \omega$ to the rectangle with endpoint sets $\{i, i'\}$ and $\{j, j'\}$ shows that the matrix

$$E_{ij} = D_{ij} + R\mathbf{1}\{i = j\}$$

has zero 2-by-2 rectangular sums. Thus $E_{ij} = a_i + b_j$; for example one may take $a_i = E_{i1}$ and $b_j = E_{1j} + E_{11}$. Assign a_i to the edge that changes coordinate r from o to e_i while coordinate t remains o, and assign b_j to the edge changing coordinate t while r remains o. Do this for each coordinate pair and pure background. Together with the preceding assignments this defines φ on every edge of L. The sum around the arm square with labels i, j is

$$D_{ij} + a_i + b_j = R\mathbf{1}\{i = j\}.$$

Every square present in L has $i \neq j$, so $d\varphi = 0$. The same identity also describes the four-edge boundary of a same-label square, although that square is not a cell of L.

To prove exactness on M_{ab}, first restrict h to the ordinary binary cube $[e_a, e_b]^5 \subseteq P$. Every square there has $R = 0$, hence $h = dU$ on its pure vertices and edges. Extend U to a one-center vertex by $U(\text{pure vertex with label } i) + \lambda_r(i)$, where $i \in \{a, b\}$. This does not depend on i, because $h(a\text{–}b) = \lambda_r(a) + \lambda_r(b)$. Extend it to a two-center vertex by adding the assigned a_i or b_j to the value at an adjacent one-center vertex. Here the fixed three-label background uses only a, b, so $R = 0$. The displayed square-boundary identity, including $i = j$, shows that all choices give the same value. These extensions satisfy $\varphi = dU$ on every edge of M_{ab}.

Finally evaluate φ on the three canonical paths. Replacing each two-arm passage by its ordinary simplex edge preserves the value, by the definition of the λ_r. Write A_r, B_r, C_r for the updates $1 \to 2$, $2 \to 3$, and $3 \to 1$ in coordinate r. The resulting loop in P has word

$$A_1 \cdots A_5 \, B_1 \cdots B_5 \, C_1 \cdots C_5.$$

Interchanging adjacent updates in distinct coordinates changes its h-value by ω on their product square. First move each B_s left across the A_t with $t > s$, obtaining $A_1 B_1 \cdots A_5 B_5$ before the C-updates. At the interchange of B_s and A_t, the three background coordinates have $s - 1$ labels equal to 3, $t - s - 1$ equal to 2, and $5 - t$ equal to 1. They are all different exactly when $(s, t) = (2, 4)$. That interchange has $|\{1, 2\} \cap \{2, 3\}| = 1$. Next move each C_s left across the later $A_t B_t$ blocks. At every such interchange the background uses only 1 and 3, so its ω-value is zero. The final word is the concatenation of the five single-coordinate triangles $A_r B_r C_r$, each with h-value zero. The original loop therefore has value one.

Any other edge path in M_{ab} with the same endpoints has the same φ-value, since φ is exact there. Cellular approximation relative to the three vertices now shows that the boundary of any proposed disk map would have value one. A cocycle evaluates to zero on a boundary, giving the asserted contradiction. □

Lemma H.20 (Five probability rows cannot stay below one)**.** *Let $Q_s : \Delta_3 \to \Delta_3$, $1 \le s \le 5$, be continuous, and suppose that for some $0 \le \eta < 1/2$,*

$$Q_{s,i}(e_i) \ge 1 - \eta, \qquad Q_{s,i}(v) \le \eta \qquad \textit{whenever } v_i = 0.$$

Then some $v \in \Delta_3$ and coordinate i satisfy

$$\sum_{s=1}^{5} 2Q_{s,i}(v)(1 - Q_{s,i}(v)) \ge 1.$$

Proof. Put $g(q) = 2q(1 - q)$ and suppose the conclusion fails. By compactness there is a $\gamma > 0$ such that every column sum is at most $1 - \gamma$. Choose

$$0 < \delta < \min\{1/6,\ 1/2 - \eta,\ \sqrt{\gamma}/2\}.$$

For a probability vector $q \in \Delta_3$, define

$$a_i(q) = (q_i - 1/2 + \delta)_+, \qquad a_0(q) = (1/2 + \delta - \max_i q_i)_+, \qquad d(q) = a_0(q) + \sum_i a_i(q),$$

and define a continuous map $T : \Delta_3 \to Y$ by

$$T_i(q) = \frac{(a_i(q) - \max_{j \neq i} a_j(q))_+}{d(q)}.$$

The denominator is at least 2δ. At most one numerator is positive, so the image lies in Y. Moreover $T(q)$ is a nonleaf exactly when $\max_i q_i < 1/2 + \delta$. In that case $a_0 > 0$, so its arm length is less than one. In the complementary case only the largest a_i is nonzero and $a_0 = 0$, giving the leaf e_i. If $T(q)$ lies on open arm i, then also $q_i > 1/2 - \delta$.

A probability vector with all coordinates at most $1/2 + \delta$ satisfies

$$\sum_i g(q_i) = 2\Big(1 - \sum_i q_i^2\Big) \geq 1 - 4\delta^2.$$

Here the sum of squares is maximized by $(1/2+\delta, 1/2-\delta, 0)$, up to permutation. Thus three nonleaf rows would have total disagreement at least $3 - 12\delta^2 > 3(1 - \gamma)$, exceeding the sum of the three column bounds. Two rows on the same open arm i would both have $q_i \in (1/2 - \delta, 1/2 + \delta)$, giving column sum greater than $1 - 4\delta^2 > 1 - \gamma$. Consequently

$$v \longmapsto \big(T(Q_1(v)), \ldots, T(Q_5(v))\big)$$

is a continuous map into L. At e_i, the inequality $Q_{s,i}(e_i) \geq 1 - \eta > 1/2 + \delta$ gives $T(Q_s(e_i)) = e_i$, so it preserves the required pure vertices. On the domain edge with missing coordinate i, $Q_{s,i} \leq \eta < 1/2 - \delta$, so $a_i = 0$ and $T_i = 0$. That edge is therefore mapped into the corresponding M_{ab}. Lemma H.19 rules out this map. □

Theorem H.21 (The exact high-confidence proper boundary)**.** *Let $\mathcal{H}_3$ consist of exactly the three point functions on $\{1, 2, 3\}$. For $K = 2^B \geq 8$, $0 < \alpha < 1/2$, and $0 < \beta < 1/(2K)$, every proper agnostic (α, β)-learner using a fixed finite sample size, at most B total shared fair bits and no other randomness, and pointwise ρ-replicable on every distribution satisfies*

$$\rho > \frac{1}{K}.$$

Conversely every $\rho > 1/K$ is achievable with exactly B shared fair bits and a fixed finite sample size. In particular,

$$\rho^*_{\mathrm{prop}}(3, K; \alpha, \beta) = \Gamma_{3,K}(\beta) = \frac{1}{K}, \tag{76}$$

and no finite-sample learner attains this infimum.

Proof. Put $g(q) = 2q(1 - q)$ and pad unused bits to obtain K deterministic seed rules. For $P_\theta(X = i, Y = +1) = \theta_i$, write

$$q_{s,i}(\theta) = \mathbb{P}\{A(S; s) = h_i\}, \qquad G_i(\theta) = \sum_{s=1}^{K} g(q_{s,i}(\theta)).$$

The rows are continuous probability vectors. At each distribution vertex, an incorrect proper output has excess risk one. Since $\beta < 1/K$, every seed outputs h_i at e_i. These distinct vertex outputs also imply that the sample size n is positive. On the opposite face $\theta_i = 0$, output h_i has excess risk at least $1/2 > \alpha$, so

$$\sum_{s=1}^{K} q_{s,i}(\theta) \leq K\beta < 1/2.$$

Suppose $\rho \leq 1/K$, so $G_i(\theta) \leq 1$ everywhere. Select any five seed rows. At a distribution vertex their disagreement sums are zero. At any other θ, its support has at least two points. If i is in that support, choose another supported point j. For every seed, the all-i and all-j sample strings imply

$$0 < \theta_i^n \leq q_{s,i}(\theta) \leq 1 - \theta_j^n < 1.$$

There is at least one unselected seed, whose disagreement at i is strictly positive. The five selected rows therefore have column sum strictly less than $G_i(\theta) \le 1$. If i is not in the support, the preceding face bound instead gives

$$\sum_{s \text{ selected}} g(q_{s,i}(\theta)) \le 2 \sum_{s=1}^{K} q_{s,i}(\theta) \le 2K\beta < 1.$$

Thus every column sum of the five selected rows is strictly less than one on the entire simplex. Their face coordinates are bounded by $\eta = K\beta < 1/2$ and they preserve every vertex, contradicting Lemma H.20. This argument also excludes the endpoint $\rho = 1/K$.

For the upper bound put $\alpha_0 = \min\{\alpha, 1/16\}$ and use the ordered proper selector from the proof of Corollary H.4, now with exactly K grid points. It selects the first of h_1, h_2, h_3 whose empirical risk is within $(\alpha_0/2)U$ of the empirical minimum, where U is uniform on the K midpoints. For $n\alpha_0^2/K^2 \ge 8$, that proof bounds pointwise disagreement by

$$\frac{1}{K} + \frac{32}{K} \exp\left(-\frac{n\alpha_0^2}{32K^2}\right).$$

Uniform empirical-risk deviation at most $\alpha_0/4$ gives accuracy for every seed, with failure probability at most $6e^{-n\alpha_0^2/8}$. Any prescribed $\rho > 1/K$ and $\beta > 0$ are therefore met by some fixed finite n, using only the B shared bits. The equality with $\Gamma_{3,K}$ follows from Theorem H.15. □

Corollary H.22 (The exact high-confidence proper bit count)**.** *For $\mathcal{H}_3$, $0 < \alpha \le 1/16$, $0 < \rho \le 1/8$, and $0 < \beta < \rho/2$, allow any fixed finite sample size and no randomness other than shared fair bits. The minimum total number of bits is*

$$\left\lfloor \log_2 \frac{1}{\rho} \right\rfloor + 1 \quad \textit{for proper learners}, \qquad \left\lfloor \log_2 \frac{1}{\rho} \right\rfloor \quad \textit{when improper outputs are allowed.}$$

Proof. Put $b = \lfloor \log_2(1/\rho) \rfloor \ge 3$ and $K = 2^b$. Then

$$\frac{1}{2K} < \rho \le \frac{1}{K}, \qquad \beta < \frac{\rho}{2} \le \frac{1}{2K}.$$

Theorem H.21 excludes every proper learner with at most b bits. The ordered selector in its proof with $2K$ grid points achieves the target using $b+1$ bits, since $\rho > 1/(2K)$. For improper learning, the radius-one construction in Theorem M.3 supplies b bits, and Proposition 3.4 excludes fewer. □

Theorem H.23 (The joint sample tradeoff at the proper boundary)**.** *Let $\mathcal{H}_3$ consist of exactly the three point functions on $\{1,2,3\}$. Suppose*

$$K = 2^B \ge 8, \qquad 0 < \alpha \le 1/16, \qquad 0 < \beta < \frac{1}{2K}, \qquad \rho = \frac{1}{K} + \Delta, \qquad 0 < \delta := K\Delta \le \frac{1}{4}.$$

Let $m^*_{\text{prop}}(\alpha, \beta, \rho; B)$ be the least fixed sample size of a proper agnostic (α, β)-learner for $\mathcal{H}_3$ that is pointwise ρ-replicable on every distribution and uses at most B total shared fair bits and no other randomness. Then, with absolute implicit constants,

$$m^*_{\text{prop}}(\alpha, \beta, \rho; B) = \Theta\left(\frac{K^2 \log(1/\delta) + \log(1/\beta)}{\alpha^2}\right). \tag{77}$$

The upper bound uses exactly B shared fair bits.

Proof. We first prove the explicit boundary lower bound

$$n \ge \frac{K^2 \log(1/\delta)}{2^{22}\alpha^2} \tag{78}$$

for every learner in the statement. Pad unused bits to obtain K deterministic seed rules, put $\eta = K\beta < 1/2$, and write $g(q) = 2q(1-q)$.

A localized family and its five-row witnesses. Let $u = (1/3, 1/3, 1/3)$, and for $\theta \in \Delta_3$ let P_θ put positive labels on all its mass, with feature probability vector

$$p(\theta) = (1 - 4\alpha)u + 4\alpha\theta.$$

Every coordinate of $p(\theta)$ is at least $1/4$. At the parameter vertex e_i, the risk of h_i is smaller than either other risk by 4α. On the opposite face $\theta_i = 0$, the excess risk of h_i is at least 2α. Writing

$$q_{s,i}(\theta) = \mathbb{P}_{Z\sim P_\theta^n}\{A(Z;s) = h_i\}, \qquad G_i(\theta) = \sum_{s=1}^{K} g(q_{s,i}(\theta)),$$

accuracy therefore implies, for every seed s,

$$q_{s,i}(e_i) \geq 1 - \eta, \qquad q_{s,i}(\theta) \leq \eta \quad \text{if } \theta_i = 0.$$

The rows q_s are continuous probability vectors: each coordinate is a polynomial in the probabilities of a finite sample alphabet. Replication gives $G_i(\theta) \leq 1 + \delta$. For every five-element set S of seeds, Lemma H.20 supplies a witness (θ_S, i_S) satisfying

$$\sum_{s\in S} g(q_{s,i_S}(\theta_S)) \geq 1, \qquad \sum_{s\notin S} g(q_{s,i_S}(\theta_S)) \leq \delta. \tag{79}$$

Many witnesses when $K \geq 64$. For now suppose $0 < \delta \leq 2^{-12}$ and put $L = \log(1/\delta)$. Let $q = K/8$, a power of two at least eight, and choose five distinct elements $x_1, \ldots, x_5$ of the finite field $\mathbb{F}_q$. Identify $5q$ of the seeds with $\{1, \ldots, 5\} \times \mathbb{F}_q$. For every $(a, b) \in \mathbb{F}_q^2$ form the set

$$S_{a,b} = \{(j, ax_j + b) : 1 \leq j \leq 5\}.$$

These $q^2 = K^2/64$ five-element sets are distinct and any two intersect in at most one seed, since two distinct affine functions agree at at most one field element. Choose one witness for each set. At least $M \geq K^2/192$ of these witnesses have the same query i.

Consider two distinct sets S, T in that group. At θ_S, if they are disjoint, some seed in S has $g(q_{s,i}) \geq 1/5$. If they intersect, their common seed contributes at most $1/2$, so some seed in $S \setminus T$ has $g(q_{s,i}) \geq 1/8$. Thus in either case there is an $r \in S \setminus T$ for which both $q_{r,i}(\theta_S)$ and $1 - q_{r,i}(\theta_S)$ are at least $1/16$. At θ_T, (79) instead gives $g(q_{r,i}(\theta_T)) \leq \delta$, and hence

$$\min\{q_{r,i}(\theta_T), 1 - q_{r,i}(\theta_T)\} \leq \delta.$$

For a sample Z, choose the deterministic event E to be either $\{A(Z; r) = h_i\}$ or its complement so that its probability under $P_{\theta_T}^n$ is at most δ. Its probability under $P_{\theta_S}^n$ is at least $1/16$. Binary data processing, and monotonicity of binary relative entropy in these two ranges, give

$$\begin{aligned} n\,\mathrm{KL}(P_{\theta_S}\|P_{\theta_T}) &\geq \mathrm{kl}(1/16, \delta) \\ &\geq \frac{L}{16} - h(1/16) \geq \frac{L}{32}, \end{aligned} \tag{80}$$

where h is binary entropy with natural logarithms. The last step uses $h(1/16) < 1/4$ and $L \geq 12 \log 2 > 8$.

For any two feature probability vectors in this family,

$$\mathrm{KL}(p\|p') \leq \sum_i \frac{(p_i - p'_i)^2}{p'_i} \leq 4\|p - p'\|_2^2.$$

Consequently the M vectors $p(\theta_S)$ at same-query witnesses have pairwise Euclidean distance at least $d = \sqrt{L/(128n)}$. They lie in the two-dimensional affine plane of the simplex, inside the disk centered at u of radius $R = 4\alpha\sqrt{2/3}$. Disjoint open disks of radius $d/2$ around the witnesses fit inside the disk of radius $R + d/2$. Comparing their areas gives $(\sqrt{M} - 1)d \leq 2R$. Since $M \geq K^2/192 \geq 4$, it follows that

$$n \geq \frac{3L(\sqrt{M} - 1)^2}{16384\alpha^2} \geq \frac{3LM}{65536\alpha^2} \geq \frac{K^2 L}{2^{22}\alpha^2}.$$

The smaller dyadic budgets. For $K \in \{8, 16, 32\}$ and $\delta \leq 2^{-12}$, choose any five-element seed set and its witness, with query i. An unselected seed r has either the event $E = \{A(Z; r) = h_i\}$ or its complement of probability at most δ there. At some parameter vertex the same event has probability

at least $1-\eta > 1/2$: use vertex i for the first event and a different vertex for its complement. The squared diameter of the localized simplex is $32\alpha^2$, so data processing gives

$$128n\alpha^2 \geq \mathrm{kl}(1/2, \delta) \geq L/4.$$

Thus $n \geq L/(512\alpha^2)$, which implies (78) for these three budgets as well.

A baseline bound for the remaining gaps. We show $n \geq K^2/(1024\alpha^2)$ throughout the stated range. Fix a query x, put all feature mass at x, and vary the positive-label probability t over $[1/2-\alpha, 1/2+\alpha]$. Let $q_s(t)$ be the probability that seed s predicts positively at x. Accuracy at the endpoints gives

$$\sum_s q_s(1/2-\alpha) \leq \eta < 1/2, \qquad \sum_s (1-q_s(1/2+\alpha)) \leq \eta < 1/2.$$

At most one seed has left endpoint value at least $1/4$, and at most one has right endpoint value at most $3/4$. At least $K-2$ rows therefore go from below $1/4$ to above $3/4$. For their open active sets $A_s = \{t \in [1/2-\alpha, 1/2+\alpha] : 1/4 < q_s(t) < 3/4\}$, clipping q_s to $[1/4, 3/4]$ and taking total variation yields

$$\sum_s \int_{A_s} |q_s'(t)|\, dt \geq \frac{K-2}{2}.$$

Each active row contributes more than $3/8$ to disagreement, whereas $\sum_s g(q_s(t)) \leq 1+\delta < 2$. Thus at most five rows are active at any t. The Bernoulli sample score has variance $n/[t(1-t)]$; applying Cauchy–Schwarz to the covariance of this score with the deterministic output indicator gives

$$|q_s'(t)| \leq \frac{1}{2}\sqrt{\frac{n}{t(1-t)}}.$$

Integration over the indicated interval now gives

$$\frac{K-2}{2} \leq 5\sqrt{n}\,\arcsin(2\alpha) \leq \frac{10\alpha\sqrt{n}}{\sqrt{1-4\alpha^2}},$$

and hence

$$n \geq \frac{(K-2)^2(1-4\alpha^2)}{400\alpha^2} \geq \frac{K^2}{1024\alpha^2}.$$

For $2^{-12} \leq \delta \leq 1/4$, this implies (78), since $\log(1/\delta) \leq 12\log 2 < 9$.

The ordinary confidence lower bound is $n \geq \log(1/\beta)/(32\alpha^2)$. Indeed, at a fixed query compare label probabilities $1/2+\alpha$ and $1/2-\alpha$. A wrong prediction has excess risk 2α, so the learner defines a test with both errors at most β. Including the independent shared seed in the test leaves the relative entropy of its inputs equal to the sample relative entropy. One-example relative entropy is at most $16\alpha^2$, whereas binary data processing gives at least $\mathrm{kl}(1-\beta, \beta) \geq \frac{1}{2}\log(1/\beta)$ for $\beta \leq 1/8$. Combining this bound with (78) proves the lower half of (77).

Finally use the ordered proper selector in the proof of Theorem H.21, with exactly K grid points, and choose

$$n = \left\lceil \max\left\{ \frac{32K^2}{\alpha^2}\log\frac{32}{\delta},\ \frac{8}{\alpha^2}\log\frac{6}{\beta} \right\} \right\rceil.$$

Its pointwise disagreement is at most $1/K + (32/K)e^{-n\alpha^2/(32K^2)} \leq (1+\delta)/K$. Its empirical-risk deviation event gives accuracy for every seed, with failure probability at most $6e^{-n\alpha^2/8} \leq \beta$. This is a proper learner using only the B shared bits. For $\delta \leq 1/4$, its sample count has the order stated in (77). □

H.8 Allowing probability on other outputs

The exact boundary and matching sample law above require exactly the three outputs in $\mathcal{H}_3$. When a larger point class permits other outputs, retaining three coordinates gives subprobability rows. The five-row obstruction still yields a quantitative lower bound in terms of their total missing mass.

Lemma H.24 (Five subprobability rows with bounded missing mass)**.** *Let $Q_s : \Delta_3 \to \{q \in [0,1]^3 : \sum_i q_i \le 1\}$, $1 \le s \le 5$, be continuous. Suppose that $0 \le \eta, \nu < 1/2$, and for every $v \in \Delta_3$,*

$$\sum_{s=1}^{5} \left(1 - \sum_{i=1}^{3} Q_{s,i}(v)\right) \le \eta.$$

Suppose also that, for every s, i,

$$Q_{s,i}(e_i) \ge 1 - \nu, \qquad Q_{s,i}(v) \le \nu \quad \text{whenever } v_i = 0.$$

Then some v and coordinate i satisfy

$$\sum_{s=1}^{5} 2Q_{s,i}(v)(1 - Q_{s,i}(v)) \ge 1 - \frac{2\eta^2}{9}.$$

Proof. Write $g(q) = 2q(1-q)$. If the conclusion fails, compactness gives a $\gamma > 2\eta^2/9$ such that every column sum is at most $1 - \gamma$. Choose $\delta > 0$ small enough that

$$\delta < \min\{1/18,\, 1/2 - \nu\}, \qquad 4\delta^2 < \gamma, \qquad \frac{2(\eta + 9\delta)^2}{9} + 6\delta^2 < \gamma.$$

The formula for T in Lemma H.20 also defines a continuous map on subprobability vectors. Its denominator remains at least 2δ; its image is a nonleaf precisely when $\max_i q_i < 1/2 + \delta$. If two images lie on the same open arm i, their column disagreement exceeds $1 - 4\delta^2 > 1 - \gamma$, as before.

Suppose three images are nonleaves. For their nine entries, put $\widetilde{q}_{s,i} = \min\{q_{s,i}, 1/2\}$. Their total mass before clipping is at least $3 - \eta$, and clipping removes at most 9δ. Some column thus has

$$\sum_{s \text{ in these three rows}} \widetilde{q}_{s,i} \ge 1 - \frac{\eta + 9\delta}{3} =: r.$$

Here $1/2 < r \le 1$. Lemma H.2 gives clipped column disagreement at least $1 - 2(\eta + 9\delta)^2/9$. Clipping can increase g on any one entry by at most $2\delta^2$, so the original column disagreement is at least

$$1 - \frac{2(\eta + 9\delta)^2}{9} - 6\delta^2 > 1 - \gamma,$$

a contradiction. Hence at most two images are nonleaves, and any two open arms have different colors. The product of the five T-images therefore lies in L. Since $1 - \nu > 1/2 + \delta$, it sends each parameter vertex to the required pure vertex. On a face with missing coordinate i, $Q_{s,i} \le \nu < 1/2 - \delta$ gives $T_i = 0$. Lemma H.19 rules out the resulting map. □

Theorem H.25 (A stronger confidence bound for larger point classes)**.** *Let $\mathcal{X}$ be a standard Borel space with at least three points, and let $\mathcal{H}$ consist of all point functions on $\mathcal{X}$, with or without the constant negative classifier. If $K = 2^B \ge 8$, $0 < \alpha < 1/3$, and $0 < \beta < 1/(2K)$, every proper agnostic (α, β)-learner using a fixed finite sample size, at most B total shared fair bits and no other randomness, and pointwise ρ-replicable on every distribution satisfies*

$$\rho > \frac{1 - \frac{2}{9}(K\beta)^2}{K}. \tag{81}$$

Proof. Pad unused bits to obtain K deterministic seed rules. Choose three domain points, put $u = (1/3, 1/3, 1/3)$, and choose $2\alpha < t < 1$. Let P_θ have only positive labels and feature probabilities $p(\theta) = (1-t)u + t\theta$, $\theta \in \Delta_3$. Write $q_{s,i}(\theta) = \mathbb{P}\{A(S; s) = h_i\}$ for the three chosen point functions. These are continuous subprobability rows. Every output outside these three functions is negative on the support and has excess risk $\max_i p_i(\theta) \ge 1/3 > \alpha$. Accuracy therefore gives the global bound

$$\sum_{s=1}^{K} \left(1 - \sum_{i=1}^{3} q_{s,i}(\theta)\right) \le K\beta.$$

On the face $\theta_i = 0$, output h_i also has excess risk at least $t/2 > \alpha$, so $q_{s,i} \le K\beta$ for every seed. At vertex e_i every incorrect output has excess risk at least $t > \alpha$, giving $q_{s,i}(e_i) \ge 1 - K\beta$. Any

five rows now satisfy Lemma H.24 with $\eta = \nu = K\beta < 1/2$. Its column lower bound, divided by K, is a lower bound on pointwise disagreement. In fact the bound is strict. For every seed s and coordinate i, the event $\{A(S; s) = h_i\}$ and its complement are both nonempty: their probabilities at e_i and at a different vertex are at least $1 - K\beta > 1/2$, respectively. Every sample in the finite positive-label alphabet has positive probability under every P_θ^n. Thus $0 < q_{s,i}(\theta) < 1$ everywhere. At a five-row witness, any of the $K - 5$ unselected rows contributes strictly positive disagreement, proving the claim. □

For $\alpha < 1/3$ and $K \geq 8$, this improves the coefficient $1/2$ in Theorem H.7 to $2/9$. The coefficient $2/9$ is sharp for the central-three-row mass estimate: the rows

$$(1/2 - \eta/3, 1/2, 0), \quad (0, 1/2 - \eta/3, 1/2), \quad (1/2, 0, 1/2 - \eta/3)$$

have total missing mass η and each column has disagreement $1 - 2\eta^2/9$. This local example supplies neither the required global maps nor a learner approaching the bound. The next result quantifies the sample cost of approaching this necessary bound.

Theorem H.26 (Finite-sample cost near the five-row bound)**.** *Let $\mathcal{X}$ be a standard Borel space with at least three points, and let $\mathcal{H}$ consist of all point functions on $\mathcal{X}$, with or without the constant negative classifier. Suppose $K = 2^B \geq 8$, $0 < \alpha \leq 1/16$, and $0 < \beta < 1/(2K)$, and put*

$$\eta = K\beta, \qquad a = 1 - 2\eta^2/9, \qquad \varepsilon = K\rho - a.$$

Every proper agnostic (α, β)-learner using a fixed finite sample size n, at most B total shared fair bits and no other randomness, and pointwise ρ-replicable on every distribution has $\varepsilon > 0$. If $\varepsilon \leq 1/4$, then

$$n \geq \frac{K^2 \log(1/\varepsilon)}{2^{24}\alpha^2}, \qquad n \geq \frac{\log(1/\beta)}{32\alpha^2}.$$

Consequently $n = \Omega\big(\alpha^{-2}[K^2 \log(1/\varepsilon) + \log(1/\beta)]\big)$ in this range.

Proof. Strict positivity follows from Theorem H.25. To obtain the quantitative bound, pad the seed space to K seeds and use the localized family

$$p(\theta) = (1 - 4\alpha)u + 4\alpha\theta, \qquad u = (1/3, 1/3, 1/3), \quad \theta \in \Delta_3,$$

with positive labels, as in Theorem H.23. Its coordinates are at least $1/4$. The excess risk of an incorrect point function at vertex e_i is at least 4α, and that of h_i on the opposite face is at least 2α. All other outputs are negative on the support and have excess risk at least $1/3 > \alpha$. Thus the continuous subprobability rows $q_{s,i}(\theta) = \mathbb{P}_{P_\theta^n}\{A(Z; s) = h_i\}$ satisfy

$$\sum_s \left(1 - \sum_i q_{s,i}(\theta)\right) \leq \eta, \quad q_{s,i}(e_i) \geq 1 - \eta, \quad q_{s,i}(\theta) \leq \eta \text{ if } \theta_i = 0.$$

Write $g(q) = 2q(1 - q)$. Replication gives $G_i(\theta) := \sum_s g(q_{s,i}(\theta)) \leq a + \varepsilon$. For each five-element seed set S, Lemma H.24 supplies a witness (θ_S, i_S) with

$$\sum_{s \in S} g(q_{s,i_S}(\theta_S)) \geq a, \qquad \sum_{s \notin S} g(q_{s,i_S}(\theta_S)) \leq \varepsilon. \tag{82}$$

First suppose $\varepsilon \leq 2^{-12}$, write $L = \log(1/\varepsilon)$, and assume $K \geq 64$. The affine five-element sets over $\mathbb{F}_{K/8}$ used in Theorem H.23 give $K^2/64$ sets with pairwise intersection at most one. At least $M \geq K^2/192$ of their witnesses have the same query i. For two distinct sets S, T in this group, there is an $r \in S \setminus T$ with $g(q_{r,i}(\theta_S)) > 1/9$. Indeed, if S, T are disjoint, one of the five contributions is at least $a/5 > 1/9$; otherwise one of the four remaining contributions is at least $(a - 1/2)/4 > 1/9$, since $a > 17/18$. Both $q_{r,i}(\theta_S)$ and its complement are therefore at least $1/18$. At θ_T, (82) gives $g(q_{r,i}(\theta_T)) \leq \varepsilon$, so one of the event $\{A(Z; r) = h_i\}$ and its complement has probability at most ε there. Binary data processing yields

$$n \,\mathrm{KL}(P_{\theta_S} \| P_{\theta_T}) \geq \mathrm{kl}(1/18, \varepsilon) \geq L/18 - h(1/18) \geq L/36.$$

Here h is binary entropy with natural logarithms; $h(1/18) < 2/9$ and $L \geq 12 \log 2 > 8$ justify the last step. Since $\mathrm{KL}(p \| p') \leq 4\|p - p'\|_2^2$, the same-query witnesses have pairwise distance at least

$d = \sqrt{L/(144n)}$. They lie in a planar disk of radius $R = 4\alpha\sqrt{2/3}$. The disjoint disks of radius $d/2$ about them imply $(\sqrt{M}-1)d \le 2R$. As $M \ge 4$,

$$n \ge \frac{3L(\sqrt{M}-1)^2}{18432\alpha^2} \ge \frac{3LM}{73728\alpha^2} \ge \frac{K^2L}{4718592\alpha^2} \ge \frac{K^2L}{2^{24}\alpha^2}.$$

For $K \in \{8, 16, 32\}$, choose any five-row witness and an unselected seed. By (82), one of its h_i event and its complement has probability at most ε at the witness. The same event has probability at least $1-\eta > 1/2$ at a parameter vertex. The squared diameter of the localized simplex is $32\alpha^2$, so

$$128n\alpha^2 \ge \mathrm{kl}(1/2, \varepsilon) \ge L/2 - \log 2 \ge L/4.$$

The resulting $n \ge L/(512\alpha^2)$ implies the claimed bound for these three budgets.

