

A Sharp Dynamic Haussler Lemma

Shihao Shao¹ and Dacheng Tao¹

¹College of Computing and Data Science, Nanyang Technological University

{shihao.shao, dacheng.tao}@ntu.edu.sg

Abstract

Haussler’s covering lemma (JCTA 1995) established the sharp $(C/\delta)^d$ metric-entropy bound for binary classes of VC dimension d under normalized ℓ_1 distance. We establish a sharp global dynamic counterpart for bounded real-valued classes in every normalized ℓ_p metric, $1 \leq p < \infty$. For every such p , every integer $d \geq 1$, and every $0 < \varepsilon < 1/4$, we prove the matching bounds

$$\left(\frac{c}{\varepsilon}\right)^{pd} \leq \sup_{\mathcal{F}, T: \text{sfat}_{\varepsilon/32}(\mathcal{F}) \leq d} N_p^{\text{dyn}}(\mathcal{F}, \varepsilon, T) \leq \left(\frac{C}{\varepsilon}\right)^{pd},$$

where $c, C > 0$ are universal constants, independent also of p . The upper bound holds for each class and each horizon, and extends to all $0 < \varepsilon < 1/2$. The cover is a single finite family of deterministic causal experts, fixed independently of the instance sequence, that approximates every target on every sequence to normalized ℓ_p error at most ε . The lower bound already holds for binary classes on a single fixed sequence, even with real-valued covering experts.

For $p = 2$, our theorem resolves the horizon-independent sequential ℓ_2 covering problem of Rakhlin, Sridharan, and Tewari (PTRF 2015) in a stronger global form. For $p = 1$, it yields the sharp $(C/\delta)^{\text{Ldim}(\mathcal{H})}$ global dynamic covering bound for binary classes. This resolves the global approximate covering problem posed by Alon, Ben-Eliezer, Dagan, Moran, Naor, and Yogev (STOC 2021), as well as the dynamic Haussler problem posed by Malliaris and Moran (Trans. AMS 2024). Our proof uses SOA-style predictions with scheduled rank-decreasing version-space updates to control large pointwise errors. Rather than keeping every resulting prediction stream as a standalone expert, we use accurate prefixes that can be extended by exact predictions after identifying a target. Each activated target has many eligible prefixes but needs only one continuation. Shared-parent routing distributes these continuation requests across eligible prefixes, avoiding the worst-case concentration of demand on individual prefixes and eliminating the logarithmic loss.

Contents

1	Introduction	1
1.1	Our Contributions	2
1.2	Historical Context and Related Work	3
2	Technical Overview	5
2.1	Outlier Reduction and Correction Schedules	5
2.2	From One Pool to Many Short Pools	6
2.3	Shared-Parent Routing and Sharp Counting	7
2.4	Simultaneous Accuracy and Global Determinization	9
2.5	The Matching Lower Bound	10
3	Definitions and Main Results	11
3.1	Sequential Fat-Shattering Dimension	11
3.2	Global Dynamic Covering Numbers	12
3.3	Sharp Upper and Lower Bounds	12
3.4	Binary Classes and Sequential Covering	13
4	Correction Schedules and Adaptive Sampling	15
4.1	Separation and Rank-Decreasing Centers	15
4.2	Correction Schedules and Pointwise Covers	17
4.3	Partial Binary Completion and Adaptive Sampling	19
5	Master Histories and Multi-Pool Accuracy	23
5.1	The Analytical Master and Uniform Accuracy	24
5.2	Tolerant Replay and Buffered Corrections	26
5.3	Simultaneous Accuracy Across Pools	28
6	Robust Identification and Eligible Parents	31
6.1	Robust Identification and Local Counting	31
6.2	Parent and Child Experts	33
6.3	Eligibility and Many Compatible Parents	34
7	Shared-Parent Routing and Load Balancing	36
7.1	The Causal Routing Protocol	36
7.2	Routing Probabilities and Load Bounds	37
8	Expert Counting and Global Determinization	39
8.1	Expert Counting and Simultaneous Coverage	39
8.2	Game Determinization and Main Upper Bound	40
9	Matching Lower Bounds	43
9.1	Finite-Field Hyperplanes and Their Dimensions	43
9.2	Fixed-Sequence Packing and Lower Bound	44

1 Introduction

How many predictors are needed to approximate an entire function class? In the classical setting, Haussler’s covering lemma [Hau95] gives a sharp answer for binary classes. For every class $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$ of VC dimension d , every finite instance sequence $x_1, \dots, x_T$, and every $0 < \delta < 1$, there is a set of at most $(C/\delta)^d$ representatives such that each target in $\mathcal{H}$ disagrees with some representative on at most δT positions. The bound is independent of the sequence length, and the exponent is exactly the combinatorial dimension. VC dimension originated in uniform-convergence theory [VC71] and became a central parameter in distribution-free PAC learnability [Blu+89]. The connection between combinatorial dimensions and metric entropy extends to real-valued function classes [MV03] and plays a central role in uniform convergence and generalization [VW96; BM02]. More broadly, metric entropy underlies chaining bounds for stochastic processes [Dud67; Tal05].

The classical cover may be chosen after the entire instance sequence is known. Such hindsight dependence is unsuitable for online prediction, where future instances are unavailable. To eliminate this dependence, one may instead maintain a bounded number of candidate prediction streams, updating their predictions as instances arrive. This leads to the notion of a *causal expert*, whose prediction at time t may depend on the observed prefix $x_1, \dots, x_t$. We ask whether a single finite family of such experts, specified in advance, can approximate every target on every possible input sequence. This global guarantee requires consistency across different sequences: whenever two sequences share a prefix, each expert must make identical predictions along that prefix.

We study this question for bounded real-valued classes, measuring approximation in normalized ℓ_p distance for any $1 \leq p < \infty$. Let $\mathcal{F} \subseteq [0, 1]^{\mathcal{X}}$ be nonempty, and fix a horizon $T \geq 1$. A deterministic causal expert is a tuple $g = (g_t)_{t=1}^T$ with $g_t : \mathcal{X}^t \rightarrow [0, 1]$. A finite family $\mathcal{G}$ is a global dynamic ℓ_p cover at accuracy ε if

$$\forall x_{1:T} \in \mathcal{X}^T \forall f \in \mathcal{F} \exists g \in \mathcal{G} : \left(\frac{1}{T} \sum_{t=1}^T |g_t(x_{1:t}) - f(x_t)|^p \right)^{1/p} \leq \varepsilon. \quad (1)$$

We write $N_p^{\text{dyn}}(\mathcal{F}, \varepsilon, T)$ for the smallest cardinality of such a family. The family may depend on $\mathcal{F}, p, \varepsilon$, and T , but is fixed before the instances are revealed. Its experts receive no target values, and the expert witnessing the approximation may depend on both the target and the entire sequence. Thus *global* refers to a single family that works on all instance sequences; *integral* refers to an actual finite family of experts.

Sequential covering has a substantial history. The Littlestone dimension characterizes the optimal deterministic mistake bound in realizable online learning [Lit88]. Extending this characterization to agnostic online learning, Ben-David, Pál, and Shalev-Shwartz [BPS09] constructed a global exact cover for every binary class of Littlestone dimension d , consisting of at most $\sum_{i=0}^{\min\{d, T\}} \binom{T}{i}$ deterministic causal experts. Allowing a positive error rate suggests that the dependence on T might disappear. In the real-valued setting, Rakhlin, Sridharan, and Tewari [RST10; RST15a; RST15b] developed sequential covering numbers and asked for an ℓ_2 bound independent of the horizon. Related sequential complexity methods also apply to objectives beyond external regret, including calibration and approachability [RST11]. Sequential covering numbers also have direct applications to online prediction. Rakhlin and Sridharan [RS14] established optimal regret rates for online nonparametric regression in terms of sequential entropy. Gaillard and Gerchinovitz [GG15] subsequently developed chaining-based forecasters with entropy-dependent regret guarantees. The global approximate covering question for binary classes was posed in [Alo+21], and Malliaris and Moran [MM24] explicitly asked for a dynamic analogue of Haussler’s lemma. Horizon-independent fractional covering bounds were established for global binary covering by Alon, Ben-Eliezer, Dagan,

Moran, Naor, and Yogev [Alo+21] and for tree-dependent real-valued sequential covering by Block, Dagan, and Rakhlin [BDR21]. The former provides one distribution over causal experts, whereas the latter allows the distribution to depend on the prescribed instance tree. Haussler’s lemma, however, guarantees an actual finite family of representatives. A sharp, horizon-independent integral dynamic analogue, together with its extension to real-valued classes, has remained open.

1.1 Our Contributions

We establish a sharp global dynamic covering theorem for bounded real-valued classes, uniformly over all finite $p \geq 1$. The controlling parameter is the *sequential fat-shattering dimension* $\text{sfat}_\eta(\mathcal{F})$, the scale-sensitive sequential analogue of VC dimension. Informally, it is the maximum depth of a complete binary instance tree on which every path can be realized by a function in $\mathcal{F}$, with the two alternatives at each node separated by a gap η around a node-dependent threshold. For binary classes, this dimension equals Littlestone dimension at every scale $0 < \eta \leq 1$. The precise definition is given in the preliminaries.

Theorem 1 (Sharp global dynamic ℓ_p covering). *There are universal constants $c > 0$ and $C \geq 1$ such that, for every $1 \leq p < \infty$, every integer $d \geq 0$, and every $0 < \varepsilon < 1/4$,*

$$\left(\frac{c}{\varepsilon}\right)^{pd} \leq \sup_{\mathcal{X}, T \geq 1, \emptyset \neq \mathcal{F} \subseteq [0,1]^\mathcal{X}: \text{sfat}_{\varepsilon/32}(\mathcal{F}) \leq d} N_p^{\text{dyn}}(\mathcal{F}, \varepsilon, T) \leq \left(\frac{C}{\varepsilon}\right)^{pd}. \quad (2)$$

More precisely, for every $1 \leq p < \infty$, every $T \geq 1$, and every $0 < \varepsilon < 1/2$, each nonempty class $\mathcal{F} \subseteq [0,1]^\mathcal{X}$ with $\text{sfat}_{\varepsilon/32}(\mathcal{F}) < \infty$ satisfies

$$N_p^{\text{dyn}}(\mathcal{F}, \varepsilon, T) \leq \left(\frac{C}{\varepsilon^p}\right)^{\text{sfat}_{\varepsilon/32}(\mathcal{F})}. \quad (3)$$

The constants are independent of p , d , T , and ε . The lower bound is witnessed by binary classes on a single fixed instance sequence, even when the covering experts are allowed real-valued predictions. If the displayed dimension is zero, one expert suffices.

The theorem identifies the exact worst-case accuracy exponent pd , with universal constants in the base. The more precise upper bound (3) retains d as the outer exponent and has a base constant independent also of p . Thus the bounds for $p = 1$ and $p = 2$ are $(C/\varepsilon)^d$ and $(C/\varepsilon^2)^d$, respectively. In every case, the dimension is measured at scale $\varepsilon/32$, corresponding to pointwise accuracy, while the mean p th-power error budget is ε^p .

The lower bound is given by indicators of affine hyperplanes over a finite field. On one fixed instance sequence, their prediction vectors form a normalized ℓ_p packing with separation greater than 2ε . Since the p th power of the normalized ℓ_p distance between binary vectors equals their normalized Hamming distance, the same construction explains the exponent pd for every finite p . The triangle inequality shows that even a real-valued expert can cover at most one vector in this packing.

Sequential covering and the $p = 2$ problem. Restricting every global expert to a fixed instance tree gives a sequential cover on that tree, with the same cardinality. Theorem 1 therefore yields horizon-independent integral sequential ℓ_p covers for every $1 \leq p < \infty$. In particular, its $p = 2$ case resolves the sequential ℓ_2 covering question of Rakhlin, Sridharan, and Tewari [RST15b] in a stronger *global* form: a single family of causal experts, fixed independently of the instance sequence, suffices to approximate every target on every possible sequence. It supplies an integral covering theorem alongside the fractional result of Block, Dagan, and Rakhlin [BDR21].

Binary classes and the $p = 1$ case. For a binary class $\mathcal{H}$ of Littlestone dimension d , the $p = 1$ case yields a family of at most $(C'/\delta)^d$ *binary* causal experts such that

$$\forall x_{1:T} \in \mathcal{X}^T \forall h \in \mathcal{H} \exists g \in \mathcal{G} : \sum_{t=1}^T \mathbf{1}\{g_t(x_{1:t}) \neq h(x_t)\} \leq \delta T, \quad (4)$$

for every $0 < \delta < 1/4$. Indeed, apply (3) with $p = 1$ and $\varepsilon = \delta/2$, then round each prediction at $1/2$. Each binary mistake contributes at least $1/2$ to the original absolute error, so the number of mistakes is at most δT . The finite-field construction gives a matching worst-case lower bound $(c'/\delta)^d$. This settles the global dynamic covering question posed in [Alo+21] and the dynamic Haussler problem of Malliaris and Moran [MM24]. A global cover also supplies a finite reference family for prediction-with-expert-advice methods [CL06], although our theorem does not address efficient construction of the experts.

Proof ideas. Our proof first establishes an outlier cover of size $(C_0 R)^d$ for a uniformly separated class, with pointwise error at least 8η on fewer than $T/(8R)$ rounds. Setting $\eta = \varepsilon/32$, passing to a maximal uniformly η -separated subclass of $\mathcal{F}$, and choosing $R = 16/\varepsilon^p$ then yields mean p th-power error at most $(9\varepsilon/32)^p + \varepsilon^p/128 < \varepsilon^p$ for the original class. This gives the bound $(16C_0/\varepsilon^p)^d$ for every finite p .

To obtain the outlier guarantee, SOA-style predictions and correction schedules implement rank-decreasing version-space updates without target labels. A common master and tolerant replays across short sampling pools give each target many accurate histories. A pool *identifies* a function when a robust certificate isolates it in a singleton version space; the function *activates* once identified by at least half the pools. Bounds on the number of functions each pool can certify within a low-rank version space, combined with this activation threshold, limit how many activated targets can be eligible for any given parent.

Rather than retaining every full schedule as an expert, we use parent schedules to cover unactivated targets or provide accurate prefixes for activated ones. Each activated target has many eligible parents but needs only one exact continuation. Shared-parent routing spreads these requests, avoiding worst-case concentration on individual parents and reducing their leading continuation capacity by a factor proportional to the number of pools. This eliminates the logarithmic overhead. A finite-horizon game argument then yields a deterministic global cover.

1.2 Historical Context and Related Work

Two distinctions clarify the relation between our theorem and earlier covering results: whether the cover may depend on a prescribed instance sequence or tree, and whether covering is witnessed by finitely many representatives or by a probability measure assigning uniformly positive mass to successful approximations.

Static, sequential, and global covers. In Haussler’s classical theorem [Hau95], the representatives may be chosen after the sample is specified. The sequential covers of Rakhlin, Sridharan, and Tewari [RST15b] are instead defined on a fixed binary instance tree $\mathbf{x}$. A cover is a finite family $V_{\mathbf{x}}$ of real-valued prediction trees such that, for every target f and every root-to-leaf path σ , some $\mathbf{v} \in V_{\mathbf{x}}$ approximates the target trace along that path. The family may depend on the entire instance tree, including nodes outside the realized path. The distinction is captured by the

following quantifier orders

$$\begin{aligned} \text{tree-dependent: } & \forall \mathbf{x} \exists V_{\mathbf{x}} \forall f \in \mathcal{F} \forall \sigma \in \{-1, +1\}^T \exists \mathbf{v} \in V_{\mathbf{x}}, \\ \text{global: } & \exists \mathcal{G} \forall x_{1:T} \in \mathcal{X}^T \forall f \in \mathcal{F} \exists g \in \mathcal{G}. \end{aligned} \tag{5}$$

In the first line, accuracy is measured on the path σ of $\mathbf{x}$; the second line has the accuracy requirement in (1). In both cases the witnessing member may depend on the target and the complete path or sequence. The global formulation additionally specifies consistent predictions on every shared instance prefix, across all input sequences. Restricting a global family to a tree gives a sequential cover of the same size. Taking a supremum of sequential covering numbers over trees does not, by itself, interchange the first two quantifiers.

Globality is already present in the exact binary expert construction of Ben-David, Pál, and Shalev-Shwartz [BPS09], whose size grows with T , and in the fractional dynamic covers of [Alo+21]. The global approximate covering questions in [Alo+21; MM24] ask for a horizon-independent finite family. Our theorem supplies such a family with the sharp dimension and accuracy dependence, and extends the guarantee to bounded real-valued classes.

These questions also connect online learning to model-theoretic stability. Malliaris and Shelah [MS14] established regularity lemmas for stable graphs, while Malliaris and Moran [MM24] subsequently studied excellent sets through agnostic online learning.

Real-valued metric entropy. Scale-sensitive dimensions have long been used to control static covering numbers for real-valued function classes, both on finite samples and under distribution-dependent metrics [Alo+97; BKP97; MV03]. Empirical entropy also yields oracle inequalities and risk bounds for nonparametric regression [RST17]. Recently, Aiyer, Mansour, Moran, Shao, and Waknine [Aiy+26] obtained sharp results for empirical ℓ_∞ covering numbers at optimal scales. In these settings, covering families may depend on the sample or probability measure defining the metric. Our theorem instead provides a single finite family of causal experts, with sharp global dynamic bounds in every finite ℓ_p metric.

Geometric refinements. Haussler’s static packing bound has also been refined for shallow geometric set systems [DEG16]. These refinements exploit additional structure of static set systems, whereas our focus is global causal covering.

Fractional and integral covering. For a target–sequence pair $(f, x_{1:T})$, let $B_\varepsilon(f, x_{1:T})$ be the set of causal experts satisfying the error bound in (1) for that pair. A global fractional cover of size $\Gamma \geq 1$ is a probability measure μ , fixed independently of the input sequence, such that

$$\forall x_{1:T} \in \mathcal{X}^T \forall f \in \mathcal{F} : \quad \mu(B_\varepsilon(f, x_{1:T})) \geq \Gamma^{-1}. \tag{6}$$

This is the probability-mass formulation used for binary dynamic covers in [Alo+21]. A tree-dependent fractional cover instead chooses a measure $\mu_{\mathbf{x}}$ on prediction trees for each fixed $\mathbf{x}$, and imposes the same mass condition for every target and path on that tree, as in Block, Dagan, and Rakhlin [BDR21].

An integral cover requires an actual finite family $\mathcal{G}$ intersecting every set $B_\varepsilon(f, x_{1:T})$. The uniform distribution on a family of size K satisfies (6) with $\Gamma = K$. The converse is not immediate: Γ controls the reciprocal of the mass of each successful set, rather than the cardinality of the support of μ . In particular, even a finitely supported fractional cover need not have support of size at most Γ . Here *integral* concerns the selection of representatives; the representatives themselves may make real-valued predictions.

Horizon-independent fractional covering bounds are already known. For binary classes, [Alo+21] establishes a global fractional bound of $(C/\delta)^d$ at normalized Hamming accuracy δ , while Block, Dagan, and Rakhlin [BDR21] obtain real-valued sequential ℓ_2 fractional bounds controlled by sequential fat-shattering dimension. These results, however, do not directly yield finite global integral covers. The difficulty is simultaneous coverage. Sampling m independent experts from μ misses a fixed target–sequence pair with probability at most $(1 - 1/\Gamma)^m \leq e^{-m/\Gamma}$. For M prescribed pairs, a union bound gives simultaneous coverage with probability at least $1 - \zeta$ using $m = \lceil \Gamma \log(M/\zeta) \rceil$ samples, where $0 < \zeta < 1$. Thus, for any prescribed finite collection of M target–sequence pairs, there exists a family of m experts covering all of them. However, this argument introduces a logarithmic factor and does not directly handle the possibly infinite collection of all target–sequence pairs. On a fixed static sample, the classical packing argument converts fractional mass into a finite cover after doubling the radius. In the sequential setting, however, maximal packings on individual paths need not yield a common covering family, while packing uniformly across paths imposes a stronger requirement that is not controlled by the fractional covering bound. Block, Dagan, and Rakhlin [BDR21] discuss this failure of the classical packing–covering argument.