For the remaining range $2^{-12} \le \varepsilon \le 1/4$, the one-query score argument in Theorem H.23 applies unchanged. More explicitly, place all feature mass at a fixed query and let the positive-label probability vary over $[1/2-\alpha, 1/2+\alpha]$. Accuracy forces at least $K-2$ seed prediction probabilities to cross $[1/4, 3/4]$. Each active row contributes more than $3/8$ disagreement, and their sum is at most $a + \varepsilon \le 5/4 < 2$, so at most five rows are active. The score estimate $|q_s'(t)| \le \frac{1}{2}\sqrt{n/[t(1-t)]}$ therefore gives

$$\frac{K-2}{2} \le 5\sqrt{n}\,\arcsin(2\alpha), \qquad n \ge \frac{K^2}{1024\alpha^2}.$$

Since $L \le 12\log 2 < 9$, this implies the required bound. Finally, the two Bernoulli label distributions with parameters $1/2 \pm \alpha$ at a fixed query give $n \ge \log(1/\beta)/(32\alpha^2)$ by the confidence test in that theorem. Additional point-function outputs still give only the two possible predictions at this query, so neither this test nor the score argument changes. □

Corollary H.27 (Necessary missing mass for subcritical replication)**.** *Under the assumptions of Theorem H.26, fix three domain points and the positive-label family $p(\theta) = (1-4\alpha)(1/3, 1/3, 1/3) + 4\alpha\theta$. For the padded deterministic seed rules define*

$$q_{s,i}(\theta) = \mathbb{P}_{P_\theta^n}\{A(Z; s) = h_i\}, \qquad r_s(\theta) = 1 - \sum_{i=1}^{3} q_{s,i}(\theta), \qquad b_* = \min_{|S|=5} \sup_{\theta\in\Delta_3} \sum_{s\in S} r_s(\theta).$$

Then $b_* \le K\beta$ and

$$\rho > \frac{1 - 2b_*^2/9}{K}.$$

In particular, if $\rho < 1/K$, every five-element seed set S satisfies

$$\sup_{\theta\in\Delta_3} \sum_{s\in S} r_s(\theta) > 3\sqrt{\frac{1-K\rho}{2}}.$$

At least $K-4$ seed rules must therefore output a function outside the three selected point functions with positive probability on this family.

Proof. Accuracy gives $\sum_s r_s(\theta) \le K\beta$, proving $b_* \le K\beta < 1/2$. Choose a five-element set attaining the minimum in b_*. Its rows satisfy Lemma H.24 with missing mass parameter b_* and face parameter $\nu = K\beta$. The lemma gives a column witness of disagreement at least $1 - 2b_*^2/9$. The full-support argument in Theorem H.25 makes every unselected seed's contribution strictly positive there, which proves the strict bound. Rearranging it gives the displayed inequality for b_* and hence for every set S. If five seed rules never output outside the selected three functions on this family, their missing mass is identically zero, forcing $\rho > 1/K$. This proves the final assertion. □

The missing-mass condition is necessary for feasibility. The required sample size diverges near the five-row lower bound; attainability remains a separate question. The subcritical quartic bound below uses the gap $\zeta = K\rho - 1 + (K\beta)^2/2$; a uniform dominance relation between the two estimates is not established. Determining the exact boundary and matching sample upper bounds after adding the negative classifier or enlarging the domain remains open.

H.9 Further necessary sample bounds for point-function classes

The following sample bounds apply to the broader point-function classes in Theorem H.7. They are conditional on attaining $\rho < 1/K$ and do not establish feasibility in that range. For exactly $\mathcal{H}_3$ with $K \geq 8$, Theorem H.21 gives the stronger infeasibility conclusion at the stated high confidence.

Lemma H.28 (A boundary contact for two seed maps)**.** *Let f, g satisfy the assumptions of Lemma H.6, with $0 \leq \eta < 1/2$. Suppose there is no $x \in \Delta_3$ and coordinate i with $f_i(x) = g_i(x) = 1/2$. Then on some edge joining e_i and e_j there is a point at which*

$$f_i, f_j, g_i, g_j \in [1/2 - \eta, 1/2].$$

Proof. The vector fields $H_i = (f_i - 1/2, g_i - 1/2)$ never vanish. As in the argument-lift proof of Lemma H.6, normalize continuous arguments ϕ_i to lie in $(-\pi, -\pi/2)$ on the opposite face. This is possible because the edge assumptions imply $f_i + g_i \leq \eta < 1/2$ there. Then $\phi_i(e_j) = -3\pi/4$ for $j \neq i$, whereas $\phi_i(e_i) = \pi/4 + 2\pi k_i$ for integers k_i. If H_i, H_j never pointed in the same direction on their edge, the difference $\phi_j - \phi_i$ would avoid $2\pi\mathbb{Z}$ there. Its two endpoint values would give $k_i + k_j = -1$. This cannot hold on all three edges.

Thus some edge contains a point where H_i, H_j point in the same direction. At that point put $a = 1 - f_i - f_j$ and $b = 1 - g_i - g_j$. The assumptions give $a, b \geq 0$ and $a + b \leq \eta$, and $H_i + H_j = -(a, b)$. Since the nonzero vectors H_i, H_j have the same direction, both have nonpositive coordinates. Hence $f_i, f_j, g_i, g_j \leq 1/2$. The identities $f_i + f_j = 1 - a$ and $g_i + g_j = 1 - b$ then bound all four coordinates below by $1/2 - \eta$. □

Theorem H.29 (A quartic proper sample lower bound)**.** *Let $\mathcal{H}$ be either point class in Theorem H.7, and let*

$$K = 2^B \geq 2, \qquad 0 < \alpha < 1/2, \qquad 0 < \beta < \frac{1}{4K}, \qquad \rho < \frac{1}{K}.$$

If a proper agnostic (α, β)-learner for $\mathcal{H}$ uses a fixed finite number n of examples, at most B total shared fair bits and no other randomness, and is ρ-pointwise replicable on every distribution, then

$$n \geq \frac{K^2(K-1)^2(1-4K\beta)^2}{576(\arcsin\alpha)^2} \geq \frac{K^2(K-1)^2(1-4K\beta)^2}{768\alpha^2}. \tag{83}$$

In particular, when $\beta \leq 1/(8K)$,

$$n \geq \frac{K^2(K-1)^2}{3072\alpha^2} = \Omega(K^4/\alpha^2). \tag{84}$$

Proof. Pad the seed to K equally likely deterministic rules. Choose three domain points and use $P_\theta(X = i, Y = +1) = \theta_i$, $\theta \in \Delta_3$. Let $q_{s,i}(\theta)$ be the probability that seed s outputs the point function h_i. These are polynomials of degree at most n; each seed row is a subprobability vector if other outputs are present. Put $\eta = K\beta < 1/4$. As in Theorem H.7, accuracy at the vertices forces every seed to output the corresponding point function. On the edge joining e_i and e_j, all outputs other than h_i, h_j have excess risk at least $1/2 > \alpha$, so

$$\sum_s (1 - q_{s,i} - q_{s,j}) \leq \eta.$$

In particular the opposite face has $\sum_s q_{s,i} \leq \eta$. Pointwise replication gives, for every θ, i,

$$G_i(\theta) := \sum_s 2q_{s,i}(\theta)(1 - q_{s,i}(\theta)) \leq K\rho < 1.$$

For each unordered pair of distinct seeds, the two rows never have a coordinate equal to $1/2$ simultaneously: such a coordinate would contribute one to G_i. Lemma H.28 therefore supplies an edge and a point where both rows assign probability in $[1/2 - \eta, 1/2]$ to each of that edge's endpoint labels. Choose one such witness for every seed pair.

Fix one of the three edges and fix an endpoint i of that edge, with the other endpoint denoted by j. Parameterize it by $\theta = te_i + (1-t)e_j$, and write $p_s(t) = q_{s,i}(\theta)$. Each p_s is a polynomial, with $p_s(0) = 0$ and $p_s(1) = 1$. Its active set

$$A_s = \{t \in (0,1) : 1/4 < p_s(t) < 3/4\}$$

is consequently a finite union of open intervals. Every component has both endpoints in $(0, 1)$ and has probability $1/4$ or $3/4$ at each endpoint. At any t, at most two seeds are active: three active seeds would contribute more than $3(3/8) > 1$ to G_i.

Consider the intersection graph of all active component intervals on this edge. It is a forest. Indeed, when intervals are ordered by their left endpoints, a new interval can intersect at most one earlier interval; two such intersections would give a common open neighborhood in three intervals. Each chosen pair witness on this edge lies in one active component of each of its two seeds, since $\eta < 1/4$. Assign an edge between these two component vertices. Different seed pairs cannot produce the same component pair. The assigned edges and their incident components form a subgraph of the forest. Thus the number of assigned edges is at most the number of used components. Over the three domain edges, the number of assigned edges is exactly $\binom{K}{2}$.

Every used component contains a witness with $p_s(t) \in [1/2 - \eta, 1/2]$. Its probability changes by at least $1/4 - \eta$ between that witness and each component endpoint. Its total variation is therefore at least

$$c_\eta = \frac{1}{2} - 2\eta > 0.$$

If V_E denotes the number of used components on the current domain edge, then

$$c_\eta V_E \le \sum_s \int_{A_s} |p_s'(t)|\, dt. \tag{85}$$

Accuracy localizes every active set. When $t < (1 - \alpha)/2$, output h_i has excess risk $1 - 2t > \alpha$. Thus $\sum_s p_s(t) \le \eta$, and every $p_s(t) \le \eta < 1/4$. When $t > (1 + \alpha)/2$, only h_i can be accurate: h_j has excess risk $2t - 1 > \alpha$, and every other output has excess risk $t > \alpha$. Hence $\sum_s (1 - p_s(t)) \le \eta$, so every $p_s(t) \ge 1 - \eta > 3/4$. Consequently all active components lie in the central interval $[(1 - \alpha)/2, (1 + \alpha)/2]$.

Finally, differentiating the probability of any fixed sample event in this Bernoulli family and applying Cauchy–Schwarz to its covariance with the likelihood score gives

$$|p_s'(t)| \le \sqrt{\frac{n p_s(t)(1 - p_s(t))}{t(1 - t)}} \le \frac{1}{2}\sqrt{\frac{n}{t(1 - t)}} \qquad (0 < t < 1).$$

Here the likelihood score is $(T - nt)/(t(1 - t))$, where $T \sim \mathrm{Bin}(n, t)$, and its variance is $n/(t(1 - t))$. At most two seeds are active at a given t. Therefore

$$\sum_s \int_{A_s} |p_s'(t)|\, dt \le \sqrt{n} \int_{(1-\alpha)/2}^{(1+\alpha)/2} \frac{dt}{\sqrt{t(1 - t)}} = 2\sqrt{n}\, \arcsin \alpha.$$

Combine this bound with (85) on the three edges:

$$\frac{K(K - 1)}{2} \le \sum_E V_E \le \frac{6 \arcsin \alpha}{c_\eta} \sqrt{n}.$$

Squaring and substituting $c_\eta = (1 - 4K\beta)/2$ gives the first bound in (83). For $0 < \alpha < 1/2$, $\arcsin \alpha \le 2\alpha/\sqrt{3}$, proving its second bound and the stated consequence. □

The strict condition $\rho < 1/K$ is essential to this argument. This is a necessary sample bound conditional on achieving such a replication parameter; it does not assert that parameters below $1/K$ are feasible for every K, nor does it determine the full proper sample frontier.

Corollary H.30 (Monotone edge responses cannot cross the boundary)**.** *Let a proper learner satisfy the assumptions of Theorem H.29, except for the condition $\rho < 1/K$. Fix three domain points. For each of their three two-point positive-label distribution families, fix one endpoint. Suppose that, for every seed, the probability of outputting that endpoint's point function is monotone in its mass. If $K \ge 8$, the learner cannot have $\rho < 1/K$.*

Proof. On each edge the chosen endpoint probability runs continuously from zero to one. Monotonicity makes its active set a single interval, so the forest in the preceding proof has exactly K vertices and at most $K - 1$ edges. Across all three edges, the required seed-pair witnesses therefore give $K(K - 1)/2 \le 3(K - 1)$, which implies $K \le 6$. □

Theorem H.31 (Confidence and distance from the proper lower bound)**.** *Let a proper learner satisfy the assumptions of Theorem H.29, with $K \geq 4$ and $0 < \beta \leq 1/(8K)$. Put*

$$\eta = K\beta, \qquad \zeta = K\rho - 1 + \frac{\eta^2}{2}.$$

Then $0 < \zeta < \eta^2/2$, and

$$n \geq \frac{K^2(K-1)^2}{3840\alpha^2} \log \frac{1}{\zeta}. \tag{86}$$

In particular,

$$n \geq \frac{K^2(K-1)^2}{1920\alpha^2} \log \frac{1}{K\beta} = \Omega\left(\frac{K^4}{\alpha^2} \log \frac{1}{K\beta}\right). \tag{87}$$

Proof. Use the three-point distribution family, seed probabilities $q_{s,i}$, and column sums G_i from the proof of Theorem H.29. For each unordered seed pair $\{s,t\}$, choose the boundary contact constructed in the proof of Lemma H.28. On its edge joining e_i and e_j, the vectors

$$H_i = (q_{s,i} - 1/2, q_{t,i} - 1/2), \qquad H_j = (q_{s,j} - 1/2, q_{t,j} - 1/2)$$

point in the same direction and satisfy

$$H_i + H_j = -(e_s, e_t), \qquad e_s = 1 - q_{s,i} - q_{s,j} \geq 0, \quad e_t = 1 - q_{t,i} - q_{t,j} \geq 0, \qquad e_s + e_t \leq \eta.$$

Choose the endpoint $\ell \in \{i,j\}$ whose vector has smaller norm. Because the two vectors have the same direction,

$$\|H_\ell\|_2 \leq \frac{1}{2}\|(e_s, e_t)\|_2 \leq \frac{\eta}{2}.$$

At query ℓ of this contact, the pair contributes at least

$$2q_{s,\ell}(1 - q_{s,\ell}) + 2q_{t,\ell}(1 - q_{t,\ell}) = 1 - 2\|H_\ell\|_2^2 \geq 1 - \frac{\eta^2}{2}$$

to $G_\ell \leq K\rho$. Consequently every other seed $r \notin \{s,t\}$ obeys

$$2q_{r,\ell}(1 - q_{r,\ell}) \leq \zeta, \qquad \min\{q_{r,\ell}, 1 - q_{r,\ell}\} \leq \zeta. \tag{88}$$

There is such a seed because $K \geq 4$. Its probability $q_{r,\ell}$ is strictly between zero and one at the contact: the contact is interior to the edge, and the all-i and all-j sample strings have positive probability and force different outputs for this seed. Thus $\zeta > 0$. The assumption $\rho < 1/K$ also gives $\zeta < \eta^2/2 \leq 1/128$.

Now take contacts of two different seed pairs $\mathcal{A}$ and $\mathcal{B}$ on the same domain edge. Parameterize this edge by the mass of either fixed endpoint, and let the two parameters be t and u. Choose $r \in \mathcal{A} \setminus \mathcal{B}$, and use the endpoint ℓ selected above for contact $\mathcal{B}$. At contact $\mathcal{A}$, for each of the two endpoints, the probability that seed r outputs its point function lies in $[1/2 - \eta, 1/2]$. This applies to ℓ even if the endpoint selected for $\mathcal{A}$ was different. At contact $\mathcal{B}$, (88) applies to r. Thus one of the two deterministic sample events $E = \{A(S;r) = h_\ell\}$ and its complement satisfies

$$P_t^n(E) \geq \frac{3}{8}, \qquad P_u^n(E) \leq \zeta.$$

The row and query may depend on the two contacts; each comparison concerns the same two underlying sample laws.

Write $\mathrm{kl}(a,b)$ for the relative entropy between Bernoulli laws with parameters a, b. Data processing and the binary entropy bound give

$$\begin{aligned} n\,\mathrm{kl}(t,u) &\geq \mathrm{kl}(P_t^n(E), P_u^n(E)) \\ &\geq \frac{3}{8}\log\frac{1}{\zeta} - \log 2 \geq \frac{1}{5}\log\frac{1}{\zeta}, \end{aligned}$$

where the last inequality uses $\zeta < 1/128$. The accuracy localization in Theorem H.29 places all contacts in an interval of length α, contained in $[1/4, 3/4]$. Hence

$$\mathrm{kl}(t,u) \leq \frac{(t-u)^2}{u(1-u)} \leq \frac{16}{3}(t-u)^2.$$

Every two different seed-pair contacts on the same edge are therefore separated by at least

$$\left(\frac{3}{80n}\log\frac{1}{\zeta}\right)^{1/2}.$$

In particular their locations cannot coincide.

Among the $\binom{K}{2}$ chosen contacts, some edge contains at least $M = \lceil K(K-1)/6 \rceil$ of them. Since $K \geq 4$, $M - 1 \geq K(K-1)/12$. Ordering those contacts in their interval of length α yields

$$\alpha \geq (M-1)\left(\frac{3}{80n}\log\frac{1}{\zeta}\right)^{1/2}.$$

This is (86). Finally, $\zeta < \eta^2/2$ gives $\log(1/\zeta) > 2\log(1/\eta)$, proving (87). □

Equivalently, the finite-sample residual satisfies $\zeta \geq \exp[-3840n\alpha^2/(K^2(K-1)^2)]$. Its strict positivity is a conclusion for finite-sample learners; it does not establish a strict gap for the variational infimum $\Gamma_{N,K}$ or identify the exact proper feasibility threshold.

I FINITE RANDOMNESS FROM AN ORDERED HYPOTHESIS CLASS

Interval selection extends from point functions to classes with a bounded number of label changes along a public hypothesis order. The order need not have a first eligible hypothesis. In that case, the learner outputs a pointwise limit of eligible hypotheses. This gives the optimal sample rate and logarithmic total randomness for every class of VC dimension one on a countable domain, under general agnostic distributions.

An *initial segment* of a linearly ordered set $(\mathcal{D}, \prec)$ is a subset C such that $h \in C$ and $g \prec h$ imply $g \in C$. Any two initial segments are nested. For a binary class $\mathcal{D}$, let $n_{\mathrm{uc}}(\mathcal{D}, \varepsilon, \delta)$ be an integer such that, for every distribution P and every sample size $n \geq n_{\mathrm{uc}}(\mathcal{D}, \varepsilon, \delta)$,

$$\mathbb{P}\left\{\sup_{h\in\mathcal{D}} |\widehat{\mathrm{R}}_S(h) - \mathrm{R}_P(h)| > \varepsilon\right\} \leq \delta, \qquad \widehat{\mathrm{R}}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{h(X_i) \neq Y_i\}.$$

For a class of finite VC dimension $d \geq 1$, the standard uniform-convergence bound used in Corollary 2.3 permits

$$n_{\mathrm{uc}}(\mathcal{D}, \varepsilon, \delta) = O\left(\frac{d\log(2/\varepsilon) + \log(2/\delta)}{\varepsilon^2}\right). \tag{89}$$

Theorem I.1 (A bounded number of cuts)**.** *Let $\mathcal{X}$ be countable, and let $\mathcal{D}$ be a nonempty countable binary class with a fixed public total order. Suppose that, for every $x \in \mathcal{X}$, there are a fixed label $b_x \in \{-1, +1\}$ and at most k fixed nonempty proper initial segments $C_{x,1}, \ldots, C_{x,k_x}$ such that*

$$h(x) = b_x(-1)^{\sum_{\ell=1}^{k_x} \mathbf{1}\{h \notin C_{x,\ell}\}} \qquad (h \in \mathcal{D}), \qquad k_x \leq k. \tag{90}$$

Here $k \geq 1$ is an integer, and an empty sum is allowed. For $0 < \alpha, \beta, \rho \leq 1/4$, put

$$B = \left\lceil \log_2 \frac{k}{\rho} \right\rceil, \qquad K = 2^B, \qquad n = \max\left\{\left\lceil \frac{32k^2}{\rho^2\alpha^2} \right\rceil, n_{\mathrm{uc}}(\mathcal{D}, \alpha/4, \beta)\right\}.$$

Whenever the displayed uniform-convergence bound is finite, there is an improper agnostic (α, β)-learner for $\mathcal{D}$ that is ρ-pointwise replicable on every distribution. It uses exactly n examples and exactly B shared fair bits, with no other randomness. Its outputs lie in the pointwise closure of $\mathcal{D}$.

Proof. **The learner.** Write $r_* = \min_{h\in\mathcal{D}} \widehat{\mathrm{R}}_S(h)$ and, for each nonempty cut C, write $r_C = \min_{h\in C} \widehat{\mathrm{R}}_S(h)$. These minima exist even when the sets are infinite: empirical risks take values in the finite set $\{0, 1/n, \ldots, 1\}$. Use the B shared bits to draw U uniformly from

$$\left\{\frac{j + 1/2}{K} : j = 0, \ldots, K-1\right\}.$$

For every cut, set $G_C = 2(r_C - r_*)/\alpha$, and output

$$H_{S,U}(x) = b_x(-1)^{\sum_{\ell=1}^{k_x} \mathbf{1}\{U < G_{C_{x,\ell}}\}}. \tag{91}$$

All orders, cuts, and baseline labels are fixed before sampling. The rule is measurable because its sample space is countable and its shared seed space is finite. The guarantee is statistical; evaluating the public order and empirical minima may require unrestricted computation.

A pointwise limit of eligible hypotheses. Define the nonempty set

$$F = \left\{ h \in \mathcal{D} : \widehat{\mathrm{R}}_S(h) \le r_* + \frac{\alpha}{2} U \right\}.$$

Attainment of each cut minimum gives the exact equivalence

$$F \cap C \neq \varnothing \quad \Longleftrightarrow \quad U \ge G_C. \tag{92}$$

In particular, the weak and strict comparisons in (91) also handle equality at a grid point.

Fix a finite set of test points $E \subseteq \mathcal{X}$ and consider all their cuts. They are finitely many nested initial segments. If some intersect F, let C_0 be the smallest such cut and choose any $h \in F \cap C_0$. This h belongs to every cut intersecting F and to no cut disjoint from F. If no cut intersects F, any $h \in F$ has the same property. Equations (90)–(92) then show that $h(x) = H_{S,U}(x)$ for every $x \in E$.

Take an increasing sequence of finite sets exhausting $\mathcal{X}$. The preceding argument supplies a sequence (h_j) in F that converges pointwise to $H_{S,U}$. For a finite domain, one can instead take a single hypothesis matching the output on the entire domain. Dominated convergence gives

$$\mathrm{R}_P(H_{S,U}) = \lim_{j \to \infty} \mathrm{R}_P(h_j). \tag{93}$$

The sequence is used only in the proof; the learner is the explicit prediction rule (91).

Accuracy. On the event of uniform deviation at most $\alpha/4$, every $h \in F$ satisfies

$$\mathrm{R}_P(h) \le \widehat{\mathrm{R}}_S(h) + \frac{\alpha}{4} \le r_* + \frac{\alpha}{2} + \frac{\alpha}{4} \le \mathrm{OPT}_{\mathcal{D}}(P) + \alpha.$$

The last inequality uses $r_* \le \mathrm{OPT}_{\mathcal{D}}(P) + \alpha/4$ and does not require attainment of the true risk infimum. Passing to the limit in (93) proves the same bound for the output, simultaneously for every seed. The definition of n_{uc} proves (α, β) accuracy.

Replication. Replacing one observation changes every empirical risk by at most $1/n$. Consequently it changes each of r_C, r_* by at most $1/n$, and changes G_C by at most $4/(n\alpha)$. Clip G_C to $V_C = \min\{1, \max\{0, G_C\}\}$. Because every midpoint lies strictly between zero and one, this clipping preserves every comparison in (91). It also preserves the bounded differences bound. Tensorization of variance and the bound of one quarter of the squared range for each conditional variance therefore give

$$\mathrm{Var}(V_C) \le \frac{n}{4} \left(\frac{4}{n\alpha} \right)^2 = \frac{4}{n\alpha^2}.$$

For two independent samples, V_C, V_C' are independent and identically distributed. Lemma G.1, including its strict-comparison version, implies

$$\mathbb{P}\{\mathbf{1}\{U < V_C\} \neq \mathbf{1}\{U < V_C'\}\} \le \frac{1}{2K} + \frac{2\sqrt{2}}{\alpha\sqrt{n}}.$$

If the predictions at a fixed x differ, at least one of its k_x cut comparisons differs. A union bound, without requiring independence among cuts, gives

$$\mathbb{P}\{H_{S,U}(x) \neq H_{S',U}(x)\} \le \frac{k}{2K} + \frac{2\sqrt{2}k}{\alpha\sqrt{n}} \le \frac{\rho}{2} + \frac{\rho}{2} = \rho.$$

This argument holds for every distribution and does not condition on the accuracy event. The midpoint index is the learner's only randomness. □

Remark I.2 (When the selector is proper). If the public order is a fixed enumeration, the set F has a first member. For every initial segment C, that member lies in C exactly when $F \cap C \neq \varnothing$. Thus (91) is the first eligible hypothesis and is proper. In particular, the point-function order has at most two label changes at each point. General public orders can lack a first eligible member, which is why the theorem allows a pointwise limit as its output.

The tree structure of VC-one classes is classical: Ben-David (2015, Theorem 4) gives a domain-side tree-order representation after flipping a reference hypothesis. The laminarity argument below is its dual four-pattern argument. We give the hypothesis-side interval order and dense-subclass construction needed for the finite-grid learner explicitly.

Lemma I.3 (A public interval order for a VC-one class)**.** *Let $\mathcal{H}$ be a nonempty binary class of VC dimension at most one on a countable domain, and fix $h_0 \in \mathcal{H}$. There is a countable subclass $\mathcal{D} \subseteq \mathcal{H}$ containing h_0 and a public total order on $\mathcal{D}$ such that:*

1. $\mathrm{OPT}_{\mathcal{D}}(P) = \mathrm{OPT}_{\mathcal{H}}(P)$ *for every distribution P;*
2. *h_0 is the least element of the order;*
3. *for each x, the set $A_x = \{h \in \mathcal{D} : h(x) \neq h_0(x)\}$ is an interval of the order, possibly empty.*

The class, order, and intervals can be fixed independently of the unknown distribution.

Proof. Fix a finite-set exhaustion of the domain. For every set in the exhaustion and every label pattern on that set realized by $\mathcal{H}$, choose one representative from $\mathcal{H}$. Let $\mathcal{D}$ be the union of these representatives and $\{h_0\}$. It is countable. Every $h \in \mathcal{H}$ is the pointwise limit of a sequence in $\mathcal{D}$, so dominated convergence implies $\mathrm{OPT}_{\mathcal{D}}(P) = \mathrm{OPT}_{\mathcal{H}}(P)$ for every P. All representative choices are public and precede sampling.

The family $(A_x)_{x \in \mathcal{X}}$ is laminar: any two of its sets are nested or disjoint. Indeed, if $A_x \cap A_y$, $A_x \setminus A_y$, and $A_y \setminus A_x$ were all nonempty, their witnesses, together with $h_0 \notin A_x \cup A_y$, would realize all four binary label patterns on $\{x, y\}$, contrary to the VC-dimension assumption. Multiplying the labels at each coordinate by the fixed value h_0 does not change whether those two points are shattered.

We give the interval-order construction explicitly as a compactness argument on finite orders. On a finite ground set, every laminar family admits an order in which its members are intervals: discard empty sets and duplicates, add the whole ground set and all singletons, and form the inclusion tree, and list its singleton leaves by a depth-first traversal. The leaves below each tree node are consecutive. If one element belongs to none of the original sets, its singleton can be visited first.

If $\mathcal{D}$ is finite, this argument applies directly and places h_0 first. Otherwise, enumerate its distinct elements as $h_0, h_1, \ldots$. At level m consider every order of $\{h_0, \ldots, h_{m-1}\}$ that places h_0 first and makes every trace of every A_x an interval. Although there may be infinitely many sets A_x, they have only finitely many distinct traces on this finite ground set. Those traces are laminar, so each level is nonempty by the preceding finite construction. Restricting a valid order to the preceding ground set preserves all the conditions.

Connect each order to its restriction. The resulting tree is finitely branching and has nonempty levels of every finite depth. Kőnig's lemma supplies an infinite branch. Its consistent finite orders define a total order on $\mathcal{D}$ with h_0 least. If two members of A_x enclose a third hypothesis in this order, all three occur in some finite level, whose interval property puts that third hypothesis in A_x. Thus every A_x is convex, as required. Fix any such branch once and for all. This is a distribution-independent existence construction; no efficient procedure for finding or evaluating the order is asserted. □

Corollary I.4 (Every VC-one class has the optimal finite-seed rate)**.** *Every nonempty binary class $\mathcal{H}$ of VC dimension at most one on a countable domain, for $0 < \alpha, \beta, \rho \leq 1/4$, admits an improper agnostic (α, β)-learner that is ρ-pointwise replicable on every distribution, uses exactly*

$$B = \left\lceil \log_2 \frac{2}{\rho} \right\rceil$$

total shared fair bits, and uses a fixed number of examples

$$n = \max\left\{\left\lceil \frac{128}{\rho^2\alpha^2}\right\rceil, n_{\mathrm{uc}}(\mathcal{D}, \alpha/4, \beta)\right\} = O\left(\frac{\rho^{-2} + \log(2/\alpha) + \log(2/\beta)}{\alpha^2}\right), \tag{94}$$

where $\mathcal{D}$ is the public subclass from Lemma I.3. There is no assumption that $\mathcal{H}$ attains Bayes risk, and the learner uses no knowledge of the atom masses or conditional label means.

Proof. Use the public class and order supplied by Lemma I.3. An interval A_x changes membership at at most two cuts of a total order. More explicitly, when A_x is nonempty, let L_x be the set of elements strictly before the interval and let $J_x = L_x \cup A_x$. Both are initial segments, L_x contains h_0, and J_x can be the entire class. Membership in A_x is the difference between membership in J_x and in L_x. Omitting a trivial cut if $J_x = \mathcal{D}$, and using no cuts if A_x is empty, gives (90) with $b_x = h_0(x)$ and $k_x \le 2$.

Apply Theorem I.1 with $k = 2$. The class $\mathcal{D}$ has VC dimension at most one, so (89) gives the stated sample bound; a VC-zero class also obeys this bound. Its class optimum equals that of $\mathcal{H}$ for every distribution. The theorem's accuracy and replication conclusions therefore apply to $\mathcal{H}$, with the displayed total seed and fixed sample count. □

Proposition I.5 (A VC-two obstruction to a bounded-cut order)**.** *On $\mathcal{X} = \mathbb{N}^2$, where $\mathbb{N} = \{1, 2, \ldots\}$, let*

$$\mathcal{H}_\square = \{h_{a,b} : a, b \in \mathbb{N}\}, \qquad h_{a,b}(x, y) = 2\mathbf{1}\{x \le a,\ y \le b\} - 1.$$

This class has VC dimension two. For every $m \ge 2$, in every total order of the m^2 hypotheses $\mathcal{H}_m = \{h_{a,b} : 1 \le a, b \le m\}$, some query point has at least

$$\left\lceil \frac{m^2 - 1}{2m - 1} \right\rceil$$

label changes along that order. Consequently, no total order on $\mathcal{H}_\square$ admits a uniform finite bound k in (90).

Proof. The points $(1, 2)$ and $(2, 1)$ are shattered by the four hypotheses with $a, b \in \{1, 2\}$. To shatter three points, each point would have to be excludable by an anchored rectangle containing the other two. Such an excluded point must have a strictly larger first coordinate than both other points, or a strictly larger second coordinate than both. At most two of the three points can satisfy this condition, because each coordinate has at most one strict maximum. Thus no three points can be shattered.

List the parameter pairs in any proposed order as $(a_1, b_1), \ldots, (a_{m^2}, b_{m^2})$. Consider the $2m - 1$ query points

$$Q_m = \{(x, 1) : 1 \le x \le m\} \cup \{(1, y) : 1 \le y \le m\}.$$

Between consecutive hypotheses, the number of changed predictions among $(x, 1)$ is $|a_{i+1} - a_i|$, and the corresponding number among $(1, y)$ is $|b_{i+1} - b_i|$. The duplicated point $(1, 1)$ never changes its label. If T_z counts the label changes at a fixed query z, then

$$\sum_{z \in Q_m} T_z = \sum_{i=1}^{m^2-1} \left(|a_{i+1} - a_i| + |b_{i+1} - b_i|\right) \ge m^2 - 1,$$

since consecutive pairs are distinct. Averaging over Q_m proves the finite-order bound.

Restricting an initial segment of any total order to $\mathcal{H}_m$ gives an initial segment of the induced order. A label represented by at most k initial cuts can change at most k times along a finite order. Applying the preceding bound for arbitrarily large m therefore excludes every uniform finite k for $\mathcal{H}_\square$. □

Remark I.6 (The obstruction persists under dense restriction). Passing to a pointwise-dense subclass does not remove this example. Each $h_{a,b}$ is the only member of $\mathcal{H}_\square$ with its labels on the three points (a, b), $(a+1, 1)$, and $(1, b+1)$; hence every such dense subclass contains it. The obstruction concerns bounded-cut hypothesis orders. Sample and randomness lower bounds for other anchored-rectangle learners require separate arguments.

Proposition I.7 (Asymptotic failure of a lexicographic grid selector)**.** *Fix an integer $K \geq 1$ and an even integer $M \geq 8K$, and put*

$$D = M^2 - 1, \qquad \alpha = \frac{3KM}{2D} \leq \frac{1}{4}.$$

Order $\mathcal{H}_M$ from Proposition I.5 lexicographically by (a, b). On n observations, let A_n draw a shared U uniformly from the K midpoints of $[0, 1]$ and return the first hypothesis satisfying

$$\widehat{\mathrm{R}}_S(h) \leq \min_{g \in \mathcal{H}_M} \widehat{\mathrm{R}}_S(g) + \frac{\alpha}{2} U.$$

There is a distribution P_M with deterministic labels and $\mathrm{OPT}_{\mathcal{H}_M}(P_M) = 1/4$, while its Bayes risk is zero, such that at the fixed point $z = (1, M/2 + 1)$,

$$\lim_{n \to \infty} \mathbb{P}\{A_n(S; U)(z) \neq A_n(S'; U)(z)\} = \frac{1}{2}, \qquad S, S' \overset{\text{iid}}{\sim} P_M^n.$$

For every fixed $\beta > 0$, the same selector is nevertheless an ordinary agnostic (α, β)-learner for $\mathcal{H}_M$ for all sufficiently large n.

Proof. Assign mass $1/4$ to $(1, 1)$ with label -1. Assign positive labels and masses

$$P_M\{X = (a, 1)\} = \frac{3M}{4D} \quad (2 \leq a \leq M), \qquad P_M\{X = (1, b)\} = \frac{3}{4D} \quad (2 \leq b \leq M).$$

These masses sum to one. Number the hypotheses by $j = (a - 1)M + b - 1$, so $j = 0, \ldots, D$ is their lexicographic order. Every hypothesis errs on the negative point $(1, 1)$, and its covered positive mass is $3j/(4D)$. Thus

$$r_j := \mathrm{R}_{P_M}(h_j) = 1 - \frac{3j}{4D}, \qquad r_D = \frac{1}{4}.$$

The labels are deterministic, establishing the Bayes-risk assertion as well.

Fix a seed $\ell \in \{0, \ldots, K - 1\}$ and put

$$u_\ell = \frac{\ell + 1/2}{K}, \qquad j_\ell = D - M(\ell + 1/2).$$

This is an integer between 0 and $D - 1$. Its parameter pair is $(M - \ell, M/2)$, and direct substitution gives

$$r_{j_\ell} - r_D = \frac{\alpha}{2} u_\ell.$$

At $z = (1, M/2 + 1)$, the consecutive hypotheses h_{j_ℓ} and $h_{j_\ell + 1}$ predict -1 and $+1$, respectively.

Keep M, K fixed while $n \to \infty$. Since the finitely many population risks are strictly decreasing, their empirical risks have the same strict ordering with probability tending to one. Convergence of their differences to their distinct population values also shows that, at seed ℓ, the first eligible hypothesis belongs to $\{h_{j_\ell}, h_{j_\ell + 1}\}$ with probability tending to one. On this event, it is $h_{j_\ell + 1}$ precisely when

$$\widehat{\mathrm{R}}_S(h_{j_\ell}) - \widehat{\mathrm{R}}_S(h_D) > \frac{\alpha}{2} u_\ell.$$

The difference on the left is the empirical fraction of positive examples missed by h_{j_ℓ}: all hypotheses have the same error on $(1, 1)$, and h_D covers every positive support point. It is a Bernoulli sample mean with mean exactly $\alpha u_\ell / 2 \in (0, 1)$. The central limit theorem therefore implies that its probability of strictly exceeding its mean tends to $1/2$. Writing $p_{n,\ell} = \mathbb{P}\{A_n(S; u_\ell)(z) = +1\}$, we obtain $p_{n,\ell} \to 1/2$ for each seed. Conditional independence of S, S' now gives

$$\mathbb{P}\{A_n(S; U)(z) \neq A_n(S'; U)(z)\} = \frac{1}{K} \sum_{\ell=0}^{K-1} 2 p_{n,\ell}(1 - p_{n,\ell}) \longrightarrow \frac{1}{2}.$$

Finally, on a uniform empirical-risk deviation event of size $\alpha/4$, every candidate eligible at any seed has risk at most $\mathrm{OPT}_{\mathcal{H}_M}(P) + \alpha$, for every distribution P. Uniform convergence for the finite class gives this event probability at least $1 - \beta$ for all sufficiently large n. The failure is therefore specific to replication of this selector, rather than ordinary accuracy. □

I.1 Products on a fixed partition

A different extension of the ordered-class construction permits independent hypothesis choices on a fixed finite partition of the domain. The partition is known, but its block probabilities need not be known or bounded below. Normalizing each empirical risk gap by the square root of its observed block count makes its variance independent of the block probability.

Lemma I.8 (Variance of a normalized block risk gap). *Let E be a fixed subset of a countable domain and let $\varnothing \neq \mathcal{C} \subseteq \mathcal{D}$ be countable binary classes on E. On $n \geq 1$ IID labeled examples from P, put*

$$N = \sum_{i=1}^{n} \mathbf{1}\{X_i \in E\}, \qquad L_{\mathcal{A}} = \min_{h \in \mathcal{A}} \sum_{i: X_i \in E} \mathbf{1}\{h(X_i) \neq Y_i\} \quad (\mathcal{A} = \mathcal{C}, \mathcal{D}).$$

Define

$$F = \begin{cases} (L_{\mathcal{C}} - L_{\mathcal{D}})/\sqrt{nN}, & N > 0, \\ 0, & N = 0. \end{cases}$$

For every distribution of the examples, $\mathrm{Var}(F) \leq 4/n$.

Proof. The minima exist because the loss counts take finitely many integer values. Write $G = L_{\mathcal{C}} - L_{\mathcal{D}}$, so $0 \leq G \leq N$. Replace one observation by an independent copy, and let R be the number of the other $n-1$ observations falling in E. If both the old and new observation lie outside E, then F does not change. If both lie inside, each minimum loss count changes by at most one, so G changes by at most two and the change of F is at most $2/\sqrt{n(R+1)}$.

If exactly one lies inside, compare the block samples of sizes R and $R+1$. Adding one observation increases each minimum loss count by a number in $[0,1]$, so the two gap values differ by at most one. For $R \geq 1$, use the smaller sample's gap $G_R \leq R$ to obtain

$$\begin{aligned} \left| \frac{G_{R+1}}{\sqrt{R+1}} - \frac{G_R}{\sqrt{R}} \right| &\leq \frac{1}{\sqrt{R+1}} + R\left(\frac{1}{\sqrt{R}} - \frac{1}{\sqrt{R+1}} \right) \\ &\leq \frac{3}{2\sqrt{R+1}} \leq \frac{2}{\sqrt{R+1}}. \end{aligned}$$

For $R = 0$, the corresponding bound follows from $G_0 = 0$ and $0 \leq G_1 \leq 1$. Thus the same replacement bound holds whenever at least one of the two observations lies in E.

Put $p = P(X \in E)$. The case $p = 0$ is immediate. Otherwise $R \sim \mathrm{Bin}(n-1, p)$ independently of the two membership indicators, whose union has probability at most $2p$. If $F^{(i)}$ denotes the replaced score, then

$$\mathbb{E}(F - F^{(i)})^2 \leq \frac{8p}{n} \mathbb{E} \frac{1}{R+1} = \frac{8(1-(1-p)^n)}{n^2} \leq \frac{8}{n^2}.$$

Here the identity follows by integrating $\mathbb{E}t^R = (1-p+pt)^{n-1}$ over $0 \leq t \leq 1$. The Efron–Stein inequality now gives $\mathrm{Var}(F) \leq \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}(F - F^{(i)})^2 \leq 4/n$. □

Theorem I.9 (A finite product of ordered classes). *Let $\mathcal{X} = \bigsqcup_{j=1}^{M} \mathcal{X}_j$ be a fixed known partition of a countable domain into $M \geq 1$ nonempty blocks. On each block let $\mathcal{D}_j$ be a nonempty countable binary class with a fixed public order and a representation of the form (90) using at most $k \geq 1$ nonempty proper initial cuts at each point. Define its full product by*

$$\mathcal{D} = \{h : h|_{\mathcal{X}_j} \in \mathcal{D}_j \text{ for } j = 1, \ldots, M\}.$$

For $0 < \alpha, \beta, \rho \leq 1/4$, put

$$B = \left\lceil \log_2 \frac{k}{\rho} \right\rceil, \qquad K = 2^B, \qquad n = \max\left\{ \left\lceil \frac{128k^2M}{\rho^2\alpha^2} \right\rceil, n_{\mathrm{uc}}(\mathcal{D}, \alpha/4, \beta) \right\}.$$

When the displayed uniform-convergence bound is finite, there is an improper agnostic (α, β)-learner for $\mathcal{D}$ that is ρ-pointwise replicable on every distribution, uses exactly n examples, and uses exactly B shared fair bits with no other randomness. Its outputs lie in the pointwise closure of $\mathcal{D}$. The seed is shared across all blocks and does not depend on M.

Proof. **The learner.** On one sample of size n, let N_j be the number of examples in block j. For a hypothesis on that block, let L_h be its number of sample errors within the block. Write

$$L_{*,j} = \min_{h \in \mathcal{D}_j} L_h, \qquad L_C = \min_{h \in C} L_h$$

for each of its nonempty initial cuts C. When $N_j = 0$, all these quantities are zero. Define the normalized score

$$G_C = \begin{cases} \dfrac{2\sqrt{M}}{\alpha} \dfrac{L_C - L_{*,j}}{\sqrt{nN_j}}, & N_j > 0, \\ 0, & N_j = 0. \end{cases}$$

Draw a single U uniformly from the K midpoints of $[0,1]$ using the B shared bits. For $x \in \mathcal{X}_j$, use the public baseline b_x and cuts belonging to that block and set

$$H(x) = b_x(-1)^{\sum_{\ell=1}^{k_x} \mathbf{1}\{U < G_{C_{x,\ell}}\}}.$$

No separate random choice is made for any block. This is a measurable prediction rule on the countable sample space.

Accuracy and closure. For each block define the nonempty eligible set

$$F_j = \left\{ h \in \mathcal{D}_j : L_h - L_{*,j} \le \frac{\alpha U}{2\sqrt{M}} \sqrt{nN_j} \right\}.$$

For $N_j > 0$, attainment of the cut minimum gives $F_j \cap C \neq \varnothing$ exactly when $U \ge G_C$. For $N_j = 0$, one has $F_j = \mathcal{D}_j$ and $G_C = 0$, so the same equivalence holds. Applying the finite-cut matching argument from Theorem I.1 within each block shows that, on every finite set of query points, H agrees with a member of $F = \prod_{j=1}^{M} F_j$. An exhaustion of the countable domain therefore gives a sequence of members of F converging pointwise to H.

The minimum empirical risk of the product class is $n^{-1} \sum_j L_{*,j}$. Every $h \in F$ consequently satisfies

$$\begin{aligned} \widehat{\mathrm{R}}_S(h) - \min_{g \in \mathcal{D}} \widehat{\mathrm{R}}_S(g) &= \frac{1}{n} \sum_{j=1}^{M} (L_{h|_{\mathcal{X}_j}} - L_{*,j}) \\ &\le \frac{\alpha U}{2\sqrt{M}} \sum_{j=1}^{M} \sqrt{N_j/n} \le \frac{\alpha}{2}. \end{aligned}$$

The last step uses $\sum_j N_j/n = 1$ and Cauchy–Schwarz. On a uniform empirical-risk deviation event of size $\alpha/4$, all members of F thus have risk at most $\mathrm{OPT}_{\mathcal{D}}(P) + \alpha$. Dominated convergence gives the same bound for H, simultaneously for every seed. The prescribed uniform-convergence sample size proves accuracy. This argument includes empty observed blocks and requires no lower bound on their probabilities.

Replication. For each cut, Lemma I.8 gives

$$\mathrm{Var}(G_C) \le \frac{16M}{n\alpha^2}.$$

Clipping to $V_C \in [0,1]$ preserves the threshold event and does not increase variance. For the latter assertion, use $\mathrm{Var}(V_C) = \frac{1}{2}\mathbb{E}(V_C - V_C')^2$ and the fact that clipping is 1-Lipschitz. The scores from independent training samples are independent and identically distributed. Lemma G.1 therefore bounds the disagreement of each strict cut comparison by

$$\frac{1}{2K} + \frac{4\sqrt{2M}}{\alpha\sqrt{n}}.$$

At a fixed query point, only the at most k cuts in its own block matter. A union bound gives

$$\mathbb{P}\{H_S(x) \neq H_{S'}(x)\} \le \frac{k}{2K} + \frac{4\sqrt{2k}\sqrt{M}}{\alpha\sqrt{n}} \le \frac{\rho}{2} + \frac{\rho}{2} = \rho.$$

This uses the same U in all blocks and in both executions. It holds for every distribution, without conditioning on an accuracy or block-count event. □

Corollary I.10 (Products of VC-one classes)**.** *On a fixed known partition into M countable blocks, let $\mathcal{H}_j$ have VC dimension one on block j. For $0 < \alpha, \beta, \rho \le 1/4$, their full product has VC dimension $d = M$ and admits an improper agnostic (α, β)-learner that is ρ-pointwise replicable on every distribution, uses exactly*

$$B = \left\lceil \log_2 \frac{2}{\rho} \right\rceil$$

total shared fair bits, and uses

$$O\left(\frac{M/\rho^2 + M \log(2/\alpha) + \log(2/\beta)}{\alpha^2} \right)$$

examples. The block probabilities are unknown and unrestricted.

Proof. On each block apply Lemma I.3 to obtain a public countable pointwise-dense subclass with at most two cuts at each point. The product of these subclasses is pointwise dense in the original product and hence has the same class optimum for every distribution. VC dimension adds across a finite partition: one shattered point from each block gives the lower bound M, and any set of more than M points has at least two points in one block, which cannot be shattered there. Apply Theorem I.9 with $k = 2$, so its replication sample requirement is $512M/(\rho^2\alpha^2)$, and use (89) for the product class of VC dimension M. □

I.2 A LIMITATION OF CHOOSING A NEAR-OPTIMAL FIBER FIRST

Proposition I.11 (A second tolerance does not stabilize the first fiber)**.** *Consider $\mathcal{H}_\square$ from Proposition I.5. On a sample S of size n, write*

$$f_S(a) = \min_{b \ge 1} \widehat{\mathrm{R}}_S(h_{a,b}), \qquad r_S^* = \min_{a \ge 1} f_S(a).$$

Let a selector draw U uniformly from the $K = 2^B$ midpoints of $[0,1]$, where $B \ge 1$, and take the smallest a with

$$f_S(a) \le r_S^* + \alpha U/4.$$

It then returns any $h_{a,b}$ satisfying $\widehat{\mathrm{R}}_S(h_{a,b}) \le f_S(a) + \alpha/4$. Thus this includes a second shared tolerance $\alpha V/4$ with any $V \in [0,1]$ and first-eligible selection within the chosen fiber.

For every $0 < \alpha \le 1/64$, put

$$M = 2 \left\lfloor \frac{1}{2} \min\left\{ K, \frac{1}{16\alpha} \right\} \right\rfloor.$$

There is a finitely supported distribution P with deterministic labels and positive class optimum for which, at $z = (1,2)$, two independent executions sharing U satisfy

$$\lim_{n \to \infty} \mathbb{P}\{H_S(z) \ne H_{S'}(z)\} = \frac{M}{2K} = \Omega\left(\min\left\{ 1, \frac{1}{\alpha K} \right\} \right). \tag{95}$$

This limit is unaffected by additional shared randomness in the second-stage choice. Consequently, for $\rho < 1/4$, requiring this selector to be ρ-pointwise replicable on every distribution for all sufficiently large sample sizes requires $K = \Omega(1/(\alpha\rho))$. This is a limitation of the stated first-fiber rule, not of arbitrary couplings or learning algorithms.

Proof. The integer M is even, $2 \le M \le K$, and $M\alpha \le 1/16$. Put $\delta = \alpha/(4K)$, $\gamma = \alpha/2$, and

$$c_M = \frac{1}{4}, \qquad c_i = \frac{1}{4} - \delta(M - i - 1/2) \quad (0 \le i < M).$$

At the points $(i+1, j)$, where $0 \le i \le M$ and $j = 1, 2$, prescribe signed masses

$$w_{0,1} = c_0, \quad w_{0,2} = -\gamma, \qquad w_{i,1} = c_i - c_{i-1} + (-1)^i\gamma, \quad w_{i,2} = 2(-1)^{i+1}\gamma \quad (i \ge 1).$$

A signed mass w means atom probability $|w|$ and deterministic label $\mathrm{sign}(w)$. The increments of c_i are δ, except for the last increment $\delta/2$. Since M is even, the total prescribed probability is

$$L = \sum_{i,j} |w_{i,j}| = \frac{1}{4} - M\delta + (3M+1)\gamma \le \frac{1}{4} + \frac{3}{32} + \frac{1}{128} < 1.$$

Put the remaining probability $1 - L$ at $(1, 3)$ with label -1. This completes a distribution with deterministic labels.

Let $R_0 = P(Y = +1)$, the risk of the constant negative function. Write $f(a) = \min_{b \geq 1} \mathrm{R}_P(h_{a,b})$ for the population profile. Summing signed masses over a rectangle gives its risk reduction relative to R_0. At row $a = i + 1$, these reductions are

$$C_{i,1} = c_i - \gamma(i \bmod 2), \qquad C_{i,2} = c_i - \gamma(1 - (i \bmod 2)).$$

Every $b \geq 3$ additionally includes the negative atom at $(1, 3)$, making it worse than $b = 2$ in population risk. Rows beyond $M + 1$ introduce no new support points. Hence

$$f(i + 1) = R_0 - c_i \quad (0 \leq i \leq M),$$

followed by a constant profile. The best column is $b = 1$ when i is even and $b = 2$ when i is odd; the gap between these two columns is always γ. In particular the optimum is attained at $h_{M+1,1}$, and

$$\mathrm{OPT}_{\mathcal{H}_\square}(P) = R_0 - \frac{1}{4} = \frac{M(3\gamma - \delta)}{2} = M\left(\frac{3\alpha}{4} - \frac{\alpha}{8K}\right) > 0.$$

Bayes risk is zero.

For a seed $u_\ell = (\ell + 1/2)/K$ with $0 \leq \ell < M$, set $i_\ell = M - 1 - \ell$. Then

$$f(i_\ell + 1) - f(M + 1) = c_M - c_{i_\ell} = \delta(\ell + 1/2) = \alpha u_\ell / 4.$$

Keep all parameters fixed and let $n \to \infty$. There are only finitely many distinct traces on the support. With probability tending to one, every empirical fiber has its stated best column, the gaps to the other column and to all $b \geq 3$ exceed $\alpha/4$, and the strictly improving part of the empirical profile preserves its order. Its other gaps converge as well. Thus the outer choice at u_ℓ is, with probability tending to one, one of the adjacent rows $i_\ell + 1$ and $i_\ell + 2$. The second row is chosen exactly when the empirical gap of the first row exceeds its displayed population mean.

On the event that the best columns agree with their population values, this gap is a sample mean of the difference between the losses of two fixed rectangles. Its variance is nonzero. For odd i_ℓ, the two rectangles disagree at the negative atom $(1, 2)$; for even i_ℓ, they disagree at the next odd row's negative first-column atom. Both agree at the positive atom $(1, 1)$, which has positive probability. The central limit theorem consequently makes the two outer choices asymptotically equiprobable.

The adjacent rows have opposite best-column parities. At $z = (1, 2)$ their best columns give opposite predictions. Because their empirical column gaps eventually exceed $\alpha/4$, every permitted second-stage choice has the prediction of that row's best column. This assertion is uniform over all additional shared random choices. The outer choices on the two data sets are independent conditional on their shared midpoint. Each active seed therefore contributes asymptotic conditional disagreement $1/2$. For $\ell \geq M$, its tolerance strictly exceeds the first row's population gap, so the outer choice is the first row and its prediction at z is asymptotically always negative. Averaging proves (95).

Finally, $M \geq \frac{1}{2} \min\{K, 1/(16\alpha)\}$ gives the displayed order bound. If $K \leq 1/(16\alpha)$, then $M = K$ and the limit is $1/2$. Otherwise $M \geq 1/(32\alpha)$, so a limit at most $\rho < 1/4$ requires $K \geq 1/(64\alpha\rho)$. Ordinary accuracy is not the issue: the two empirical tolerances together give excess empirical risk at most $\alpha/2$, and VC uniform convergence gives agnostic (α, β) accuracy for any fixed $\beta > 0$ and all sufficiently large n. □

J ONE-BOUNDARY SELECTION ON A VC-ONE TREE

The improper learner below improves the finite-bit feasibility threshold in Corollary I.4 by selecting a union of certified tree intervals. Although the selected intervals can change when the sample changes, their membership at a fixed point is controlled by one empirical score, with a narrow indifference band. This gives the single-boundary term $1/(2K)$ for a seed with K outcomes. We derive the domain-side tree representation used in the proof, following the structure of VC-one classes in Ben-David (2015, Theorem 4).

Lemma J.1 (A grid with a local indifference band)**.** *Let S consist of n independent observations, and let F_S be a real-valued statistic whose replace-one sensitivity is at most $2/n$. Let $A_{S,w} \in \{0,1\}$ be a deterministic function of S, w. For $a > 0$ and an integer $K \geq 1$, put*

$$h = \frac{a}{K}, \qquad \eta = \frac{h}{4}, \qquad w_j = (j + 1/2)h \quad (0 \leq j < K).$$

Suppose that, for every sample and every grid value,

$$\mathbf{1}\{F_S \geq w_j + \eta\} \leq A_{S,w_j} \leq \mathbf{1}\{F_S \geq w_j - \eta\}. \tag{96}$$

If S' is an independent copy of S and J is independent and uniform on $\{0, \dots, K-1\}$, then, whenever $na^2/(32K^2) \geq \log 2$,

$$\mathbb{P}\{A_{S,w_J} \neq A_{S',w_J}\} \leq \frac{1}{2K} + \frac{8}{K} \exp\left(-\frac{na^2}{32K^2}\right). \tag{97}$$

No accuracy event is assumed in this conclusion.

Proof. Set $c = \mathbb{E}F_S$. Bounded differences gives each of the one-sided inequalities

$$\mathbb{P}\{F_S - c \geq t\}, \quad \mathbb{P}\{F_S - c \leq -t\} \leq e^{-nt^2/2} \qquad (t > 0).$$

For a fixed seed j, write $p_j = \mathbb{P}\{A_{S,w_j} = 1\}$. The two output bits are independent and identically distributed, so their disagreement probability is $2p_j(1-p_j) \leq 1/2$. If $|w_j - c| > \eta$, the appropriate side of (96) also gives

$$2p_j(1-p_j) \leq 2\exp\Big(-\frac{n}{2}(|w_j - c| - \eta)^2\Big).$$

Choose a grid point w_{j_0} closest to c, resolving a tie arbitrarily. Its contribution to the average is at most $1/(2K)$. For each $r \geq 1$, there are at most two other indices with $|j - j_0| = r$, and each satisfies

$$|w_j - c| - \eta \geq h(r - 3/4).$$

This also holds when the nearest grid point is an endpoint and c lies outside the grid interval. Consequently

$$\mathbb{P}\{A_{S,w_J} \neq A_{S',w_J}\} \leq \frac{1}{2K} + \frac{4}{K} \sum_{r \geq 1} e^{-nh^2(r-3/4)^2/2}.$$

For integers $r \geq 1$, $(r - 3/4)^2 \geq r/16$. Writing $b = nh^2/32 \geq \log 2$, the sum is at most $e^{-b}/(1 - e^{-b}) \leq 2e^{-b}$, proving the claim. □

A rooted tree order is a partial order with a least element such that the predecessors of each element form a chain. It need not have immediate successors or predecessors. A *tree interval* is an order-convex chain, with either endpoint permitted to be absent. The empty set is included. On a finite tree, a nonempty interval has the form $[v, u]$ for comparable vertices $v \preceq u$.

Lemma J.2 (Certified prefixes on a finite tree)**.** *Let T be a finite rooted tree with root r, empirical vertex weights $\widehat{a}_v$, and true vertex weights a_v. Let $E \subseteq T$ contain every vertex with a nonzero empirical weight, and give E a single fixed total order. Write*

$$\widehat{G}(v,u) = \sum_{z \in [v,u]} \widehat{a}_z, \qquad \widehat{F}_v = \max\big(\{0\} \cup \{\widehat{G}(v,u) : u \in E,\ v \preceq u\}\big).$$

Whenever $\widehat{F}_v > 0$, let $u(v)$ be the first maximizing endpoint in the order on E. Suppose $W > \eta > 0$. Call v strong if $\widehat{F}_v \geq W + \eta$ and weak if $\widehat{F}_v \geq W - \eta$. For every strong v, let P_v be the prefix of $[v, u(v)]$ ending immediately before its first nonweak vertex, or the whole path if there is no such vertex. Set $B = \bigcup_{v \text{ strong}} P_v$. Then

$$\{v : \widehat{F}_v \geq W + \eta\} \subseteq B \subseteq \{v : \widehat{F}_v \geq W - \eta\}. \tag{98}$$

If

$$\left|\sum_{v \in I} (a_v - \widehat{a}_v)\right| \leq \eta \quad \text{for every tree interval } I \subseteq T, \tag{99}$$

then

$$\sum_{v \in B} a_v \geq 0, \qquad \sum_{v \in B} a_v \geq \widehat{F}_r - W - \eta. \tag{100}$$

The set B is a disjoint union of finitely many tree intervals, each of strictly positive true gain.

Proof. Every strong vertex is weak and belongs to its own prefix, and a selected prefix contains only weak vertices. This proves (98).