Our proof instead uses shared-parent routing to share continuation capacity across compatible sampled histories. This yields a sharp bound on the global integral covering number in terms of sequential fat-shattering dimension, without relying on a general fractional-to-integral conversion principle.

2 Technical Overview

We follow the standard optimal algorithm (SOA) principle: predict a center of the current *version space*, the functions surviving earlier restrictions. A fixed *correction schedule* prescribes when and in which direction to restrict this space; the candidate streams need no observed target labels. We restrict updates to short randomly sampled lists of rounds, called *pools*. Multiple pools give each target many accurate histories, but an activated target needs only one accurate past followed by exact predictions. Our *shared-parent routing* distributes these one-per-target continuations among eligible histories, reducing the leading capacity required per history by a factor of the number of pools. This cancels the multiplicative cost of maintaining multiple pools, removing the remaining logarithmic loss. A final game argument makes the family deterministic and global.

2.1 Outlier Reduction and Correction Schedules

We first control the number of rounds with large pointwise error. Call such a round an *outlier*, and write $\|h - h'\|_\infty = \sup_{x \in \mathcal{X}} |h(x) - h'(x)|$.

Proposition 2 (Global outlier cover). *There is a universal constant $C_0 \geq 1$ with the following property. Let $d \geq 0$ be an integer, $0 < \eta < 1$, $R > 32$, $T \geq 1$, and let $\emptyset \neq \mathcal{H} \subseteq [0, 1]^\mathcal{X}$ satisfy $\text{sfat}_\eta(\mathcal{H}) \leq d < \infty$ and $\|h - h'\|_\infty > \eta$ for distinct $h, h' \in \mathcal{H}$. There is a family $\mathcal{G}$ of at most $(C_0 R)^d$ deterministic causal experts such that, for every instance sequence $x_{1:T}$ and every $h \in \mathcal{H}$, some $g \in \mathcal{G}$ satisfies*

$$\#\{t \in [T] : |g_t(x_{1:t}) - h(x_t)| \geq 8\eta\} < \frac{T}{8R}. \quad (7)$$

The family is fixed independently of the instance sequence.

To obtain the ℓ_p theorem, set $\eta = \varepsilon/32$ and $d = \text{sfat}_\eta(\mathcal{F})$. Choose a maximal uniformly greater-than- η separated subclass $\mathcal{H} \subseteq \mathcal{F}$: distinct members have uniform distance greater than η , and

maximality makes every $f \in \mathcal{F}$ lie within uniform distance η of some $h \in \mathcal{H}$. Its dimension is at most d . Applying Proposition 2 with $R = 16/\varepsilon^p$ gives

$$\frac{1}{T} \sum_{t=1}^T |g_t(x_{1:t}) - f(x_t)|^p \leq (9\eta)^p + \frac{1}{8R} = \varepsilon^p \left[\left(\frac{9}{32} \right)^p + \frac{1}{128} \right] < \varepsilon^p. \quad (8)$$

Here the error is below 9η outside the outliers and at most one everywhere. The family has size at most $(16C_0/\varepsilon^p)^d$. Thus the ℓ_p upper bound reduces to Proposition 2, with p entering only through the outlier budget.

Corrections that decrease dimension. For a version space $V \subseteq \mathcal{H}$ consisting of functions remaining after restrictions, define its *rank* as $\text{sfat}_\eta(V)$ (rank -1 if empty). Distinct functions in the separated class would shatter a depth-one tree; thus rank zero means V is a singleton, which predicts exactly. Assume $d \geq 1$ and $|\mathcal{H}| \geq 2$ henceforth.

For a nonempty V of rank r and current instance x , choose a center $a(V, x) \in [0, 1]$ such that both sign-update classes

$$V_x^- := \{f \in V : f(x) \leq a(V, x) - \eta\}, \quad V_x^+ := \{f \in V : f(x) \geq a(V, x) + \eta\} \quad (9)$$

have rank at most $r - 1$. Such a center exists because two rank- r subclasses separated by a gap η at x would form a shattered tree of depth $r + 1$; Lemma 12 gives the precise choice. For a target h , a reference procedure starts from $V = \mathcal{H}$. On each pool-selected round, while $|a(V, x) - h(x)| > \eta$, it restricts V to the sign class containing h and recomputes the center. Each update keeps h and decreases rank, so at most d updates occur; after them, selected-round error is at most η . Unselected rounds cause no update.

Encoding updates without target values. Experts receive no target values, so we enumerate all correction lists. On a pool of at most m selected rounds, a schedule is

$$J = ((s_1, \sigma_1), \dots, (s_i, \sigma_i)), \quad 1 \leq s_1 \leq \dots \leq s_i \leq m, \quad \sigma_\ell \in \{-, +\}, \quad i \leq d.$$

The depth i counts corrections; an *ordinal* s_ℓ is a position on the selected list, not in $[T]$. Repeated ordinals allow successive corrections at one instance. The empty schedule is allowed; an invalid update produces zero thereafter. A schedule updates before predicting. For each target and realized input-and-selection history, recording the correction times and signs of the target-dependent reference procedure yields a witnessing schedule in this fixed family. Only ordinals and signs are encoded: no expert reads target values or discretizes them.

There are $n_i(m) = 2^i \binom{m+i-1}{i}$ depth- i schedules for $i \geq 1$, and $n_0(m) = 1$. Their total through depth r satisfies

$$Q_r(m) := \sum_{i=0}^r n_i(m) \leq \left[2e \left(1 + \frac{m}{r} \right) \right]^r \quad (r \geq 1). \quad (10)$$

For a fixed sampling outcome, these streams use only the observed instance-and-selection history. Selecting every round yields a deterministic pointwise cover of size $Q_d(T)$, sufficient for $T = O(dR)$.

2.2 From One Pool to Many Short Pools

Put $b = T/R$; we permit fewer than $b/8$ outliers. A Bernoulli- γ pool selects each round independently with probability γ , after x_t is chosen. Inputs may depend on past random choices, but not on the current sampling bits.

The cost of a direct single-pool analysis. A pool with $\gamma b = \Theta(d)$ and cap $m = \lceil 2\gamma T \rceil = \Theta(dR)$ admits only $Q_d(m) \leq (CR)^d$ schedules. But accuracy on selected rounds does not control errors on unselected rounds. We need a single sampling outcome on which, simultaneously for every target, some schedule has few outliers over the entire horizon.

Lemma 17 bounds the probability that many errors occur but few are selected, uniformly over targets. Applying it to each fixed schedule and taking a union bound gives, for $\gamma \leq 1/2$ and γb a sufficiently large multiple of d , a principal term of the form

$$\exp\left\{O\left(d \log\left(1 + \frac{\gamma T}{d}\right)\right) - c\gamma b\right\}, \quad (11)$$

plus exponentially small terms accounting for excessive sample lengths. With $\gamma b = \Theta(d)$, the displayed bound exceeds 1 as R grows and thus fails to establish simultaneous accuracy. Taking $\gamma b = Ad \log R$ for a sufficiently large constant A makes it useful, but increases the schedule count to $(CR \log R)^d$. Selecting every round ($m := T$) handles the range where the required rate exceeds $1/2$. This direct analysis thus leaves an extra factor of $(\log R)^d$.

Several short pools and a common reference. Fix a sufficiently large universal α . The all-round cover handles $b \leq 32\alpha d$; otherwise set

$$k = 4\lceil \log_2 R \rceil, \quad \beta = \frac{\alpha d}{b} < \frac{1}{32}, \quad m = \lceil 2\beta T \rceil = \Theta(dR). \quad (12)$$

Use independent Bernoulli- β bits $\xi_{t,j}$ across rounds $t \in [T]$ and pools $j \in [k]$. Each pool retains only its first m selections. We call the uncapped lists *raw* pools; their independent bits are used in the probability estimates. Let **Cap** be the event that no raw list exceeds m ; on this event, raw and capped simulations agree.

To exploit all pools without lengthening their codes, for each h run a target-dependent reference procedure, its *master*, on the union of raw selections if $k\beta \leq 1/8$, and on every round otherwise. This larger sample lets us control the master's errors uniformly over targets. Every pool selection is also a master selection, so the master's final value y_t^h is within η of $h(x_t)$ on those rounds. For each target h , the analytical master provides a common reference prediction history for controlling the errors of its pool replays.

A *replay* in pool j maintains its own version space. On that pool's selected rounds it repeatedly corrects toward y_t^h until its center is within 4η , using the same sign updates. For example, $y_t^h > a + 4\eta$ implies $h(x_t) > a + 3\eta$, whereas the retained constraint is only $f(x_t) \geq a + \eta$. Thus h lies more than 2η beyond each retained inequality's boundary; this distance is the constraint's *slack*. There are at most d corrections, and selected-round error is at most 5η . Replays are analytical witnesses, not additional experts. On **Cap**, each is reproduced by a formal schedule on its capped pool.

Proposition 21 shows that, with probability above $15/16$, **Cap** holds and every target has at least $3k/4$ replays with fewer than $b/8$ outliers of size at least 8η over the whole horizon; see Section 2.4. Retaining every pool's full schedules would cost $K_{\text{naive}} = kQ_d(m) \leq (CR)^d \log R$ experts. To eliminate this remaining factor of k , we exploit the redundancy of accurate histories across pools.

2.3 Shared-Parent Routing and Sharp Counting

Identifying a function permits exact continuation. Once a schedule's version space is $\{h\}$, the function h is known and future predictions can equal $h(x_t)$ exactly; this is a *continuation*. It cannot retroactively repair earlier predictions, so it must extend a stream with an accurate past.

A schedule is *born* once all its prescribed updates have occurred; its version space is fixed thereafter. A *robust certificate* for h is a valid born schedule whose version space is $\{h\}$ and whose update constraints hold for h with slack at least 2η , as in the replay inequality above. This margin will make different functions certified in one pool differ by more than 2η on some selected instance, allowing us to bound how many functions the pool can identify. A pool *identifies* h when it has such a certificate; h *activates* once at least $k/2$ pools identify it. Identification concerns a candidate function, not the unknown true target: its past values can be checked without labels.

Parents preserve the past; children supply continuations. For every pool and every schedule of depth at most $d-1$, retain its prediction stream as a *parent*. There are $P = kQ_{d-1}(m)$ parents. After a depth- i parent v is born valid, its version space V_v is fixed and has rank at most $d-i$. Preallocate c_i additional streams, called *children*, to every depth- i parent. Each child copies its parent's outputs until assigned to an activated function h , then predicts h exactly from the assignment round onward. All children exist from the start, so assignment never changes their past outputs.

At the activation of h on round t , a parent is *eligible* if it is valid and born by $t-1$, and

$$h \in V_v, \quad |V_v| \geq 2, \quad \sum_{s < t} \mathbf{1}\{|p_s^v - h(x_s)| \geq 8\eta\} < b/8,$$

where p_s^v is its actual output. Retain one eligible parent per pool in a fixed order, and call the resulting set $\mathcal{E}(h, t)$. Eligibility is testable at activation using only known past outputs and the now-identified function, not future accuracy.

On the simultaneous-accuracy event, at least $3k/4$ replays for h are accurate, while fewer than $k/2$ pools identified it by $t-1$. Thus more than $k/4$ accurate replay prefixes still have nonsingleton version spaces: a singleton replay prefix would already be a robust certificate. Such a space has rank at least one, so its prefix uses at most $d-1$ corrections and coincides with a preallocated parent. Hence $|\mathcal{E}(h, t)| \geq k/4$; any of these parents has an accurate past that a child can extend by predicting h exactly. If h never activates, an accurate unidentified pool supplies a full replay of depth at most $d-1$, already among the parents. Depth- d schedules serve identification but are not retained.

Robust certificates bound the demand at each parent. Two distinct functions identified in one pool must differ by more than 2η on a selected instance: one violates a constraint of the other's singleton certificate, whereas the certified function satisfies that constraint with slack at least 2η . For any rank- r subclass V , the pointwise radius- η schedule cover on that selected list has size $Q_r(m)$. No covering vector can approximate two such identified functions, so the pool identifies at most $Q_r(m)$ members of V .

Let $\mathcal{A}$ be the set of activated functions. Each is identified in at least $k/2$ pools, and hence

$$\frac{k}{2}|\mathcal{A} \cap V| \leq kQ_r(m), \quad |\mathcal{A} \cap V| \leq 2Q_r(m). \quad (13)$$

A depth- i parent can therefore be eligible for at most $2Q_{d-i}(m)$ activations. This bound holds on every outcome, even for version spaces produced during the run, and ensures finitely many activations when $\mathcal{H}$ is infinite.

Routing balances the continuation requests. Without routing, a depth- i parent might need $2Q_{d-i}(m)$ children: all eligible activations could choose it. Instead, each activation with at least

$k/4$ eligible parents chooses one uniformly using fresh randomness; an activation with fewer is left unassigned. Any given parent is then chosen with conditional probability at most $4/k$ per request, and its expected load is at most $8Q_{d-i}(m)/k$. Full parents remain eligible to preserve the $4/k$ selection bound, while a separate no-overload event controls capacity violations. An exponential tail bound, followed by a union bound over all P parents, justifies capacities

$$c_i = \left\lceil \frac{16Q_{d-i}(m)}{k} + \log(4P) \right\rceil, \quad 0 \leq i \leq d-1. \quad (14)$$

The logarithm covers simultaneous load fluctuations, and the probability of any overload is at most $1/4$ even with adaptive inputs, without conditioning on the accuracy event.

There are $kn_i(m)$ depth- i parents. Counting all parents and children, including invalid schedules and unused streams, gives

$$\begin{aligned} K &= P + \sum_{i=0}^{d-1} kn_i(m)c_i \\ &\leq 16 \sum_{i=0}^{d-1} n_i(m)Q_{d-i}(m) + P(\log(4P) + 2) \\ &\leq 16Q_d(2m) + P(\log(4P) + 2). \end{aligned} \quad (15)$$

Thus the k parents-per-schedule factor cancels with the $1/k$ capacity per parent. The final inequality follows by distributing at most d signed corrections between two lists of length m (Lemma 14). No child is assigned to two targets; the saving comes from choosing among many eligible parents. Since $m = O(dR)$, the first term is at most $(CR)^d$. Parents have depth at most $d-1$, so their overhead is at most $Cd(\log R)^2(CR)^{d-1}$. The saved power of R absorbs the logarithmic factors, since $(\log R)^2 \leq R$; the factor $d \leq 2^d$ is absorbed into the universal base constant. Thus the overhead is at most $(C'R)^d$. Overall, the counts improve as

$$\underbrace{(CR \log R)^d}_{\text{one long pool}} \longrightarrow \underbrace{(CR)^d \log R}_{\text{many short pools}} \longrightarrow \underbrace{(CR)^d}_{\text{shared-parent routing}}.$$

2.4 Simultaneous Accuracy and Global Determinization

We control four events: **Cap** bounds pool lengths; **Master** bounds each target's large master errors; **Joint** excludes too many replays with frequent master-replay disagreements but few sampled disagreements; and **NoOverload** prevents routing beyond child capacities. All four hold together with positive probability.

Cap: bounded sampling. A Chernoff bound makes **Cap** highly probable. On this event, every target's master code lies in a fixed family of $Q_d(4km)$ codes. We use the independent raw bits without conditioning on **Cap**.

Master: uniform master accuracy. For fixed master code J , let a_t^J denote its *provisional prediction*, computed before the current pool bits are revealed. Lemma 17 bounds, uniformly over targets, the probability of many master errors of size at least 2η but few sampled errors exceeding η , in either direction. In the union regime, $u = 1 - (1 - \beta)^k = \Theta(k\beta)$ and $ub = \Theta(\alpha kd)$ make this bound small enough for a union bound over fixed codes and both directions. Outside these union-bound failures, the d -correction bound implies **Master** on **Cap**: every target has fewer than

$b/32$ such provisional errors. In the all-round regime, each such error forces a correction, so **Master** holds directly.

Joint: simultaneous replay accuracy. Let $a_{t,j}^J$ denote pool j 's provisional replay prediction. Set $D_j^J := \sum_t \mathbf{1}\{|a_t^J - a_{t,j}^J| > 6\eta\}$ and $H_j^J := \sum_t \xi_{t,j} \mathbf{1}\{|a_t^J - a_{t,j}^J| > 6\eta\}$, counting all and sampled disagreements. Their indicators are fixed before current independent pool bits. For a fixed J and any $k/4$ pools, a joint exponential supermartingale bounds the probability that all have $D_j^J \geq b/16$ and $H_j^J \leq 2d$ by $e^{-\alpha kd/256}$. A union bound over $Q_d(4km) \binom{k}{k/4}$ code-subset pairs yields **Joint** with high probability; the replay histories need not be independent.

For a target's realized code J^h on **Cap**, every sampled disagreement forces a master or replay correction, so $H_j^{J^h} \leq 2d$. Hence **Joint** gives at least $3k/4$ pools with $D_j^{J^h} < b/16$. Together with **Master**, each has fewer than $b/32 + b/16 < b/8$ provisional errors of size at least 8η : each is a large master error or disagreement. Final predictions obey the same bound, since selected-round errors are at most 5η and unselected predictions are unchanged.

NoOverload: successful routing. The load bound gives $\Pr(\text{NoOverload}^c) \leq 1/4$. For sufficiently large α , the first three events hold with probability above $15/16$, providing eligible parents for activated targets and covering parents for unactivated ones. On the intersection of these three events with **NoOverload**, each activation receives an unused child. A union bound then gives simultaneous coverage with probability greater than $1 - 1/16 - 1/4 = 11/16$, even against adaptive input rules observing earlier sampling and routing choices.

A global strategy, not a seed for each sequence. Positive success probability against each adaptive input rule does not give one seed for all sequences. Instead, consider a finite-horizon perfect-information game: the adversary chooses x_t , then the constructor chooses sampling and routing actions, both observing past moves. The constructor wins when its K streams cover all targets. If no deterministic constructor strategy always won, backward induction would give an adversary forcing failure against every constructor strategy, contradicting the probability bound. Fixing a winning strategy makes each stream's predictions depend only on $x_{1:t}$, yielding K deterministic causal experts. This proves Proposition 2; (8) gives the general- p upper bound.

2.5 The Matching Lower Bound

For a prime power q , let $\mathbb{F}_q$ be a field with q elements. Take the hyperplane indicators $h_{a,c}(x) = \mathbf{1}\{a \cdot x = c\}$ on $\mathbb{F}_q^d$, where $a \in \mathbb{F}_q^d \setminus \{0\}$ and $c \in \mathbb{F}_q$. The resulting class contains $q(q^d - 1)/(q - 1) \geq q^d$ distinct functions and has Littlestone dimension exactly d . For the upper bound on dimension, predict one precisely on the affine span of previously observed positive instances. This span stays inside the target hyperplane. Each mistake adds a positive point outside the span and increases its affine dimension by one, from -1 for the initially empty span to at most $d - 1$. Thus at most d mistakes occur. For the reverse bound, the d standard basis vectors are shattered.