The fixed endpoint order makes the maximizing suffix consistent. Suppose $x \in [v, u(v)]$ and $\widehat{F}_x > 0$. Every competing endpoint below x can also be used below v. Subtracting the common prefix before x shows that $u(v)$ maximizes the gain from x. That suffix gain must be positive: otherwise a positive-gain endpoint for x would improve the choice for v. A tie at x is also a tie at v. The same endpoint order therefore gives $u(x) = u(v)$.

If two selected prefixes intersect, their starting vertices are comparable. The later starting vertex lies in the earlier prefix and has positive score. By the preceding argument the two endpoints agree, and the weak-vertex stopping rule agrees on their common suffix. Thus the later prefix is contained in the earlier one. The inclusion-maximal prefixes are consequently pairwise disjoint and their union is B.

If P_v stops before a nonweak vertex z, then

$$\widehat{G}(P_v) = \widehat{F}_v - \widehat{G}(z, u(v)) \geq \widehat{F}_v - \widehat{F}_z > 2\eta.$$

If it reaches its endpoint, then $\widehat{G}(P_v) = \widehat{F}_v \geq W + \eta > 2\eta$. By (99), every maximal prefix has true gain greater than η. Summing their disjoint gains proves the first inequality in (100).

If r is strong, its prefix is maximal. A truncated root prefix has empirical gain greater than $\widehat{F}_r - W + \eta$, and hence true gain greater than $\widehat{F}_r - W$. A complete root prefix has true gain at least $\widehat{F}_r - \eta$. The other maximal prefixes have positive true gains, so in either case the second inequality in (100) follows. If r is not strong, then $\widehat{F}_r < W + \eta$, and that inequality follows from the nonnegative total gain already proved. □

For a countable rooted tree, the class of all tree intervals has VC dimension at most two. Three incomparable or partly incomparable vertices cannot be shattered, since a chain cannot contain two incomparable vertices. Three comparable vertices cannot be shattered, since an interval cannot contain the first and third while omitting the second. In particular, its number of traces on $m \geq 2$ points is at most $1 + m + \binom{m}{2}$ by Sauer's lemma.

Let $n_{\mathrm{int}}(\varepsilon, \delta)$ be a universal sample bound that guarantees, for every countable rooted tree, every distribution of a vertex V and a sign $Z \in \{-1, +1\}$,

$$\mathbb{P}\left\{ \sup_{I \text{ tree interval}} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i \mathbf{1}\{V_i \in I\} - \mathbb{E}[Z\mathbf{1}\{V \in I\}] \right| > \varepsilon \right\} \leq \delta \quad (n \geq n_{\mathrm{int}}(\varepsilon, \delta)). \tag{101}$$

An additional cemetery value for V, belonging to no interval, is allowed. Such a bound exists with

$$n_{\mathrm{int}}(\varepsilon, \delta) = O\left(\frac{\log(2/\delta)}{\varepsilon^2} \right). \tag{102}$$

The empirical-process bound removes the logarithmic factor in $1/\varepsilon$.

Fix the sample and let $\mathcal{V} = \{(Z_i \mathbf{1}\{V_i \in I\})_{i=1}^n : I \text{ a tree interval}\}$, equipped with the normalized Euclidean metric $\|v - w\|_n^2 = n^{-1} \sum_i (v_i - w_i)^2$. Multiplication by the fixed signs Z_i does not change distances, and squared distance between two indicator vectors is their normalized Hamming distance. Haussler's packing bound for a binary class of VC dimension at most two (Haussler, 1995, Theorem 1) therefore gives internal u-nets of $\mathcal{V}$ with at most Cu^{-4} elements, for $0 < u \leq 1$ and a universal C. The bound applies to the finite trace class, so no separability condition is needed here.

Choose a 2^{-j}-net $\mathcal{V}_j$ for each $j \geq 1$, put $\mathcal{V}_0 = \{0\}$, and let $\pi_j(v)$ be a nearest net point. The zero vector is available because the empty interval is included. We may take $\pi_j(v) = v$ for all sufficiently large j, since $\mathcal{V}$ is finite. Then

$$v = \sum_{j \geq 1} \big(\pi_j(v) - \pi_{j-1}(v)\big), \qquad \|\pi_j(v) - \pi_{j-1}(v)\|_n \leq 3 \cdot 2^{-j}.$$

There are at most $|\mathcal{V}_j||\mathcal{V}_{j-1}|$ different increments at level j, whose logarithm is $O(j + 1)$. For independent Rademacher signs σ_i, the exponential-moment bound for $n^{-1} \sum_i \sigma_i w_i$ gives

$$\mathbb{E}_\sigma \sup_{v \in \mathcal{V}} \left| \frac{1}{n} \sum_i \sigma_i v_i \right| \leq \frac{C}{\sqrt{n}} \sum_{j \geq 1} 2^{-j} \sqrt{j + 1} \leq \frac{C'}{\sqrt{n}}.$$

Symmetrization bounds the expected supremum in (101) by $2C'/\sqrt{n}$. That supremum has replace-one sensitivity at most $2/n$, so bounded differences gives an additional $\sqrt{2\log(1/\delta)/n}$ with failure probability at most δ. This proves (102). Countability ensures measurability of the sample-dependent suprema, even though the interval class can be uncountable.

Theorem J.3 (VC-one learning above the single-boundary threshold)**.** *Let $\mathcal{H}$ be a nonempty binary class of VC dimension at most one on a countable domain. Let $0 < \alpha, \beta, \rho \le 1/4$, let $B \ge 0$ be an integer, and put $K = 2^B$. Suppose*

$$\Delta = \rho - \frac{1}{2K} > 0.$$

There is an improper agnostic (α, β)-learner for $\mathcal{H}$ that is ρ-pointwise replicable, uses exactly B shared fair bits and no other randomness, and uses exactly

$$n = \max\left\{\left\lceil \frac{32K^2}{\alpha^2}\log\left(2 + \frac{8}{K\Delta}\right)\right\rceil, n_{\mathrm{int}}\left(\frac{\alpha}{4K}, \beta\right)\right\} \tag{103}$$

examples. Thus the sample bound is

$$O\left(\frac{K^2}{\alpha^2}\left[\log\left(2 + \frac{1}{K\Delta}\right) + \log\frac{2}{\beta}\right]\right). \tag{104}$$

All structural choices are fixed independently of the unknown distribution. The theorem is a statistical existence result and does not assert efficient computation of the class representation.

Proof. **A countable public tree.** Fix $f_0 \in \mathcal{H}$ and put $A_x = \{h \in \mathcal{H} : h(x) \ne f_0(x)\}$. The nonempty sets A_x form a laminar family. Otherwise three hypotheses from $A_x \cap A_y$, $A_x \setminus A_y$, and $A_y \setminus A_x$, together with f_0, would realize all four patterns on $\{x, y\}$. Ignore coordinates with $A_x = \varnothing$, since every comparator agrees there with f_0. Identify coordinates with equal nonempty A_x, and write $q(x)$ for the resulting equivalence class. There are countably many classes. Order them by reverse inclusion of their profiles:

$$v \preceq w \quad \Longleftrightarrow \quad A_w \subseteq A_v.$$

After equality has been quotiented out this is a partial order. The ancestors of any vertex are linearly ordered, because their profiles all contain one nonempty profile and laminarity makes them comparable. Add an artificial least root r. This produces a countable rooted tree T. Fix a public enumeration of its vertices.

For every $h \in \mathcal{H}$, its flipped vertices form a downward closed chain in $T \setminus \{r\}$: membership in a profile implies membership in its supersets, and two profiles containing h cannot be disjoint. After adjoining r, this is an initial chain and hence a tree interval. The chain property is essential: a branching downward-closed set is not a path.

Transform observations by $Z = -Y f_0(X)$ and $V = q(X)$, sending ignored coordinates to the cemetery value. The artificial root has zero probability. Define signed vertex masses and their empirical counterparts by

$$a_v = \mathbb{E}[Z\mathbf{1}\{V = v\}], \qquad \widehat{a}_v = \frac{1}{n}\sum_{i=1}^{n} Z_i \mathbf{1}\{V_i = v\}.$$

The grouping of coordinates is legitimate since $\sum_{v \in T} |a_v| \le 1$. Flipping f_0 on a set of vertices C reduces its risk by $a(C) = \sum_{v \in C} a_v$. All these sums are absolutely convergent.

A finite tree determined by the sample. Let S_0 be the set of noncemetery vertices observed in the sample. If S_0 is empty, output f_0; the score defined below is then zero everywhere. Otherwise, for $v \in T$ put

$$E_v = \{s \in S_0 : v \preceq s\}.$$

The distinct nonempty sets E_v form a finite laminar family $\mathcal{L}$, containing $S_0 = E_r$. Indeed, vertices with a common observed descendant are comparable, and vertices without a common observed descendant have disjoint E sets. For $E \in \mathcal{L}$, define its cell

$$C_E = \{v \in T : E_v = E\}.$$

Each cell is an order-convex chain: all its vertices precede any fixed member of E, and every vertex between two of them has the same set of observed descendants. It contains at most one point of S_0. If it contains such a point s, that point is its greatest vertex, because $s \in E_v$ for every $v \in C_E$.

Order the cells by reverse inclusion of their E sets. They form a finite rooted tree with root C_{S_0}. A laminar family of distinct nonempty subsets of a set of size s has at most $2s-1$ members: adjoin the whole set and all singletons, form its inclusion tree, and note that every internal vertex then has at least two children. Consequently this sample tree has at most $2|S_0|-1$ vertices. Its empirical vertex weight is $\widehat{a}(C_E)$.

If $E' \subsetneq E$ are nonempty cells, every vertex of C_E precedes every vertex of $C_{E'}$: both precede a common member of E', and the opposite order would reverse the inclusion of their observed-descendant sets. It follows that a directed interval of the finite cell tree expands to an order-convex chain of T. Any vertex between two expanded cells has a nonempty observed-descendant set between theirs, so no intermediate vertex is omitted. Thus every interval used below is a genuine original-tree interval, including when a physical endpoint is absent.

Scores and the learner. For every original-tree vertex v, define

$$\widehat{F}_v = \max\left\{0, \sup_{u \succeq v} \frac{1}{n}\sum_{i=1}^{n} Z_i \mathbf{1}\{V_i \in [v,u]\}\right\}. \tag{105}$$

If this score is positive, its maximum is attained at some $u \in S_0$ below v. In fact, a positive-gain path contains an observed vertex, and its last observed vertex has the same empirical gain. Only finitely many observed endpoints need be considered. Let $u(v)$ be the first maximizer in the public enumeration of S_0.

For $v \in C_E$, the empirical interval gain to an observed endpoint $u \in E$ is

$$\sum_{s \in S_0:\, v \preceq s \preceq u} \widehat{a}_s.$$

It depends only on E and u. It is also the gain along the corresponding cell-tree path. Thus $\widehat{F}_v$ and its canonical maximizing endpoint are constant throughout each cell and equal the corresponding finite-tree quantities. Vertices with $E_v = \varnothing$ have score zero.

Use the B shared bits to draw J uniformly from $\{0, \dots, K-1\}$ and set

$$W = (J + 1/2)\frac{\alpha}{K}, \qquad \eta = \frac{\alpha}{4K}.$$

Apply the prefix rule of Lemma J.2 on the finite cell tree, using the cells that contain observed vertices as its endpoint set. They include every cell with nonzero empirical weight, and their endpoint order is inherited from the public vertex enumeration. Let $B_{S,W}$ be the union of the selected cells, expanded to the original tree. Output

$$H_{S,J}(x) = \begin{cases} -f_0(x), & A_x \neq \varnothing \text{ and } q(x) \in B_{S,W}, \\ f_0(x), & \text{otherwise.} \end{cases}$$

This is a deterministic function of the sample and the single finite seed. In particular, neither the cell tree nor any tie is randomized. Sample-dependent cells need not agree between runs. Since the labeled sample space is countable and the seed space is finite, the learner is measurable.

Accuracy. Use (101) with error η and failure probability β. On its good event, every interval of the sample cell tree satisfies (99), because its expansion is an original-tree interval. There are only finitely many selected maximal prefixes. Lemma J.2 therefore gives

$$a(B_{S,W}) \geq \widehat{F}_r - W - \eta. \tag{106}$$

This assertion also holds when S_0 is empty.

For each $h \in \mathcal{H}$, let D_h be its flipped vertices together with r. This is an initial chain. If it meets S_0, its finite set of observed vertices has a last element, and its empirical gain equals that of the root path to that element. If it does not meet S_0, its empirical gain is zero. Hence

$$a(D_h) \leq \widehat{a}(D_h) + \eta \leq \widehat{F}_r + \eta.$$

Combining this with (106) gives, simultaneously for every comparator and every seed,

$$\mathrm{R}_P(H_{S,J}) - \mathrm{R}_P(h) = a(D_h) - a(B_{S,W}) \le W + 2\eta \le \alpha.$$

The last inequality uses the largest grid midpoint $W \le \alpha - \alpha/(2K)$. Taking the true-risk infimum over h proves the agnostic guarantee. The union of selected intervals can improve on every single path; no properness claim is made.

Replication. Fix a test point whose profile is nonempty and let $v = q(x)$. The finite-tree sandwich and the equality of scores on cells imply, for every sample,

$$\mathbf{1}\{\widehat{F}_v \ge W + \eta\} \le \mathbf{1}\{v \in B_{S,W}\} \le \mathbf{1}\{\widehat{F}_v \ge W - \eta\}.$$

It holds also for unobserved vertices with empty E_v, since their score is zero and $W - \eta > 0$. Replacing one observation changes each empirical signed interval gain in (105) by at most $2/n$. The maximum with zero and the supremum preserve this bound, so the original score has replace-one sensitivity at most $2/n$ regardless of how the sample cells change. Lemma J.1, with $a = \alpha$, gives

$$\mathbb{P}\{H_{S,J}(x) \ne H_{S',J}(x)\} \le \frac{1}{2K} + \frac{8}{K} e^{-n\alpha^2/(32K^2)} \le \frac{1}{2K} + \Delta = \rho.$$

The sample bound in (103) ensures the lemma's logarithmic condition and the final inequality. Coordinates with an empty profile always receive $f_0(x)$ and have disagreement zero. The proof is unconditional on the accuracy event, so no additional β term enters replication. Finally, the only random choice is the index J, which is drawn using exactly B fair shared bits. □

Remark J.4 (Feasibility and sample dependence). The theorem works for every prescribed finite seed budget satisfying the strict inequality $\rho > 1/(2K)$. In particular, choosing $K = \Theta(1/\rho)$ with constant slack gives $O(\log(2/\beta)/(\rho^2\alpha^2))$ samples and $O(\log(1/\rho))$ total shared fair bits for every VC-one class. This upper bound includes confidence dependence. It leaves the joint optimum in α, β, ρ, B unspecified.

Corollary J.5 (The VC-one tradeoff at fixed confidence)**.** *Fix $\beta_0 \in (0, 1/4]$. For every binary class $\mathcal{H}$ of VC dimension exactly one on a countable domain, every $0 < \alpha \le 1/4$, every integer $B \ge 2$, and $K = 2^B$, uniformly for $0 < K\Delta \le 1/4$,*

$$m^*_{\mathcal{H}}(\alpha, \beta_0, 1/(2K) + \Delta; B) = \Theta_{\beta_0}\left(\frac{K^2}{\alpha^2} \log \frac{1}{K\Delta}\right). \tag{107}$$

Here the minimum ranges over fixed-sample learners using at most B total shared fair bits and no other randomness, and improper outputs are allowed. The constants are independent of the class, its domain, α, B, Δ; they may depend on the fixed confidence level β_0.

Proof. Theorem J.3 gives the upper bound, since $\log(1/(K\Delta)) \ge \log 4$ absorbs the fixed confidence term. Proposition F.5 gives the lower bound. In the displayed range, $\log(1/(2K\Delta))$ and $\log(1/(K\Delta))$ differ by at most a factor of two. □

J.1 An additive confidence term

The confidence factor in (104) can be removed from the term multiplied by K^2. For this purpose, reserve a constant fraction of the accuracy tolerance by shortening the threshold grid. The remaining argument controls the total negative gain of the selected intervals, rather than certifying each interval separately at confidence $1 - \beta$.

Lemma J.6 (Unions of tree intervals)**.** *For any rooted tree order and any integer $m \ge 1$, the class $\mathcal{U}_m$ of unions of at most m tree intervals has VC dimension at most $2m$.*

Proof. Fix a set Q of $2m + 1$ distinct vertices. Form a finite forest on Q by making the nearest strict ancestor in Q, when one exists, the parent of each vertex. The parent exists uniquely whenever the set of strict ancestors in Q is nonempty, because that set is a finite chain. Properly color this forest with two colors, and let Q_+ be a color class of size at least $m + 1$. No tree interval avoiding $Q \setminus Q_+$ can contain two points of Q_+. Incomparable points cannot belong to one chain. Between two comparable points of Q_+, the ancestor path in the forest contains a point of the other color; order convexity would include that point in the interval. Thus any union of at most m intervals that avoids $Q \setminus Q_+$ contains at most m points of Q_+. The labeling with positive set Q_+ cannot be realized, so Q is not shattered. □

Lemma J.7 (Deviation with a cost per interval)**.** *Let T be a countable rooted tree, let (V, Z) have any distribution on $T \times \{-1, +1\}$, with an optional cemetery value belonging to no interval, and write*

$$a(U) = \mathbb{E}[Z\mathbf{1}\{V \in U\}], \qquad \widehat{a}(U) = \frac{1}{n}\sum_{i=1}^{n} Z_i \mathbf{1}\{V_i \in U\}.$$

There is a universal constant C_0 such that, for all $\eta, t > 0$ and $n\eta^2 \geq C_0$,

$$\mathbb{P}\left\{\exists m \geq 1,\ U \in \mathcal{U}_m : \widehat{a}(U) - a(U) > 2\eta m + t\right\} \leq e^{-nt^2/2}. \tag{108}$$

Proof. Put

$$D_m = \sup_{U \in \mathcal{U}_m} \big(\widehat{a}(U) - a(U)\big).$$

The same packing and chaining argument used for (102), now with VC dimension at most $2m$ from Lemma J.6, gives a universal constant c such that

$$\mathbb{E}D_m \leq c\sqrt{\frac{m}{n}}. \tag{109}$$

Indeed, Haussler's packing bound gives normalized Euclidean covering numbers with $\log N(u) \leq c' m \log(c'/u)$ for $0 < u \leq 1$. Multiplication by the observed signs preserves the metric. Dyadic chaining therefore bounds the conditional Rademacher average by $c''\sqrt{m/n}\sum_{j\geq 1} 2^{-j}\sqrt{j+1}$; symmetrization proves (109). All constants here are independent of m.

Replacing one observation changes D_m by at most $2/n$. Choose $C_0 \geq \max\{c^2, 2\log 2\}$. If $n\eta^2 \geq C_0$, then $\mathbb{E}D_m \leq \eta\sqrt{m} \leq \eta m$. Bounded differences and a union bound over the positive integers give

$$\begin{aligned}
\mathbb{P}\{\exists m \geq 1 : D_m > 2\eta m + t\} &\leq \sum_{m\geq 1} e^{-n(t+\eta m)^2/2} \\
&\leq e^{-nt^2/2} \sum_{m\geq 1} e^{-n\eta^2 m^2/2} \\
&\leq e^{-nt^2/2} \frac{e^{-n\eta^2/2}}{1 - e^{-n\eta^2/2}} \leq e^{-nt^2/2}.
\end{aligned}$$

The labeled sample space is countable, so all these suprema and events are measurable, including when the interval classes themselves are uncountable. □

Remark J.8 (A general range-family version). The same deviation argument applies to any range family containing the empty set, on a countable domain, with VC dimension at most an integer $d \geq 1$. The standard union-class bound gives VC dimension $O(dm\log(em))$ for unions of at most m ranges (Blumer et al., 1989); it also follows by bounding their growth function by the mth power of the original growth function and applying Sauer's lemma. Thus the corresponding signed deviations satisfy $\mathbb{E}D_m \leq c\sqrt{dm\log(em)/n}$. Since $\log(em) \leq m$ for $m \geq 1$, the condition $n\eta^2 \geq Cd$ again implies $\mathbb{E}D_m \leq \eta m$, and (108) follows unchanged. Using this confidence argument in learning still requires a construction of disjoint selected ranges with the empirical gain margin, together with suitable comparator and replication bounds.

Theorem J.9 (VC-one learning with additive confidence)**.** *Let $\mathcal{H}$ be a nonempty binary class of VC dimension at most one on a countable domain. Let $0 < \alpha, \beta, \rho \leq 1/4$, let $B \geq 0$ be an integer, put $K = 2^B$, and suppose $\Delta = \rho - 1/(2K) > 0$. There is a universal constant C for which an improper agnostic (α, β)-learner for $\mathcal{H}$ is ρ-pointwise replicable on every distribution, uses exactly B shared fair bits and no other randomness, and uses exactly*

$$n = \max\left\{\left\lceil \frac{128K^2}{\alpha^2}\log\left(2 + \frac{8}{K\Delta}\right)\right\rceil, \left\lceil \frac{CK^2}{\alpha^2}\right\rceil, \left\lceil \frac{C}{\alpha^2}\log\frac{2}{\beta}\right\rceil\right\} \tag{110}$$

examples. In particular, the sample bound is

$$O\left(\frac{1}{\alpha^2}\left[K^2\log\left(2 + \frac{1}{K\Delta}\right) + \log\frac{2}{\beta}\right]\right). \tag{111}$$

All structural choices are fixed independently of the unknown distribution, and replication is unconditional on the accuracy event.

Proof. Use the public tree, empirical scores, sample cell tree, and certified-prefix construction from the proof of Theorem J.3. Change only the thresholds to

$$W = (J + 1/2)\frac{\alpha}{2K}, \qquad \eta = \frac{\alpha}{8K}, \qquad J \sim \mathrm{Unif}\{0, \dots, K-1\}.$$

As before, J is drawn with exactly B shared fair bits; all subsequent operations are deterministic.

Write $\varepsilon = \alpha/8$ and $t = \alpha/4$. Increase the universal constant C in (110) so that $n\eta^2 \geq C_0$, $n \geq n_{\mathrm{int}}(\varepsilon, \beta/2)$, and $e^{-nt^2/2} \leq \beta/2$. Equations (101) and (108) then imply, with probability at least $1 - \beta$, both

$$|\widehat{a}(I) - a(I)| \leq \varepsilon \qquad \text{for every tree interval } I, \tag{112}$$

$$\widehat{a}(U) - a(U) \leq 2\eta m + t \qquad \text{for every } m \geq 1 \text{ and } U \in \mathcal{U}_m. \tag{113}$$

Work on this common event, which does not depend on the seed. Let $P_1, \dots, P_M$ be the inclusion-maximal selected prefixes, expanded from the sample cells into the original tree. The purely empirical part of Lemma J.2 shows that these are disjoint tree intervals and that $\widehat{a}(P_i) > 2\eta$ for every i; no uniform convergence assumption is needed for those facts.

Let $\mathcal{N} = \{i : a(P_i) < 0\}$ and put $L = -\sum_{i \in \mathcal{N}} a(P_i)$. If $m = |\mathcal{N}| > 0$, their disjoint union $U = \bigcup_{i \in \mathcal{N}} P_i$ belongs to $\mathcal{U}_m$ and has $\widehat{a}(U) > 2\eta m$. Thus (113) gives

$$L = -a(U) \leq \widehat{a}(U) - a(U) - 2\eta m \leq t.$$

If $\mathcal{N}$ is empty, the same conclusion holds because $L = 0$. In particular, the total selected true gain satisfies $a(B_{S,W}) \geq -L$.

If the root is strong, its prefix P_r is one of the maximal prefixes. If it stops before a nonweak vertex z, the endpoint consistency in the finite-tree proof gives

$$\widehat{a}(P_r) = \widehat{F}_r - \widehat{G}(z, u(r)) \geq \widehat{F}_r - \widehat{F}_z > \widehat{F}_r - W + \eta.$$

If it reaches its endpoint, then $\widehat{a}(P_r) = \widehat{F}_r$. In either case (112) implies

$$a(B_{S,W}) \geq a(P_r) - L \geq \widehat{F}_r - W - \varepsilon - L.$$

If the root is not strong, then $\widehat{F}_r < W + \eta \leq W + \varepsilon$, so the same bound follows from $a(B_{S,W}) \geq -L$. It also holds for an empty observed tree, when both $\widehat{F}_r$ and $a(B_{S,W})$ are zero.

Every comparator's flipped initial chain D_h has empirical gain at most $\widehat{F}_r$, exactly as in the original proof. By (112), $a(D_h) \leq \widehat{F}_r + \varepsilon$. Consequently, simultaneously for every comparator and every seed,

$$\mathrm{R}_P(H_{S,J}) - \mathrm{R}_P(h) = a(D_h) - a(B_{S,W}) \leq W + 2\varepsilon + L \leq \frac{\alpha}{2} + \frac{\alpha}{4} + \frac{\alpha}{4} = \alpha.$$

This proves agnostic accuracy, including arbitrary misspecification. The events above range over the original tree, so they apply to the sample-dependent cells and to intervals containing unobserved vertices.

For replication, the pointwise sandwich still holds for every sample with these new W, η, and each original score has replace-one sensitivity at most $2/n$. Apply Lemma J.1 with $a = \alpha/2$. For every fixed test point, independently of either accuracy event,

$$\mathbb{P}\{H_{S,J}(x) \neq H_{S',J}(x)\} \leq \frac{1}{2K} + \frac{8}{K} e^{-n\alpha^2/(128K^2)} \leq \frac{1}{2K} + \Delta = \rho.$$

The first sample term ensures the lemma's logarithmic condition and the last inequality. Points with an empty profile still receive the fixed prediction $f_0(x)$. The same countability argument as before proves measurability and covers all points outside the observed support. No additional random choice has been introduced. □

Remark J.10 (Why the grid is shortened). The earlier grid reaches within $O(\alpha/K)$ of the full tolerance α. Even on a one-point domain, its largest threshold can then be only $O(\alpha/K)$ from a distribution at which the base prediction has excess risk greater than α. Shortening the grid leaves a constant multiple of α for the two aggregate accuracy deviations. The cost per selected interval remains of order α/K, as required by the replication argument.

Corollary J.11 (The full VC-one tradeoff)**.** *For every binary class $\mathcal{H}$ of VC dimension exactly one on a countable domain, every $0 < \alpha, \beta \le 1/4$, every integer $B \ge 2$, and $K = 2^B$, uniformly for $0 < K\Delta \le 1/4$,*

$$m^*_{\mathcal{H}}(\alpha, \beta, 1/(2K) + \Delta; B) = \Theta\left(\frac{1}{\alpha^2}\left[K^2 \log\frac{1}{K\Delta} + \log\frac{1}{\beta}\right]\right). \tag{114}$$

The constants are absolute and independent of the class and its domain. The minimum ranges over fixed-sample learners using at most B total shared fair bits and no other randomness, and improper outputs are allowed.

Proof. Theorem J.9 gives the upper bound; in the displayed range, $\rho = 1/(2K) + \Delta \le 3/(4K) \le 1/4$ and $\log(1/(K\Delta)) \ge \log 4$. Since $\mathrm{VC}(\mathcal{H}) = 1$, some point admits both comparator labels. Restricting the feature distribution to that point turns any learner for $\mathcal{H}$, including an improper learner, into a learner for the two hypotheses on one point with the same sample size, bits, confidence, and pointwise replication. Proposition F.7 therefore gives the matching joint lower bound. □

K Capped certificates for tree intervals

The capped recursion extends tree-prefix selection to comparators starting at any vertex. Disjoint selected intervals account for the saturation surplus. This gives a VC-two family outside the bounded-cut hypothesis-order method, including the class of all finite intervals on $\mathbb{N}$. Extending this construction to general VC-two classes, including anchored rectangles on $\mathbb{N}^2$, remains open.

Lemma K.1 (Capped path certificates)**.** *Let T be a finite rooted tree with real vertex weights $\widehat{a}_v$, and fix $\lambda > 0$. Compute, from the leaves toward the root,*

$$\begin{aligned} M_v &= \max\big(\{0\} \cup \{F_c : c \text{ is a child of } v\}\big), \\ F_v = \min\{\lambda, \max\{0, \widehat{a}_v + M_v\}\}, &\qquad e_v = (\widehat{a}_v + M_v - \lambda)_+. \end{aligned} \tag{115}$$

If $M_v > 0$, choose a child attaining M_v, resolving ties by a fixed public order, and call it the successor of v. If $M_v = 0$, give v no successor. Suppose $0 < \eta < W$ and $W + \eta \le \lambda$. Call v strong if $F_v \ge W + \eta$ and weak if $F_v \ge W - \eta$. From every strong vertex, follow successors until just before the first nonweak vertex, or until there is no successor. Let B be the union of these prefixes, and let $\mathcal{P}$ be its inclusion-maximal selected prefixes. Then:

1. *The members of $\mathcal{P}$ are pairwise disjoint tree intervals, and*
$$\{v : F_v \ge W + \eta\} \subseteq B \subseteq \{v : F_v \ge W - \eta\}. \tag{116}$$
2. *Every v with $e_v > 0$ lies in B. For each $P \in \mathcal{P}$, writing $E_P = \sum_{v \in P} e_v$, one has*
$$\widehat{a}(P) > 2\eta + E_P, \qquad \sum_{P \in \mathcal{P}} E_P = \sum_{v \in T} e_v. \tag{117}$$
3. *For every tree interval $I \subseteq T$,*
$$\widehat{a}(I) \le \lambda + \sum_{v \in I} e_v \le \lambda + \sum_{v \in T} e_v. \tag{118}$$

Proof. There is a single chosen successor at every vertex. Thus, if two selected prefixes meet, their starting vertices are comparable, and the later prefix is contained in the earlier one: their successor choices and stopping rule agree after the later starting point. Their maximal members are therefore disjoint. The strong and weak definitions give (116). Also, $e_v > 0$ implies $F_v = \lambda$, so every positive-surplus vertex is strong.