List all $T = q^d$ domain points once. The normalized Hamming distance between distinct hyperplane indicators is at least $2(q - 1)/q^2$. For binary vectors, this Hamming distance is the p th power of the normalized ℓ_p distance. For $0 < \varepsilon < 1/4$, choose q as the largest power of two not exceeding $(2\varepsilon)^{-p}$. The traces are then pairwise more than 2ε apart, and

$$q^d > \left(\frac{1}{2^{p+1}\varepsilon^p} \right)^d \geq \left(\frac{1}{4\varepsilon} \right)^{pd}.$$

The triangle inequality lets even an arbitrary real-valued covering vector approximate at most one trace. Since binary sequential fat-shattering dimension equals Littlestone dimension at every scale in $(0, 1]$, this proves the matching exponent pd on a single fixed sequence, without using causality or adaptive inputs.

3 Definitions and Main Results

We now formalize the notions used in the preceding overview and state the precise covering theorems and their consequences.

Throughout, $\mathcal{X}$ is a nonempty instance domain, and function classes are nonempty unless stated otherwise. For an integer $T \geq 1$, write $[T] = \{1, \dots, T\}$ and $x_{1:t} = (x_1, \dots, x_t)$. For a sign sequence σ , write $\sigma_{<t} = (\sigma_1, \dots, \sigma_{t-1})$; the prefix at $t = 1$ is empty. The exponent p always lies in $[1, \infty)$. All prediction rules are arbitrary set-theoretic maps; no computability or measurability assumption is imposed.

3.1 Sequential Fat-Shattering Dimension

Fat-shattering extends VC dimension to real-valued function classes [KS94; BLW96]. Its sequential analogue is a central complexity measure in online learning [RST10; RST15a]. An $\mathcal{X}$ -valued binary tree of depth r is a collection $\mathbf{x} = (\mathbf{x}_t)_{t=1}^r$ of maps $\mathbf{x}_t : \{-1, +1\}^{t-1} \rightarrow \mathcal{X}$. A real-valued tree $\mathbf{s} = (\mathbf{s}_t)_{t=1}^r$ is defined in the same way, with range $\mathbb{R}$. Thus the instance and threshold at level t may depend on the preceding signs, but not on the sign selected at that level.

Definition 3 (Sequential fat-shattering dimension). *Let $\eta > 0$ and $\mathcal{F} \subseteq [0, 1]^\mathcal{X}$. The class $\mathcal{F}$ sequentially fat-shatters an instance tree $\mathbf{x}$ of depth r at scale η if there is a real-valued threshold tree $\mathbf{s}$ such that*

$$\forall \sigma \in \{-1, +1\}^r \exists f \in \mathcal{F} \forall t \in [r] : \quad \sigma_t(f(\mathbf{x}_t(\sigma_{<t})) - \mathbf{s}_t(\sigma_{<t})) \geq \frac{\eta}{2}. \quad (16)$$

The sequential fat-shattering dimension $\text{sfat}_\eta(\mathcal{F})$ is the largest such depth, and is ∞ if these depths are unbounded. A nonempty class has dimension at least zero; we set $\text{sfat}_\eta(\emptyset) = -1$.

Our convention uses a margin $\eta/2$ on each side of a threshold, so the two branches are separated by a gap η . Instances may repeat, including along a single path. The dimension is monotone under inclusion of classes and nonincreasing in the scale: if $\mathcal{F}_0 \subseteq \mathcal{F}$ and $0 < \eta \leq \eta'$, then

$$\text{sfat}_{\eta'}(\mathcal{F}_0) \leq \text{sfat}_\eta(\mathcal{F}).$$

For a binary class $\mathcal{H} \subseteq \{0, 1\}^\mathcal{X}$, the *Littlestone dimension* $\text{Ldim}(\mathcal{H})$ is the largest depth of an instance tree on which every sign path σ has a witness $h \in \mathcal{H}$ satisfying $h(\mathbf{x}_t(\sigma_{<t})) = (1 + \sigma_t)/2$ at every level. We use the same conventions for infinite dimension and the empty class. For binary classes, sequential fat-shattering dimension coincides with Littlestone dimension:

$$\text{sfat}_\eta(\mathcal{H}) = \text{Ldim}(\mathcal{H}), \quad 0 < \eta \leq 1.$$

Indeed, a Littlestone-shattered tree is η -fat-shattered with thresholds $1/2$, while fat-shattering of a binary class at any positive scale forces the two branches at each node to have labels 0 and 1, respectively.

3.2 Global Dynamic Covering Numbers

For $u, v \in \mathbb{R}^T$, define the normalized ℓ_p distance by

$$\rho_{p,T}(u, v) := \left(\frac{1}{T} \sum_{t=1}^T |u_t - v_t|^p \right)^{1/p}. \quad (17)$$

A *deterministic causal expert* at horizon T is a tuple $g = (g_t)_{t=1}^T$ with $g_t : \mathcal{X}^t \rightarrow [0, 1]$. The prediction $g_t(x_{1:t})$ may use the current instance x_t and all preceding instances. The expert receives no target values. For a fixed sequence, write

$$g[x_{1:T}] := (g_t(x_{1:t}))_{t=1}^T, \quad f[x_{1:T}] := (f(x_t))_{t=1}^T.$$

Definition 4 (Global dynamic ℓ_p cover). *Let $\mathcal{F} \subseteq [0, 1]^\mathcal{X}$, $T \geq 1$, and $\varepsilon > 0$. A finite family $\mathcal{G}$ of deterministic causal experts is a global dynamic ℓ_p cover of $\mathcal{F}$ at accuracy ε and horizon T if*

$$\forall x_{1:T} \in \mathcal{X}^T \quad \forall f \in \mathcal{F} \quad \exists g \in \mathcal{G} : \quad \rho_{p,T}(g[x_{1:T}], f[x_{1:T}]) \leq \varepsilon. \quad (18)$$

The global dynamic covering number $N_p^{\text{dyn}}(\mathcal{F}, \varepsilon, T)$ is the least cardinality of such a finite family, or ∞ if none exists.

The entire family, including its predictions on every possible prefix, is fixed before the sequence is chosen. The witnessing expert may depend on both the target and the complete sequence, but it is a single member of the family used throughout that sequence. There is no requirement that this member be identifiable online. Experts need not arise from functions in $\mathcal{F}$. In particular, *integral* covering means selecting finitely many experts; it imposes no integrality constraint on their predictions.

The family may depend on $\mathcal{F}, p, \varepsilon$, and T . A bound independent of the horizon therefore asserts a uniform bound on its size for all T , while allowing a different family at each horizon.

3.3 Sharp Upper and Lower Bounds

Theorem 5 (Global dynamic ℓ_p upper bound). *There is a universal constant $C \geq 1$ such that, for every $1 \leq p < \infty$, every integer $T \geq 1$, every $0 < \varepsilon < 1/2$, and every nonempty $\mathcal{F} \subseteq [0, 1]^\mathcal{X}$ with $d = \text{sfat}_{\varepsilon/32}(\mathcal{F}) < \infty$,*

$$N_p^{\text{dyn}}(\mathcal{F}, \varepsilon, T) \leq \left(\frac{C}{\varepsilon^p} \right)^d. \quad (19)$$

The constant C is independent of p, d, T, ε , and the class. If $d = 0$, the covering number is exactly one.

The zero-dimensional case follows directly from the definition. If $\text{sfat}_\eta(\mathcal{F}) = 0$, no two values at any one instance can differ by at least η , since they would shatter a depth-one tree. Hence

$$\sup_{f \in \mathcal{F}} f(x) - \inf_{f \in \mathcal{F}} f(x) \leq \eta \quad (x \in \mathcal{X}).$$

The expert that predicts the midpoint of this interval at the current instance has pointwise error at most $\eta/2$ against every target. Taking $\eta = \varepsilon/32$ proves the assertion.

Theorem 6 (Matching ℓ_p lower bound). *For every $1 \leq p < \infty$, every integer $d \geq 1$, and every $0 < \varepsilon < 2^{-1-1/p}$, there are a finite domain $\mathcal{X}$, a binary class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ with $\text{sfat}_{\varepsilon/32}(\mathcal{H}) = d$, and a horizon $T \geq 1$ such that*

$$N_p^{\text{dyn}}(\mathcal{H}, \varepsilon, T) > \left(\frac{1}{2^{p+1}\varepsilon^p} \right)^d. \quad (20)$$

The lower bound holds even when covering is required only on one fixed instance sequence and the covering predictions are arbitrary vectors in $[0,1]^T$.

The construction uses indicators of affine hyperplanes in $\mathbb{F}_q^d$ and lists the points of this domain once. For a suitable power of two q , the class has Littlestone dimension exactly d and at least q^d members, whose traces are pairwise more than 2ε apart in $\rho_{p,T}$. Each covering vector can therefore approximate at most one trace. The class and the sequence may depend on p, d , and ε ; the lower bound is a worst-case assertion, not a lower bound for every class of dimension d .

For a nonnegative integer d , define the worst-case quantity

$$\mathfrak{M}_{d,p}(\varepsilon) := \sup_{T \geq 1} \sup_{\substack{\mathcal{X}, \emptyset \neq \mathcal{F} \subseteq [0,1]^{\mathcal{X}}: \\ \text{sfat}_{\varepsilon/32}(\mathcal{F}) \leq d}} N_p^{\text{dyn}}(\mathcal{F}, \varepsilon, T). \quad (21)$$

Theorems 5 and 6 give, for $d \geq 1$ and $0 < \varepsilon < 2^{-1-1/p}$,

$$\left(\frac{1}{2^{p+1}\varepsilon^p} \right)^d < \mathfrak{M}_{d,p}(\varepsilon) \leq \left(\frac{C}{\varepsilon^p} \right)^d. \quad (22)$$

In particular, since $2^{1+1/p} \leq 4$ and $C^{1/p} \leq C$, for every $1 \leq p < \infty$, $d \geq 1$, and $0 < \varepsilon < 1/4$,

$$\left(\frac{1}{4\varepsilon} \right)^{pd} < \mathfrak{M}_{d,p}(\varepsilon) \leq \left(\frac{C}{\varepsilon} \right)^{pd}. \quad (23)$$

For $d = 0$, $\mathfrak{M}_{0,p}(\varepsilon) = 1$ on the same range. Thus (23), with a non-strict lower inequality, also includes $d = 0$ and yields Theorem 1 with $c = 1/4$.

3.4 Binary Classes and Sequential Covering

We record two consequences with their proofs. The first gives binary-valued experts for binary classes; the second restricts the global family to a prescribed instance tree.

Binary Hamming covers. For $u, v \in \{0,1\}^T$, let

$$d_{H,T}(u, v) := \frac{1}{T} \sum_{t=1}^T \mathbf{1}\{u_t \neq v_t\}$$

be normalized Hamming distance. For $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$, define $N_{\text{bin}}^{\text{dyn}}(\mathcal{H}, \delta, T)$ as in Definition 4, with binary-valued experts $g_t : \mathcal{X}^t \rightarrow \{0,1\}$ and with $d_{H,T}(g[x_{1:T}], h[x_{1:T}]) \leq \delta$ in place of the ℓ_p requirement. Requiring binary outputs distinguishes this number from N_1^{dyn} , which permits real outputs.

Corollary 7 (Sharp binary dynamic covering). *There is a universal constant $C_{\text{bin}} \geq 1$ such that, for every binary class $\mathcal{H}$ with $d = \text{Ldim}(\mathcal{H}) < \infty$, every $T \geq 1$, and every $0 < \delta < 1$,*

$$N_{\text{bin}}^{\text{dyn}}(\mathcal{H}, \delta, T) \leq \left(\frac{C_{\text{bin}}}{\delta} \right)^d. \quad (24)$$

Conversely, for every integer $d \geq 1$ and $0 < \delta < 1/4$, some binary class of Littlestone dimension d and some horizon satisfy

$$N_{\text{bin}}^{\text{dyn}}(\mathcal{H}, \delta, T) > \left(\frac{1}{4\delta} \right)^d. \quad (25)$$

Proof. Apply Theorem 5 with $p = 1$ and $\varepsilon = \delta/2$. The sequential fat-shattering dimension at scale $\delta/64$ equals d , so the resulting real-valued family has size at most $(2C/\delta)^d$. Round each prediction to $\bar{g}_t(x_{1:t}) = \mathbf{1}\{g_t(x_{1:t}) \geq 1/2\}$. For a binary target,

$$\mathbf{1}\{\bar{g}_t(x_{1:t}) \neq h(x_t)\} \leq 2|g_t(x_{1:t}) - h(x_t)|.$$

The rounded family is causal and has normalized Hamming error at most δ , proving (24) with $C_{\text{bin}} = 2C$. For the lower bound, apply Theorem 6 with $p = 1$ and $\varepsilon = \delta$, and use $N_{\text{bin}}^{\text{dyn}} \geq N_1^{\text{dyn}}$. ■

In particular, Corollary 7 resolves the global approximate covering question posed by Alon, Ben-Eliezer, Dagan, Moran, Naor, and Yogev [Alo+21] and the dynamic Haussler problem of Malliaris and Moran [MM24]. For binary vectors, $\rho_{p,T}(u, v)^p = d_{H,T}(u, v)$. For $0 < \varepsilon < 1$, taking $\delta = \varepsilon^p$ in (24) therefore also gives a global cover by *binary* experts with normalized ℓ_p accuracy ε and size at most $(C_{\text{bin}}/\varepsilon^p)^d$.

Covers on a fixed instance tree. For a tree $\mathbf{x} = (\mathbf{x}_t)_{t=1}^T$ and a sign path $\sigma \in \{-1, +1\}^T$, write $\mathbf{x}(\sigma) = (\mathbf{x}_t(\sigma_{<t}))_{t=1}^T$. If $\mathbf{x}$ is an instance tree, $f(\mathbf{x}(\sigma))$ denotes the vector obtained by applying f to each coordinate. The normalized *sequential ℓ_p covering number* $N_p^{\text{seq}}(\mathcal{F}, \varepsilon, \mathbf{x})$ is the least cardinality of a finite family V of $[0, 1]$ -valued prediction trees of depth T such that

$$\forall f \in \mathcal{F} \forall \sigma \in \{-1, +1\}^T \exists \mathbf{v} \in V : \quad \rho_{p,T}(\mathbf{v}(\sigma), f(\mathbf{x}(\sigma))) \leq \varepsilon. \quad (26)$$

As before, the covering number is ∞ if no finite family exists. Here V may depend on the full tree $\mathbf{x}$, as in the sequential covering definition of Rakhlin, Sridharan, and Tewari [RST15b]. Set

$$N_p^{\text{seq}}(\mathcal{F}, \varepsilon, T) := \sup_{\mathbf{x}} N_p^{\text{seq}}(\mathcal{F}, \varepsilon, \mathbf{x}),$$

where the supremum ranges over all instance trees of depth T .

Corollary 8 (Sequential ℓ_p covering). *For every $1 \leq p < \infty$, $T \geq 1$, and $\varepsilon > 0$,*

$$N_p^{\text{seq}}(\mathcal{F}, \varepsilon, T) \leq N_p^{\text{dyn}}(\mathcal{F}, \varepsilon, T). \quad (27)$$

Consequently, if $0 < \varepsilon < 1/2$ and $d = \text{sfat}_{\varepsilon/32}(\mathcal{F}) < \infty$, then

$$N_p^{\text{seq}}(\mathcal{F}, \varepsilon, T) \leq \left(\frac{C}{\varepsilon^p} \right)^d. \quad (28)$$

The worst-case accuracy exponent pd is sharp for these sequential covering numbers as well.

Proof. Fix an instance tree $\mathbf{x}$ and a global family $\mathcal{G}$. For each $g \in \mathcal{G}$, define a prediction tree by restriction:

$$\mathbf{v}_t^g(\sigma_{<t}) := g_t(\mathbf{x}_1(\emptyset), \mathbf{x}_2(\sigma_1), \dots, \mathbf{x}_t(\sigma_{<t})).$$

For every target f and path σ , the global covering property applied to the instance sequence $\mathbf{x}(\sigma)$ supplies a member of this family satisfying (26). Its cardinality is at most $|\mathcal{G}|$. Taking the minimum over global covers and then the supremum over instance trees proves (27); if no finite global cover exists, the inequality is immediate. Theorem 5 gives (28).

For sharpness, take the fixed sequence from Theorem 6 and set $\mathbf{x}_t(\sigma_{<t}) = x_t$ at every node of level t . Fixing any one sign path turns each prediction tree into one covering vector for that sequence. The fixed-sequence lower bound therefore gives

$$N_p^{\text{seq}}(\mathcal{H}, \varepsilon, T) > \left(\frac{1}{2^{p+1}\varepsilon^p} \right)^d$$

for the class in that theorem and its stated parameter range. ■

In particular, the case $p = 2$ resolves the horizon-independent sequential ℓ_2 covering question posed by Rakhlin, Sridharan, and Tewari [RST15b].

4 Correction Schedules and Adaptive Sampling

We establish the deterministic and probabilistic ingredients used to prove Proposition 2. We first justify the reduction to a separated class and construct finite families of rank-decreasing correction schedules. We then encode prediction errors by partial binary functions and prove a sampling estimate that is uniform over such functions, even when the inputs depend on past sampling choices. The choice of sampling parameters and the construction of the master and replay histories are deferred to Section 5.

4.1 Separation and Rank-Decreasing Centers

The reduction in Section 2.1 uses a uniformly separated subclass, chosen independently of the instance sequence. We first give the existence argument, including the case of an infinite function class.

Lemma 9 (Uniformly separated reduction). *For every nonempty $\mathcal{F} \subseteq [0, 1]^{\mathcal{X}}$ and every $\eta > 0$, there is a nonempty subclass $\mathcal{H} \subseteq \mathcal{F}$ such that*

$$\|h - h'\|_{\infty} > \eta \quad (h \neq h', h, h' \in \mathcal{H}), \quad \forall f \in \mathcal{F} \exists h \in \mathcal{H} : \|f - h\|_{\infty} \leq \eta. \quad (29)$$

Moreover, $\text{sfat}_{\eta}(\mathcal{H}) \leq \text{sfat}_{\eta}(\mathcal{F})$.

Proof. Order the uniformly greater-than- η separated subclasses of $\mathcal{F}$ by inclusion. The union of a chain is still separated: any two members of the union belong together to one class in the chain. Zorn's lemma therefore supplies a maximal separated subclass $\mathcal{H}$. It is nonempty, since a singleton could otherwise be added. If some $f \in \mathcal{F}$ had distance greater than η from every member of $\mathcal{H}$, adjoining f would contradict maximality. This proves the approximation property. The dimension inequality follows from monotonicity under inclusion. ■

Together with the calculation in (8), Lemma 9 reduces the main upper bound to Proposition 2. For the proof of that proposition, fix a nonempty separated class $\mathcal{H}$ and parameters η, R, T, d satisfying its hypotheses. The *rank* of a class V means $\text{sfat}_\eta(V)$. Separation is needed for the following consequence of rank zero, rather than for the schedule construction itself.

Lemma 10 (Rank-zero classes). *If $\emptyset \neq V \subseteq [0, 1]^\mathcal{X}$ has rank zero, then*

$$\sup_{f \in V} f(x) - \inf_{f \in V} f(x) \leq \eta \quad (x \in \mathcal{X}). \quad (30)$$

If, in addition, V is a subclass of a uniformly greater-than- η separated class, then V is a singleton.

Proof. Two functions whose values at an instance differ by at least η would shatter a depth-one tree, using the midpoint of their values as the threshold. Thus every pairwise difference at a fixed instance is less than η . Taking the supremum of these differences gives (30); the supremum need not be attained. If distinct $h, h' \in V$ had $\|h - h'\|_\infty > \eta$, some instance would satisfy $|h(x) - h'(x)| > \eta$, contradicting the same argument. ■

In particular, a separated class of rank zero is covered exactly by its unique member. This proves the $d = 0$ case of Proposition 2. The remaining local lemmas apply to arbitrary finite-rank classes, so they may also be used for version spaces arising later in the construction.

Rank-decreasing centers and sign updates. We need a prediction rule determined by the current version space and instance, with the property that a sufficiently separated correction strictly lowers rank. The elementary operation behind this rule is to attach two shattered trees below a common root.

Lemma 11 (Gluing shattered trees). *Let $V \subseteq [0, 1]^\mathcal{X}$, $x \in \mathcal{X}$, and $z_1 - z_0 \geq \eta$. If both classes*

$$\{f \in V : f(x) \leq z_0\}, \quad \{f \in V : f(x) \geq z_1\}$$

have rank at least an integer $r \geq 0$, then $\text{sfat}_\eta(V) \geq r + 1$.

Proof. Attach depth- r shattered trees of the two subclasses below a new root at x with threshold $(z_0 + z_1)/2$; the assumption $z_1 - z_0 \geq \eta$ guarantees the required margin $\eta/2$ at the root. ■

For a nonempty class V of finite rank $r = \text{sfat}_\eta(V)$, define the two critical thresholds and their center by

$$\begin{aligned} \ell(V, x) &:= \inf\{z \in [0, 1] : \text{sfat}_\eta(\{f \in V : f(x) \leq z\}) = r\}, \\ u(V, x) &:= \sup\{z \in [0, 1] : \text{sfat}_\eta(\{f \in V : f(x) \geq z\}) = r\}, \\ a(V, x) &:= \frac{1}{2}(\ell(V, x) + u(V, x)). \end{aligned} \quad (31)$$

Both defining sets are nonempty: the lower fiber equals V at $z = 1$, and the upper fiber equals V at $z = 0$. Hence the thresholds and the center belong to $[0, 1]$. We do not assume that the infimum or supremum is attained.

Lemma 12 (Rank-decreasing center). *For every such V and x , the two sign-update classes*

$$V_x^- := \{f \in V : f(x) \leq a(V, x) - \eta\}, \quad V_x^+ := \{f \in V : f(x) \geq a(V, x) + \eta\} \quad (32)$$

have rank at most $r - 1$, with the convention $\text{sfat}_\eta(\emptyset) = -1$.

Proof. Abbreviate the thresholds by ℓ, u and the center by a . Suppose first that $u - \ell > \eta$. Choose $\ell < z_0 < z_1 < u$ with $z_1 - z_0 \geq \eta$. By the definition of ℓ , some admissible lower threshold is smaller than z_0 ; monotonicity therefore makes the lower fiber at z_0 have rank r . The corresponding argument for u makes the upper fiber at z_1 have rank r . Lemma 11 would give rank at least $r + 1$ for V , a contradiction. Consequently $u - \ell \leq \eta$, and

$$a - \eta \leq \ell - \eta/2 < \ell, \quad a + \eta \geq u + \eta/2 > u.$$

If $a - \eta < 0$, the lower sign-update class is empty. Otherwise, its threshold is strictly below ℓ , so its rank cannot be r . Similarly, the upper class is either empty because $a + \eta > 1$, or has rank smaller than r because its threshold exceeds u . Ranks are integers, which proves both bounds. ■

A *sign update* replaces the current nonempty version space V by V_x^- or V_x^+ . A nonempty update is called *valid*. Starting from a class of rank at most r , any sequence of i valid updates leaves a version space of rank at most $r - i$; in particular, there can be at most r valid updates. The update is specified by its sign and uses the fixed center rule in (31), not a target value or a discretized approximation to that value.

4.2 Correction Schedules and Pointwise Covers

We next turn prescribed sign updates into label-free causal simulations. Fix an initial nonempty version space V_0 of rank at most r and a positive integer m . A selection history specifies which rounds are selected. A selected round has an *ordinal* s , meaning that it is the s th selected round, regardless of its position in the original instance sequence.

A *correction schedule* of depth i is a list

$$J = ((s_1, \sigma_1), \dots, (s_i, \sigma_i)), \quad 1 \leq s_1 \leq \dots \leq s_i \leq m, \quad \sigma_j \in \{-, +\}. \quad (33)$$

The empty list is the depth-zero schedule. Repeated ordinals are allowed because more than one rank-decreasing correction may be needed at a single instance. The classical Sauer–Shelah–Perles bound [Sau72; She72] has dynamic analogues [BPS09; MM24]. Our encoding similarly uses dimension-decreasing updates, but accommodates signed corrections and repeated ordinals for real-valued prediction.

Simulation and prediction timing. Every simulation starts at V_0 . On round t , after observing x_t but before the current selection is revealed, its *provisional prediction* is $a(V, x_t)$, using the version space left by the preceding round. If this round has selected ordinal s , the simulation executes all entries of J with $s_j = s$, in their listed order. Each entry applies the indicated sign update, and the center is recomputed from the new version space before the next entry is executed. The simulation emits its *final prediction* only after all these updates. On an unselected round it makes no update, and its final prediction equals its provisional prediction. We denote the provisional and final predictions by a_t^J and y_t^J , respectively.

An update to an empty fiber makes the simulation invalid. From that update onward, it remains in an absorbing invalid state and emits zero, including on the current round. If the selection history ends before a prescribed ordinal is reached, the corresponding update is not executed. The same rule is also defined on selection histories longer than m : there are simply no prescribed updates at ordinals larger than m . These conventions ensure that every formal schedule has an output on every outcome, not only on outcomes where the intended sampling cap is respected.

The simulation uses only the instance history, the selection history, and its fixed schedule. If the selection policy is a deterministic causal function of the instances, the simulation defines

a deterministic causal expert. Under random selection, its outputs additionally depend on the selection history, while remaining causal. We will first analyze these simulations and later obtain deterministic experts by the argument in Section 8.

A valid schedule is *born* once all its prescribed corrections have occurred. The empty schedule is born at time zero. After birth, its version space remains fixed, although its predictions may still vary with the current instance. A born depth- i schedule has residual rank at most $r - i$. Moreover, let q_t be the number of selected rounds up to round t . Removing from a schedule J all entries with selection ordinals greater than q_t leaves its version spaces and predictions unchanged on rounds $1, \dots, t$. The resulting prefix schedule has completed all its prescribed corrections by round t , provided they are valid.

The formal schedule family. The number of weakly increasing ordinal lists of length i in $[m]$ is $\binom{m+i-1}{i}$, and every list admits 2^i sign assignments. Thus, writing $n_0(m) = 1$, the exact counts are

$$n_i(m) := 2^i \binom{m+i-1}{i}, \quad Q_r(m) := \sum_{i=0}^r n_i(m). \quad (34)$$

These counts include invalid schedules, schedules that never become born, and schedules whose outputs coincide. Consequently $Q_r(m)$ is an upper bound on the number of distinct prediction rules, with no need to recognize or remove such duplications.

Lemma 13 (Pointwise schedule cover). *Let $\emptyset \neq V_0 \subseteq [0, 1]^X$ satisfy $\text{sfat}_\eta(V_0) \leq r$. With every round selected, the schedules of depth at most r form one family of at most $Q_r(m)$ deterministic causal experts such that, on every instance sequence of length at most m , each target in V_0 is approximated by one expert with pointwise error at most η . More generally, on any selection history with at most m selected rounds, a schedule in the same formal family attains error at most η on every selected round.*

Proof. Fix a target $f \in V_0$ and an instance-and-selection history. Consider the following reference procedure, which is used only to find a witnessing schedule. It starts from V_0 and does not update on unselected rounds. At a selected instance x , while $|f(x) - a(V, x)| > \eta$, it updates toward $f(x)$: it uses the plus fiber if $f(x) > a(V, x) + \eta$ and the minus fiber otherwise. It recomputes the center and repeats at the same instance until the error is at most η .

Every update retains f , and Lemma 12 decreases the rank by at least one. Since the version space stays nonempty, at most r updates occur over the entire history. Thus, the correction loop at each selected round terminates with prediction error at most η . Record the selected ordinal and sign of each update, retaining their order when an ordinal occurs repeatedly. The resulting list is a schedule of depth at most r in (33). Its ordinary simulation has exactly the reference procedure's states and final predictions, by induction on rounds and on updates within a round. It therefore has error at most η at every selected round.

When all rounds are selected, this is the asserted pointwise cover. The formal family is fixed independently of both the target and the instance sequence. Only the choice of its witnessing member used the target values; no expert in the family receives those values. $\blacksquare$

We record both the size estimate and a convolution identity for these counts. The latter will be used when counting the continuation capacity in Section 8.

Lemma 14 (Schedule-count estimates). *For integers $m, r \geq 1$,*

$$Q_r(m) \leq 2^r \binom{m+r}{r} \leq \left[2e \left(1 + \frac{m}{r} \right) \right]^r. \quad (35)$$

For every integer $\ell \geq 0$,

$$\sum_{i=0}^{\ell} n_i(m) n_{\ell-i}(m) = n_{\ell}(2m). \quad (36)$$

Consequently, for every integer $d \geq 1$,

$$\sum_{i=0}^{d-1} n_i(m) Q_{d-i}(m) \leq Q_d(2m). \quad (37)$$

Proof. The hockey-stick identity gives

$$Q_r(m) \leq 2^r \sum_{i=0}^r \binom{m+i-1}{i} = 2^r \binom{m+r}{r}.$$

Using $\binom{N}{r} \leq (eN/r)^r$ proves (35). For the convolution, each of the m selection ordinals can carry any number $j \geq 0$ of corrections, with 2^j possible sign assignments. Thus each ordinal contributes the generating function

$$\sum_{j=0}^{\infty} 2^j z^j = (1 - 2z)^{-1},$$

and multiplying over all m ordinals gives

$$\sum_{i=0}^{\infty} n_i(m) z^i = (1 - 2z)^{-m}.$$

Squaring this identity and comparing the coefficient of z^{ℓ} gives (36). Summing it over $\ell \leq d$ yields

$$\sum_{\substack{i,j \geq 0 \\ i+j \leq d}} n_i(m) n_j(m) = Q_d(2m).$$

The sum in (37) contains all of these nonnegative terms except the one indexed by $(i, j) = (d, 0)$. ■

In particular, the all-round cover has size $Q_d(T) \leq [2e(1 + T/d)]^d$ when $d \geq 1$. This will handle the short-horizon regime directly. For larger horizons, the same schedule family will be used on shorter sampled lists. In either case, the count depends on selected ordinals and signs, without an additional factor for discretizing the target values at scale η .

4.3 Partial Binary Completion and Adaptive Sampling

The sampling argument encodes one-sided prediction errors using partial binary labels: discrepancies at most η receive label 0, those at least 2η receive label 1, and intermediate values receive the undefined symbol $*$. This gap lets us control the resulting partial binary classes by sequential fat-shattering dimension at scale η . We first define these classes, then construct the finite families of binary rules needed to analyze their sampled values.

Let $\mathcal{Z}$ be a domain. A *partial binary class* is a subset $\mathcal{A} \subseteq \{0, 1, *\}^{\mathcal{Z}}$. Such a class shatters a binary instance tree $z_t : \{0, 1\}^{t-1} \rightarrow \mathcal{Z}$, $t \in [s]$, if

$$\forall b \in \{0, 1\}^s \exists \phi \in \mathcal{A} \forall t \in [s] : \phi(z_t(b_{<t})) = b_t.$$

Thus $*$ realizes neither edge of the tree. Define $\text{Ldim}(\mathcal{A})$ to be the largest shattered depth, or infinity when these depths are unbounded. As before, a nonempty class has dimension at least zero, and the empty class has dimension -1 . On total binary classes this is ordinary Littlestone dimension.

Partial error classes. For $f \in \mathcal{H}$ and $(x, a) \in \mathcal{X} \times [0, 1]$, define

$$\phi_f^+(x, a) := \begin{cases} 0, & f(x) - a \leq \eta, \\ 1, & f(x) - a \geq 2\eta, \\ *, & \eta < f(x) - a < 2\eta, \end{cases} \quad \phi_f^-(x, a) := \begin{cases} 0, & a - f(x) \leq \eta, \\ 1, & a - f(x) \geq 2\eta, \\ *, & \eta < a - f(x) < 2\eta. \end{cases} \quad (38)$$

Write $\mathcal{A}^\pm := \{\phi_f^\pm : f \in \mathcal{H}\}$. A definite one records a one-sided error of size at least 2η ; a nonzero value, including $*$, records an error strictly greater than η in that direction.

Lemma 15 (Dimension of the partial error classes). *Both partial classes satisfy*

$$\text{Ldim}(\mathcal{A}^+), \quad \text{Ldim}(\mathcal{A}^-) \leq \text{sfat}_\eta(\mathcal{H}). \quad (39)$$

Proof. Suppose a depth- s tree of pairs (x, a) is shattered by $\mathcal{A}^+$. Replacing each node (x, a) with instance x and threshold $a + 3\eta/2$ yields an η -fat-shattered tree, since labels 0 and 1 imply $f(x) \leq a + \eta$ and $f(x) \geq a + 2\eta$, respectively. The argument for $\mathcal{A}^-$ is identical, using threshold $a - 3\eta/2$ and interchanging the two branches at each node. ■

Completion by finitely many causal rules. For a partial class W and $b \in \{0, 1\}$, write $W_z^b := \{\phi \in W : \phi(z) = b\}$. For nonempty W of finite Littlestone dimension, the standard optimal algorithm (SOA) [Lit88] prediction $\text{soa}(W, z)$ is the label maximizing $\text{Ldim}(W_z^b)$, with ties broken in favor of zero. As in Lemma 11, the two defined-label fibers cannot both have the rank of W . Since SOA chooses a maximum-rank fiber,

$$\text{Ldim}(W_z^{1-\text{soa}(W, z)}) \leq \text{Ldim}(W) - 1. \quad (40)$$

Undefined values belong to neither fiber.

Lemma 16 (Partial SOA completion). *Let $\mathcal{A} \subseteq \{0, 1, *\}^{\mathcal{Z}}$ have Littlestone dimension at most $r \geq 0$, and let $n \geq 1$ be an integer. There is a family of at most*

$$S_r(n) := \sum_{i=0}^{\min\{r, n\}} \binom{n}{i}$$

deterministic causal prediction rules $v = (v_t)_{t=1}^n$, with $v_t : \mathcal{Z}^t \rightarrow \{0, 1\}$, such that, on every sequence of length at most n , each $\phi \in \mathcal{A}$ is matched at all its defined positions by one rule.

Proof. If $\mathcal{A}$ is empty, a constant rule suffices. Otherwise, for each $J \subseteq [n]$ with $|J| \leq r$, define a rule starting from $W = \mathcal{A}$. Outside J , it outputs $\text{soa}(W, z)$ and makes no update. At a round in J , it outputs the opposite label and restricts W to that opposite-label fiber. An empty restriction makes the rule invalid, with output zero from that round onward. Thus every J specifies a total binary causal rule without access to partial target values. There are $S_r(n)$ such sets J .

To find a rule for a particular target ϕ and sequence, use a reference procedure that restricts its version space only when $\phi(z)$ is defined and differs from the current SOA prediction. It records each such round and restricts to $W_z^{\phi(z)}$. It leaves W unchanged when the prediction is correct or when $\phi(z) = *$. Every restriction retains ϕ and, by (40), strictly decreases rank. The recorded set J therefore has size at most r . Induction on rounds shows that the rule indexed by this J has exactly the reference procedure's states and matches every defined target value. ■

The lemma constructs causal completions along input histories; it does not choose one total static completion of each partial target. Its codes record at most one binary reversal per round and require no sign, which explains the count $S_r(n)$ rather than $Q_r(n)$. We will also use the elementary estimate

$$S_r(n) \leq \left(\frac{en}{r}\right)^r \quad (1 \leq r \leq n). \quad (41)$$

Indeed, with $x = r/n \leq 1$,

$$x^r S_r(n) \leq \sum_{i=0}^r \binom{n}{i} x^i \leq (1+x)^n \leq e^{nx} = e^r.$$

For $r = 0$ the count is one, and for $r \geq n$ it is 2^n .

Adaptive low-hit sampling. Sampling against inputs that react to earlier sample selections has been studied in [BY20; Alo+21]. We now bound the probability that a partial target has many definite ones but few sampled nonzero values. A completion rule must match all defined labels but may output either label at an undefined position. Consequently, the number of sampled ones predicted by the completion is at most the number of sampled nonzero values of the partial target.

Order of information. Let $\xi_1, \dots, \xi_T$ be independent Bernoulli- γ bits. The input z_t is chosen before ξ_t is revealed and may depend on $\xi_{<t}$ and an independent finite-support auxiliary seed ω . Equivalently, z_t is a deterministic function of $(\xi_{<t}, \omega)$. We call this a *predictable input stream* for the Bernoulli sample. Fixing ω leaves the product law of the sampling bits unchanged.

Lemma 17 (Adaptive low-hit sampling). *Let $\mathcal{A} \subseteq \{0, 1, *\}^{\mathbb{Z}}$ have Littlestone dimension at most an integer $r \geq 0$. Let $z_1, \dots, z_T$ be a predictable input stream for independent Bernoulli- γ bits $\xi_1, \dots, \xi_T$, where $0 < \gamma \leq 1/2$. For each $\phi \in \mathcal{A}$, define*

$$L_\phi := \sum_{t=1}^T \mathbf{1}\{\phi(z_t) = 1\}, \quad H_\phi := \sum_{t=1}^T \xi_t \mathbf{1}\{\phi(z_t) \neq 0\}. \quad (42)$$

Let $q \geq 0$ and $B \geq 2q$ satisfy $\gamma B \geq 16 \log 2$. With $M = \lceil 4\gamma T \rceil$ and $c_0 = \log(3/2)/4$, one has

$$\Pr\{\exists \phi \in \mathcal{A} : L_\phi \geq B, H_\phi \leq q\} \leq 2 \left(S_r(M) 2^q e^{-c_0 \gamma B} + e^{-2\gamma T/3} \right). \quad (43)$$

Proof. Fix a value of the auxiliary seed ω in its support. The sampling bits remain independent Bernoulli- γ variables, and each z_t is now a deterministic function of $\xi_{<t}$. All probabilities below are conditional on this fixed seed until the final averaging step. Call a round *red* when $\xi_t = 1$, and let E denote the event on the left of (43).

A second sample for comparison. We introduce a second sample to detect definite ones outside the red sample. After the red bits have been generated, independently color each nonred round *blue* with probability $\gamma/(1-\gamma)$. This choice makes the marginal probability of blue equal to γ , the same as that of red.

For each red-bit outcome in E , choose a witnessing ϕ . There are only finitely many red-bit outcomes, so this choice raises no measurability issue even when $\mathcal{A}$ is infinite. Conditional on that red outcome, both the input sequence and this witness are fixed. The witness has at least

$B - q \geq B/2$ nonred definite-one positions, since its red definite ones are included among its at most q red nonzero values. Its number of blue definite ones is therefore binomial with mean

$$\nu \geq \frac{\gamma}{1-\gamma}(B - q) \geq \gamma B/2.$$

Since $\gamma B/4 \leq \nu/2$, a standard Chernoff bound gives, conditional on the red outcome,

$$\Pr\{\text{at least } \gamma B/4 \text{ blue definite ones}\} \geq 1 - e^{-\nu/8} \geq 1 - e^{-\gamma B/16} \geq \frac{1}{2}.$$

Let E' be the event that some partial target has at most q red nonzero values and at least $\gamma B/4$ blue definite ones. Thus $\Pr(E' \mid \xi_{1:T}) \geq 1/2$ on E , and hence

$$\Pr(E) \leq 2 \Pr(E'). \quad (44)$$

Conditioning on the sampled union. The three possible colors—red, blue, and neither—have probabilities $\gamma, \gamma, 1 - 2\gamma$, independently across rounds. Equivalently, we may first generate independent union bits $U_t \sim \text{Bernoulli}(2\gamma)$ and then give each included round an independent fair red-or-blue color. Setting $\xi_t = 1$ exactly on red rounds and reconstructing z_t from $\xi_{<t}$ gives the original joint distribution of the red bits and inputs.