Put $d_v = (-\widehat{a}_v - M_v)_+$. The recursion gives

$$\widehat{a}_v = F_v - M_v + e_v - d_v. \tag{119}$$

On every selected prefix, $F_v \ge W - \eta > 0$, so $d_v = 0$. For a prefix P starting at s and ending at t, the chosen-successor equalities telescope to

$$\widehat{a}(P) = F_s - M_t + E_P.$$

If t has a successor, that successor is nonweak, and hence $M_t < W - \eta$. If it has none, $M_t = 0 < W - \eta$. Since $F_s \geq W + \eta$, the first inequality in (117) follows. Surplus coverage and disjointness give its equality.

For an arbitrary interval $I = [s, t]$, consecutive vertices need not follow the chosen successors. Nevertheless, $M_v \geq F_c$ for its next vertex c. Dropping the nonnegative d_v in (119) and telescoping gives

$$\widehat{a}(I) \leq F_s - M_t + \sum_{v \in I} e_v \leq \lambda + \sum_{v \in I} e_v.$$

The empty interval satisfies the same bound. □

The score in (115) admits a public-path description that is useful when the finite tree depends on the sample. For $a \in \mathbb{R}$, set $\phi_a(t) = \min\{\lambda, \max\{0, a + t\}\}$. For a path $[v, u]$, compose the maps $\phi_{\widehat{a}_z}$ along the path, starting with zero at its far end. The maximum of these capped path values is F_v. Indeed, ϕ_a is monotone, so maximizing over the next vertex commutes with applying ϕ_a. Zero weights can be inserted or removed, since ϕ_0 is the identity on $[0, \lambda]$. Finally, changing the path weights by δ_z changes its value by at most $\sum_z |\delta_z|$, by the nonexpansiveness of clipping.

Theorem K.2 (Tree intervals with a prescribed finite seed)**.** *Let T be a countable rooted tree order and let*

$$\mathcal{H}_{\mathrm{int}}(T) = \{2\mathbf{1}_I - 1 : I \text{ is a tree interval in } T\}.$$

The empty interval is permitted. Let $0 < \alpha, \beta, \rho \leq 1/4$, let $B \geq 0$ be an integer, and put $K = 2^B$. If $\Delta = \rho - 1/(2K) > 0$, there is an improper agnostic (α, β)-learner that is ρ-pointwise replicable, uses exactly B total shared fair bits and no other randomness, and uses a fixed number of examples satisfying

$$n = O\left(\frac{1}{\alpha^2}\left[K^2 \log\left(2 + \frac{1}{K\Delta}\right) + \log\frac{2}{\beta}\right]\right). \tag{120}$$

In particular, choosing $K = \Theta(1/\rho)$ with constant slack gives $O(\alpha^{-2}[\rho^{-2} + \log(2/\beta)])$ samples and $O(\log(1/\rho))$ total shared bits. All choices are fixed independently of the unknown distribution, and no assumption is made on its misspecification error.

Proof. Use the constant negative classifier as the baseline, and write

$$a_v = \mathbb{E}[Y\mathbf{1}\{X = v\}], \qquad \widehat{a}_v = \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbf{1}\{X_i = v\}.$$

Flipping the baseline on a set A reduces its risk by $a(A)$. Set

$$\lambda = \frac{\alpha}{2}, \qquad \eta = \frac{\lambda}{4K}, \qquad W_j = (j + 1/2)\frac{\lambda}{K} \quad (0 \leq j < K).$$

These choices ensure $\eta < W_j$ and $W_j + \eta < \lambda$ for every grid value.

Sample cells and public scores. Let S_0 be the finite set of observed vertices. If it is empty, output the baseline. Otherwise define $E_v = \{s \in S_0 : v \preceq s\}$ and, for each distinct nonempty such set E, let $C_E = \{v : E_v = E\}$. As in the sample-cell construction in the proof of Theorem J.3, these cells are order-convex chains, each containing at most one observed vertex, which is its greatest vertex when it exists. The cells, ordered by reverse inclusion of their E sets, form a finite rooted tree. Every directed interval of this cell tree expands to an original-tree interval. Give each cell weight $\widehat{a}(C_E)$ and run the capped recursion on this finite tree. Fix its successor ties deterministically, for example by the least index in a public enumeration of T among observed vertices below the candidate child.

At every original vertex v, define the score by

$$F_{S,v} = \sup_{u \succeq v} \Gamma_S(v, u), \tag{121}$$

where $\Gamma_S(v, u)$ is the capped backward composition of the empirical weights at the observed vertices in $[v, u]$, and is zero if there are none. Only finitely many observed endpoints need be considered: an endpoint after the last observed vertex has the same path value as that observed endpoint. The maximum is therefore attained whenever it is positive. For v in a cell C_E, the observed vertices on

a path to a fixed observed endpoint depend only on E. The preceding path description shows that $F_{S,v}$ is constant on the cell and equals the score of the finite cell-tree recursion. When E_v is empty, $F_{S,v} = 0$.

Replacing one observation changes the empirical signed weights on any fixed path by total absolute amount at most $2/n$. Insert zero weights to align the two finite observed lists before applying the nonexpansiveness bound. Every $\Gamma_S(v, u)$, and hence its supremum $F_{S,v}$, has replace-one sensitivity at most $2/n$. This statement is about fixed public paths and therefore remains valid when the sample cells change.

Draw J uniformly from $\{0, \dots, K-1\}$ using exactly B shared fair bits. Apply the strong/weak prefix rule of Lemma K.1 on the cell tree, using W_J, and expand its selected union to a set $B_{S,J} \subseteq T$. Output $2\mathbf{1}_{B_{S,J}} - 1$. This is a deterministic function of the labeled sample and the finite seed; their countable sample space also ensures measurability.

Accuracy. Write $E_* = \sum_C e_C$ for the sum of cell surpluses and let $\mathcal{P}_J$ be the expanded maximal selected prefixes. For $P \in \mathcal{P}_J$, let E_P be its cell surplus. Lemma K.1 gives

$$\widehat{a}(P) > 2\eta + E_P, \qquad \sum_{P \in \mathcal{P}_J} E_P = E_*.$$

Consequently, if

$$L_J = \sum_{P \in \mathcal{P}_J} (E_P - a(P))_+,$$

then $a(B_{S,J}) \geq E_* - L_J$. For any $m \geq 1$, let $\mathcal{U}_m$ consist of unions of at most m tree intervals, and put

$$D_m = \sup_{A \in \mathcal{U}_m} \big(\widehat{a}(A) - a(A)\big), \qquad Z_\eta = \max\{0, \sup_{m \geq 1} (D_m - 2\eta m)\}.$$

Taking the union of those selected components for which $E_P > a(P)$ proves $L_J \leq Z_\eta$, simultaneously for every seed. Lemma J.7 gives, for a universal constant $C > 0$,

$$n\eta^2 \geq C \quad \Longrightarrow \quad \mathbb{P}\{Z_\eta > t\} \leq \exp(-nt^2/2) \qquad (t > 0). \tag{122}$$

It follows by taking $n \geq C'(\eta^{-2} + \alpha^{-2} \log(2/\beta))$ that $L_J \leq \alpha/4$ for all seeds with probability at least $1 - \beta/2$.

Ordinary signed uniform convergence over single tree intervals, as in (102), gives $a(I) \leq \widehat{a}(I) + \alpha/4$ simultaneously for all comparator intervals with probability at least $1 - \beta/2$, once $n \geq C''\alpha^{-2} \log(2/\beta)$. If a comparator interval contains observations, its empirical trace is the cell-tree interval between its first and last observed vertices; it therefore obeys (118). If it contains none, its empirical gain is zero and the same upper bound holds. Thus, on the intersection of these two events, for every comparator I and every seed,

$$a(I) - a(B_{S,J}) \leq \lambda + \frac{\alpha}{4} + L_J \leq \alpha.$$

This is the agnostic accuracy guarantee, including when the population optimum is not attained.

Replication. At every fixed original vertex v, the cell construction and (116) give

$$\mathbf{1}\{F_{S,v} \geq W_J + \eta\} \leq \mathbf{1}\{v \in B_{S,J}\} \leq \mathbf{1}\{F_{S,v} \geq W_J - \eta\}.$$

This also holds when E_v is empty. The score has sensitivity at most $2/n$, so Lemma J.1, with grid length $a = \lambda$, gives

$$\mathbb{P}\{H_{S,J}(v) \neq H_{S',J}(v)\} \leq \frac{1}{2K} + \frac{8}{K} \exp\left(-\frac{n\lambda^2}{32K^2}\right).$$

Taking

$$n \geq \frac{32K^2}{\lambda^2} \log\left(2 + \frac{8}{K\Delta}\right)$$

makes this at most $1/(2K) + \Delta = \rho$ and also ensures the logarithmic condition of the local-grid lemma. Increasing the universal constant if necessary ensures the accuracy conditions above. Since $\eta = \alpha/(8K)$, the resulting fixed sample count has the form (120). Replication is unconditional on the accuracy events. □

Remark K.3 (A known baseline and comparator subclasses). The same upper bound holds for any nonempty comparator class whose members have the form $h_I(v) = f_0(v)(-1)^{\mathbf{1}\{v\in I\}}$, where f_0 is a known fixed binary function and each permitted I is a tree interval. Transform the observed label to $Z = -Yf_0(X)$, apply the capped construction to its signed weights, and flip f_0 on the selected union. The risk reduction is still $\mathbb{E}[Z\mathbf{1}\{X \in A\}]$ for a flip set A. The proof compares the output with every tree interval simultaneously, so restricting the comparator family preserves its conclusion. The empty interval need not be permitted, and f_0 need not belong to the comparator class.

Corollary K.4 (The full tree-interval tradeoff)**.** *For every nonempty countable rooted tree order T, every $0 < \alpha, \beta \le 1/4$, every integer $B \ge 2$, and $K = 2^B$, uniformly for $0 < K\Delta \le 1/4$,*

$$m^*_{\mathcal{H}_{\mathrm{int}}(T)}(\alpha, \beta, 1/(2K) + \Delta; B) = \Theta\left(\frac{1}{\alpha^2}\left[K^2 \log\frac{1}{K\Delta} + \log\frac{1}{\beta}\right]\right). \tag{123}$$

The constants are absolute and independent of T. The minimum ranges over fixed-sample learners using at most B total shared fair bits and no other randomness; improper outputs are allowed.

Proof. Theorem K.2 gives the upper bound, since $\log(1/(K\Delta)) \ge \log 4$ and $\rho = 1/(2K) + \Delta \le 3/(4K) \le 1/4$. Fix any $v \in T$. The empty interval and the singleton $\{v\}$ give both comparator labels at v. Restricting the feature distribution to v turns any learner for $\mathcal{H}_{\mathrm{int}}(T)$ into a learner for the two hypotheses on one point, with the same sample count, seed budget, confidence, and pointwise replication. Proposition F.7 gives the matching joint lower bound, including for improper learners. □

Proposition K.5 (An interval class beyond bounded hypothesis cuts)**.** *The countable class of empty and finite intervals on $\mathbb{N}$ has VC dimension two and is covered by Theorem K.2. No public total order of this class has a uniformly bounded number of label changes at every domain point. The same holds for every pointwise-dense subclass.*

Proof. Two points are shattered, while three ordered points cannot be shattered because an interval containing the first and last must contain the middle one. On $\{1, \dots, m\}$, the empty interval and its $m(m+1)/2$ nonempty intervals are distinct hypotheses with distinct traces. Restrict any public order to these hypotheses. Between each two consecutive hypotheses at least one of the m labels changes. If T_x denotes the number of changes at x, then

$$\sum_{x=1}^{m} T_x \ge \frac{m(m+1)}{2}, \qquad \max_{1\le x\le m} T_x \ge \left\lceil \frac{m+1}{2} \right\rceil.$$

These bounds rule out a uniform finite cut count. Every nonempty finite interval $[a, b]$ is isolated by its positive labels at a, b and its negative labels at $b+1$ and, when $a > 1$, at $a-1$. Consequently, a pointwise-dense subclass must contain all these intervals. Repeating the count without the empty interval gives $\sum_{x=1}^{m} T_x \ge m(m+1)/2 - 1$, which still rules out every uniform finite cut bound. The theorem also applies to the finite-interval comparator subclass: its proof only uses interval upper bounds, and the learner is allowed to output an improper union. □

K.1 Products with unknown block masses

For a product of d tree-interval classes on disjoint public blocks, the replication cost grows linearly in d. The thresholds are scaled by the square roots of the empirical block masses. A population-centered concentration bound handles the resulting changes in normalization, while conditioning on the block counts allows the confidence deviations to be summed.

Lemma K.6 (Contraction in the prefix gains)**.** *Fix a cap $c > 0$. For a finite ordered path with weights $a_1, \dots, a_k$, let*

$$\Gamma_c(a_1, \dots, a_k) = \phi_{a_1} \circ \cdots \circ \phi_{a_k}(0), \qquad \phi_a(t) = \min\{c, \max\{0, a + t\}\}.$$

For another set of weights $b_1, \dots, b_k$ on the same path,

$$|\Gamma_c(a_1, \dots, a_k) - \Gamma_c(b_1, \dots, b_k)| \le \max_{1\le i\le k} \left|\sum_{r=1}^{i} (a_r - b_r)\right|. \tag{124}$$

For an absolutely summable signed weight vector a on a countable rooted tree, define $\Phi_c(a)_v$ as the supremum over endpoints $u \succeq v$ of the corresponding capped backward path values. These values are well-defined even when paths have no immediate neighbors, and

$$|\Phi_c(a)_v - \Phi_c(b)_v| \le \sup_{I \text{ tree interval}} |a(I) - b(I)| \tag{125}$$

for any two absolutely summable signed weight vectors.

Proof. Write $s_i = \sum_{r=1}^{i} a_r$ and $s_0 = 0$. Distributing each addition through the subsequent minima and maxima expands the backward composition into a fixed expression of minima and maxima whose leaves are among s_i and $c + s_i$, $0 \le i \le k$. The same expression, with the same order of operations, holds for the b weights. Minima and maxima are 1-Lipschitz in the maximum norm of their arguments. Every pair of corresponding leaves differs by at most the right side of (124), which proves that inequality.

For a fixed public path in a countable tree, restrict an absolutely summable weight vector to finite sets increasing to the tree, and compose the retained weights in path order. The elementary bound by the sum of the absolute weight changes, proved after (119), makes these values Cauchy and their limit independent of the chosen finite exhaustion. This defines the path value. For two finite restrictions, insert zero weights to align their supports. Each prefix sum in (124) is then the gain of an original-tree interval. Passing to the limit is legitimate because the omitted absolute masses tend to zero. This proves the corresponding bound for every public path; taking the supremum over endpoints proves (125). □

These clipped compositions are related to two-sided reflection on an interval and its continuity properties (Kruk et al., 2007). The prefix estimate above uses the backward order of the path composition.

Lemma K.7 (Concentration of a normalized capped score)**.** *Suppose each of n independent observations enters a fixed block with probability p. Conditional on entering, let (V, Z) have any distribution on a countable rooted tree and signs $Z \in \{-1, +1\}$, with an optional cemetery value belonging to no interval. Let N be the block count, $q = N/n$, and*

$$b_v = \mathbb{E}[Z\mathbf{1}\{V = v\} \mid \textit{block}], \qquad \widehat{b}_v = \frac{1}{N} \sum_{i \text{ in block}} Z_i \mathbf{1}\{V_i = v\} \quad (N > 0).$$

When $p = 0$, set $b = 0$. For any cap $c > 0$ and fixed vertex v, put

$$\widehat{F}_v = \begin{cases} \Phi_c(\sqrt{q}\,\widehat{b})_v, & N > 0, \\ 0, & N = 0, \end{cases} \qquad f_v = \Phi_c(\sqrt{p}\,b)_v.$$

There is a universal constant A such that, for every $s > 0$,

$$\mathbb{P}\{|\widehat{F}_v - f_v| \ge A/\sqrt{n} + s\} \le 3e^{-ns^2/8}. \tag{126}$$

The constants do not depend on the cap, the block mass, or the tree.

Proof. For $N > 0$, write $D_N = \sup_I |\widehat{b}(I) - b(I)|$ over tree intervals. The empirical-process bound proving (102) gives $\mathbb{E}[D_N \mid N] \le A/\sqrt{N}$. Conditioning on N, bounded differences also gives

$$\mathbb{P}\{\sqrt{q}\, D_N \ge A/\sqrt{n} + t \mid N\} \le e^{-nt^2/2} \qquad (t > 0).$$

Give $\sqrt{q}\, D_N$ the value zero when $N = 0$; the displayed tail then holds unconditionally as well. By Lemma K.6 and $|b(I)| \le 1$,

$$|\widehat{F}_v - f_v| \le \sqrt{q}\, D_N + |\sqrt{q} - \sqrt{p}|.$$

This also holds when $N = 0$. The square-root binomial bound (9) gives

$$\mathbb{P}\{|\sqrt{q} - \sqrt{p}| \ge t\} \le 2e^{-nt^2}.$$

Using the two bounds at $t = s/2$ proves (126), including zero counts and zero population block mass. □

Theorem K.8 (Tree-interval products with one finite seed)**.** *Let $d \geq 1$, let $T_1, \ldots, T_d$ be nonempty countable rooted tree orders, and consider the class*

$$\mathcal{H}_\times = \{H : H(j, v) = 2\mathbf{1}\{v \in I_j\} - 1,\ I_j \text{ a tree interval of } T_j\}$$

on the disjoint union of these public blocks. Empty intervals are permitted. Let $0 < \alpha, \beta, \rho \leq 1/4$, let $B \geq 0$ be an integer, put $K = 2^B$, and suppose $\Delta = \rho - 1/(2K) > 0$. There is a universal constant C such that an improper agnostic (α, β)-learner for $\mathcal{H}_\times$ is ρ-pointwise replicable on every distribution and uses exactly

$$n = \max\left\{\left\lceil \frac{2048dK^2}{\alpha^2} \log\left(2 + \frac{24}{K\Delta}\right)\right\rceil, \left\lceil \frac{CdK^2}{\alpha^2}\right\rceil, \left\lceil \frac{C}{\alpha^2} \log \frac{2}{\beta}\right\rceil\right\} \tag{127}$$

examples and exactly B total shared fair bits, with no other randomness. The block masses are unknown to the learner. In particular, the sample bound is

$$O\left(\frac{1}{\alpha^2}\left[dK^2 \log\left(2 + \frac{1}{K\Delta}\right) + \log \frac{2}{\beta}\right]\right). \tag{128}$$

Proof. **The learner.** Let N_j be the number of observations in block j and put $q_j = N_j/n$. Set

$$c = \frac{\alpha}{2\sqrt{d}}, \qquad \lambda_j = c\sqrt{q_j}, \qquad \eta_j = \frac{\lambda_j}{4K}.$$

Draw one index J uniformly from $\{0, \ldots, K-1\}$ using the B shared bits. On a block with $N_j > 0$, run the sample-cell capped-prefix construction with raw weights

$$\widehat{a}_{j,v} = \frac{1}{n} \sum_{i : X_i = (j,v)} Y_i,$$

cap λ_j, and thresholds $W_{j,J} = (J + 1/2)\lambda_j/K$ and η_j. Output positively on its selected union and negatively elsewhere. On a block with $N_j = 0$, output negatively everywhere. All trees, enumerations, and tie rules are public. The same index J is used on every block, and the learner makes no further random choice.

Accuracy conditional on the counts. For the analysis, let p_j be the population block mass and let P_j be the conditional labeled distribution in that block, chosen arbitrarily if $p_j = 0$. Write

$$b_j(I) = \mathbb{E}_{P_j}[Y\mathbf{1}\{X \in I\}], \qquad \widehat{b}_j(I) = \widehat{a}_j(I)/q_j \quad (N_j > 0).$$

Conditional on the count vector, the grouped observations are independent, with N_j observations from each P_j. For each nonempty empirical block put

$$C_j = \sup_{I \text{ tree interval}} |\widehat{b}_j(I) - b_j(I)|,$$

$$D_{j,m} = \sup_{U \text{ a union of at most } m \text{ intervals}} \big(\widehat{b}_j(U) - b_j(U)\big),$$

$$\eta'_j = \frac{\eta_j}{q_j} = \frac{\alpha}{8K\sqrt{dq_j}}, \qquad Z_j = \max\{0, \sup_{m \geq 1}(D_{j,m} - 2\eta'_j m)\}.$$

Only terms with $N_j > 0$ are used in these definitions. Since

$$N_j(\eta'_j)^2 = \frac{n\alpha^2}{64dK^2},$$

a sufficiently large universal C in (127) ensures the condition of Lemma J.7 on every such block. Consequently, conditional on the counts,

$$\mathbb{P}\{Z_j > t \mid (N_1, \ldots, N_d)\} \leq e^{-N_j t^2/2}, \qquad \mathbb{E}[Z_j \mid (N_1, \ldots, N_d)] \leq \sqrt{\frac{\pi}{2N_j}}.$$

Also $\mathbb{E}[C_j \mid (N_1, \ldots, N_d)] \leq A/\sqrt{N_j}$ for a universal A. Thus, with

$$V = \sum_{j : N_j > 0} q_j(C_j + Z_j),$$

there is a universal A_1 such that

$$\mathbb{E}[V \mid (N_1, \dots, N_d)] \le A_1 \sum_j \frac{\sqrt{N_j}}{n} \le A_1 \sqrt{\frac{d}{n}}.$$

When the counts are fixed, replacing one grouped observation changes C_j and Z_j by at most $2/N_j$ each. The conditional penalties η'_j are fixed. Hence V has bounded differences $4/n$ per observation, and, after averaging over the counts,

$$\mathbb{P}\{V > A_1\sqrt{d/n} + t\} \le e^{-nt^2/8}. \tag{129}$$

Let Q_N be the distribution with block masses q_j and conditional laws P_j. For a selected maximal prefix P in block j, let E_P be its raw cell surplus from (117), and let E_j be the total raw surplus in that block. The same equation gives

$$\widehat{a}_j(P) > E_P + 2\eta_j, \qquad \sum_P E_P = E_j.$$

Take the disjoint union of those selected prefixes for which $E_P > q_j b_j(P)$. Dividing their empirical margins by q_j and applying the definition of Z_j shows

$$\sum_P (E_P - q_j b_j(P))_+ \le q_j Z_j.$$

Therefore the selected union in block j has gain under Q_N at least $E_j - q_j Z_j$. For every comparator interval I_j, its gain under Q_N is at most

$$q_j b_j(I_j) \le \widehat{a}_j(I_j) + q_j C_j \le \lambda_j + E_j + q_j C_j.$$

The last inequality is the capped comparator bound, applied to its empirical cell-tree trace as before. Blocks with zero empirical mass contribute zero to both gains. Summing the risk differences gives, for every product comparator h and every seed,

$$\mathrm{R}_{Q_N}(H_{S,J}) - \mathrm{R}_{Q_N}(h) \le \sum_j \lambda_j + V \le \frac{\alpha}{2} + V. \tag{130}$$

To return to the true distribution, put $M = \sum_j |q_j - p_j|$. The multinomial variances and Cauchy–Schwarz give $\mathbb{E}M \le \sqrt{d/n}$. Replacing one observation changes M by at most $2/n$, so

$$\mathbb{P}\{M > \sqrt{d/n} + t\} \le e^{-nt^2/2}.$$

For any two classifiers, each conditional risk difference lies in $[-1, 1]$. Changing the block masses from q_j to p_j therefore changes their total risk difference by at most M, simultaneously for all classifiers. Increasing C in (127) ensures $V \le \alpha/4$ and $M \le \alpha/4$ with probability at least $1 - \beta$, by (129) and the last display. Equation (130) then proves agnostic α-accuracy for all seeds simultaneously. In particular, zero-count blocks are covered by the same mass discrepancy bound. No confidence union bound over the d blocks is used.

Replication of a fixed query. Fix a vertex v in block j. On a sample with $N_j > 0$, divide the raw capped score by $\sqrt{q_j}$. Homogeneity of clipping makes the result

$$\widehat{F}_{j,v} = \Phi_c(\sqrt{q_j}\,\widehat{b}_j)_v.$$

Give it value zero when $N_j = 0$. Its deterministic population center is $f_{j,v} = \Phi_c(\sqrt{p_j}\,b_j)_v$. Lemma K.7 gives

$$\mathbb{P}\{|\widehat{F}_{j,v} - f_{j,v}| \ge A/\sqrt{n} + s\} \le 3e^{-ns^2/8}.$$

Put $h = c/K$, $u_\ell = (\ell + 1/2)h$, and $\eta_0 = h/4$. The construction satisfies, for every sample and seed,

$$\mathbf{1}\{\widehat{F}_{j,v} \ge u_J + \eta_0\} \le \mathbf{1}\{H_{S,J}(j, v) = +1\} \le \mathbf{1}\{\widehat{F}_{j,v} \ge u_J - \eta_0\}.$$

For $N_j = 0$ this follows from $\widehat{F}_{j,v} = 0$ and $u_J - \eta_0 > 0$.

Choose a grid point nearest to $f_{j,v}$. At this seed, the two independent samples give disagreement at most $1/2$. A grid point at index distance $r \geq 1$ is at distance at least $h(r - 1/2)$ from the center. A sufficiently large C ensures $A/\sqrt{n} \leq h/8$. At each such other seed, the sandwich and the score tail bound its disagreement by

$$6 \exp\left(-\frac{nh^2}{8}(r - 7/8)^2\right).$$

There are at most two grid points at each index distance. Since $(r - 7/8)^2 \geq r/64$ for integers $r \geq 1$, the same geometric summation as in the local-grid lemma gives, whenever $nh^2/512 \geq \log 2$,

$$\begin{aligned}\mathbb{P}\{H_{S,J}(j,v) \neq H_{S',J}(j,v)\} &\leq \frac{1}{2K} + \frac{12}{K}\sum_{r\geq 1} e^{-nh^2(r-7/8)^2/8} \\ &\leq \frac{1}{2K} + \frac{24}{K}e^{-nh^2/512} \\ &= \frac{1}{2K} + \frac{24}{K}e^{-n\alpha^2/(2048dK^2)} \leq \frac{1}{2K} + \Delta.\end{aligned}$$

The first term in (127) ensures the logarithmic condition and the last inequality. This argument includes random changes in the block count and sample cells, and is unconditional on every accuracy event. Countability ensures measurability throughout, including at vertices outside the observed support. □

Corollary K.9 (The full joint tradeoff for tree products)**.** *For the product class in Theorem K.8, uniformly for* $0 < \alpha \leq 1/16$, $0 < \beta \leq 1/4$, $B \geq 4$, $K = 2^B$, *and* $0 < K\Delta \leq 1/4$,

$$m^*_{\mathcal{H}_\times}(\alpha, \beta, 1/(2K) + \Delta; B) = \Theta\left(\frac{1}{\alpha^2}\left[dK^2 \log\frac{1}{K\Delta} + \log\frac{1}{\beta}\right]\right). \tag{131}$$

The constants are absolute and independent of the trees. The minimum allows improper outputs, uses a fixed sample size, and permits at most B total shared fair bits with no other randomness.

Proof. Theorem K.8 gives the upper bound. Choose one vertex in each nonempty tree. Independently choosing the empty interval or that singleton in each block shatters these d points. Proposition F.8 supplies the term $dK^2\alpha^{-2}\log(1/(K\Delta))$ in the displayed parameter range. Concentrating the feature distribution on any one chosen point and applying Proposition F.7 supplies the confidence term $\alpha^{-2}\log(1/\beta)$. The maximum of the two lower bounds is at least half their sum, and the different constant numerators inside the logarithms change only universal factors here. □

Corollary K.10 (Products of VC-one classes)**.** *Let $\mathcal{H}_1, \ldots, \mathcal{H}_d$ be binary classes of VC dimension exactly one on disjoint countable public domains, and let $\mathcal{H}_\times$ be their full product. Then (131) holds for this class in the same parameter range. Its upper bound also holds when each factor is nonempty and has VC dimension at most one, under the upper-bound parameters of Theorem K.8.*

Proof. Fix a baseline from each nonempty factor. Apply the public profile quotient from the proof of Theorem J.3 within each block. The flipped set of every comparator becomes an initial chain in a countable rooted tree, and hence a tree interval. Coordinates on which all comparators agree become a cemetery value with the fixed baseline prediction. Transform labels to the signed baseline gains.

The product construction applies to these public trees, using the original block counts, including observations sent to the cemetery. Empty observed cell trees receive the baseline prediction. All interval deviations and normalized score bounds permit the cemetery value, and the capped comparator bound applies to the initial-chain subclass. Thus the same proof gives the upper bound with unknown block masses and exactly the same total seed. This argument covers arbitrary misspecification and any number of original coordinates with the same profile.

If every factor has VC dimension exactly one, choose one point admitting both labels in each factor. The full product shatters these d points. The dimension lower bound and the one-point confidence lower bound therefore apply exactly as in the preceding corollary. □

K.2 Real intervals and mixed public blocks

The interval construction also works on the real line with its Borel sigma-field. Here an interval may have open, closed, or unbounded endpoints; the empty set and singletons are included. Two points require care beyond the countable-domain argument: the deviation suprema must be measurable, and a population capped score must account for nonatomic signed mass.

Lemma K.11 (Measurable deviations for real intervals)**.** *Let Q be any Borel probability law of $(X, Z) \in \mathbb{R} \times \{-1, +1\}$, and write $b(A) = \mathbb{E}_Q[Z\mathbf{1}\{X \in A\}]$ and $\widehat{b}_N(A) = N^{-1}\sum_{i=1}^N Z_i\mathbf{1}\{X_i \in A\}$. For each $m \geq 1$, the signed deviation suprema over unions of at most m real intervals are Borel functions of the sample. The bounds (102), (109), and (108) hold for these classes with universal constants. In particular, for some universal A,*

$$\mathbb{E} \sup_{I \text{ real interval}} |\widehat{b}_N(I) - b(I)| \leq \frac{A}{\sqrt{N}}. \tag{132}$$

These assertions include distributions with arbitrary atoms.

Proof. Let $\mathcal{I}_0$ contain the empty set and every bounded open interval with rational endpoints. Every real interval is a pointwise limit of a sequence from $\mathcal{I}_0$. At a finite included endpoint, take rational endpoints approaching from outside the target interval; at an excluded endpoint, approach from inside. For a singleton, approach from outside on both sides. At an infinite endpoint, let the corresponding rational endpoint diverge. These choices give the correct limiting membership at the endpoints themselves, as well as at every other real point.

Approximating each of finitely many constituent intervals in this way proves the same assertion for unions of at most m intervals, using the countable family $\mathcal{U}_{m,0}$ of unions of at most m members of $\mathcal{I}_0$. For any fixed sample and any such approximating sequence, both its empirical signed gain and its true signed gain converge. The first fact follows from the finite sample; the second follows from dominated convergence, since $|Z| \leq 1$. Consequently each signed or absolute deviation supremum equals its supremum over $\mathcal{U}_{m,0}$. This is a countable supremum of Borel functions. The penalized supremum over the positive integers m is measurable as well. No distribution-dependent list of atoms is needed.

Unions of m real intervals have VC dimension at most $2m$, by the alternating-label argument on $2m + 1$ ordered points. The finite-trace packing, chaining, and symmetrization arguments proving the displayed bounds therefore apply to these countable determining families. Their bounded-differences constants remain $2/N$. The expectation and concentration estimates, including the union over m with its per-interval penalty, are exactly the ones in Lemma J.7. □

Lemma K.12 (Measurable population centers for capped real paths)**.** *For a finitely supported signed measure μ on $\mathbb{R}$, define $\Phi_c(\mu)(x)$ by the capped backward path rule with cap $c > 0$, maximizing over paths $[x, u]$ and using the atom weights of μ. Let $Q, b, \widehat{b}_N$ be as in Lemma K.11. There is a Borel function $G_{Q,c} : [0, 1] \times \mathbb{R} \to [0, c]$ such that, with $\|\nu\|_{\mathcal{I}} = \sup_{I \text{ real interval}} |\nu(I)|$,*

$$|\Phi_c(t\mu)(x) - G_{Q,c}(t, x)| \leq t\|\mu - b\|_{\mathcal{I}}, \tag{133}$$

$$|G_{Q,c}(t, x) - G_{Q,c}(t', x)| \leq |t - t'|, \qquad G_{Q,c}(0, x) = 0, \tag{134}$$

for every finitely supported μ of total variation at most one and $t, t' \in [0, 1]$. In particular, Lemma K.7 holds on $\mathbb{R}$, with population center $G_{Q,c}(\sqrt{p}, x)$ and the same form of the bound.

Proof. Zero-weight locations may be inserted or removed in a capped composition. Aligning two finite supports and applying Lemma K.6 therefore gives

$$\sup_x |\Phi_c(t\mu)(x) - \Phi_c(t\nu)(x)| \leq t\|\mu - \nu\|_{\mathcal{I}}. \tag{135}$$

Only observed endpoints need be considered when evaluating an empirical score. Sort the distinct observations, aggregate their signed weights at ties, and use finite maxima and capped compositions. This makes $(S, t, x) \mapsto \Phi_c(t\widehat{b}_N)(x)$ jointly Borel: there are finitely many possible weak orderings of a sample, and on each ordering the formula uses Borel comparisons, finite sums, minima, and maxima. Grouping observations at the same location is essential; the atom's total signed weight is used in one clipping operation.

Take independent empirical measures of sizes N and M. Equation (135) and (132) imply

$$\sup_{t\in[0,1],x}\left|\mathbb{E}\Phi_c(t\widehat{b}_N)(x)-\mathbb{E}\Phi_c(t\widehat{b}_M)(x)\right|\le A\left(\frac{1}{\sqrt{N}}+\frac{1}{\sqrt{M}}\right).$$

Define the uniform limit

$$G_{Q,c}(t,x)=\lim_{M\to\infty}\mathbb{E}\Phi_c(t\widehat{b}_M)(x).$$

Each expectation is Borel in (t,x): integration of bounded jointly Borel functions preserves measurability, as follows first for measurable rectangles and then by the monotone-class argument. Thus the limit is Borel. For a fixed finite measure μ, the same contraction and the triangle inequality give

$$|\Phi_c(t\mu)(x)-\mathbb{E}\Phi_c(t\widehat{b}_M)(x)|\le t\left(\|\mu-b\|_{\mathcal{I}}+\frac{A}{\sqrt{M}}\right).$$

Letting M increase proves (133). Since every $\widehat{b}_M$ has total variation at most one, the finite scores are 1-Lipschitz in their scaling parameter t, by (135). Passing to expectations and limits proves (134).

For the normalized-score conclusion, put $q=N/n$ and $D_N=\|\widehat{b}_N-b\|_{\mathcal{I}}$ when $N>0$. The two inequalities just proved give

$$|\Phi_c(\sqrt{q}\,\widehat{b}_N)(x)-G_{Q,c}(\sqrt{p},x)|\le\sqrt{q}\,D_N+|\sqrt{q}-\sqrt{p}|.$$

When $N=0$, give the empirical score and $\sqrt{q}\,D_N$ the value zero. The inequality still holds by (134). Conditional on $N>0$, (132) and bounded differences give the same tail for $\sqrt{q}\,D_N$ as in Lemma K.7. Combining it with (9) proves

$$\mathbb{P}\{|\widehat{F}_x-G_{Q,c}(\sqrt{p},x)|\ge A/\sqrt{n}+s\}\le 3e^{-ns^2/8}\qquad(s>0).$$

The center is defined using the entire signed measure $b(A)=\mathbb{E}_Q[Z\mathbf{1}\{X\in A\}]$. Using the entire measure accounts for both atomic and nonatomic mass. □

Theorem K.13 (Intervals on real lines and mixed public blocks)**.** *Let $\mathcal{X}$ be the Borel disjoint union of $d\ge 1$ public blocks, and let $\mathcal{H}=\prod_{j=1}^{d}\mathcal{H}_j$. Each factor is nonempty and is of one of the following types:*

1. *A class on a countable rooted tree, contained in the class of flips of a fixed known baseline on tree intervals.*
2. *A class on a Borel copy of $\mathbb{R}$ with a fixed public Borel coordinate identification, contained in the class of flips of a fixed known Borel baseline on real intervals of any endpoint type.*
3. *A binary class of VC dimension at most one on a countable domain.*

For $0<\alpha,\beta,\rho\le 1/4$, an integer $B\ge 0$, $K=2^B$, and $\Delta=\rho-1/(2K)>0$, an improper agnostic (α,β)-learner for $\mathcal{H}$ is ρ-pointwise replicable on every Borel distribution, uses exactly B total shared fair bits with no other randomness, and uses

$$O\left(\frac{1}{\alpha^2}\left[dK^2\log\left(2+\frac{1}{K\Delta}\right)+\log\frac{2}{\beta}\right]\right)\tag{136}$$

examples. Its evaluation map is jointly Borel in the sample, seed, and query. The block masses are unknown and unrestricted. Baseline membership in the comparator class is not required in the first two cases.

Proof. Use the label transform $Z=-Yf_0(X)$ on each baseline block. For a factor of the third type, use the public tree quotient, baseline, and cemetery construction from Corollary K.10. Every comparator then flips a tree interval. These maps are measurable on their countable domains.

Use the block counts, cap, thresholds, and common seed from Theorem K.8:

$$q_j=N_j/n,\qquad c=\frac{\alpha}{2\sqrt{d}},\qquad \lambda_j=c\sqrt{q_j},\qquad \eta_j=\frac{\lambda_j}{4K},\qquad W_{j,J}=(J+1/2)\frac{\lambda_j}{K}.$$

Take the fixed sample size in (127), with its universal constant increased if needed. Countable tree blocks use the existing sample-cell algorithm. An unobserved block always receives its baseline prediction.

Here is the explicit rule on an observed real block. Sort its distinct observations as $z_1 < \cdots < z_r$, and form the cells

$$C_1 = (-\infty, z_1], \qquad C_i = (z_{i-1}, z_i] \quad (2 \le i \le r).$$

The right tail (z_r, ∞) is left at the baseline. Give cell C_i the raw signed weight $\widehat{a}_i = n^{-1} \sum_{\ell : X_\ell = (j, z_i)} Z_\ell$. Set $F_{r+1} = 0$ and compute backwards

$$F_i = \min\{\lambda_j, \max\{0, \widehat{a}_i + F_{i+1}\}\}.$$

Starting with $b_0 = 0$, scan the cells in increasing order and set

$$b_i = \mathbf{1}\{F_i \ge W_{j,J} + \eta_j \text{ or } (b_{i-1} = 1 \text{ and } F_i \ge W_{j,J} - \eta_j)\}.$$

Flip the baseline on exactly the cells with $b_i = 1$. This is the strong/weak selected-prefix rule on a chain: weak cells have positive scores, so every continuation uses the chosen successor. Its maximal selected runs are disjoint Borel intervals.

All operations in this finite rule are Borel, including sorting, grouping ties, and testing membership in $(z_{i-1}, z_i]$. Thus its evaluation map is jointly Borel, also at sample locations and in unobserved gaps or tails. Multiplication by a known Borel baseline preserves this property. The same holds on the countable blocks, and there are only finitely many public blocks. In particular, the output's population risk is a measurable function of the sample and seed, by integrating its jointly Borel loss.

For the analysis, write p_j for the true block mass and P_j for the conditional law of its coordinate and transformed sign; if $p_j = 0$, choose P_j arbitrarily, since then $N_j = 0$ almost surely and the population center is zero. When $N_j > 0$, let $\widehat{b}_j$ be its empirical conditional signed measure, as in the tree-product proof. For a fixed real query x, the raw score is the maximum of the capped empirical compositions on the public paths $[x, u]$. It is constant on the displayed cells and zero to the right of the sample. It can be evaluated using only observed endpoints $u \ge x$, even when x is unobserved. Thus its normalized version is exactly $\Phi_c(\sqrt{q_j}\,\widehat{b}_j)(x)$, with value zero when $N_j = 0$. The deterministic center for the replication proof is $G_{P_j,c}(\sqrt{p_j}, x)$ from Lemma K.12. The normalized tail and the pointwise sandwich therefore have exactly the forms used in the tree-product proof. The grid summation there applies unchanged, including at atoms of either the data law or the score law. No sample-cell agreement between the two runs is required.

For accuracy, define true gains as signed measures on intervals, rather than as sums over individual real points. Every selected run is an original real interval. Every comparator interval has either an empty empirical trace or a consecutive segment $z_a, \ldots, z_b$ of the observed locations, regardless of its endpoint conventions. Its empirical gain is therefore the gain of a cell-chain interval, to which the capped comparator inequality applies. Lemma K.11 supplies all single-interval and penalized-union deviation bounds used in the conditional-count accuracy argument. Their suprema are measurable even for data-dependent selected runs. The conditional aggregate-deviation and multinomial mass-discrepancy bounds in that argument are consequently unchanged. They give accuracy simultaneously for all seeds and all product comparators, including zero-count blocks.

This proves (136) and the exact total bit count. The proof covers every interval on the first two types of block, so restricting either comparator family to a nonempty subclass preserves the upper bound. No restriction on approximation error, atomic mass, or nonatomic mass was used. □

Corollary K.14 (The joint tradeoff for nontrivial mixed products)**.** *Assume that every factor in Theorem K.13 contains at least two distinct hypotheses. Uniformly for* $0 < \alpha \le 1/16$, $0 < \beta \le 1/4$, $B \ge 4$, $K = 2^B$, *and* $0 < K\Delta \le 1/4$,

$$m^*_{\mathcal{H}}(\alpha, \beta, 1/(2K) + \Delta; B) = \Theta\left(\frac{1}{\alpha^2}\left[dK^2 \log \frac{1}{K\Delta} + \log \frac{1}{\beta}\right]\right). \tag{137}$$

The constants are absolute and independent of the factors and their block masses. This includes full real-interval and tree-interval factors, as well as countable factors of VC dimension exactly one.

Proof. The upper bound follows from Theorem K.13. Two distinct binary hypotheses disagree at some point, so choose one such point in each block. The full product shatters these d points. The proof of Proposition F.8 uses distributions on a finite shattered set and therefore applies to this Borel domain without change. Concentrating the distribution on one of the chosen points also gives the confidence lower bound from Proposition F.7. Taking the maximum of the two bounds proves the assertion. □

Corollary K.15 (A single real interval factor)**.** *For one nontrivial real-interval factor as in Theorem K.13, the sharper single-factor parameter range $0 < \alpha, \beta \le 1/4$, $B \ge 2$, $K = 2^B$, and $0 < K\Delta \le 1/4$ also satisfies*

$$m^*_{\mathcal{H}}(\alpha, \beta, 1/(2K) + \Delta; B) = \Theta\left(\frac{1}{\alpha^2}\left[K^2 \log \frac{1}{K\Delta} + \log \frac{1}{\beta}\right]\right).$$

Proof. Use the upper bound with $d = 1$ and the singleton testing lower bound at any point on which two comparators disagree. □

L Unknown partitions: unions of intervals

The comparator intervals are unknown and need not follow a public partition. A square-root mass cost and a fixed charge per component give linear dependence on the number of intervals. The learner uses an interval supremum independent of the chosen optimal certificate. Stability is proved for this score, rather than for the certificate itself. Regularization and certificates connect this construction to submodular optimization (Bach, 2013, Sections 6.2 and 10.1). We prove the finite certificate and derive a local score and rounding rule for overlapping intervals.

Lemma L.1 (Interval certificates and a local score)**.** *Let $1, \dots, m$ be a finite line, with weights $p_i \ge 0$ and $|a_i| \le p_i$, and fix $\tau, \lambda > 0$. Write $p(I) = \sum_{i \in I} p_i$ and $a(I) = \sum_{i \in I} a_i$. For a subset A, put*

$$\begin{aligned} g_p(A) &= \sum_{C \text{ a maximal interval of } A} \sqrt{p(C)}, \\ J_{\tau,\lambda}(A) &= a(A) - \tau g_p(A) - \lambda c(A), \\ V_{\tau,\lambda}(I) &= \max_{A \subseteq I} J_{\tau,\lambda}(A), \end{aligned}$$

where $c(A)$ is the number of nonempty maximal interval components. In particular, a nonempty zero-mass component still incurs the charge λ. The empty set is permitted in each maximum. There exist weights $e_i \ge 0$ and $q_i = a_i - e_i$ such that

$$\begin{aligned} q(I) &\le \tau\sqrt{p(I)} + \lambda \quad \text{for every nonempty interval } I, \\ E &:= \sum_i e_i = V_{\tau,\lambda}([1, m]). \end{aligned} \tag{138}$$

For any such weights, define

$$F_x = \max\left\{0, \max_{I \ni x:\, p(I) > 0} \frac{q(I) - \lambda}{\sqrt{p(I)}}\right\}. \tag{139}$$

An empty inner maximum contributes zero. The score is independent of these weights and satisfies the local formula

$$F_x = \max\left\{0, \max_{I \ni x:\, p(I) > 0} \frac{a(I) - V_{\tau,\lambda}(I) - \lambda}{\sqrt{p(I)}}\right\} \in [0, \tau]. \tag{140}$$

If $\tau/\sqrt{2} < w < \tau$, let $B_w = \{x : F_x \ge w\}$. Every maximal interval C of B_w satisfies

$$\begin{aligned} a(C) &\ge e(C) + \lambda + (w - \tau/\sqrt{2})\sqrt{p(C)}, \\ p(C) &\ge w^2, \qquad \sum_{C \subseteq B_w} e(C) = E. \end{aligned} \tag{141}$$

If $\sum_i p_i \le 1$, every union A of at most k intervals satisfies

$$a(A) \le E + \tau\sqrt{k} + \lambda k. \tag{142}$$

All assertions allow zero-mass sites.

Proof. Set $V_t = V_{\tau,\lambda}([1,t])$, with $V_0 = 0$, and take $e_i = V_i - V_{i-1}$. These increments are nonnegative. For $I = [s,t]$, adjoining I to a maximizing subset of $[1, s-1]$ gives

$$V_t \ge V_{s-1} + a(I) - \tau\sqrt{p(I)} - \lambda.$$

If two components become adjacent, merging them only decreases both the square-root cost and the component charge. Thus $q_i = a_i - e_i$ satisfies (138). More generally, for every A and any weights satisfying that display,

$$J_{\tau,\lambda}(A) = e(A) + \sum_{C \text{ a component of } A} \big(q(C) - \tau\sqrt{p(C)} - \lambda\big) \le e(A) \le E.$$

Consequently, for any maximizer A_* of $J_{\tau,\lambda}$, equality holds throughout: $e_i = 0$ off A_*, and each component C of A_* is tight, meaning $q(C) = \tau\sqrt{p(C)} + \lambda$. A zero-mass component cannot be tight, since $q \le a$ and $a = 0$ on such a component.

Put $L(I) = a(I) - V_{\tau,\lambda}(I)$. Applying the preceding inequality to subsets of I gives $V_{\tau,\lambda}(I) \le e(I)$, and hence $q(I) \le L(I)$. Also $L(I) - \lambda \le \tau\sqrt{p(I)}$, by taking $A = I$ in its defining maximum. To prove the reverse inequality between the two scores, fix an interval I with $L(I) > \lambda$. Let $C_1, \dots, C_r$ be the components of A_* that meet I, put $K_j = C_j \cap I$, and enlarge I to the interval $U = I \cup \bigcup_j C_j$. Since $U \setminus A_* = I \setminus A_*$, tightness gives

$$q(U) = a(I \setminus A_*) + \tau \sum_j \sqrt{p(C_j)} + \lambda r.$$

The candidate $A_* \cap I$ in the definition of $V_{\tau,\lambda}(I)$ gives $L(I) \le a(I \setminus A_*) + \tau \sum_j \sqrt{p(K_j)} + \lambda r$. Here each K_j is nonempty, even when its mass is zero; this is why its component charge is retained. Concavity of the square root implies

$$\sum_j \big(\sqrt{p(C_j)} - \sqrt{p(K_j)}\big) \ge \sqrt{p(U)} - \sqrt{p(I)}.$$

Indeed, add the disjoint sets $C_j \setminus I$ successively; the square-root increment from an individual K_j is at least the increment from the larger current total mass. It follows that

$$q(U) \ge L(I) + \tau\big(\sqrt{p(U)} - \sqrt{p(I)}\big).$$

Since $0 < (L(I) - \lambda)/\sqrt{p(I)} \le \tau$, this yields $(q(U) - \lambda)/\sqrt{p(U)} \ge (L(I) - \lambda)/\sqrt{p(I)}$ and proves (140).

The set B_w is the union of intervals I satisfying $q(I) \ge \lambda + w\sqrt{p(I)}$ with $p(I) > 0$. For a component C of this union, choose an irredundant cover $I_1, \dots, I_r$ by such intervals, ordered from left to right. Their left and right endpoints both strictly increase. No point belongs to three members: otherwise a middle interval is contained in the union of its two neighbors and is redundant. Write $O_j = I_j \cap I_{j+1}$, allowing an empty intersection. These overlaps are pairwise disjoint. If $s \le r - 1$ of them are nonempty, then

$$q(C) = \sum_j q(I_j) - \sum_j q(O_j).$$

Each I_j contains its two disjoint neighboring overlaps, so $\sqrt{p(I_j)} \ge (\sqrt{p(O_{j-1})} + \sqrt{p(O_j)})/\sqrt{2}$, where the missing overlaps have mass zero. Therefore

$$\begin{aligned} q(C) &\ge \lambda(r - s) + w \sum_j \sqrt{p(I_j)} - \tau \sum_j \sqrt{p(O_j)} \\ &\ge \lambda + (w - \tau/\sqrt{2}) \sum_j \sqrt{p(I_j)} \ge \lambda + (w - \tau/\sqrt{2})\sqrt{p(C)}. \end{aligned}$$

Every positive-surplus site lies in a positive-mass tight component of A_*, and hence has score τ. This proves the surplus equality in (141). Finally, a qualifying $I \subseteq C$ has $\lambda + w\sqrt{p(I)} \le q(I) \le a(I) \le p(I)$, proving its mass bound. For a union of at most k disjoint intervals, (138) and Cauchy–Schwarz give (142). □

Scores on the original line. Let $\mathcal{X}$ be either a countable linearly ordered set or $\mathbb{R}$ with its Borel sets. An interval means an order-convex subset; either endpoint may be included or excluded, and unbounded intervals are allowed. For a sample $S = ((X_i, Y_i))_{i=1}^n$, write

$$\widehat{p}_S(I) = \frac{1}{n}\sum_i \mathbf{1}\{X_i \in I\}, \qquad \widehat{a}_S(I) = \frac{1}{n}\sum_i Y_i \mathbf{1}\{X_i \in I\}.$$

Use these empirical weights to define $V_{S,\tau,\lambda}(I)$ and $F_{S,\tau,\lambda}(x)$ by (140), with a supremum over original intervals. The maximum in $V_{S,\tau,\lambda}(I)$ ranges over finite unions of intervals contained in I. A zero empirical mass contributes score zero. Although the definition uses all intervals, it is a finite computation: if $z_1 < \cdots < z_m$ are the distinct observations, partition the line into their singletons, the nonempty open gaps between them, and the two nonempty exterior rays. These at most $2m+1$ cells carry weights $|\widehat{a}| \le \widehat{p}$; all nonsingleton cells have zero empirical mass. Consecutive cell unions are original intervals, and every original interval has the same empirical trace as such a union. For intervals containing a fixed query, the trace can be represented by a cell union also containing that query. A maximizing subset may be filled across any unsampled separation between its components: this keeps its empirical signed gain and decreases both penalties. Thus its trace has a representation by whole cells with no larger cost. Conversely, any such cell set can be restricted back to a local interval that cuts a gap cell: this only truncates or removes zero-mass portions, preserves the empirical trace, and cannot increase the cost. On the whole line, every cell set is a finite union of original intervals. The local values and scores therefore agree exactly with the finite definitions on these cells.

In particular, the score is constant on each cell. In an interior gap (z_j, z_{j+1}) it equals $\max\{F_{S,\tau,\lambda}(z_j), F_{S,\tau,\lambda}(z_{j+1})\}$; on an exterior ray it equals the adjacent observed score. Thus a thresholded score is a finite union of original intervals, and Lemma L.1 applies to its actual components, including their unsampled gaps and tails. These finite maxima and comparisons also show that the resulting prediction map is measurable. A countable linear order may equivalently be embedded in the real line by a fixed public order embedding.

Pointwise concentration of the local score. The concentration argument uses a population mass only to define an auxiliary random variable. The learner still computes the empirical local score and does not need the distribution.

Lemma L.2 (A replacement entropy bound)**.** *Let $f(S)$ be a bounded measurable function of independent observations. Let $S^{(i)}$ replace observation i by an independent observation with the same law, and let $\mathbb{E}'_i$ average only over this replacement, conditionally on S. Suppose that, for some $v, b > 0$,*

$$V_+(S) := \sum_i \mathbb{E}'_i[(f(S) - f(S^{(i)}))_+^2] \le v, \qquad |f(S) - f(S^{(i)})| \le b.$$

Then, for every $t > 0$,

$$\mathbb{P}\{|f(S) - \mathbb{E}f(S)| \ge t\} \le 2\exp\left\{-\frac{t^2}{2(2\mathrm{e}v + bt)}\right\}. \tag{143}$$

The variance assumption uses positive replacement differences; it does not require either sign of a deletion difference.

Proof. Write $\mathrm{Ent}(U) = \mathbb{E}[U\log U] - \mathbb{E}U\log\mathbb{E}U$. Conditionally on all coordinates other than i, Jensen's inequality gives $\mathrm{Ent}_i(U) \le \mathrm{Cov}_i(U, \log U)$. Use entropy tensorization, as in the product-space entropy method of (Boucheron et al., 2009), and symmetrize each covariance with an independent replacement. For $\theta \ne 0$, write $d_i = f(S) - f(S^{(i)})$. Sorting each pair so that $d_i > 0$ gives

$$\begin{aligned}\mathrm{Ent}(e^{\theta f}) &\le \sum_i \mathbb{E}\left[e^{\theta f}(1 - e^{-\theta d_i})\theta d_i \mathbf{1}\{d_i > 0\}\right] \\ &\le \theta^2 e^{|\theta| b}\mathbb{E}[e^{\theta f} V_+(S)].\end{aligned}$$

The last inequality holds for both signs of θ. For $\theta < 0$ its relevant factor is $|\theta| d_i(e^{|\theta| d_i} - 1)$, which is at most $\theta^2 d_i^2 e^{|\theta| b}$; it is still weighted by $e^{\theta f}$ at the larger value of f. Thus, for $|\theta| b \le 1$, the entropy

is at most $ev\theta^2\mathbb{E}[e^{\theta f}]$. The resulting differential inequality for the logarithmic moment generating function, integrated from zero, gives

$$\log \mathbb{E}e^{\theta(f-\mathbb{E}f)} \le ev\theta^2, \qquad |\theta|b \le 1.$$

For negative θ the direction of integration and the subsequent multiplication by θ both reverse. Apply Chernoff's inequality with $|\theta| = t/(2ev + bt) < 1/b$ to obtain (143). □

Lemma L.3 (Relative masses around a fixed query)**.** *There are universal constants $C, c > 0$ such that, for every distribution P_X on the original line, every fixed query x, every $n \ge 1$, and $0 < \varepsilon \le 1$,*

$$\mathbb{P}\{\exists I \ni x : p(I) \ge \varepsilon,\ \widehat{p}_S(I) < p(I)/2\} \le C\exp(-cn\varepsilon). \tag{144}$$

Atoms and zero-mass query points are allowed.

Proof. First consider either of the two nested families of intervals anchored immediately to the left or to the right of x. For $t > 0$ put $r_j = t(6/5)^j$, $j \ge 0$. For each j for which there is a member of mass at least r_j, intersect all such members to obtain J_j. An anchored interval family has a countable cofinal subfamily for this intersection, so continuity of measure gives $p(J_j) \ge r_j$. Also J_j is contained in every member of mass at least r_j. This description retains an atomic boundary point exactly when it is necessary to reach that mass.

If a member J has $p(J) \in [r_j, 6r_j/5)$ and $\widehat{p}_S(J) < 3p(J)/4$, then $\widehat{p}_S(J_j) < 9r_j/10$. The binomial lower-tail Chernoff bound therefore gives probability at most $\exp(-nr_j/200)$ for such a failure in this mass bin. When $nt \ge 1$, summing these bounds is at most $C\exp(-cnt)$: indeed, $(6/5)^j \ge 1 + j/5$, and the resulting geometric-series denominator is bounded away from zero by a universal constant. When $nt < 1$, enlarge C so that the asserted probability bound is at least one. Consequently, with failure probability at most $C\exp(-cnt)$, every member of the anchored family of mass at least t has empirical mass at least three quarters of its population mass.

For an interval $I \ni x$, split it into the disjoint pieces $I \cap (-\infty, x)$ and $I \cap [x, \infty)$, and apply this assertion to the two corresponding nested families with $t = \varepsilon/8$. The pieces of mass less than t have total mass at most $2t$. On the two successful events, if $p(I) \ge \varepsilon$, then

$$\widehat{p}_S(I) \ge \frac{3}{4}(p(I) - 2t) \ge \frac{9}{16}p(I) > \frac{1}{2}p(I).$$

A union bound proves (144). A countable linear order is treated through its public embedding in the real line. □

Lemma L.4 (Stability of a fixed interval under replacement)**.** *Assume $n\tau^2 \ge 2$. For a fixed interval I and a fixed number $r \ge 0$, put*

$$T_S(I) = \widehat{a}_S(I) - V_{S,\tau,\lambda}(I) - \lambda, \quad u_S = \max\{\widehat{p}_S(I), r\}, \quad Y_S = \max\left\{\frac{3}{4}\tau, \frac{(T_S(I))_+}{\sqrt{u_S}}\right\}.$$

A ratio with denominator zero is defined to be zero. Replacing one observation gives

$$|Y_S - Y_{S^{(i)}}| \le \frac{7}{n\sqrt{\max\{u_S, u_{S^{(i)}}\}}}, \tag{145}$$

where a zero denominator means that the difference is zero. The difference also vanishes unless the old or the new observation belongs to I.

Proof. Delete one observation while retaining the normalization $1/n$, or equivalently replace it by a cemetery observation that belongs to no interval. Every component of a maximizing set for V can be assumed to have strictly positive objective, by removing other components. Its empirical mass u is then greater than τ^2, since its gain is at most u whereas its cost is $\tau\sqrt{u} + \lambda$. Deleting an observation from such a component changes its square-root cost by at most

$$\tau(\sqrt{u} - \sqrt{u - 1/n}) = \frac{\tau}{n(\sqrt{u} + \sqrt{u - 1/n})} \le \frac{1}{n},$$

using $u - 1/n \ge \tau^2/2$. Conversely, inserting an observation into a component of a maximizing set after deletion changes that cost by at most $1/n$. Compare the same original set of intervals in

each direction. Its component charge $\lambda c(A)$ does not change, and its gain changes by at most $1/n$. Thus deletion changes V by at most $2/n$ and T by at most $3/n$. Passing between two replacements through their common deleted sample changes T by at most $6/n$. Using original sets in these comparisons avoids any dependence on changes to the empirical cell partition.

The empty-set and full-interval candidates for V give

$$0 \le (T_S(I))_+ \le \min\{\widehat{p}_S(I),\, \tau\sqrt{\widehat{p}_S(I)}\}.$$

In particular, its ratio in the statement is at most $\min\{\sqrt{u_S}, \tau\}$. Suppose $u_S \ge u_{S^{(i)}} > 0$. Positive-part contraction bounds the difference of the unclipped ratios by

$$\frac{6}{n\sqrt{u_S}} + \left(1 - \sqrt{\frac{u_{S^{(i)}}}{u_S}}\right)\sqrt{u_{S^{(i)}}} \le \frac{7}{n\sqrt{u_S}},$$

because $|u_S - u_{S^{(i)}}| \le 1/n$. The zero-denominator case follows from the same ratio bound, and interchanging the two samples treats the opposite ordering. Clipping at the floor preserves the bound. □

Lemma L.5 (A pointwise score center without a covering factor)**.** *There are universal constants $C, c > 0$ such that, whenever $0 < \tau \le 1$, $\lambda > 0$, and $n\tau^2 \ge 2$, for every distribution P and every fixed query x there is a deterministic $c_{P,n,\tau,\lambda}(x) \in [3\tau/4, \tau]$ satisfying*

$$\mathbb{P}\left\{|\max\{3\tau/4, F_{S,\tau,\lambda}(x)\} - c_{P,n,\tau,\lambda}(x)| \ge s\right\} \le C\exp(-cns^2), \qquad s > 0. \tag{146}$$

The center may depend on the sample size. The assertion is unconditional on any accuracy event.