Write $N = \sum_{t=1}^T U_t$. Since $\mathbb{E}N = 2\gamma T$ and $M \geq 4\gamma T$, another standard Chernoff bound gives

$$\Pr\{N > M\} \leq e^{-2\gamma T/3}. \quad (45)$$

Fix a union pattern with $N \leq M$, and enumerate its included rounds as $\tau_1 < \dots < \tau_N$. Conditional on this pattern, the colors at these rounds remain independent and fair. Since z_t depends only on the preceding sampling bits $\xi_{<t}$, fixing the union pattern makes z_{τ_j} a function of the colors at $\tau_1, \dots, \tau_{j-1}$. Indeed, $\xi_t = 1$ exactly when round t is included and colored red. Thus the compressed stream $z_{\tau_1}, \dots, z_{\tau_N}$ is predictable with respect to the independent fair colors. Conditioning on the union has not fixed this input stream; it may still depend on the preceding colors.

Use the fixed family $\mathcal{V}$ of at most $S_r(M)$ deterministic binary causal rules supplied by Lemma 16 at horizon M . On E' , some rule in $\mathcal{V}$ matches all defined labels of a witnessing target on the compressed stream. It has at least $\gamma B/4$ blue ones, because it matches every blue definite one. It has at most q red ones, because it must output zero at every red position where the target is zero; all other red positions, including undefined ones, are included in the target's red nonzero count.

Bounding the red–blue discrepancy. Still conditioning on the union pattern, fix a rule $v \in \mathcal{V}$. Let A_j and C_j count its red and blue ones, respectively, among the first j included rounds. The rule's prediction at the next included round is determined before that round's color is revealed. Therefore

$$Z_j := 2^{-A_j} (3/2)^{C_j}, \quad 0 \leq j \leq N,$$

is a nonnegative martingale with $Z_0 = 1$. When the prediction is zero, its multiplier is one; when the prediction is one, the conditional expected multiplier is

$$\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{3}{2} = 1.$$

On $\{A_N \leq q, C_N \geq \gamma B/4\}$, one has $Z_N \geq 2^{-q} (3/2)^{\gamma B/4}$. Markov's inequality gives

$$\Pr\{A_N \leq q, C_N \geq \gamma B/4 \mid U_{1:T}\} \leq 2^q (3/2)^{-\gamma B/4} = 2^q e^{-c_0 \gamma B}.$$

Take a union bound over $\mathcal{V}$, average over union patterns with $N \leq M$, and add (45). We obtain

$$\Pr(E') \leq S_r(M)2^q e^{-c_0 \gamma B} + e^{-2\gamma T/3}.$$

Combining this with (44) proves the asserted bound conditional on each value of ω . Averaging over ω completes the proof. $\blacksquare$

For the application to prediction errors, fix a formal schedule J and set $z_t = (x_t, a_t^J)$, where a_t^J is its provisional prediction. Under the stated information order, this stream is predictable. For either partial error class in (38), a sampled nonzero value records a one-sided provisional error exceeding η . When J is the schedule produced for a target h by the reference procedure in Lemma 13, each such round requires at least one correction, so $H_{\phi_h^\pm} \leq d$.

Section 5 will apply Lemma 17 to each fixed schedule and take a union bound over all schedules, so the resulting guarantee also holds for the witnessing schedule of every target.

5 Master Histories and Multi-Pool Accuracy

The preceding section gives sign schedules that are accurate on selected rounds. We now introduce multiple sampling pools to obtain a simultaneous guarantee: for every target, a fixed positive fraction of the pools provide schedules with few large prediction errors over the entire horizon, including unselected rounds. We will also ensure that the target satisfies every constraint recorded by these schedules with a positive margin.

Parameters and capped sampling pools. Fix $\mathcal{H}, \eta, R, T, d$ as in Proposition 2. Recall that the rank of a version space is its sequential fat-shattering dimension at scale η . The case $d = 0$ was settled in Section 4.1, so assume $d \geq 1$ and put $b = T/R$. Fix a universal constant $\alpha \geq 128$; we will specify sufficient conditions on its size at the end of this section.

If $b \leq 32\alpha d$, select every round. Lemmas 13 and 14 give a global pointwise radius- η cover of size

$$Q_d(T) \leq \left[2e \left(1 + \frac{T}{d} \right) \right]^d \leq (66e\alpha R)^d. \quad (46)$$

Its witnessing experts have no errors of size at least 8η . This proves Proposition 2 in this regime, after increasing its universal constant if necessary.

For the rest of the section assume $b > 32\alpha d$ and set

$$k = 4\lceil \log_2 R \rceil, \quad \beta = \frac{\alpha d}{b} < \frac{1}{32}, \quad m = \lceil 2\beta T \rceil = \lceil 2\alpha d R \rceil. \quad (47)$$

In particular, $k \geq 24$ and k is divisible by four. All logarithms below are natural unless a base is indicated.

For each $t \in [T]$ and $j \in [k]$, let $\xi_{t,j} \sim \text{Bernoulli}(\beta)$, independently over all rounds and pools. The *raw* selection list of pool j contains the rounds with $\xi_{t,j} = 1$. Its *capped* list retains only the first m such rounds, with indicators

$$\widehat{\xi}_{t,j} := \xi_{t,j} \mathbf{1} \left\{ \sum_{s < t} \xi_{s,j} < m \right\}.$$

The capped list therefore has length at most m on every outcome. The retained indicators are not themselves independent Bernoulli variables; all independence arguments below use the raw bits.

Define

$$\text{Cap} := \left\{ \sum_{t=1}^T \xi_{t,j} \leq m \text{ for every } j \in [k] \right\}. \quad (48)$$

Each raw list has expected length $\beta T = \alpha d R$. Since $m \geq 2\beta T$, a Chernoff bound and a union bound give

$$\Pr(\text{Cap}^c) \leq k e^{-\alpha d R/3}. \quad (49)$$

On **Cap**, the raw and capped selection indicators agree at every round. Thus, running the same schedule with raw or capped selections on the same realized input history produces identical states and predictions. We prove the sampling bounds under the original product law of the raw bits, without conditioning on **Cap**, and use this agreement to transfer the guarantees to the capped simulations on **Cap**.

Adaptive inputs. As in Lemma 17, we allow adaptive input sequences. Write $\xi_t = (\xi_{t,1}, \dots, \xi_{t,k})$ and $\xi_{<t} = (\xi_1, \dots, \xi_{t-1})$. More precisely, we allow

$$x_t = F_t(\xi_{<t}, \omega_0), \quad (50)$$

where F_t are arbitrary deterministic maps and ω_0 is an independent finite-support auxiliary random seed.

Conditional on ω_0 and $\xi_{<t}$, the instance x_t and the provisional predictions of all fixed schedule simulations are determined, while the coordinates of ξ_t remain independent Bernoulli- β variables.

All probability bounds in this section hold uniformly over input rules of the form (50).

5.1 The Analytical Master and Uniform Accuracy

We first specify a common selection process for the reference histories, containing every round selected by any pool. The choice depends on $k\beta = \alpha k d R/T$. When $k\beta \leq 1/8$, we use the union of the pool selections, whose sampling rate is comparable to $k\beta$. When $k\beta > 1/8$, we have $T < 8\alpha k d R$, so selecting every round is affordable. Accordingly, define

$$\zeta_t := \begin{cases} \mathbf{1}\{\xi_t \neq \mathbf{0}\}, & k\beta \leq 1/8, \\ 1, & k\beta > 1/8. \end{cases} \quad (51)$$

We call these cases the *union regime* and the *all-round regime*, respectively. In either case, $\xi_{t,j} = 1$ implies $\zeta_t = 1$.

In the union regime, the ζ_t are independent Bernoulli variables with rate

$$u = 1 - (1 - \beta)^k, \quad \frac{k\beta}{2} \leq u \leq k\beta \leq \frac{1}{8}. \quad (52)$$

The upper bound follows from a union bound, while the lower bound follows from $(1 - \beta)^k \leq e^{-k\beta}$ and $1 - e^{-z} \geq z/2$ for $0 \leq z \leq 1$.

The analytical master. For each target $h \in \mathcal{H}$, define its *analytical master* by the following target-dependent procedure. Initialize $V = \mathcal{H}$. At round t , before revealing ξ_t , form the provisional prediction $a_t^h = a(V, x_t)$. If $\zeta_t = 0$, leave V unchanged and set $y_t^h = a_t^h$. If $\zeta_t = 1$, while $|a(V, x_t) - h(x_t)| > \eta$, apply the sign update toward $h(x_t)$ as in Lemma 13, recomputing the center after each update. Set $y_t^h = a(V, x_t)$ after the correction loop.

Every update retains h and decreases the sequential fat rank. Hence the procedure makes at most d corrections over the entire horizon, and

$$\zeta_t = 1 \implies |y_t^h - h(x_t)| \leq \eta. \quad (53)$$

The procedure uses target values only to find a witnessing schedule; it is not a label-free prediction rule. Record the selected ordinal and sign of each correction, and denote the resulting code by J^h . This code is defined on every raw outcome and always has depth at most d , although its ordinals need not satisfy a fixed cap.

A fixed family of master codes. Set

$$M_* := 4km, \quad (54)$$

and let $\mathcal{J}_*$ be the family of all formal sign schedules of depth at most d on ordinals $1, \dots, M_*$. It is fixed before the experiment, and $|\mathcal{J}_*| = Q_d(M_*)$. In the union regime, **Cap** bounds the number of master selections by km . In the all-round regime,

$$k\beta > \frac{1}{8} \implies T < 8\alpha kdR \leq 4km = M_*.$$

Thus, on **Cap**, $J^h \in \mathcal{J}_*$ for every target. There is no truncation of the master: M_* is a bound used to specify the finite code family.

For every fixed $J \in \mathcal{J}_*$, simulate its prescribed updates on the master selection list, starting from $\mathcal{H}$, and denote its provisional and final predictions by a_t^J and y_t^J . Use the conventions of Section 4.2: an empty-fiber update makes the simulation invalid with subsequent predictions zero, and no updates are prescribed at ordinals greater than M_* . Every fixed-code simulation is therefore defined on every raw outcome, including outcomes outside **Cap**. All simulations use the same realized inputs and selection lists; varying J does not change the input rule. On an outcome with $J^h \in \mathcal{J}_*$, the simulation of J^h has exactly the target-driven master's states and predictions, by induction on rounds and on corrections within each round.

Accuracy before the current selection. The selected-round guarantee (53) concerns final predictions. We also need control of the provisional predictions on all rounds. Define

$$\text{Master} := \left\{ \text{for every } h \in \mathcal{H}, \quad \#\{t \in [T] : |a_t^h - h(x_t)| \geq 2\eta\} < \frac{b}{32} \right\}. \quad (55)$$

Lemma 18 (Uniform master accuracy). *In the all-round regime, **Master** holds on every outcome. In the union regime, put $M = \lceil 4uT \rceil$ and let $c_0 = \log(3/2)/4$ be the constant in Lemma 17. Then*

$$\Pr(\text{Cap} \cap \text{Master}^c) \leq 4Q_d(M_*) \left(S_d(M) 2^d e^{-c_0 u b / 64} + e^{-2uT/3} \right). \quad (56)$$

Proof. In the all-round regime, every provisional error of size at least 2η forces a correction. There are at most d such rounds, and $d < b/32$ because $b > 32\alpha d$. This proves the first claim.

Now consider the union regime and fix $J \in \mathcal{J}_*$. We will apply Lemma 17 to the stream (x_t, a_t^J) and the Bernoulli- u master selections. The lemma requires the stream to be determined by the past sampling bits and an auxiliary seed independent of those bits. Although ζ_t are independent Bernoulli- u variables, the inputs may depend on the full preceding pool vectors $\xi_{<t}$, which are not determined by $\zeta_{<t}$ and ω_0 alone.

An equivalent way to generate the raw vectors is to draw independent Bernoulli- u bits ζ_t and independent nonzero vectors W_t , where each W_t has the Bernoulli- β product law conditioned on being nonzero, and to set

$$\xi_t = \zeta_t W_t.$$

The entire sequence $(W_t)_{t=1}^T$ is independent of the union bits and of ω_0 . This generates exactly the original joint law of the raw pool vectors. Fixing the finite-support auxiliary seed $\omega = (\omega_0, W_1, \dots, W_T)$ leaves the union bits independent Bernoulli- u variables. Once this auxiliary seed is fixed, each preceding pool vector is determined by the corresponding union bit through $\xi_s = \zeta_s W_s$, for $s < t$. Consequently, the stream (x_t, a_t^J) is determined by $\zeta_{<t}$ and ω , as required by Lemma 17. For each fixed ω , Lemma 17 applies and gives the same probability bound, independently of ω . Averaging this conditional bound over ω yields the corresponding unconditional bound. Since the reconstructed pool vectors have the original joint distribution, the bound holds for the original sampling process.

For either partial error class $\mathcal{A}^+$ or $\mathcal{A}^-$ from (38), apply Lemma 17 with

$$\gamma = u, \quad r = q = d, \quad B = b/64.$$

The dimension hypothesis follows from Lemma 15. The numerical hypotheses follow from $u \leq 1/8$ and

$$\frac{b}{64} > \frac{\alpha d}{2} \geq 2d, \quad \frac{ub}{64} \geq \frac{\alpha kd}{128} \geq 16 \log 2,$$

using $k \geq 24$ and $\alpha \geq 128$. For this fixed code $J \in \mathcal{J}_*$ and either sign, the probability that some partial target has at least $b/64$ definite ones but at most d sampled nonzero values is at most

$$2 \left(S_d(M) 2^d e^{-c_0 ub/64} + e^{-2uT/3} \right).$$

Take a union bound over the two signs and all $Q_d(M_*)$ fixed codes.

Outside this union of bad events, consider an outcome in Cap and any target h . Its actual code J^h belongs to $\mathcal{J}_*$. A sampled nonzero value of either $\phi_h^+(x_t, a_t^{J^h})$ or $\phi_h^-(x_t, a_t^{J^h})$ means that the corresponding one-sided provisional error exceeds η . It therefore forces a correction in the target-driven master. For each sign, there are at most d such sampled rounds, so the one-sided definite-error count is less than $b/64$. The sum of these two counts is exactly $\#\{t : |a_t^h - h(x_t)| \geq 2\eta\}$, which is consequently less than $b/32$. This proves (56). The estimates were applied to fixed codes before selecting any J^h ; no conditioning on a target's realized code was used. ■

5.2 Tolerant Replay and Buffered Corrections

We next use the master's final predictions as reference values for version space updates in each individual raw pool. Each pool maintains its own version space and applies the same center rule and sign-update rule as the master, but to its own version space. Unlike the master, which corrects when its center differs from $h(x_t)$ by more than η , a pool replay corrects when its center differs from the master's final prediction by more than 4η .

The replay rule. For each fixed master code $J \in \mathcal{J}_*$ and each pool j , start a separate version space at $\mathcal{H}$. Its provisional prediction at round t is the center of the version space left by round $t-1$ at x_t . After revealing ξ_t , first compute the master's final value y_t^J . If $\xi_{t,j} = 0$, leave the replay version space unchanged. If $\xi_{t,j} = 1$, repeatedly update its current center a toward y_t^J while $|a - y_t^J| > 4\eta$: use the plus fiber in (32) when $y_t^J > a + 4\eta$, and the minus fiber otherwise.

Recompute the center after each update and emit the final prediction when the loop stops. We call this procedure the *tolerant replay* of J in pool j , and denote its provisional and final predictions by $a_{t,j}^J$ and $y_{t,j}^J$.

For an arbitrary fixed formal code $J \in \mathcal{J}_*$, the replay need not retain any particular target, since the master's final predictions y_t^J need not approximate any target on that outcome. An update may therefore lead to an empty fiber, in which case the replay enters an absorbing invalid state and outputs zero from that round onward. Every nonempty update strictly decreases rank, so at most d such updates can occur. Hence each correction loop terminates either when the current center is within 4η of the reference value or when an update produces an empty fiber. In either case, the replay has a well-defined final prediction on every round of every raw outcome.

For a target h , apply exactly the same replay rule to its target-driven master values y_t^h , and write $a_{t,j}^h$ and $y_{t,j}^h$ for the resulting predictions. These target-driven replays are defined directly from the master for h , not by conditioning on any formal code. On an outcome with $J^h \in \mathcal{J}_*$, they coincide with the replays of J^h .

Lemma 19 (Buffered replay). *For every target $h \in \mathcal{H}$ and pool j , the target-driven replay always retains h and makes at most d corrections. More precisely, whenever a correction is performed at center a , we have*

$$\begin{cases} h(x_t) > a + 3\eta, & \text{for a plus update,} \\ h(x_t) < a - 3\eta, & \text{for a minus update.} \end{cases} \quad (57)$$

Moreover, on every round selected by pool j ,

$$|y_{t,j}^h - h(x_t)| \leq 5\eta. \quad (58)$$

Proof. Every pool-selected round is also selected by the master, so $|y_t^h - h(x_t)| \leq \eta$ on such rounds.

If the replay performs a plus update at center a , then $y_t^h > a + 4\eta$, and hence

$$h(x_t) \geq y_t^h - \eta > a + 3\eta.$$

The minus case is symmetric. Since the corresponding sign-update fibers are defined by the thresholds $a + \eta$ and $a - \eta$, respectively, each correction retains h . Every correction strictly decreases rank, so at most d corrections occur.

On a selected round, the replay therefore terminates with $|y_{t,j}^h - y_t^h| \leq 4\eta$. The triangle inequality gives the claimed error bound. $\blacksquare$

On **Cap**, the target-driven replay in pool j makes at most d corrections, all at selected ordinals in $[m]$. Recording their ordinals and signs gives a formal schedule J_j^h of depth at most d . Simulating J_j^h on the same instance sequence and pool j 's capped selection history reproduces exactly the replay's version spaces and predictions.

From disagreements to prediction errors. For each target and pool, define

$$\begin{aligned} D_j^h &:= \sum_{t=1}^T \mathbf{1}\{|a_t^h - a_{t,j}^h| > 6\eta\}, \\ H_j^h &:= \sum_{t=1}^T \xi_{t,j} \mathbf{1}\{|a_t^h - a_{t,j}^h| > 6\eta\}. \end{aligned} \quad (59)$$

Thus D_j^h counts disagreements between the provisional predictions, and H_j^h counts those disagreements selected by pool j . On a round selected by pool j with $|a_t^h - a_{t,j}^h| > 6\eta$, if the master makes at least one correction, charge the round to the master. Otherwise, $y_t^h = a_t^h$, so $|y_t^h - a_{t,j}^h| > 6\eta > 4\eta$, and the replay must make at least one correction; charge the round to the replay. Since each procedure makes at most d corrections in total, at most $2d$ distinct rounds can be charged. Hence

$$H_j^h \leq 2d \quad (h \in \mathcal{H}, j \in [k]). \quad (60)$$

If $|a_{t,j}^h - h(x_t)| \geq 8\eta$ and $|a_t^h - h(x_t)| < 2\eta$, then $|a_t^h - a_{t,j}^h| > 6\eta$ by the triangle inequality. Consequently, on **Master**,

$$\#\{t : |a_{t,j}^h - h(x_t)| \geq 8\eta\} < \frac{b}{32} + D_j^h. \quad (61)$$

By (58), a final replay error of size at least 8η can occur only on an unselected round. There the provisional and final predictions agree, so final outliers are a subset of provisional outliers. In particular, on **Master**, the condition $D_j^h < b/16$ makes both outlier counts less than $3b/32 < b/8$.