Proof. Put $a = 3/4$. For the analysis, define the auxiliary score:

$$\begin{aligned} u_S(I) &= \max\{\widehat{p}_S(I), p(I)/2\}, \\ \widetilde{Y}_S(I) &= \max\{a\tau, (T_S(I))_+/\sqrt{u_S(I)}\}, \\ f(S) &= \sup_{I \ni x} \widetilde{Y}_S(I). \end{aligned}$$

The zero-denominator convention remains in force. In particular, $a\tau \le f(S) \le \tau$.

These suprema are measurable, including on the real line. Indeed, if $I \ni x$ has nonempty empirical trace, let H be the closed convex hull of that trace together with x. Then $H \subseteq I$ and the empirical traces coincide. Hence V, T, and $\widehat{p}$ are unchanged, whereas $p(H) \le p(I)$; the auxiliary ratio cannot decrease. An empty trace contributes only the floor. Consequently $f(S)$ is a finite maximum over closed intervals with endpoints among $x, X_1, \ldots, X_n$ that contain x. Their population masses are Borel functions of the endpoints, by the distribution function and its left limits. This also proves joint measurability in the sample and query. The same reasoning applies after embedding a countable order.

We bound its positive replacement variance. If $f(S) = a\tau$, that variance is zero. Otherwise choose an interval I with $\widetilde{Y}_S(I) \ge f(S) - \epsilon > a\tau$. We may use arbitrarily small $\epsilon > 0$, so the argument does not rely on attainment of the supremum. For every independent replacement,

$$(f(S) - f(S^{(i)}))_+ \le (\widetilde{Y}_S(I) - \widetilde{Y}_{S^{(i)}}(I))_+ + \epsilon.$$

The fixed interval in this comparison is an original interval and remains a candidate after replacement. Lemma L.4 and its support condition give

$$\sum_i \mathbb{E}_i'[(\widetilde{Y}_S(I) - \widetilde{Y}_{S^{(i)}}(I))_+^2] \le \frac{49(\widehat{p}_S(I) + p(I))}{nu_S(I)} \le \frac{147}{n}.$$

Here the old observations contribute $n\widehat{p}_S(I)$ possible hits and the independent replacements contribute $np(I)$ expected hits. Minkowski's inequality gives $\sqrt{V_+(S)} \le \sqrt{147/n} + \sqrt{n}\,\epsilon$. Letting ϵ decrease to zero proves $V_+(S) \le 147/n$.

There is also a uniform bound on every replacement jump. At the larger of two unequal values of f, take the same kind of approximate maximizing interval above the floor. Its unclipped ratio is at most $\sqrt{u_S(I)}$, so $u_S(I) \ge a^2\tau^2$. Lemma L.4, followed by $\epsilon \downarrow 0$, gives

$$|f(S) - f(S^{(i)})| \le \frac{7}{an\tau} = \frac{28}{3n\tau}.$$

Lemma L.2 now implies

$$\mathbb{P}\{|f(S) - \mathbb{E}f(S)| \geq s\} \leq 2\exp(-cns^2), \qquad 0 < s \leq \tau.$$

It remains to compare the auxiliary and actual scores. Apply Lemma L.3 with $\varepsilon = a^2\tau^2$. Outside an event of probability at most $C\exp(-cn\tau^2)$, every interval through x of population mass at least ε has $p(I) \leq 2\widehat{p}_S(I)$. Any interval whose original ratio exceeds $a\tau$ has $\widehat{p}_S(I) \geq a^2\tau^2$. If it had $p(I) > 2\widehat{p}_S(I)$, it would contradict that successful mass event. Thus its auxiliary denominator equals its empirical denominator. All other intervals contribute at most the floor in both suprema, since the auxiliary denominator is larger. Consequently

$$f(S) = \max\{a\tau, F_{S,\tau,\lambda}(x)\}$$

on the successful event. Set $c_{P,n,\tau,\lambda}(x) = \mathbb{E}f(S)$. For $0 < s \leq \tau$, combine the two failure probabilities and use $s^2 \leq \tau^2$ to obtain (146). For $s > \tau$ the event is empty. Population masses have entered only this proof, so the learner acquires neither distributional information nor additional random bits. □

Lemma L.6 (The finite grid and replication)**.** *There are universal constants $C, c > 0$ with the following property. Let $0 < \tau \leq 1$, $\lambda > 0$, $K \geq 1$, and use the grid (153). If $n\tau^2 \geq CK^2$, then the learner that thresholds $F_{S,\tau,\lambda}$ at a uniform grid point satisfies, at every fixed query x,*

$$\mathbb{P}_{S,S',j}\{\mathbf{1}\{F_{S,\tau,\lambda}(x) \geq w_j\} \neq \mathbf{1}\{F_{S',\tau,\lambda}(x) \geq w_j\}\} \leq \frac{1}{2K} + \frac{C}{K}\exp\left(-c\frac{n\tau^2}{K^2}\right). \tag{147}$$

The samples are independent and use the same uniform j. In particular, for $\Delta > 0$, the right side is at most $1/(2K) + \Delta$ whenever

$$n \geq \frac{CK^2}{\tau^2}\log\left(2 + \frac{1}{K\Delta}\right). \tag{148}$$

No accuracy event is used in this assertion.

Proof. Put $h = \tau/(10K)$ and choose a grid point nearest to $c_{P,n,\tau,\lambda}(x)$, resolving a tie by the smaller index. For this one seed the two predictions are independent Bernoulli variables with the same parameter, so their disagreement probability is at most $1/2$. There are at most two other grid points at each positive index distance r, and each is at distance at least $h(r-1/2)$ from the center. All grid points exceed the clipping floor $3\tau/4$, so clipping does not change a prediction. A disagreement at another grid point requires at least one of the two clipped scores to deviate from the center by its distance to that point. By (146), summing over the grid bounds the total disagreement by

$$\frac{1}{2K} + \frac{C}{K}\sum_{r\geq 1}\exp\{-cnh^2(r-1/2)^2\}.$$

Once nh^2 exceeds a universal constant, this series is bounded by $C\exp(-cnh^2)$, proving (147). Increasing the universal constant in (148) makes its exponential term at most $K\Delta/C$, and also ensures the preceding lower bound on nh^2. □

Lemma L.7 (A fixed charge and additive confidence)**.** *For fixed $\lambda > 0$, define*

$$Z_S = \max\left\{0, \sup_{\mathcal{C}}\left[\sum_{I\in\mathcal{C}}(\widehat{a}_S(I) - a(I)) - \lambda|\mathcal{C}|\right]\right\}, \tag{149}$$

where $\mathcal{C}$ ranges over finite pairwise disjoint families of nonempty intervals. There is a universal constant C such that, for all $n \geq 1$, $\lambda > 0$, and $t > 0$,

$$\mathbb{E}Z_S \leq \frac{C}{n\lambda}, \qquad \mathbb{P}\left\{Z_S > \frac{C}{n\lambda} + t\right\} \leq e^{-nt^2/2}. \tag{150}$$

For every $\tau/\sqrt{2} < w < \tau$ and $B_w = \{x : F_{S,\tau,\lambda}(x) \geq w\}$, one has

$$a(B_w) \geq V_{S,\tau,\lambda}(\mathcal{X}) - Z_S. \tag{151}$$

Proof. Let $\mathcal{U}_m$ be the unions of at most m intervals and put $D_m = \sup_{U\in\mathcal{U}_m}(\widehat{a}_S(U) - a(U))$. The VC dimension of $\mathcal{U}_m$ is at most $2m$: among $2m+1$ ordered points, the labeling that is positive exactly at the odd-indexed points has $m+1$ separated positive runs. The packing and chaining argument for (109) therefore gives

$$\mathbb{E}D_m \le c_0\sqrt{m/n}.$$

This bound applies equally on a countable order and on the real line, and its constant is independent of m. Replacing one observation changes D_m by at most $2/n$.

Choose a universal $A \ge 4c_0^2$ and let $t \ge A/(n\lambda)$. Young's inequality yields

$$\mathbb{E}D_m \le \lambda m/4 + c_0^2/(n\lambda) \le (\lambda m + t)/4.$$

Bounded differences and a union bound give

$$\begin{aligned}\mathbb{P}\{Z_S > t\} &\le \sum_{m\ge 1}\exp\{-9n(t+\lambda m)^2/32\}\\ &\le \sum_{m\ge 1}\exp\{-9n\lambda t m/8\} \le 2\exp\{-9n\lambda t/8\},\end{aligned}$$

after enlarging A so that the geometric-series ratio is at most $1/2$. Integrating this bound above $A/(n\lambda)$ proves $\mathbb{E}Z_S \le C/(n\lambda)$. This argument allows a small positive expected deficit; it does not require every interval's deviation to be smaller than its charge.

Disjointness gives $0 \le Z_S \le 2$. Replacing one observation changes the signed gain of the union of a fixed family by at most $2/n$; its fixed charge does not change. Taking a supremum and a positive part preserves this bound. McDiarmid's inequality proves the second part of (150).

These suprema are measurable on the real line. Fix the countable set consisting of the rationals and all atoms of P_X, and allow both open and closed endpoints from that set, together with infinities. Its finite disjoint interval families give the same supremum. For any sample and any finite family, retain atomic endpoints exactly and approach non-atomic endpoints from appropriate sides. One can preserve all finitely many sample memberships, the number of intervals, and their disjointness, while their population symmetric differences tend to zero. The resulting objectives converge to the original one. This also proves measurability of D_m and applies after embedding a countable order.

Finally, the empirical certificate gives $\widehat{a}_S(C) \ge e(C) + \lambda$ on each selected component C, and their surplus sums to $V_{S,\tau,\lambda}(\mathcal{X})$. Apply (149) to this particular disjoint family and sum to obtain (151). □

Theorem L.8 (Unions of intervals with a prescribed finite seed)**.** *Let $k \ge 1$, and let $\mathcal{H}_k$ consist of the classifiers $2\mathbf{1}_A - 1$ for unions A of at most k intervals on a countable linear order or on $\mathbb{R}$. The empty set is allowed. Suppose $0 < \alpha, \beta, \rho \le 1/4$, $B \ge 0$ is an integer, $K = 2^B$, and $\Delta = \rho - 1/(2K) > 0$. There is an improper agnostic (α,β)-learner that is ρ-pointwise replicable, uses exactly B total shared fair bits and no other randomness, and uses a fixed number of examples satisfying*

$$n = O\left(\frac{1}{\alpha^2}\left[kK^2\log\left(2+\frac{1}{K\Delta}\right) + \log\frac{2}{\beta}\right]\right). \tag{152}$$

In particular, constant slack in $K = \Theta(1/\rho)$ gives $O(\alpha^{-2}[k\rho^{-2} + \log(2/\beta)])$ samples and $O(\log(1/\rho))$ total shared bits. The comparator's intervals and their masses are unknown, and the distribution may have arbitrary agnostic misspecification.

Proof. Use the constant negative classifier as baseline, so that flipping on A improves its risk by $a(A)$. Fix

$$\tau = \frac{\alpha}{4\sqrt{k}}, \qquad \lambda = \frac{\alpha}{4k}, \qquad w_j = \frac{4}{5}\tau + \left(j+\frac{1}{2}\right)\frac{\tau}{10K}, \qquad 0 \le j < K. \tag{153}$$

The B shared fair bits choose a uniform j, and the learner outputs $2\mathbf{1}\{F_{S,\tau,\lambda} \ge w_j\} - 1$. The preceding cell construction defines this measurable classifier on the whole original domain.

For accuracy, the same VC packing argument used in Lemma L.7, now for the absolute signed deviation, gives expected uniform deviation $C\sqrt{k/n}$. Bounded differences therefore imply, with probability at least $1 - \beta/2$,

$$\sup_{A \in \mathcal{U}_k} |\widehat{a}_S(A) - a(A)| \le \alpha/4$$

whenever $n \ge C\alpha^{-2}[k + \log(2/\beta)]$. The same sample bound, with a sufficiently large constant, makes $C/(n\lambda) \le \alpha/8$ and, by Lemma L.7, ensures $Z_S \le \alpha/4$ with probability at least $1 - \beta/2$. On the intersection of these events, the certificate implies, simultaneously for all seeds,

$$\sup_{A \in \mathcal{U}_k} a(A) \le V_{S,\tau,\lambda}(\mathcal{X}) + \tau\sqrt{k} + \lambda k + \alpha/4 \le a(B_{w_j}) + \alpha.$$

Here $\tau\sqrt{k} + \lambda k = \alpha/2$.

For replication, apply Lemma L.6 to the same score and the grid (153). It bounds the disagreement at each fixed query by

$$\frac{1}{2K} + \frac{C}{K} \exp\left(-c\frac{n\tau^2}{K^2}\right).$$

A sufficiently large constant in (152) makes the second term at most Δ and satisfies the accuracy conditions. All choices, including the sample size, are fixed independently of the unknown distribution. □

Proposition L.9 (Preprocessing and query complexity)**.** *Let $N \le n$ be the number of distinct feature values in a nonempty sample. The learner of Theorem L.8 admits $O(n \log n + N^3)$ preprocessing and $O(\log(N + 1))$ time per query. After sorting and aggregating the sample, its working space and stored representation each use $O(N)$ real registers. All score preprocessing is independent of the shared seed and is reused for every grid threshold. Drawing the shared index and forming its threshold takes a separate $O(B + 1)$ operations. These bounds use an ordered-comparison, real-arithmetic RAM with unit-cost square roots and exact real comparisons; they do not assert polynomial exact bit complexity for algebraic intermediate values.*

Proof. Sort the observations and aggregate equal feature values as $z_1 < \cdots < z_N$, with masses p_t and signed gains a_t. Keep the original normalization $1/n$, including when computing a local subrange. Thus $p_t > 0$, and prefix sums evaluate $p(l, j)$ and $a(l, j)$ in constant time. As in the construction on the original line above, filling unsampled gaps within a selected trace run preserves its gain and decreases both penalties. Conversely, each trace run is realized by the interval between its observed endpoints. An isolated zero-mass component can be removed to save its charge λ. Local optimal values therefore equal those of the compressed line of N observed sites.

For each fixed i, put $v_i(r) = 0$ for $r < i$ and compute, in increasing order of $j \ge i$,

$$v_i(j) = \max\left\{v_i(j-1), \max_{i \le l \le j}\left[v_i(l-2) + a(l, j) - \tau\sqrt{p(l, j)} - \lambda\right]\right\}.$$

This is exactly $V_{S,\tau,\lambda}([z_i, z_j])$. Indeed, either site j is unselected, or the last selected run is $[l, j]$. In the latter case site $l-1$ is unselected, and the preceding selected set lies in $[i, l-2]$; its objective adds to that of the last run. This also proves the recurrence at $l = i$ and $l = i + 1$ under the empty-prefix convention. All rows together require $O(N^3)$ operations.

The same row gives

$$s_{i,j} = \max\left\{0, \frac{a(i, j) - v_i(j) - \lambda}{\sqrt{p(i, j)}}\right\}, \qquad F_t = \max_{i \le t \le j} s_{i,j}.$$

The local-score formula identifies F_t with $F_{S,\tau,\lambda}(z_t)$. Initialize every F_t to zero. After computing row i, scan $j = N, N-1, \ldots, i$, maintain $r = \max_{q \ge j} s_{i,q}$, and replace F_j by $\max\{F_j, r\}$. These suffix scans take $O(N^2)$ operations altogether. Each completed row can then be discarded, leaving only the prefix sums, one working row, its scores, and the final F_t array.

At a query, binary search locates an observed site or its neighboring gap. The original-line extension proved above assigns a gap the maximum of its two neighboring scores and an exterior ray its adjacent score. One comparison with the seed's threshold then gives the prediction. This proves the query and storage bounds, and the whole score array serves all shared seeds. □

Corollary L.10 (The full joint tradeoff for interval unions)**.** *Suppose the domain is $\mathbb{R}$, or a countable linear order containing at least $2k$ points. Uniformly for $0 < \alpha \le 1/16$, $0 < \beta \le 1/4$, $B \ge 4$, $K = 2^B$, and $0 < K\Delta \le 1/4$,*

$$m^*_{\mathcal{H}_k}(\alpha, \beta, 1/(2K) + \Delta; B) = \Theta\left(\frac{1}{\alpha^2}\left[kK^2 \log\frac{1}{K\Delta} + \log\frac{1}{\beta}\right]\right). \tag{154}$$

The constants are absolute. The minimum permits improper outputs, a fixed sample size, and at most B total shared fair bits, with no other randomness.

Proof. Theorem L.8 gives the upper bound. Any $2k$ ordered points are shattered: every labeling on them has at most k positive runs, each realized by an interval. Thus Proposition F.8, with $d = 2k$, gives the boundary term in (154). Concentrating the feature distribution on any one of these points and invoking Proposition F.7 gives the confidence term $\alpha^{-2}\log(1/\beta)$. The maximum of the two lower bounds is at least half their sum. On the stated range, the different fixed constants inside the logarithms affect only universal factors. □

Proposition L.11 (Failure of the square-root-only selector)**.** *Set $\lambda = 0$ in (140), and write $F^{(0)}_{S,\tau}$ for the resulting score. Let $M \ge 256$ and $n \ge 2M$ be integers, and put $\tau = M^{-1/2}$, $\alpha = 4M^{-1/2}$, $D = n/M$, and $r = (8\mathrm{e})^{-2D}$. There is a distribution P on $\mathbb{N} \times \{-1, +1\}$ such that*

$$\mathbb{P}_S\left\{\forall w < \tau : \mathrm{R}_P(2\mathbf{1}\{F^{(0)}_{S,\tau} \ge w\} - 1) - \inf_{h \in \mathcal{H}_1} \mathrm{R}_P(h) \ge r/4\right\} \ge 1 - \exp\{-Mr^2/(8D)\}.$$

If $M > 65536$ and $D \le \log M/[8\log(8\mathrm{e})]$, its accuracy failure probability is at least $1 - \exp\{-\sqrt{M}/(8D)\}$, which tends to one in this regime. The failure thus persists up to sample sizes of order $\alpha^{-2}\log(1/\alpha)$ for this particular selector. It is not a lower bound for arbitrary learners; the single-interval learner of Corollary K.15 does not incur this logarithmic accuracy cost.

Proof. Take X uniform on $\{1, \dots, M\}$ and $\mathbb{P}(Y = +1 \mid X) = 1/4$. The constant negative classifier is optimal, and flipping mass u increases risk by $u/2$. Put $d = \lceil D \rceil$. Call an atom favorable if it occurs exactly d times, all with positive labels. At such an atom, $q = \widehat{p} = \widehat{a} = d/n \ge \tau^2$ and the singleton value is $V_{S,\tau,0} = q - \tau\sqrt{q}$. Its local score is therefore τ, so every stated threshold flips it. A fixed atom is favorable with probability

$$\pi = \binom{n}{d}(4M)^{-d}(1 - 1/M)^{n-d} \ge 8^{-d}e^{-2n/M} \ge (8\mathrm{e})^{-2D} = r.$$

Here $\binom{n}{d} \ge (n/d)^d$, $d \le 2D$, and $\log(1 - 1/M) \ge -2/M$ were used. If G is the fraction of favorable atoms, then $\mathbb{E}G = \pi$; replacing one observation changes G by at most $2/M$. McDiarmid's inequality gives $\mathbb{P}\{G < r/2\} \le \exp\{-Mr^2/(8D)\}$. On its complement every threshold has excess risk at least $r/4$. Finally, the stated regime gives $r \ge M^{-1/4}$ and $r/4 > \alpha$. □

Remark L.12 (Other baselines and comparison classes). A known measurable baseline f_0 can replace the constant negative classifier: use signed labels $-Yf_0(X)$ and flip f_0 on the selected set. The same proof compares against any subclass of flips on at most k intervals, whether or not it contains f_0. The output may have more than k components. The construction requires no public partition. Its guarantee applies to interval flips; extending it to anchored rectangles and general VC-two classes remains open.

M Finite randomness for sparse changes to a fixed classifier

We give an order-free learner that changes the baseline at finitely many observed points. The same sample rate follows from Theorem L.8 by embedding the domain in the real line and treating singleton changes as intervals. The direct proof instead combines a squared-deviation identity with a shifted normalized signal, using one sample for both discovery and estimation.

Throughout, $\mathcal{X}$ is a standard Borel space and $f_0 : \mathcal{X} \to \{-1, +1\}$ is known and measurable. This includes every countable discrete domain. Transform labels to $Z = -Y f_0(X)$ and write

$$p_x = P(X = x), \qquad b_x = \mathbb{E}[Z\mathbf{1}\{X = x\}],$$
$$q_x = \frac{1}{n}\sum_i \mathbf{1}\{X_i = x\}, \quad \widehat{b}_x = \frac{1}{n}\sum_i Z_i\mathbf{1}\{X_i = x\}.$$

Flipping f_0 on a finite set A reduces its risk by $b(A) = \sum_{x\in A} b_x$.

Lemma M.1 (Squared deviations of atomic gains)**.** *Put $\delta_x = \widehat{b}_x - b_x$ and*

$$D_d = \sup_{A:\,|A|\le d} |\delta(A)|, \qquad Z_\lambda = \max\left\{0,\ \sup_{A\ \text{finite}} [\delta(A) - \lambda|A|]\right\}.$$

For $d \ge 1$ and $\lambda > 0$,

$$\mathbb{E}\sum_x \delta_x^2 = \frac{1 - \sum_x b_x^2}{n} \le \frac{1}{n}, \qquad \mathbb{E}D_d \le \sqrt{\frac{d}{n}}, \qquad \mathbb{E}Z_\lambda \le \frac{1}{4n\lambda}. \tag{155}$$

Both D_d and Z_λ change by at most $2/n$ upon replacing one observation. In particular, for $t > 0$,

$$\mathbb{P}\left\{D_d > \sqrt{d/n} + t \text{ or } Z_\lambda > 1/(4n\lambda) + t\right\} \le 2e^{-nt^2/2}. \tag{156}$$

Proof. The population atom set $\mathcal{A} = \{x : p_x > 0\}$ is countable, and every nonzero δ_x lies in $\mathcal{A}$ or among the observations. These sums and suprema are measurable: enumerate the fixed atoms together with the observed points and remove repetitions using equality, which is a Borel relation on a standard Borel space. Finite subsets of this countable enumeration give the same suprema. Thus the notation does not require an uncountable sum of nonzero terms.

For two independent observations, equality of their feature values occurs only at population atoms, almost surely. Consequently

$$\mathbb{E}[Z_iZ_j\mathbf{1}\{X_i = X_j\}] = \sum_x b_x^2 \qquad (i \ne j).$$

Expanding $\sum_x \widehat{b}_x^2 = n^{-2}\sum_{i,j} Z_iZ_j\mathbf{1}\{X_i = X_j\}$, and using $\mathbb{E}\sum_x \widehat{b}_x b_x = \sum_x b_x^2$, proves the first identity in (155). Cauchy–Schwarz gives $D_d \le \sqrt{d}\,(\sum_x \delta_x^2)^{1/2}$. Also $\sum_x |\delta_x| \le 2$, so only finitely many δ_x exceed λ, and

$$Z_\lambda = \sum_x (\delta_x - \lambda)_+ \le \frac{1}{4\lambda}\sum_x \delta_x^2.$$

The last inequality is $(u - \lambda)_+ \le u^2/(4\lambda)$. Taking expectations proves the remaining bounds.

For every fixed finite set, replacing one observation changes its empirical signed gain by at most $2/n$. Taking an absolute value, a supremum, or a positive part preserves this bound; the fixed charge does not change. McDiarmid's inequality and a union bound give (156). □

Lemma M.2 (A shifted normalized atom)**.** *Fix $\tau, \lambda > 0$ and define, at every point,*

$$F_S(x) = \begin{cases} \min\{\tau, (\widehat{b}_x - \lambda)_+/\sqrt{q_x}\}, & q_x > 0, \\ 0, & q_x = 0, \end{cases}$$

$$F_P(x) = \begin{cases} \min\{\tau, (b_x - \lambda)_+/\sqrt{p_x}\}, & p_x > 0, \\ 0, & p_x = 0. \end{cases}$$

For every fixed x and $u > 0$,

$$\mathbb{P}\{|F_S(x) - F_P(x)| \ge u\} \le 4e^{-nu^2/16}. \tag{157}$$

This assertion is unconditional on any accuracy event.

Proof. Suppose $p_x > 0$ and write $\nu_x = b_x/p_x$. For $\nu \in [-1, 1]$, the function

$$g_\nu(t) = \mathrm{clip}_{[0,\tau]}(\nu t - \lambda/t), \qquad t > 0, \qquad g_\nu(0) = 0,$$

is 2-Lipschitz. If $\nu \le 0$, it is zero. On its positive, uncapped part when $\nu > 0$, one has $\lambda/t^2 < \nu \le 1$, and hence $g'_\nu(t) = \nu + \lambda/t^2 \le 2$. Continuity at the clipping transitions and at zero proves the assertion everywhere.

Let $\widehat{\nu}_x$ be the average transformed label at x when its count is positive. Lipschitzness of clipping and the preceding bound imply

$$|F_S(x) - F_P(x)| \le \sqrt{q_x}\,|\widehat{\nu}_x - \nu_x| + 2|\sqrt{q_x} - \sqrt{p_x}|,$$

where the first term is zero if the count is zero. Conditionally on that count, Hoeffding's inequality bounds the tail of the first term at v by $2e^{-nv^2/2}$. The square-root binomial bound (9) bounds the tail of $|\sqrt{q_x} - \sqrt{p_x}|$ at v by $2e^{-nv^2}$, including the zero-count endpoint. Use $v = u/2$ in the first bound and $v = u/4$ in the second to prove (157). If $p_x = 0$, then $q_x = 0$ almost surely and both scores are zero. □

Theorem M.3 (Sparse changes at a prescribed bit budget)**.** *Let $d \ge 1$ be an integer, and let $\mathcal{H}$ be a nonempty class of measurable classifiers on $\mathcal{X}$ satisfying*

$$|\{x : h(x) \ne f_0(x)\}| \le d \qquad (h \in \mathcal{H}).$$

The baseline need not belong to $\mathcal{H}$. Suppose $0 < \alpha, \beta, \rho \le 1/4$, $B \ge 0$ is an integer, $K = 2^B$, and $\Delta_B = \rho - 1/(2K) > 0$. There is an improper agnostic (α, β)-learner that is ρ-pointwise replicable on every distribution, uses exactly B total shared fair bits and no other randomness, and uses a fixed number of examples satisfying

$$n = O\left(\frac{1}{\alpha^2}\left[dK^2 \log\left(2 + \frac{1}{K\Delta_B}\right) + \log\frac{2}{\beta}\right]\right). \tag{158}$$

Its output differs from f_0 only on finitely many observed points. The bound uses the displayed Hamming radius d, not the VC dimension of an arbitrary subclass. Unknown atom masses and a nonatomic part are allowed.

Proof. Use one sample of size n, and put

$$\tau = \frac{\alpha}{4\sqrt{d}}, \qquad \lambda = \frac{\alpha}{4d}, \qquad w_j = \frac{4}{5}\tau + \left(j + \frac{1}{2}\right)\frac{\tau}{10K}, \quad 0 \le j < K.$$

The B shared bits choose a uniform j, and the learner flips $f_0(x)$ exactly when $F_S(x) \ge w_j$. Only points occurring in the sample can be flipped. Empirical equality groups, their sums, and the threshold comparison are jointly Borel in the sample and query, so this specifies a measurable classifier on the whole domain without a discovery stage or an ordering.

For its empirical certificate, write

$$e_x = (\widehat{b}_x - \tau\sqrt{q_x} - \lambda)_+, \qquad E = \sum_x e_x, \qquad B_w = \{x : F_S(x) \ge w\}.$$

All positive e_x are selected whenever $w < \tau$. At every selected point, $\widehat{b}_x \ge e_x + \lambda + w\sqrt{q_x}$. Consequently

$$\widehat{b}(B_w) \ge E + \lambda|B_w|, \qquad b(B_w) \ge E - Z_\lambda.$$

On the other hand, for every A with $|A| \le d$,

$$\widehat{b}(A) \le E + \tau\sum_{x\in A}\sqrt{q_x} + \lambda|A| \le E + \tau\sqrt{d} + \lambda d.$$

It follows, simultaneously for every shared seed, that

$$\sup_{A:\,|A|\le d} b(A) - b(B_{w_j}) \le \tau\sqrt{d} + \lambda d + D_d + Z_\lambda = \alpha/2 + D_d + Z_\lambda.$$

Lemma M.1 gives $D_d, Z_\lambda \le \alpha/4$ with probability at least $1-\beta$ as soon as $n \ge C\alpha^{-2}[d+\log(2/\beta)]$. For example, make both displayed expectation bounds at most $\alpha/8$ and use $t = \alpha/8$ in (156). Every comparator in $\mathcal{H}$ flips a set of size at most d, so this proves the required agnostic accuracy.

For replication, fix x and choose a grid point nearest to the deterministic center $F_P(x)$. At that seed, two independent predictions with the same Bernoulli parameter disagree with probability at most $1/2$. Every other grid point at index distance r is at least $h(r-1/2)$ from the center, where $h = \tau/(10K)$; there are at most two such points. Lemma M.2 bounds a disagreement there by $8\exp[-nh^2(r-1/2)^2/16]$. Summing the geometric tail, for $n\tau^2 \ge CK^2$, gives

$$\mathbb{P}\{H_S(x) \ne H_{S'}(x)\} \le \frac{1}{2K} + \frac{C}{K}\exp\left(-c\frac{n\tau^2}{K^2}\right).$$

This calculation includes centers outside the grid and uses no accuracy event. It is at most $1/(2K) + \Delta_B$ whenever $n \ge CK^2\tau^{-2}\log(2 + 1/(K\Delta_B))$. Together with the accuracy condition and $\tau^2 = \alpha^2/(16d)$, this proves (158). □

Corollary M.4 (Learning sparse changes with a finite seed)**.** *Under the hypotheses of Theorem M.3, for $0 < \alpha, \beta, \rho \le 1/4$, there is an improper agnostic (α, β)-learner that is ρ-pointwise replicable, uses exactly $B = \lceil \log_2(2/\rho) \rceil$ total shared fair bits and no other randomness, and uses*

$$O\left(\frac{1}{\alpha^2}\left[\frac{d}{\rho^2} + \log\frac{2}{\beta}\right]\right)$$

examples. Its output makes only finitely many point changes to the baseline.

Proof. For this choice, $K = \Theta(1/\rho)$ and $K\Delta_B = K\rho - 1/2$ is bounded above and below by positive universal constants. Apply Theorem M.3. □

Corollary M.5 (Full Hamming balls and the joint tradeoff)**.** *For an integer $d \ge 1$, on a standard Borel domain containing at least d points, the full Hamming ball*

$$\mathcal{H}_{f_0,d} = \{h : |\{x : h(x) \ne f_0(x)\}| \le d\}$$

has VC dimension d. Uniformly for $0 < \alpha \le 1/16$, $0 < \beta \le 1/4$, $B \ge 4$, $K = 2^B$, and $0 < K\Delta_B \le 1/4$,

$$m^*_{\mathcal{H}_{f_0,d}}(\alpha, \beta, 1/(2K) + \Delta_B; B) = \Theta\left(\frac{1}{\alpha^2}\left[dK^2 \log\frac{1}{K\Delta_B} + \log\frac{1}{\beta}\right]\right). \tag{159}$$

The minimum uses a fixed sample size and permits improper outputs and at most B total shared fair bits, with no other randomness.

Proof. Any d points are shattered by choosing which baseline predictions to flip. On $d+1$ points the labeling opposite to the baseline everywhere cannot be realized, so the VC dimension is d. Theorem M.3 supplies the upper bound. Proposition F.8 on these d shattered points supplies the boundary term. Concentrating the distribution on one such point and using Proposition F.7 supplies $\alpha^{-2}\log(1/\beta)$. The maximum of these two lower bounds is at least half their sum, proving (159). □

Remark M.6 (Approximation error and radius). The result allows arbitrary approximation error above Bayes risk. For example, with $f_0 \equiv -1$, uniform mass on $N > d$ points, and always-positive labels, the full ball has optimum risk $1 - d/N$ whereas Bayes risk is zero. For a proper subclass of the ball, only the radius-based upper bound is asserted; the matching lower bound above uses the full ball's shattered points.

N A STATISTICAL OBSTRUCTION FOR THE WEIGHTED-CDF COMPILER

Proposition F.10 exhibits amplification for prescribed ensembles. Here independent training samples of a fixed deterministic agnostic learner produce the same effect. The class and the learner are fixed, the domain is countable, and every target distribution has finite support. The lower bound applies to the weighted-CDF compiler of Theorem F.3, including its independent empirical validation step.

Theorem N.1 (A statistical lower bound for the weighted-CDF compiler)**.** *There are a countable discrete domain $\mathcal{X}$, a countable class $\mathcal{H}$ of VC dimension at most three, a deterministic agnostic learner A for $\mathcal{H}$, a point $x_* \in \mathcal{X}$, and a universal constant $c_* > 0$ with the following property. For every rational $0 < \alpha \le 1/16$, there is a finitely supported distribution P_α with $P_{\alpha,X}(x_*) = 0$ such that the following holds.*

Let $\alpha/8 \le a \le \alpha/4$, let $T, K \ge 1$ be integers, and let $n \ge 1024/\alpha^2$ be an integer. Run A at parameters $(\alpha/8, \alpha/8)$ on T fresh independent sample blocks. Form all T integer-threshold candidates, their positive-part risk weights from an independent validation sample of size n, and the inverse-CDF selection using a uniform shared midpoint of the K-point grid, as in Theorem F.3. Every base output is an exact minimizer of R_{P_α} over $\mathcal{H}$. Nevertheless, the outputs H, H' on two independent training and validation samples, using the same grid midpoint, satisfy

$$\mathbb{P}\{H(x_*) \ne H'(x_*)\} \ge c_* \min\left\{1, \frac{1}{\alpha\sqrt{T}}\right\} - \frac{1}{K}. \tag{160}$$

In particular, if $\rho + 1/K < c_*$, this compiler can be ρ-pointwise replicable on all distributions only if

$$T \ge \frac{c_*^2}{\alpha^2(\rho + 1/K)^2}. \tag{161}$$

For $K = \Theta(1/\rho)$ this is $T = \Omega(1/(\alpha^2\rho^2))$. The example is unaffected by increasing the number of examples in each base call. Thus the extra α^{-2} ensemble cost persists for this compiler paired with a fixed valid agnostic learner. The lower bound concerns this pair, rather than learning the class by arbitrary methods.

N.1 One fixed class and one deterministic learner

Write $\mathbb{Q}_\circ = \mathbb{Q}/\mathbb{Z}$ for the rational circle, oriented by increasing coordinates modulo one, and set

$$\mathcal{X} = \{x_*\} \cup \{(r, \ell, u, b) : r, u \in \mathbb{Q}_\circ, \quad \ell \in \mathbb{Q} \cap (0,1), \quad b \in \{0,1\}\}.$$

Equip this countable set with its discrete sigma-field. For $s \in \mathbb{Q}_\circ$ and $e \in \{0,1\}$, define

$$h_{s,e}(r, \ell, u, b) = 2\mathbf{1}\{s \in [r, r+\ell) \pmod 1\} - 1, \qquad h_{s,e}(x_*) = 2e - 1,$$

and let $\mathcal{H} = \{h_{s,e} : s \in \mathbb{Q}_\circ, e \in \{0,1\}\}$.

Lemma N.2 (The class and the sample-generated ensemble)**.** *The class $\mathcal{H}$ has VC dimension at most three and admits a deterministic exact empirical-risk minimizer A. For every rational $0 < \alpha \le 1/16$, there is a finitely supported P_α for which every member of $\mathcal{H}$ has risk $1/2 - 2\alpha$, and the following holds for every positive sample size. The output of A has a shift uniform on a finite grid and an independent extension that predicts $+1$ at x_* with probability $p = 1/4 + 2\alpha$. Fresh independent sample blocks give independent copies of these two parameters.*

Proof. For m points other than x_*, the at most $2m$ rational endpoints $r_i, r_i + \ell_i$ partition the parameter circle into at most $2m$ half-open cells. The prediction vector is constant on each cell, including its initial endpoint. Consequently three such points have at most six prediction vectors and cannot be shattered. The restriction of $\mathcal{H}$ to $\mathcal{X} \setminus \{x_*\}$ has VC dimension at most two. Adding the one independent value at x_* increases this upper bound by at most one.

Here is a deterministic empirical-risk minimizer on every nonempty sample. Read an anchor (u, b) from the first example if its feature is different from x_*, and otherwise use $(u, b) = (0, 0)$. Among all globally empirical-risk-minimizing pairs (s, e), first choose $e = b$ if any minimizer has this value, and choose $e = 1 - b$ otherwise. For this chosen e, select the first minimizing s encountered when traversing the circle from u in its fixed orientation, including u itself.

For each e, empirical risk as a function of s is a right-continuous step function with finitely many rational endpoints. Its minimum is attained, and its minimizer set is a nonempty union of half-open cells whenever this e is globally optimal. The first minimizing point from u therefore exists: it is either u or the initial endpoint of one of these cells. In particular it is rational. This defines a deterministic exact empirical-risk minimizer for every sample. All required maps are measurable

because their domains are countable and discrete. The usual VC uniform-convergence bound makes this one rule an ordinary agnostic learner, by using a fixed sample schedule

$$m_A(\varepsilon, \delta) = O\left(\frac{\log(2/\varepsilon) + \log(2/\delta)}{\varepsilon^2}\right).$$

Fix a rational $0 < \alpha \le 1/16$. Choose an integer L such that $L/4$ and $4\alpha L$ are integers, and write

$$\Gamma_L = \{0, 1/L, \dots, (L-1)/L\}, \qquad q_0 = \frac{1}{4}, \qquad q_1 = \frac{1}{4} + 4\alpha, \qquad p = \frac{1}{4} + 2\alpha.$$

Under P_α, choose r, u independently and uniformly on Γ_L, choose $b \sim \mathrm{Bernoulli}(p)$ independently, and choose $\ell = q_0, q_1$ with probabilities $1/2, 1/2$ independently of these coordinates. Give the example label -1 when $\ell = q_0$ and label $+1$ when $\ell = q_1$. Assign zero mass to x_*. This distribution has finite support.

For every $s \in \mathbb{Q}_\circ$, exactly $q_i L$ of the grid points r satisfy $s \in [r, r + q_i)$, since $q_i L$ is an integer and the arcs are half-open. Hence every $h_{s,e} \in \mathcal{H}$ has risk

$$c_0 := \frac{1}{2} q_0 + \frac{1}{2}(1 - q_1) = \frac{1}{2} - 2\alpha = \mathrm{OPT}_{\mathcal{H}}(P_\alpha). \tag{162}$$

Almost surely on a sample from P_α, all empirical endpoints and the anchor belong to Γ_L. Thus the selected shift belongs to Γ_L. Moreover, adding any k/L modulo one to every sample coordinate r and to every ancillary coordinate u preserves the sample law. It also adds k/L to the selected shift: empirical losses translate by this amount, and the order rooted at the anchor translates with them. The law of the selected shift is therefore invariant under every rotation of Γ_L, and is exactly uniform on that grid.

Almost surely the sample contains no x_*. Thus the globally minimizing shifts are the same for both values of e, so the rule selects $e = b$ from the first example. The selected shift depends only on the effective sample $(r_i, \ell_i, y_i)_i$ and on the first anchor u, and is independent of this b. It follows that e is an independent $\mathrm{Bernoulli}(p)$ variable. These conclusions hold for every positive sample size, and applying a deterministic rule to independent sample blocks gives independent outputs. □

N.2 Weights and the independent nullpoint score

Fix the distribution and ensemble from Lemma N.2. Write the independent output parameters as $(s_t, e_t)_{t=1}^T$. For $r \in \Gamma_L$, put

$$S_i(r) = \sum_{t=1}^{T} \mathbf{1}\{s_t \in [r, r+q_i) \pmod 1\} \quad (i = 0, 1), \qquad D_j = \frac{1}{L} \sum_{r \in \Gamma_L} \mathbf{1}\{S_0(r) < j \le S_1(r)\}.$$

The two arcs have the same initial point and $q_0 < q_1$, so $S_0(r) \le S_1(r)$. If R_j denotes the true risk of integer threshold j, then

$$R_j = \frac{1}{2}(1 - D_j), \qquad \sum_{j=1}^{T} D_j = \frac{1}{L} \sum_r (S_1(r) - S_0(r)) = 4\alpha T. \tag{163}$$

The last equality holds for every ensemble, because each shift covers exactly $q_i L$ grid points at length q_i.

Use a validation sample of size n independent of all training blocks, and let r_j be its empirical threshold risks. Distinguish the empirical weights from the oracle weights by writing

$$\hat{c} = \frac{1}{T} \sum_j r_j, \qquad w_j = \frac{1}{2}(\hat{c} + a - r_j)_+, \qquad W = \sum_j w_j,$$

$$w_j^\circ = \frac{1}{2}(c_0 + a - R_j)_+ = \frac{1}{4}(D_j - 4\alpha + 2a)_+, \qquad W^\circ = \sum_j w_j^\circ.$$

Since $a < 2\alpha$, (163) gives

$$\frac{aT}{2} \le W^\circ \le \alpha T, \qquad W \ge \frac{aT}{2}. \tag{164}$$

For the empirical lower bound, use $(v)_+ \ge v$ and $\hat{c} = T^{-1}\sum_j r_j$. For the oracle upper bound, use $w_j^\circ \le D_j/4$.

Put

$$E_n = \frac{1}{T}\sum_j |r_j - R_j|.$$

Since $T^{-1}\sum_j R_j = c_0$, positive-part Lipschitzness yields

$$|\hat{c} - c_0| \le E_n, \qquad \frac{1}{T}\sum_j |w_j - w_j^\circ| \le E_n, \qquad \frac{W}{T} \le \alpha + E_n. \tag{165}$$

Conditionally on the ensemble, each r_j is an average of n Bernoulli losses. Therefore

$$\mathbb{E}[E_n \mid (s_t)_{t=1}^T] \le \frac{1}{2\sqrt{n}}. \tag{166}$$

The vector $(w_j)_{j=1}^T$ is a function of the support restrictions of the ensemble and the independent validation sample. The latter contains no x_* almost surely, and these support restrictions depend only on $(s_t)_{t=1}^T$. Consequently the nullpoint vote count

$$N := \sum_{t=1}^T e_t \sim \mathrm{Bin}(T, p) \tag{167}$$

is independent of all the empirical weights. Define

$$F_w(m) = \frac{1}{W}\sum_{j=1}^m w_j, \qquad V = F_w(N).$$

The prediction at x_* is positive exactly when $U \le V$, where U is the independent shared midpoint used by the compiler.

N.3 From conditional score spread to grid disagreement

Lemma N.3 (Conditional score spread and the midpoint grid)**.** *Let w be any random nonnegative weight vector with positive total mass, independent of $N \sim \mathrm{Bin}(T, p)$. For $V = F_w(N)$, let V' be an independent copy including fresh weights. Set $Q_j = \mathbb{P}\{N \ge j\}$. Then*

$$\mathbb{E}|V - V'| \ge \mathbb{E}_w \sum_{j=1}^T \frac{w_j}{W}\, 2Q_j(1 - Q_j). \tag{168}$$

If U is an independent uniform midpoint of the K-point grid, shared by the two predictions, then

$$\mathbb{P}\{\mathbf{1}\{U \le V\} \ne \mathbf{1}\{U \le V'\}\} \ge \mathbb{E}|V - V'| - \frac{1}{K}. \tag{169}$$

Proof. First hold w fixed, and take independent counts $N, \widetilde{N}$. As F_w is nondecreasing,

$$|F_w(N) - F_w(\widetilde{N})| = \sum_{j=1}^T \frac{w_j}{W}\mathbf{1}\{\min(N, \widetilde{N}) < j \le \max(N, \widetilde{N})\}.$$

Taking expectations gives the sum on the right side of (168).

For a law ν on $[0, 1]$ with distribution function G_ν, its Gini mean difference is

$$\mathcal{G}(\nu) := \mathbb{E}_{Z, Z' \,\mathrm{iid}\sim \nu}|Z - Z'| = 2\int_0^1 G_\nu(t)(1 - G_\nu(t))\, dt.$$

The function $v \mapsto v(1-v)$ is concave. The unconditional distribution function is the average of the conditional distribution functions given w, so $\mathcal{G}(\mathcal{L}(V)) \geq \mathbb{E}_w \mathcal{G}(\mathcal{L}(V \mid w))$. This proves (168) for two completely independent executions, even though the auxiliary conditional calculation used the same weights.

For fixed v, v', disagreement occurs precisely at grid points in $(\min(v,v'), \max(v,v')]$. An interval of length $|v-v'|$ contains at least $K|v-v'|-1$ of the K midpoints. Averaging gives (169). □

Lemma N.4 (Explicit score fluctuation for large ensembles)**.** *In the construction above, if $T, n \geq 16/\alpha^2$, then*

$$\mathbb{E}|V - V'| \geq \frac{1}{512\alpha\sqrt{T}}, \qquad \mathrm{Var}(V) \geq \frac{1}{262144\alpha^2 T}. \tag{170}$$

Proof. Write $\mu = pT$ and $I = [\mu - \alpha T, \mu + \alpha T] \cap \mathbb{Z}$. For each validation observation, mark it bad if its label is negative and $S_0(r) \geq \mu - \alpha T$, or if its label is positive and $S_1(r) < \mu + \alpha T$. Let Z be the fraction of bad validation observations. For every integer j satisfying

$$\mu - \alpha T < j \leq \mu + \alpha T, \tag{171}$$

all validation errors of threshold j are marked bad, so $r_j \leq Z$. For each fixed r, the marginals of $S_i(r)$ are $\mathrm{Bin}(T, q_i)$. The two marking thresholds are at distance αT from their respective means. Hoeffding's inequality, followed by averaging over validation observations, therefore gives

$$\mathbb{E}Z \leq e^{-2\alpha^2 T}, \qquad \mathbb{P}\{Z > 1/8\} \leq 8e^{-2\alpha^2 T}.$$

By (166) and Markov's inequality, $\mathbb{P}\{E_n > \alpha\} \leq 1/8$ when $n \geq 16/\alpha^2$. Conditionally on the ensemble, $\hat{c}$ is an average of n independent $[0,1]$-valued observations with mean $c_0 \geq 3/8$. Hence

$$\mathbb{P}\{\hat{c} < 1/4\} \leq e^{-n/32}.$$

It follows that the event

$$\mathcal{E} = \{Z \leq 1/8,\ E_n \leq \alpha,\ \hat{c} \geq 1/4\}$$

has probability at least $1 - 8e^{-2\alpha^2 T} - 1/8 - e^{-n/32} \geq 1/2$. On this event, every threshold in (171) has $w_j \geq 1/16$, and (165) gives $W \leq 2\alpha T$. Thus

$$F_w(j) - F_w(j-1) \geq \frac{1}{32\alpha T} \quad \text{for every } j \text{ in (171)}. \tag{172}$$

Take independent $N, \widetilde{N} \sim \mathrm{Bin}(T, p)$, independent of all weights, and put $X = N - \widetilde{N}$. Since $p \in [1/4, 3/8]$, writing $s = 2p(1-p) \geq 3/8$ gives

$$\mathbb{E}X^2 = Ts \geq 3T/8, \qquad \mathbb{E}X^4 = Ts + 3T(T-1)s^2 \leq 6(\mathbb{E}X^2)^2.$$

Hölder's inequality implies

$$\mathbb{E}|X| \geq \frac{(\mathbb{E}X^2)^{3/2}}{(\mathbb{E}X^4)^{1/2}} \geq \frac{\sqrt{T}}{4}.$$

Hoeffding and Cauchy–Schwarz also give

$$\mathbb{E}\big[|X|\mathbf{1}\{N \notin I \text{ or } \widetilde{N} \notin I\}\big] \leq \sqrt{T/2}\sqrt{4e^{-2\alpha^2 T}} = \sqrt{2T}\, e^{-\alpha^2 T} \leq \frac{\sqrt{T}}{8}.$$

Whenever both counts belong to I, every increment between them is covered by (172). Consequently, conditionally on any support and validation realization in $\mathcal{E}$,

$$\mathbb{E}_{N,\widetilde{N}}|F_w(N) - F_w(\widetilde{N})| \geq \frac{1}{32\alpha T}\frac{\sqrt{T}}{8} = \frac{1}{256\alpha\sqrt{T}}. \tag{173}$$

The concavity argument in Lemma N.3, together with $\mathbb{P}(\mathcal{E}) \geq 1/2$, proves the first assertion.

For the variance assertion, condition on the support shifts and validation sample. These variables are independent of N, and their realization determines both w and $\mathcal{E}$. Jensen's inequality and (173) give, on $\mathcal{E}$,

$$\mathrm{Var}_N(F_w(N)) = \frac{1}{2}\mathbb{E}_{N,\widetilde{N}}(F_w(N) - F_w(\widetilde{N}))^2 \geq \frac{1}{131072\alpha^2 T}.$$

The law of total variance and $\mathbb{P}(\mathcal{E}) \geq 1/2$ finish the proof. □

N.4 A BOUND FOR ALL ENSEMBLE SIZES

Lemma N.5 (An elementary uniform binomial tail bound)**.** *For every fixed $A_0 \geq 0$, there is a constant $\kappa_{A_0} > 0$ such that, for every integer $m \geq 1$, every $p \in [1/4, 3/8]$, and every integer $j \in \{1, \dots, m\}$ with $|j - mp| \leq A_0\sqrt{m}$, the variable $B_m \sim \mathrm{Bin}(m, p)$ satisfies*

$$2\mathbb{P}\{B_m \geq j\}\mathbb{P}\{B_m < j\} \geq \kappa_{A_0}.$$

Proof. First put $X = B_m - mp$ and $v = mp(1-p)$. Direct expansion gives

$$v \geq 3m/16, \qquad \mathbb{E}X^4 = 3v^2 + v(1 - 6p(1-p)) \leq 3v^2.$$

Hölder's inequality yields $\mathbb{E}|X| \geq \sqrt{m}/4$. Since $\mathbb{E}X = 0$, this gives $\mathbb{E}X_+ = \mathbb{E}(-X)_+ \geq \sqrt{m}/8$. Cauchy–Schwarz implies

$$\mathbb{E}X_+ \leq \frac{\sqrt{m}}{16} + \frac{\sqrt{m}}{2}\sqrt{\mathbb{P}\{X \geq \sqrt{m}/16\}}.$$

Thus $\mathbb{P}\{X \geq \sqrt{m}/16\} \geq 1/64$; applying the same argument to $-X$ gives $\mathbb{P}\{X \leq -\sqrt{m}/16\} \geq 1/64$.

Choose an integer $M \geq 512(A_0 + 1)^2$ and set $\varepsilon = 64^{-M}$. If $m \geq 2M$, divide its Bernoulli trials into M independent blocks, each of length at least $m/(2M)$. With probability at least 64^{-M}, every block has positive centered fluctuation at least the square root of its length divided by 16. The resulting total fluctuation is at least

$$\frac{M}{16}\sqrt{\frac{m}{2M}} \geq (A_0 + 1)\sqrt{m}.$$

The same probability bound holds for a negative fluctuation of this magnitude. It follows that both $\mathbb{P}\{B_m \geq j\}$ and $\mathbb{P}\{B_m < j\}$ are at least ε at every stated j. If $m < 2M$, the all-one and all-zero outcomes instead give both bounds, since $1 \leq j \leq m$ and

$$p^m \geq 4^{-2M} \geq \varepsilon, \qquad (1-p)^m \geq 4^{-2M} \geq \varepsilon.$$

Finally, two complementary probabilities both at least ε have twice their product at least $2\varepsilon(1-\varepsilon) \geq \varepsilon$. Thus one may take $\kappa_{A_0} = \varepsilon$. □

Lemma N.6 (Score spread for smaller ensembles)**.** *In the construction above, if $\alpha\sqrt{T} \leq 4$ and $n \geq 1024/\alpha^2$, then*

$$\mathbb{E}|V - V'| \geq \frac{\kappa_{13}}{2},$$

where $\kappa_{13} > 0$ is the constant from Lemma N.5.

Proof. Let $\mu = pT$ and define

$$\mathcal{J} = \{j \in \{1, \dots, T\} : |j - \mu| \leq 2\alpha T + 4\sqrt{T} + 1\}.$$

The identity $S_0(r) \leq S_1(r)$ and the binomial marginals give

$$\mathbb{E}D_j = \mathbb{P}\{\mathrm{Bin}(T, q_1) \geq j\} - \mathbb{P}\{\mathrm{Bin}(T, q_0) \geq j\} = T\int_{q_0}^{q_1} \mathbb{P}\{\mathrm{Bin}(T-1, q) = j - 1\}\, dq.$$

The last equality follows by differentiating the finite binomial sum; consecutive terms telescope to $T\mathbb{P}\{\mathrm{Bin}(T-1, q) = j-1\}$. For every $q \in [q_0, q_1]$, the variable $J_q = 1 + \mathrm{Bin}(T-1, q)$ satisfies

$$|\mathbb{E}J_q - \mu| = |(T-1)(q-p) + 1 - p| \leq 2\alpha T + 1, \qquad \mathrm{Var}(J_q) \leq T/4.$$

Chebyshev's inequality therefore gives $\mathbb{P}\{J_q \notin \mathcal{J}\} \leq 1/64$. Integrating over q yields

$$\mathbb{E}\sum_{j \notin \mathcal{J}} D_j \leq \frac{4\alpha T}{64}. \tag{174}$$

Use $W \geq aT/2$, $w_j^\circ \leq D_j/4$, and (165) to obtain

$$\mathbb{E}\sum_{j\notin\mathcal{J}}\frac{w_j}{W} \leq \frac{2}{aT}\mathbb{E}\left[\frac{1}{4}\sum_{j\notin\mathcal{J}} D_j + TE_n\right] \leq \frac{\alpha}{32a} + \frac{1}{a\sqrt{n}} \leq \frac{1}{4}+\frac{1}{4} = \frac{1}{2}.$$

Consequently

$$\mathbb{E}\sum_{j\in\mathcal{J}}\frac{w_j}{W} \geq \frac{1}{2}.$$

For every $j \in \mathcal{J}$, the assumption $\alpha\sqrt{T} \leq 4$ implies

$$|j - pT| \leq 12\sqrt{T} + 1 \leq 13\sqrt{T}.$$

Lemma N.5 thus bounds $2Q_j(1-Q_j)$ below by κ_{13} throughout $\mathcal{J}$. Applying (168) proves the claim. □

Proof of Theorem N.1. The fixed domain, class, learner, and distributions are those of Lemma N.2. All base risks are exactly optimal by (162). The independence needed for Lemma N.3 follows from (167) and the subsequent discussion. If $\alpha\sqrt{T} \leq 4$, use Lemma N.6. Otherwise $T > 16/\alpha^2$, and Lemma N.4 applies. Taking

$$c_* = \min\{\kappa_{13}/2,\ 1/512\} > 0$$

gives, in both cases,

$$\mathbb{E}|V - V'| \geq c_* \min\left\{1, \frac{1}{\alpha\sqrt{T}}\right\}.$$

Equation (169) proves (160). □

Remark N.7 (Oracle weights and larger base samples). The all-T conclusion also holds with true-risk weights and no validation sample. Indeed, the smaller-ensemble proof then has no weight-estimation term, and the expected weight fraction in $\mathcal{J}$ is at least $3/4$. For $T \geq 16/\alpha^2$, the explicit bounds improve to

$$\mathbb{E}|V - V'| \geq \frac{1}{256\alpha\sqrt{T}}, \qquad \mathrm{Var}(V) \geq \frac{1}{65536\alpha^2 T}.$$

To see the improvement, the support fraction on which either central endpoint errs has expectation at most $2e^{-2\alpha^2 T}$. With probability at least $1/2$, it is at most $1/2$, so every central $D_j \geq 1/2$ and $w_j^\circ \geq 1/16$. The deterministic bound $W^\circ \leq \alpha T$ doubles the increment used in (172); the remainder of that proof is unchanged.

The same deterministic empirical-risk minimizer generates the ensemble for every block size, so larger base samples leave the obstruction intact. This is a property of the specified weighted-CDF transformation and base learner. The ordered construction below learns $\mathcal{H}$ at a better rate.

Corollary N.8 (A faster learner for the same fixed class)**.** *The fixed class $\mathcal{H}$ in Theorem N.1 admits, for every $0 < \alpha, \beta, \rho \leq 1/4$, a proper agnostic (α, β)-learner that is ρ-pointwise replicable on every distribution and uses exactly*

$$B = \left\lceil \log_2 \frac{4}{\rho} \right\rceil$$

shared fair bits, with no other randomness, and exactly

$$n = \max\left\{\left\lceil \frac{512}{\rho^2\alpha^2}\right\rceil, n_{\mathrm{uc}}(\mathcal{H}, \alpha/4, \beta)\right\} = O\left(\frac{\rho^{-2} + \log(2/\alpha) + \log(2/\beta)}{\alpha^2}\right)$$

examples. Thus the ensemble obstruction above can be overcome on this same class by changing the learning rule.

Proof. Represent each $s \in \mathbb{Q}_\circ$ in $[0,1)$ and fix the public total order that lists the entire $e = 0$ block before the entire $e = 1$ block, ordering shifts increasingly within each block. We verify the hypothesis of Theorem I.1 with $k = 4$.

For a point $x = (r, \ell, u, b) \neq x_*$, let $f_x(s) = h_{s,0}(x) = h_{s,1}(x)$ and let E_x be the set of nonzero endpoints among r and $r + \ell \pmod 1$, represented in $[0, 1)$. These endpoints are distinct, so $m := |E_x|$ is either one or two. The half-open arc convention means that f_x changes sign exactly at each $t \in E_x$, with the new sign already holding at t. For $e \in \{0, 1\}$ and $t \in E_x$, define the nonempty proper initial segment

$$C_{e,t} = \{h_{s',e'} : e' < e \text{ or } (e' = e \text{ and } s' < t)\}.$$

The strict inequality $s' < t$ makes the corresponding parity change occur at $s = t$, including both endpoints' equality cases.

Take baseline label $f_x(0)$, and use all $2m$ cuts $C_{e,t}$. If m is odd, also use the block cut $C_* = \{h_{s,0} : s \in \mathbb{Q}_\circ\}$. Indeed, the eventual sign as $s \uparrow 1$ in a block is $f_x(0)(-1)^m$, whereas the next block starts with $f_x(0)$. The extra cut is therefore needed exactly when m is odd. Equivalently, at a hypothesis in the second block, the m cuts already crossed in the first block together with this possible extra cut have even total parity. These initial segments give (90) with

$$k_x = 2m + \mathbf{1}\{m \text{ is odd}\} \leq 4.$$

At x_*, take baseline -1 and the single cut C_*. All these cuts are public, nonempty, and proper, and the definitions apply directly to the dense rational order.

The class and domain are countable, and Lemma N.2 gives $\mathrm{VC}(\mathcal{H}) \leq 3$. Theorem I.1 with $k = 4$ now supplies the statistical guarantees and total bit count. For every finite sample and seed, empirical risk within each block is a finite right-continuous step function with rational endpoints, including zero as a cell boundary, so the first nonempty block of the eligible set F has a first eligible rational shift. Thus F has a first member, and the argument of Remark I.2 shows that the theorem's prediction rule equals this member and is proper. Finally, (89) gives the stated sample bound. □