5.3 Simultaneous Accuracy Across Pools

Although the replay histories need not be independent, the current disagreement indicators are determined by the past, while the fresh pool selection bits remain independent Bernoulli- β variables even after conditioning on the past. We use this fact to establish a joint bound.

A joint bound for fixed master codes. For a fixed $J \in \mathcal{J}_*$ and its replays defined above, set

$$\Delta_{t,j}^J := \mathbf{1}\{|a_t^J - a_{t,j}^J| > 6\eta\}, \quad D_j^J := \sum_{t=1}^T \Delta_{t,j}^J, \quad H_j^J := \sum_{t=1}^T \xi_{t,j} \Delta_{t,j}^J. \quad (62)$$

Every current indicator $\Delta_{t,j}^J$ is determined before ξ_t is revealed. We do not assert $H_j^J \leq 2d$ for arbitrary formal codes; that bound is needed only for target-driven histories, as in (60).

Lemma 20 (Joint disagreement bound). *For every fixed $J \in \mathcal{J}_*$ and $I \subseteq [k]$,*

$$\Pr \left\{ D_j^J \geq \frac{b}{16}, H_j^J \leq 2d \text{ for every } j \in I \right\} \leq e^{-\alpha|I|d/64}. \quad (63)$$

Proof. Work under the original joint law of the raw pool bits. Let $D_{j,t}^J$ and $H_{j,t}^J$ denote the corresponding counts through round t , and define

$$Z_t := \exp \left(\frac{\beta}{2} \sum_{j \in I} D_{j,t}^J - (\log 2) \sum_{j \in I} H_{j,t}^J \right), \quad Z_0 = 1.$$

Conditional on ω_0 and $\xi_{<t}$, all current disagreement indicators are fixed and the current pool bits are independent Bernoulli- β variables. Hence

$$\mathbb{E} \left[\frac{Z_t}{Z_{t-1}} \middle| \omega_0, \xi_{<t} \right] = \prod_{j \in I: \Delta_{t,j}^J = 1} e^{\beta/2(1-\beta/2)} \leq 1,$$

where the last inequality uses $1 - z \leq e^{-z}$. Thus (Z_t) is a nonnegative supermartingale for the filtration generated by $\omega_0, \xi_1, \dots, \xi_t$, and $\mathbb{E} Z_T \leq 1$. On the event in (63),

$$Z_T \geq \exp \left[|I|d \left(\frac{\alpha}{32} - 2 \log 2 \right) \right],$$

since $\beta b = \alpha d$. Markov's inequality gives the upper bound $\exp[-|I|d(\alpha/32 - 2\log 2)]$. Finally, $\alpha \geq 128$ implies $\alpha/32 - 2\log 2 \geq \alpha/64$. $\blacksquare$

Define **Joint** to be the event that there are no $J \in \mathcal{J}_*$ and $I \subseteq [k]$ with $|I| = k/4$ such that $D_j^J \geq b/16$ and $H_j^J \leq 2d$ for every $j \in I$. A union bound in Lemma 20 gives

$$\Pr(\text{Joint}^c) \leq Q_d(M_*) \binom{k}{k/4} e^{-\alpha kd/256}. \quad (64)$$

This estimate does not condition on **Cap** or **Master**, and it assumes no independence between replays.

Proposition 21 (Simultaneous replay accuracy). *For a sufficiently large universal choice of α , in the regime $b > 32\alpha d$, the event*

$$\text{Good} := \text{Cap} \cap \text{Master} \cap \text{Joint}$$

has probability greater than 15/16 for every input rule (50). On this event, for every $h \in \mathcal{H}$, at least $3k/4$ pools j satisfy

$$\max \left\{ \#\{t : |a_{t,j}^h - h(x_t)| \geq 8\eta\}, \#\{t : |y_{t,j}^h - h(x_t)| \geq 8\eta\} \right\} < \frac{b}{8}. \quad (65)$$

Proof. We first prove the accuracy assertion on **Good**. Fix an outcome in this event and a target $h \in \mathcal{H}$. Since **Cap** holds, the master's correction code J^h belongs to $\mathcal{J}_*$. On this realized history, the simulations associated with J^h reproduce exactly the target-driven master and its pool replays. Hence

$$D_j^{J^h} = D_j^h, \quad H_j^{J^h} = H_j^h \quad (j \in [k]).$$

By (60), $H_j^h \leq 2d$ for every pool j . Since **Joint** holds, fewer than $k/4$ pools can satisfy $D_j^h \geq b/16$. Consequently, the set

$$I_h := \{j \in [k] : D_j^h < b/16\}$$

has cardinality at least $3k/4$.

For every $j \in I_h$, the error transfer (61) and **Master** give

$$\#\{t : |a_{t,j}^h - h(x_t)| \geq 8\eta\} < \frac{b}{32} + D_j^h < \frac{3b}{32} < \frac{b}{8}.$$

Every final replay outlier is also a provisional outlier, as established above. Thus (65) holds for every $j \in I_h$. Since the outcome and h were arbitrary, this conclusion holds for all targets on every outcome in **Good**.

It remains to show that $\Pr(\text{Good}) > 15/16$. We verify that a sufficiently large universal α makes each of the three failure probabilities small enough.

Recall that $R > 32$, $d \geq 1$, and $k = 4\lceil \log_2 R \rceil$. The parameter definitions give

$$24 \leq k \leq R, \quad \log R \leq \frac{k \log 2}{4}, \quad m \leq 3\alpha d R, \quad M_* \leq 12\alpha k d R. \quad (66)$$

In the union regime, the bounds $k\beta/2 \leq u \leq k\beta$ further imply

$$d \leq M = \lceil 4uT \rceil \leq 5\alpha k d R, \quad ub \geq \frac{\alpha k d}{2}, \quad uT \geq \frac{\alpha k d R}{2}.$$

Lemma 14 and (41) therefore yield

$$Q_d(M_*) \leq (26e\alpha kR)^d, \quad S_d(M) \leq (5e\alpha kR)^d, \quad (67)$$

where the second bound is needed only in the union regime.

Set $A_\alpha := \log(100e\alpha)$. Since $k \leq R$ and $\log R \leq k \log 2/4$, we have

$$\log(kR) \leq 2 \log R \leq \frac{k \log 2}{2}.$$

Together with (67), $d \geq 1$, and $\binom{k}{k/4} \leq 2^k$, this gives

$$\begin{aligned} \log(4Q_d(M_*)S_d(M)2^d) &\leq (2A_\alpha + 3)kd, \\ \log(4Q_d(M_*)) &\leq (A_\alpha + 2)kd, \\ \log\left(Q_d(M_*)\binom{k}{k/4}\right) &\leq (A_\alpha + 2)kd. \end{aligned}$$

Choose $\alpha \geq 128$ sufficiently large that

$$\frac{c_0\alpha}{128} \geq 2A_\alpha + 4, \quad \frac{\alpha}{256} \geq A_\alpha + 3, \quad (68)$$

where $c_0 = \log(3/2)/4$. Such a universal choice exists because $A_\alpha = O(\log \alpha)$.

In the union regime, (56) and the preceding estimates imply

$$\Pr(\text{Cap} \cap \text{Master}^c) \leq 2e^{-kd}.$$

Indeed, the first term in (56) is at most e^{-kd} because

$$\frac{c_0ub}{64} \geq \frac{c_0\alpha kd}{128} \geq (2A_\alpha + 4)kd,$$

while the second is at most e^{-kd} because

$$\frac{2uT}{3} \geq \frac{\alpha kdR}{3} \geq \frac{\alpha kd}{256} \geq (A_\alpha + 3)kd.$$

In the all-round regime, **Master** holds on every outcome, so the same probability bound holds in both regimes.

Similarly, (64) and (68) give

$$\Pr(\text{Joint}^c) \leq Q_d(M_*)\binom{k}{k/4}e^{-\alpha kd/256} \leq e^{-kd}.$$

Finally, by (49), $k \leq R$, and $\alpha \geq 128$,

$$\Pr(\text{Cap}^c) \leq ke^{-\alpha dR/3} \leq e^{-R}.$$

Combining these bounds,

$$\begin{aligned} \Pr(\text{Good}^c) &\leq \Pr(\text{Cap}^c) + \Pr(\text{Cap} \cap \text{Master}^c) + \Pr(\text{Joint}^c) \\ &\leq e^{-R} + 3e^{-kd} < \frac{1}{16}, \end{aligned}$$

where the last inequality uses $R > 32$ and $kd \geq 24$. The choice of α and all probability bounds are independent of the input rule and the law of its independent auxiliary seed. ■

By Lemma 19, on **Cap** each target-driven replay can be reproduced by a sign schedule of depth at most d on the corresponding pool's capped selection history. Thus Proposition 21 provides at least $3k/4$ accurate schedules for every target on the same event **Good**.

6 Robust Identification and Eligible Parents

We use the parameters fixed in Section 5, in the regime $b > 32\alpha d$. A singleton class is covered exactly by its unique member, so assume $|\mathcal{H}| \geq 2$. All simulations in this section use the capped pools. Each selected list therefore has length at most m , including on outcomes outside **Cap**.

The preceding section provides accurate replay histories for every target. We now consider sign schedules whose version spaces reduce to singletons. To obtain useful counting bounds, we additionally require the surviving function to satisfy every correction inequality with a prescribed margin.

6.1 Robust Identification and Local Counting

For each integer $0 \leq r \leq d$, let $\mathcal{J}_r(m)$ denote the formal sign schedules of depth at most r with selected ordinals in $[m]$. Thus $|\mathcal{J}_r(m)| = Q_r(m)$. In each capped pool j , simulate every $L \in \mathcal{J}_d(m)$ from the initial version space $\mathcal{H}$, using the rules of Section 4.2. Write $V_t^{j,L}$ for the version space after round t and $p_t^{j,L}$ for the final prediction on that round. An invalid simulation has $V_t^{j,L} = \emptyset$ and $p_t^{j,L} = 0$ from the invalidating update onward. These are simulations of prescribed sign schedules: they do not use master reference values or target values.

The inequalities imposed by a schedule. Consider a schedule $L = ((s_1, \sigma_1), \dots, (s_i, \sigma_i))$ that is valid and born by round t . Recall that this means that all its prescribed updates have occurred without producing an empty version space. For each update $\ell \in [i]$, let $t_\ell \leq t$ be its actual round and let a_ℓ be the center immediately before that update. Several t_ℓ may coincide; their centers are recomputed after each update. By the definition of the sign-update fibers,

$$V_t^{j,L} = \left\{ f \in \mathcal{H} : \begin{array}{ll} f(x_{t_\ell}) \geq a_\ell + \eta & \text{for every } \ell \text{ with } \sigma_\ell = +, \\ f(x_{t_\ell}) \leq a_\ell - \eta & \text{for every } \ell \text{ with } \sigma_\ell = - \end{array} \right\}. \quad (69)$$

Here f ranges over candidate functions in $\mathcal{H}$. We call these inequalities the *update constraints*. Their thresholds are computed from the instance history, the pool selections, and L ; they are not additional entries in the code.

Definition 22 (Robust certificate). *The simulation of L in pool j is a robust certificate for $h \in \mathcal{H}$ by round t if it is valid and born by t , its version space is $V_t^{j,L} = \{h\}$, and every update satisfies*

$$\begin{cases} h(x_{t_\ell}) \geq a_\ell + 3\eta, & \text{if } \sigma_\ell = +, \\ h(x_{t_\ell}) \leq a_\ell - 3\eta, & \text{if } \sigma_\ell = -. \end{cases} \quad (70)$$

Thus, for a plus update, $h(x_{t_\ell})$ exceeds the retained boundary $a_\ell + \eta$ by at least 2η ; for a minus update it lies at least 2η below $a_\ell - \eta$. The test first examines the simulation's version space. If this space is a singleton, its unique member specifies the function h whose values are used in (70). No externally supplied target labels are needed. These conditions are defined set-theoretically; we do not require them to be efficiently computable.

Identification and activation. Say that pool j *identifies* h by round t if one of its simulations is a robust certificate for h by that round. Let

$$\mathcal{I}_j(t) := \{h \in \mathcal{H} : \text{pool } j \text{ identifies } h \text{ by round } t\}. \quad (71)$$

Identification is tested after the schedule updates on round t . Once a schedule is born, its version space and update constraints no longer change. Its certificate therefore persists. In particular, $\mathcal{I}_j(t) \subseteq \mathcal{I}_j(t+1)$, and each schedule can certify at most one function over the entire run. Hence $|\mathcal{I}_j(t)| \leq Q_d(m)$. At time zero only the empty schedule is born, and $|\mathcal{H}| \geq 2$, so $\mathcal{I}_j(0) = \emptyset$.

A function h *activates* on the first round when at least $k/2$ pools have identified it. Formally, define

$$\tau(h) := \min \{t \in [T] : \#\{j \in [k] : h \in \mathcal{I}_j(t)\} \geq k/2\}, \quad (72)$$

with $\tau(h) = \infty$ if the set is empty. Write $\mathcal{A}_t := \{h \in \mathcal{H} : \tau(h) \leq t\}$ and $\mathcal{A} := \mathcal{A}_T$. These sets are finite, since $\mathcal{A}_t \subseteq \bigcup_{j=1}^k \mathcal{I}_j(t)$.

Singleton replay states. Write $U_{t,j}^h$ for the version space of the target-driven replay for h in raw pool j after round t . On **Cap**, record only its corrections through round t , obtaining a prefix schedule in $\mathcal{J}_d(m)$. In capped pool j , this prefix schedule is valid and born by t and reproduces the replay through that round. Lemma 19 gives the strict versions of (70) for all its updates. Consequently,

$$\text{on Cap,} \quad U_{t,j}^h = \{h\} \implies h \in \mathcal{I}_j(t). \quad (73)$$

This implication uses the buffered update inequalities, not the accuracy of the replay on unselected rounds.

Local identification bounds. The separation in (70) limits how many functions a pool can identify within a subclass of small $\text{sfat}_\eta(V)$. The bound holds for every realized capped history and does not require **Good**.

Lemma 23 (Local identification bound). *On every outcome, for every $t \in \{0, \dots, T\}$, pool $j \in [k]$, and nonempty subclass $V \subseteq \mathcal{H}$ with $\text{sfat}_\eta(V) \leq r$, where $r \in \{0, \dots, d\}$,*

$$|\mathcal{I}_j(t) \cap V| \leq Q_r(m). \quad (74)$$

Proof. Fix the outcome, t , j , and V . We first show that any two distinct functions $h, h' \in \mathcal{I}_j(t)$ differ by more than 2η at some round selected by capped pool j by time t . Choose a certificate for h . Since it is born by round t and its version space is $\{h\}$, the function h' must violate an update constraint imposed at some round $s \leq t$ selected by capped pool j .

If this is a plus update at center a , then

$$h'(x_s) < a + \eta, \quad h(x_s) \geq a + 3\eta,$$

so $h(x_s) - h'(x_s) > 2\eta$. The minus case is symmetric.

Let $z_1, \dots, z_n$ be the instances in that list through round t , in order, so $n \leq m$. If $n = 0$, no pool schedule other than the empty schedule is born, and $\mathcal{I}_j(t) = \emptyset$. Otherwise, apply Lemma 13 to the initial class V and the sequence $z_1, \dots, z_n$, selecting every position. It gives at most $Q_r(m)$ prediction vectors such that every $f \in V$ is within η of one vector at every position. Two distinct functions in $\mathcal{I}_j(t) \cap V$ cannot use the same vector: the triangle inequality would make their values differ by at most 2η at every selected position, contrary to the separation just proved. This proves (74). $\blacksquare$

Corollary 24 (Local activation bound). *Under the same hypotheses on V and r , on every outcome and for every $t \in \{0, \dots, T\}$,*

$$|\mathcal{A}_t \cap V| \leq 2Q_r(m). \quad (75)$$

In particular, $|\mathcal{A}| \leq 2Q_d(m)$.

Proof. Every $h \in \mathcal{A}_t$ has been identified by at least $k/2$ pools by time t , since identification persists after activation. Counting the pairs (h, j) with $h \in \mathcal{A}_t \cap V$ and $h \in \mathcal{I}_j(t)$ gives

$$\frac{k}{2}|\mathcal{A}_t \cap V| \leq \sum_{j=1}^k |\mathcal{I}_j(t) \cap V| \leq kQ_r(m).$$

The first assertion follows by division, and the second by taking $t = T$, $V = \mathcal{H}$, and $r = d$. $\blacksquare$

In particular, V may be a version space produced during the run; it need not be fixed before the experiment.

6.2 Parent and Child Experts

We preallocate two types of prediction streams: *parents*, which follow fixed sign schedules, and *children*, which may later be assigned to activated functions. For every pool and every schedule of depth at most $d - 1$, retain its prediction stream as a parent. Thus the set of parent indices is

$$\mathcal{P} := [k] \times \mathcal{J}_{d-1}(m), \quad P := |\mathcal{P}| = kQ_{d-1}(m). \quad (76)$$

For each $v = (j, L) \in \mathcal{P}$, let p_t^v denote the prediction of schedule L in capped pool j at round t , and call the depth of L the depth of v . The parent always follows this schedule. It outputs zero if the schedule becomes invalid.

If a depth- i parent becomes valid and born, let V_v be its version space after birth. This space is fixed for the rest of the run, and the rank-drop property gives

$$\text{sfat}_\eta(V_v) \leq d - i. \quad (77)$$

Identification considers all schedules in $\mathcal{J}_d(m)$, whereas parent indices are allocated only to schedules in $\mathcal{J}_{d-1}(m)$.

Children and their outputs. For each $i \in \{0, \dots, d - 1\}$, fix a nonnegative integer c_i before the experiment. Attach c_i additional prediction indices to every depth- i parent v , and call them its *children*, indexed by (v, ℓ) for $\ell \in \{1, \dots, c_i\}$. Each child initially copies its parent's outputs. On a round when a function h activates, at most one unused child may be assigned to h . From that round onward it predicts $h(x_s)$ on every subsequent instance x_s , and it is never reassigned.

Assignments on round t are made after the capped schedule updates and identification tests, but before emitting the child predictions on that round. Thus

$$g_t^{v,\ell} := \begin{cases} h(x_t), & \text{if this child has been assigned to } h \text{ by round } t, \\ p_t^v, & \text{if this child is still unassigned.} \end{cases} \quad (78)$$

In particular, assignment does not change any past prediction.

The capacities c_i and the rule choosing the assignments will be specified in the next section. Here we define which choices are permitted and prove their deterministic properties. For any such capacities, the number of parent and child indices is

$$K = P + \sum_{i=0}^{d-1} kn_i(m)c_i. \quad (79)$$

This counts invalid schedules and unused children as well. All these indices are present from the start. The other schedule simulations are internal calculations, not additional members of the covering family.

6.3 Eligibility and Many Compatible Parents

An unused child can cover a newly activated function h if its parent has already made few errors against h : after assignment, the child predicts h exactly. We therefore test the parent's actual predictions before the activation round. For $v \in \mathcal{P}$, $h \in \mathcal{H}$, and $t \in [T]$, write

$$E_{t-1}(v, h) := \sum_{s=1}^{t-1} \mathbf{1}\{|p_s^v - h(x_s)| \geq 8\eta\}. \quad (80)$$

The sum is zero at $t = 1$.

Definition 25 (Eligible parent). *Suppose h activates on round t . A parent v is eligible for this activation if its schedule is valid and born by round $t - 1$, and*

$$h \in V_v, \quad |V_v| \geq 2, \quad E_{t-1}(v, h) < b/8. \quad (81)$$

Let $\mathcal{E}_j(h, t)$ be the set of eligible parents in pool j . Fix an arbitrary ordering of the sign schedules in $\mathcal{J}_{d-1}(m)$. For each pool j with $\mathcal{E}_j(h, t) \neq \emptyset$, choose the eligible parent whose schedule comes first in this ordering. Let $\mathcal{E}(h, t)$ be the set of these chosen parents.

We permit an activation of h on round t to use only a parent from $\mathcal{E}(h, t)$, and only an unused child of that parent.

The sets in this definition can be determined on round t . The function h is known from its certificates at activation; its values on earlier instances can therefore be evaluated. The parent states and all predictions used in (80) are already available. Neither the test nor the choice of a representative from each pool uses future instances or future accuracy. In particular, **Good** is not part of the eligibility test. Eligibility also does not depend on whether a parent's children have already been used; a parent with no unused child still belongs to the same eligible sets.

If an unused child of $v \in \mathcal{E}(h, t)$ is assigned to h on round t , then (78) gives

$$\sum_{s=1}^T \mathbf{1}\{|g_s^{v,\ell} - h(x_s)| \geq 8\eta\} = E_{t-1}(v, h) < b/8. \quad (82)$$

Indeed, before t the child copied its parent, and from t onward it predicts h exactly. Requiring assignments to use only the information just described makes these prediction rules causal in the instances and the revealed random choices.

There is also a uniform bound on how many distinct activations can have a given parent among their choices. Fix a valid, born parent v of depth i . Every function h that activates and has $v \in \mathcal{E}(h, \tau(h))$ belongs to $\mathcal{A} \cap V_v$, since eligibility requires $h \in V_v$. By the local activation bound, the number of such functions is at most $2Q_{d-i}(m)$. Hence (77) and Corollary 24 give

$$\#\{h \in \mathcal{A} : v \in \mathcal{E}(h, \tau(h))\} \leq |\mathcal{A} \cap V_v| \leq 2Q_{d-i}(m). \quad (83)$$

A parent that never becomes valid and born is never eligible. This bound holds on every outcome, not only on **Good**, and allows the later input instances to depend on earlier assignment decisions.

Many eligible parents at activation. We now use the simultaneous replay accuracy from Proposition 21 to show that the eligible sets are large on **Good**.

Lemma 26 (Many eligible parents). *On **Good**, every function h that activates on a round $t \in [T]$ satisfies*

$$|\mathcal{E}(h, t)| \geq k/4. \quad (84)$$

Proof. Fix an outcome in **Good** and a function h activating on round t . Let $G_h \subseteq [k]$ be the pools whose replays satisfy (65). By Proposition 21, $|G_h| \geq 3k/4$. Before round t , fewer than $k/2$ pools had identified h ; otherwise it would already have activated. Therefore

$$\#\{j \in G_h : h \notin \mathcal{I}_j(t-1)\} \geq |G_h| - \#\{j : h \in \mathcal{I}_j(t-1)\} > k/4.$$

Fix a pool j in this set, and let L consist of all corrections made by its target-driven replay through round $t-1$. On **Cap** these have ordinals in $[m]$, and the capped-pool simulation of L is valid, born by $t-1$, and has version space $U_{t-1,j}^h$. This space contains h . It cannot equal $\{h\}$, since (73) would imply $h \in \mathcal{I}_j(t-1)$.

Since $\mathcal{H}$ is uniformly greater-than- η separated, every nonsingleton subclass of $\mathcal{H}$ has rank at least one. If L has depth i , the rank-drop property consequently gives

$$1 \leq \text{sfat}_\eta(U_{t-1,j}^h) \leq d - i, \quad \text{so } i \leq d - 1.$$

Thus $v = (j, L)$ is one of the preallocated parents. Its fixed version space is $V_v = U_{t-1,j}^h$, which contains h and is not a singleton. By the prefix property in Section 4.2,

$$p_s^v = y_{s,j}^h \quad (1 \leq s < t).$$

Since $j \in G_h$, it follows that

$$E_{t-1}(v, h) = \sum_{s < t} \mathbf{1}\{|y_{s,j}^h - h(x_s)| \geq 8\eta\} \leq \sum_{s=1}^T \mathbf{1}\{|y_{s,j}^h - h(x_s)| \geq 8\eta\} < b/8.$$

All eligibility conditions hold, so $\mathcal{E}_j(h, t)$ is nonempty. This applies to more than $k/4$ distinct pools. Since $\mathcal{E}(h, t)$ contains one eligible parent from each such pool, we obtain (84). ■

The parent used in this proof is the separate schedule containing only the replay's corrections before round t . Its version space remains fixed, even if the full replay makes further corrections or identifies h on round t . Only their predictions before t must agree: a child assigned on round t switches to h before emitting its prediction on that round.

Corollary 27 (Targets that do not activate). *On Good, every $h \in \mathcal{H} \setminus \mathcal{A}$ has a parent $v \in \mathcal{P}$ with*

$$\sum_{t=1}^T \mathbf{1}\{|p_t^v - h(x_t)| \geq 8\eta\} < b/8. \quad (85)$$

Proof. Fewer than $k/2$ pools identify h by time T , while at least $3k/4$ replays satisfy (65). Choose a pool j with an accurate replay that has not identified h . Its final replay version space is not a singleton, by (73). The same rank argument as above shows that its entire replay code J_j^h has depth at most $d-1$. On **Cap** this code has ordinals in $[m]$, so $v = (j, J_j^h)$ is a parent and reproduces the replay on all T rounds. Its final-prediction accuracy gives (85). ■

Thus a target that never activates is covered by a permanent parent. A target that activates has at least $k/4$ eligible parents on **Good**, and assigning an unused child of any one of them gives (82). It remains to choose the capacities and distribute the activations among these parents so that, with positive probability jointly with **Good**, every activation receives an unused child and the count in (79) is at most $(CR)^d$.

7 Shared-Parent Routing and Load Balancing

We now specify how an activating function receives a child of one of its eligible parents. The rule chooses one parent uniformly from the retained set $\mathcal{E}(h, t)$ when this set has at least $k/4$ members. We bound how often a parent is chosen and use this bound to set the child capacities c_i .

7.1 The Causal Routing Protocol

For now, let the capacities $c_0, \dots, c_{d-1}$ be any fixed nonnegative integers, as in Section 6.2. By Corollary 24, the total number of functions that activate during a run is at most

$$A_* := 2Q_d(m). \quad (86)$$

We allow A_* routing slots on each round, where one slot processes one newly activated function; slots with no function do nothing. Fix an arbitrary total order on $\mathcal{H}$. In each round, process the newly activated functions in this order.

For $t \in [T]$ and $a \in [A_*]$, let $\pi_{t,a}$ be a uniformly random permutation of the finite parent set $\mathcal{P}$. All these permutations are mutually independent and independent of the raw pool bits. A permutation specifies an ordering of all parents. For an activation with $|\mathcal{E}(h, t)| \geq k/4$, conditional on the preceding history, each member of $\mathcal{E}(h, t)$ is selected with probability $1/|\mathcal{E}(h, t)|$. The permutation $\pi_{t,a}$ is revealed only in slot a of round t , including when that slot has no function to process.

Write $\boldsymbol{\pi}_t = (\pi_{t,1}, \dots, \pi_{t,A_*})$ and $\boldsymbol{\pi}_{<t} = (\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_{t-1})$. We allow inputs of the form

$$x_t = \tilde{F}_t(\boldsymbol{\xi}_{<t}, \boldsymbol{\pi}_{<t}, \omega_{\text{in}}), \quad (87)$$

where the maps $\tilde{F}_t$ are deterministic and ω_{in} is a finite-support input seed independent of all pool bits and routing permutations. Thus x_t may depend on all earlier selections, assignments, and predictions, but is chosen before the current pool vector and current routing permutations are revealed.

Initialize a counter $L_v = 0$ for each parent $v \in \mathcal{P}$. A parent remains eligible after all its children have been assigned. If routing selects such a parent, its load counter increases and an overload is declared, but no child is assigned. At round t , reveal $\boldsymbol{\xi}_t$, perform the capped schedule updates, and determine the newly activated functions $\mathcal{A}_t \setminus \mathcal{A}_{t-1}$. List them in the fixed order as $h_{t,1}, \dots, h_{t,n_t}$, where $n_t \leq A_*$. The eligible set $\mathcal{E}(h_{t,a}, t)$ does not depend on the current routing permutation $\pi_{t,a}$.

In each slot $a \in [A_*]$, reveal $\pi_{t,a}$ and proceed as follows. If $a > n_t$, make no change. Otherwise put $h = h_{t,a}$. If $|\mathcal{E}(h, t)| < k/4$, leave h unassigned. If $|\mathcal{E}(h, t)| \geq k/4$, choose the first parent v in $\mathcal{E}(h, t)$ according to $\pi_{t,a}$ and increase L_v by one. Write ℓ for its updated value and i for the depth of v . When $\ell \leq c_i$, assign child (v, ℓ) to h . When $\ell > c_i$, make no child assignment; this is an *overload* at v .

The first c_i choices of a depth- i parent use its children in order, so every assigned child is unused at assignment. Each function is processed only on its activation round and only once. After all slots of round t , emit the parent predictions and the child predictions specified in (78).

We continue the selections and counters after an overload, and also after an activation is left unassigned. Thus L_v counts routing choices, not just successful child assignments. On Good, the case $|\mathcal{E}(h, t)| < k/4$ cannot occur, by Lemma 26.

The sampling guarantee from the preceding sections applies to (87). Indeed, the auxiliary seed

$$\omega_0 := (\omega_{\text{in}}, (\pi_{t,a})_{t \in [T], a \in [A_*]})$$

has finite support and is independent of all raw pool bits. Given this seed, every earlier assignment is determined by the past raw vectors. Hence the input rule has the form (50), and Proposition 21 gives $\Pr(\text{Good}) > 15/16$ for the routing protocol just defined.

7.2 Routing Probabilities and Load Bounds

Fix a parent v . Consider a routing step (t, a) that processes a newly activated function $h_{t,a}$. Define

$$B_{t,a}^v := \mathbf{1}\{|\mathcal{E}(h_{t,a}, t)| \geq k/4 \text{ and } v \in \mathcal{E}(h_{t,a}, t)\}, \quad (88)$$

and let $X_{t,a}^v$ be the indicator that the routing rule chooses v in this slot, whether or not a child is assigned. For an empty slot, set both indicators to zero. Thus

$$N_v := \sum_{t=1}^T \sum_{a=1}^{A_*} B_{t,a}^v \quad \text{and} \quad L_v = \sum_{t=1}^T \sum_{a=1}^{A_*} X_{t,a}^v$$

count, respectively, the activations at which v could be chosen and the activations at which it is chosen.

If v has depth i , then on every outcome

$$N_v \leq \#\{h \in \mathcal{A} : v \in \mathcal{E}(h, \tau(h))\} \leq 2Q_{d-i}(m). \quad (89)$$

The first inequality holds because each function activates only once; the size test in (88) can only remove activations from the count. The second inequality is (83). A parent that never becomes valid and born has $N_v = L_v = 0$.

To describe the information available at a routing choice, let

$$\begin{aligned} \mathcal{F}_{t,0} &:= \sigma(\omega_{\text{in}}, \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_t, \boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_{t-1}), \\ \mathcal{F}_{t,a} &:= \sigma(\mathcal{F}_{t,0}, \pi_{t,1}, \dots, \pi_{t,a}) \quad (1 \leq a \leq A_*). \end{aligned}$$

Also put $\mathcal{F}_{0,A_*} = \sigma(\omega_{\text{in}})$. These sigma-fields reveal the raw vector first on each round and then the routing permutations one at a time. The functions $h_{t,a}$ and their eligible sets are already determined by $\mathcal{F}_{t,0}$. In particular, $B_{t,a}^v$ is known given $\mathcal{F}_{t,a-1}$, while $\pi_{t,a}$ remains an independent uniform permutation. Consequently,

$$\Pr\{X_{t,a}^v = 1 \mid \mathcal{F}_{t,a-1}\} = \begin{cases} |\mathcal{E}(h_{t,a}, t)|^{-1}, & B_{t,a}^v = 1, \\ 0, & B_{t,a}^v = 0 \end{cases} \leq \frac{4}{k} B_{t,a}^v. \quad (90)$$

For the first case, each member of the eligible set is equally likely to appear first in the fresh permutation. This remains true when v has no unused children, since routing still chooses from the same set and increments L_v .

Load supermartingale and child capacities. We combine the bound on how often v can be considered with the conditional probability that it is chosen. The routing indicators $X_{t,a}^v$ need not be independent across rounds, since later inputs and eligible sets may depend on earlier routing choices.

Lemma 28 (Exponential moment of a parent load). *For every fixed parent v of depth i , every fixed choice of capacities, and every input rule (87),*

$$\mathbb{E} e^{L_v} \leq \exp\left(\frac{8(e-1)Q_{d-i}(m)}{k}\right). \quad (91)$$

Proof. Let $N_v(t, a)$ and $L_v(t, a)$ be the two counts through slot a of round t , including all earlier rounds. Both counts are initially zero. Here $a = 0$ denotes the start of round t , before any routing steps, and $a = A_*$ denotes the end of that round, after all routing steps. Thus, for $t \geq 2$,

$$N_v(t, 0) = N_v(t-1, A_*), \quad L_v(t, 0) = L_v(t-1, A_*).$$

Initially, $N_v(1, 0) = L_v(1, 0) = 0$. Set $\lambda := 4(e-1)/k$ and define

$$Z_{t,a} := \exp(L_v(t, a) - \lambda N_v(t, a)). \quad (92)$$

Its initial value is one. At a raw-vector reveal it is unchanged. At a routing slot, the increments of L_v and N_v are $X_{t,a}^v$ and $B_{t,a}^v$, respectively, so

$$\frac{Z_{t,a}}{Z_{t,a-1}} = \exp(X_{t,a}^v - \lambda B_{t,a}^v).$$

Since $X_{t,a}^v \in \{0, 1\}$ and $B_{t,a}^v$ is $\mathcal{F}_{t,a-1}$ -measurable, (90) gives

$$\begin{aligned} \mathbb{E}[e^{X_{t,a}^v} \mid \mathcal{F}_{t,a-1}] &= 1 + (e-1) \Pr\{X_{t,a}^v = 1 \mid \mathcal{F}_{t,a-1}\} \\ &\leq 1 + \lambda B_{t,a}^v \leq e^{\lambda B_{t,a}^v}. \end{aligned}$$

It follows that

$$\mathbb{E}\left[\frac{Z_{t,a}}{Z_{t,a-1}} \mid \mathcal{F}_{t,a-1}\right] \leq 1.$$

Thus Z is a nonnegative supermartingale along the finite sequence of raw-vector reveals and routing slots, with respect to the sigma-fields just defined. In particular, $\mathbb{E}Z_{T,A_*} \leq 1$.

By (89), $N_v \leq 2Q_{d-i}(m)$ on every outcome. Therefore

$$\begin{aligned} \mathbb{E}e^{L_v} &= \mathbb{E}[Z_{T,A_*} e^{\lambda N_v}] \\ &\leq e^{2\lambda Q_{d-i}(m)} \mathbb{E}Z_{T,A_*} \\ &\leq \exp\left(\frac{8(e-1)Q_{d-i}(m)}{k}\right). \end{aligned}$$

■

We can now choose the number of children for each parent. The exponential moment bound turns an additive $\log(4P)$ in the capacity into a failure probability of at most $1/(4P)$ for each parent. Since $e-1 < 2$, set

$$c_i := \left\lceil \frac{16Q_{d-i}(m)}{k} + \log(4P) \right\rceil, \quad 0 \leq i \leq d-1. \quad (93)$$

These values depend only on the parameters fixed before the experiment.

Proposition 29 (No overload). *With capacities (93), define*

$$\text{NoOverload} := \{L_v \leq c_i \text{ for every depth-}i \text{ parent } v \in \mathcal{P}, 0 \leq i \leq d-1\}. \quad (94)$$

For every input rule (87),

$$\Pr(\text{NoOverload}^c) \leq \frac{1}{4}. \quad (95)$$

Proof. Fix a parent v of depth i . Markov's inequality and Lemma 28 imply

$$\begin{aligned} \Pr\{L_v > c_i\} &\leq e^{-c_i} \mathbb{E} L_v \\ &\leq \exp\left(-c_i + \frac{8(e-1)Q_{d-i}(m)}{k}\right) \\ &\leq \exp\left(-c_i + \frac{16Q_{d-i}(m)}{k}\right) \leq \frac{1}{4P}. \end{aligned}$$

Taking a union bound over the P parent indices proves (95). ■

The overload bound is under the full joint law of sampling and routing, without conditioning on **Cap** or **Good**. On **Good** $\cap$ **NoOverload**, every activating function receives an unused child on its activation round: its eligible set has at least $k/4$ members, and the selected parent's counter never exceeds its capacity.

8 Expert Counting and Global Determinization

For the counting and probability arguments, retain the separated class $\mathcal{H}$ and the regime $d \geq 1$, $|\mathcal{H}| \geq 2$, and $b > 32\alpha d$ from the preceding section. The child capacities are now fixed by (93).

8.1 Expert Counting and Simultaneous Coverage

Lemma 30 (Expert count). *There is a universal constant $C_{\text{cnt}} \geq 1$ such that the number K of preallocated parent and child indices satisfies*

$$K \leq (C_{\text{cnt}} R)^d. \quad (96)$$

Proof. There are $kn_i(m)$ parents of depth i , each with c_i children. Using (79), (93), and $\lceil z \rceil \leq z+1$, we obtain

$$\begin{aligned} K &= P + \sum_{i=0}^{d-1} kn_i(m)c_i \\ &\leq 16 \sum_{i=0}^{d-1} n_i(m)Q_{d-i}(m) + P(\log(4P) + 2) \\ &\leq 16Q_d(2m) + P(\log(4P) + 2). \end{aligned} \quad (97)$$

The factor k from the $kn_i(m)$ depth- i parents cancels with the $1/k$ in the leading term of c_i . The term $P(\log(4P) + 2)$ accounts for the permanent parents, the logarithmic capacity overhead, and the rounding overhead. The last inequality is (37).

Put $C_1 := 14e\alpha$. Since $m \leq 3\alpha dR$, (35) gives

$$Q_d(2m) \leq [2e(1 + 6\alpha R)]^d \leq (C_1 R)^d.$$

For $d \geq 2$, the inequality $d/(d-1) \leq 2$ similarly gives

$$Q_{d-1}(m) \leq \left[2e \left(1 + \frac{3\alpha dR}{d-1}\right)\right]^{d-1} \leq (C_1 R)^{d-1}.$$

For $d = 1$, the same final bound holds because $Q_0(m) = 1$. Thus, in all cases,

$$P = kQ_{d-1}(m) \leq k(C_1R)^{d-1}. \quad (98)$$

We next bound the second term in (97). The definition $k = 4\lceil \log_2 R \rceil$ and $R > 32$ imply $k \leq 8 \log R$, and (66) gives $k \leq R$. Consequently,

$$\log(4P) + 2 \leq \log(4k) + 2 + (d-1)\log(C_1R) \leq C_2d \log R,$$

where we may take the universal constant $C_2 := 1 + (\log(4C_1) + 2)/\log 32$, using $\log k \leq \log R$ and $\log R > \log 32$. It follows that

$$\begin{aligned} P(\log(4P) + 2) &\leq 8C_2d(\log R)^2(C_1R)^{d-1} \\ &\leq 8C_2dR(C_1R)^{d-1} \\ &\leq (C_3R)^d \end{aligned} \quad (99)$$

for a universal C_3 . The second inequality uses $(\log R)^2 \leq R$ for $R > 32$; the last follows from $d \leq 2^d$, for example with $C_3 = 2 \max\{C_1, 8C_2\}$. Combining these bounds in (97), we may take $C_{\text{cnt}} = 16C_1 + C_3$. All these constants are universal because α was fixed universally in (68). ■

Simultaneous coverage. Number the K preallocated parent and child indices by $\nu \in [K]$, and let w_t^ν denote the prediction of index ν at round t under the randomized sampling and routing protocol.

Lemma 31 (Simultaneous coverage). *For every input rule (87), with probability greater than $1/2$, the outputs satisfy*

$$\forall h \in \mathcal{H} \exists \nu \in [K] : \sum_{t=1}^T \mathbf{1}\{|w_t^\nu - h(x_t)| \geq 8\eta\} < b/8. \quad (100)$$

Proof. The application of Proposition 21 in Section 7.1 and Proposition 29 give

$$\begin{aligned} \Pr(\text{Good} \cap \text{NoOverload}) &\geq 1 - \Pr(\text{Good}^c) - \Pr(\text{NoOverload}^c) \\ &> 1 - \frac{1}{16} - \frac{1}{4} = \frac{11}{16} > \frac{1}{2}. \end{aligned} \quad (101)$$

Fix an outcome in this intersection and $h \in \mathcal{H}$. If h activates, the routing protocol assigns it an unused child of a parent in $\mathcal{E}(h, \tau(h))$, as proved at the end of Section 7. By (82), this child's index supplies the required witness for h . If h does not activate, Corollary 27 supplies a parent with the same error bound. In either case, the witnessing index belongs to the preallocated family. Since h was arbitrary, the one event in (101) gives the stated coverage for every target. ■

8.2 Game Determinization and Main Upper Bound

Lemma 31 gives successful random choices for each fixed input rule. Choosing one successful realization separately for each instance sequence, however, need not give the same predictions on shared prefixes. To obtain a global family, we need one deterministic rule for choosing the pool selections and routing permutations after each observed input prefix. The guarantee against adaptive input rules allows us to obtain such a rule by the following game argument.

Lemma 32 (Finite-horizon determinization). *Consider a game of T rounds between an adversary and a constructor. In each round the adversary first chooses an input from a nonempty set, and the constructor then chooses an action from a nonempty set. The legal choices may depend on the preceding moves. Both players observe all earlier moves, and the constructor also observes the current input. Each terminal history is designated winning or losing for the constructor.*

If one randomized constructor strategy wins with positive probability against every deterministic adversary strategy, then there is a deterministic constructor strategy that wins against every adversary strategy.

Proof. Label each terminal history according to the prescribed winning condition. Working backwards through nonterminal histories, label a constructor decision node as winning if some legal action leads to a winning successor. Label a history at which the adversary is to move as winning if every legal input leads to a winning history. All other histories are losing. The recursion is well-defined because the game has finite depth, even if its branching is infinite.

Suppose that the initial history were losing. At each losing history where the adversary moves, there is an input leading to a losing history. Choose one such input at each of these histories, and choose arbitrary legal inputs at the adversary's other histories. At a losing history where the constructor moves, every legal action leads to a losing history. The chosen adversary strategy therefore forces a losing terminal history for every realization of the constructor's random choices. Its winning probability against this deterministic adversary would be zero, a contradiction.

The initial history is consequently winning. At each winning history where the constructor moves, fix an action leading to a winning history; specify arbitrary legal actions at its other histories. At every winning history where the adversary moves, all legal inputs preserve the winning status. These choices define a deterministic constructor strategy that wins against all adversary moves. The choices over arbitrary collections of histories may use the axiom of choice. ■

Remark 33 (Role of the information structure). *The determinization argument relies on perfect information: before choosing the next input, the adversary observes the constructor's realized actions, not merely their distributions. This differs from distribution-observing protocols in randomized online learning with set-valued or partial feedback [RST24; Sha+26]. Without access to the realized actions, the conclusion of Lemma 32 can fail, even in a two-round game. For example, suppose the constructor secretly chooses a fair bit and wins if it equals the adversary's second-round input. The constructor then wins with probability $1/2$ against every deterministic adversary, but no deterministic constructor strategy can guarantee a win.*

Proof of Proposition 2. The case $d = 0$ was proved in Section 4.1; a singleton $\mathcal{H}$ is covered exactly by its unique member. For $d \geq 1$ and $b \leq 32\alpha d$, the deterministic family in (46) already proves the proposition. We therefore consider the remaining regime and use the randomized construction counted in Lemma 30.

Let $\mathfrak{S}(\mathcal{P})$ be the finite set of permutations of $\mathcal{P}$. Define a game in which the adversary chooses $x_t \in \mathcal{X}$ on round t , and the constructor then chooses

$$(\xi_t, \pi_t) \in \{0, 1\}^k \times \mathfrak{S}(\mathcal{P})^{A*}.$$

Apply the capped schedule updates and routing protocol to these choices, processing the routing slots in their prescribed order, and emit the K predictions before the next adversary move. All these operations are determined by the inputs and chosen actions. Designate a terminal history as winning precisely when (100) holds.

Every deterministic adversary strategy in this game has the form (87) with a constant input seed: its earlier inputs, and hence its current choice, can be recovered recursively from the preceding

pool vectors and routing permutations. The randomized protocol therefore wins with probability greater than $1/2$ against every such strategy, by Lemma 31. Lemma 32 supplies a deterministic winning constructor strategy. Fix this strategy.

For any input prefix $x_{1:t}$, run the fixed strategy through round t . By induction on rounds, all pool vectors, routing permutations, assignments, and predictions through that round are determined by $x_{1:t}$. For each $\nu \in [K]$, define

$$g_{\nu,t}(x_{1:t}) := \text{the output of index } \nu \text{ on round } t \text{ in this run.}$$

Thus $g_{\nu,t} : \mathcal{X}^t \rightarrow [0, 1]$ is a well-defined deterministic causal map. The family

$$\mathcal{G} := \{(g_{\nu,t})_{t=1}^T : \nu \in [K]\}$$

is fixed independently of the instance sequence. Since the strategy wins against every sequence of adversary inputs,

$$\forall x_{1:T} \in \mathcal{X}^T \forall h \in \mathcal{H} \exists \nu \in [K] : \sum_{t=1}^T \mathbf{1}\{|g_{\nu,t}(x_{1:t}) - h(x_t)| \geq 8\eta\} < \frac{T}{8R}. \quad (102)$$

Finally, $|\mathcal{G}| \leq K \leq (C_{\text{cnt}}R)^d$. Choosing one universal $C_0 \geq \max\{1, C_{\text{cnt}}, 66e\alpha\}$ also includes the short-horizon bound (46), and proves the proposition in all regimes. $\blacksquare$

Proof of Theorem 5. Fix $1 \leq p < \infty$, $T \geq 1$, and $0 < \varepsilon < 1/2$, and put $\eta = \varepsilon/32$ and $d = \text{sfat}_\eta(\mathcal{F})$. If $d = 0$, the midpoint rule

$$g_t(x_{1:t}) := \frac{1}{2} \left(\inf_{f \in \mathcal{F}} f(x_t) + \sup_{f \in \mathcal{F}} f(x_t) \right)$$

has pointwise error at most $\eta/2$ against every $f \in \mathcal{F}$, by Lemma 10. Its normalized ℓ_p error is therefore at most $\eta/2 < \varepsilon$, and one expert suffices.

Suppose $d \geq 1$. Lemma 9 gives a separated subclass $\mathcal{H} \subseteq \mathcal{F}$ of rank at most d such that

$$\forall f \in \mathcal{F} \exists h \in \mathcal{H} : \|f - h\|_\infty \leq \eta.$$

Apply Proposition 2 with

$$R := \frac{16}{\varepsilon^p} > 32. \quad (103)$$

The resulting global family has size at most $(16C_0/\varepsilon^p)^d$.

Fix an instance sequence $x_{1:T}$ and a target $f \in \mathcal{F}$. Choose $h \in \mathcal{H}$ with $\|f - h\|_\infty \leq \eta$ and a witnessing expert g for h from Proposition 2. Writing

$$B := \{t \in [T] : |g_t(x_{1:t}) - h(x_t)| \geq 8\eta\},$$

we have $|B|/T < 1/(8R)$. For $t \notin B$,

$$|g_t(x_{1:t}) - f(x_t)| \leq |g_t(x_{1:t}) - h(x_t)| + |h(x_t) - f(x_t)| < 9\eta.$$

For $t \in B$, the error is at most one because both values lie in $[0, 1]$. Consequently,

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T |g_t(x_{1:t}) - f(x_t)|^p &\leq (9\eta)^p + \frac{|B|}{T} \\ &< \varepsilon^p \left[\left(\frac{9}{32} \right)^p + \frac{1}{128} \right] \\ &\leq \frac{37}{128} \varepsilon^p < \varepsilon^p. \end{aligned} \quad (104)$$

Here $(9/32)^p \leq 9/32$ because $p \geq 1$. Taking the p th root gives the required normalized ℓ_p error bound. The separated subclass and the expert family were fixed independently of f and $x_{1:T}$, so the same family satisfies Definition 4. Taking $C = 16C_0$ proves (19), with C independent of p , d , T , and ε . $\blacksquare$

The upper-bound assertions of Corollaries 7 and 8 now follow from the rounding and tree-restriction arguments in Section 3.4. Moreover, $C \geq 1$ and $p \geq 1$ give

$$\left(\frac{C}{\varepsilon^p}\right)^d \leq \left(\frac{C}{\varepsilon}\right)^{pd},$$

which also proves the upper-bound part of Theorem 1.

9 Matching Lower Bounds

We prove Theorem 6 using indicators of affine hyperplanes over a finite field. The instance sequence lists every point of the domain once. Related finite-field hyperplane classes have been studied through their static VC dimension [Asc+25]; here we require the Littlestone dimension of the full affine-hyperplane class and a fixed-sequence ℓ_p packing.

9.1 Finite-Field Hyperplanes and Their Dimensions

Fix an integer $d \geq 1$ and a prime power $q \geq 2$, and let $\mathbb{F}_q$ be a field with q elements. On the domain $\mathcal{X}_q := \mathbb{F}_q^d$, define

$$H_{a,b} := \{x \in \mathbb{F}_q^d : a \cdot x = b\}, \quad h_{a,b}(x) := \mathbf{1}\{x \in H_{a,b}\}, \quad (105)$$

where $a \in \mathbb{F}_q^d \setminus \{0\}$, $b \in \mathbb{F}_q$, and $a \cdot x = \sum_{j=1}^d a_j x_j$ is computed in $\mathbb{F}_q$. Let

$$\mathcal{H}_q := \{h_{a,b} : a \in \mathbb{F}_q^d \setminus \{0\}, b \in \mathbb{F}_q\}.$$

Each $H_{a,b}$ is an affine subspace of dimension $d - 1$ and has q^{d-1} points: if $a_j \neq 0$, choosing the other $d - 1$ coordinates freely determines x_j uniquely.

Two parameter pairs (a, b) and (a', b') define the same hyperplane exactly when $(a', b') = (\lambda a, \lambda b)$ for some $\lambda \in \mathbb{F}_q \setminus \{0\}$. Indeed, equal hyperplanes have the same direction subspace, so the nonzero linear forms $a \cdot x$ and $a' \cdot x$ have the same kernel and are proportional. Evaluating them at any point of the hyperplane then gives $b' = \lambda b$. The converse follows by multiplying the defining equation by λ . Thus each distinct hyperplane has exactly $q - 1$ parameter representations, and

$$|\mathcal{H}_q| = \frac{q(q^d - 1)}{q - 1} = q \sum_{j=0}^{d-1} q^j \geq q^d. \quad (106)$$

Lemma 34 (Dimension of the hyperplane class). *For every prime power $q \geq 2$ and integer $d \geq 1$,*

$$\text{Ldim}(\mathcal{H}_q) = \text{sfat}_\gamma(\mathcal{H}_q) = d \quad (0 < \gamma \leq 1). \quad (107)$$

Proof. For the upper bound on Littlestone dimension, consider a binary predictor that observes the true label after each prediction. It maintains the affine span S of all previously observed instances with label one, initially $S = \emptyset$, and predicts one exactly when the current instance belongs to S . After a positive label it replaces S by the affine span of S and the current instance; after a negative label it leaves S unchanged.

On any sequence labeled by the indicator of a fixed hyperplane H , we always have $S \subseteq H$. The predictor therefore never predicts one at a point labeled zero. Every mistake is a positive point outside S , whose addition increases the affine dimension by one. Assigning dimension -1 to the initially empty span, and using $\dim S \leq \dim H = d - 1$, we obtain at most d mistakes.

If $\mathcal{H}_q$ shattered a Littlestone tree of depth $d+1$, we could follow, at each node, the edge opposite the predictor's current prediction and provide that edge's label before proceeding to the next node. Some member of $\mathcal{H}_q$ realizes the resulting path. The predictor would then make $d+1$ mistakes on a sequence labeled by that member, contradicting the bound just proved. Hence $\text{Ldim}(\mathcal{H}_q) \leq d$.

For the lower bound, let $e_1, \dots, e_d$ be the standard basis vectors in $\mathbb{F}_q^d$. Every labeling of these points is realized by a hyperplane indicator. To see this, let $S \subseteq [d]$ be the indices to receive label one. If $S \neq [d]$, take $b = 0$ and

$$a_j = \begin{cases} 0, & j \in S, \\ 1, & j \notin S. \end{cases}$$

Then $a \neq 0$ and $h_{a,0}(e_j) = 1$ exactly for $j \in S$. If $S = [d]$, take $a = (1, \dots, 1)$ and $b = 1$; every e_j belongs to this hyperplane. The complete binary instance tree that queries e_j at every node of level j is therefore shattered. This gives $\text{Ldim}(\mathcal{H}_q) \geq d$.

Finally, $\mathcal{H}_q$ is binary, so the equality of sequential fat-shattering and Littlestone dimensions at scales $0 < \gamma \leq 1$, proved in Section 3.1, gives (107). $\blacksquare$

9.2 Fixed-Sequence Packing and Lower Bound

List all points of $\mathcal{X}_q$ once, in any fixed order, as $x_1, \dots, x_T$, where $T = q^d$. For distinct hyperplanes H, H' and their indicators h, h' , the normalized metric from (17) satisfies

$$\rho_{p,T}(h[x_{1:T}], h'[x_{1:T}])^p = \frac{|H \triangle H'|}{q^d} \quad (1 \leq p < \infty), \quad (108)$$

where $H \triangle H'$ is the set of points belonging to exactly one of the two hyperplanes. This identity holds because the absolute difference of two binary values, raised to the power p , is one precisely when they differ.

If the normal vectors of H and H' are proportional, the two distinct hyperplanes are parallel and disjoint. Each has q^{d-1} points, so the right-hand side of (108) is $2/q$. Otherwise the two normal vectors are linearly independent, which can occur only when $d \geq 2$. Their two defining equations have q^{d-2} common solutions, and hence

$$\frac{|H \triangle H'|}{q^d} = \frac{2q^{d-1} - 2q^{d-2}}{q^d} = \frac{2(q-1)}{q^2}.$$

For $d = 1$, hyperplanes are singleton sets, so the value $2/q$ applies to every distinct pair. Thus, for all $d \geq 1$,

$$\rho_{p,T}(h[x_{1:T}], h'[x_{1:T}])^p \geq \frac{2(q-1)}{q^2} \quad (h \neq h', h, h' \in \mathcal{H}_q). \quad (109)$$

Choice of the field size. To obtain pairwise $\rho_{p,T}$ -distance greater than 2ε , fix $1 \leq p < \infty$ and $0 < \varepsilon < 2^{-1-1/p}$, and choose

$$q := 2^{\lceil \log_2((2\varepsilon)^{-p}) \rceil}. \quad (110)$$

Since $(2\varepsilon)^{-p} > 2$, this is a power of two with $q \geq 2$. It is the largest power of two not exceeding $(2\varepsilon)^{-p}$, so $q \leq (2\varepsilon)^{-p} < 2q$. Equivalently,

$$\frac{1}{2^{p+1}\varepsilon^p} < q \leq \frac{1}{2^p\varepsilon^p}. \quad (111)$$

The separation is strict in both possible cases:

$$\begin{aligned} q > 2 : \quad & \frac{2(q-1)}{q^2} > \frac{1}{q} \geq (2\varepsilon)^p, \\ q = 2 : \quad & \frac{2(q-1)}{q^2} = \frac{1}{2} > (2\varepsilon)^p. \end{aligned}$$

The first line uses (111), and the second uses $\varepsilon < 2^{-1-1/p}$. By (109), we conclude that

$$\rho_{p,T}(h[x_{1:T}], h'[x_{1:T}]) > 2\varepsilon \quad (h \neq h', h, h' \in \mathcal{H}_q). \quad (112)$$

Proof of Theorem 6. Take q as in (110), the class $\mathcal{H}_q$, and the fixed sequence of all $T = q^d$ domain points. Lemma 34 gives $\text{sfat}_{\varepsilon/32}(\mathcal{H}_q) = d$.

An arbitrary real-valued prediction vector $w \in [0, 1]^T$ cannot approximate two distinct members $h, h' \in \mathcal{H}_q$ to normalized ℓ_p error at most ε . Otherwise, the triangle inequality would give

$$\rho_{p,T}(h[x_{1:T}], h'[x_{1:T}]) \leq \rho_{p,T}(h[x_{1:T}], w) + \rho_{p,T}(w, h'[x_{1:T}]) \leq 2\varepsilon,$$

contrary to (112). Thus any family of vectors covering all target vectors on this fixed sequence has at least $|\mathcal{H}_q|$ members.

Every deterministic causal expert g produces one such vector $g[x_{1:T}]$ on the fixed sequence. Consequently, (106) and (111) yield

$$N_p^{\text{dyn}}(\mathcal{H}_q, \varepsilon, q^d) \geq |\mathcal{H}_q| \geq q^d > \left(\frac{1}{2^{p+1}\varepsilon^p} \right)^d. \quad (113)$$

The argument already allowed arbitrary vectors in $[0, 1]^T$; hence it also proves the theorem's fixed-sequence assertion. $\blacksquare$

Taking the supremum in (21) gives the lower bound in (22). The comparisons in Sections 3.3 and 3.4 then complete Theorem 1 and the lower-bound assertions of Corollaries 7 and 8.

AI Disclosure

Several technical ideas underlying this work, particularly the detailed design of the routing mechanism and the tuning of its parameters, were developed through extensive discussions and iterative refinement involving the authors and ChatGPT 6 Pro. The authors also used ChatGPT 6 Pro to discuss the organization of the manuscript and to improve its clarity and presentation. The authors take full responsibility for all content of the paper, including the correctness of its mathematical results and proofs.

References

- [Aiy+26] Shashaank Aiyer, Yishay Mansour, Shay Moran, Han Shao, and Tom Waknine. *Scale-Sensitive Shattering: Learnability and Evaluability at Optimal Scale*. 2026. URL: <https://arxiv.org/abs/2605.13684>.
- [Alo+21] Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. “Adversarial laws of large numbers and optimal regret in online classification”. In: *Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing*. 2021, pp. 447–455.
- [Alo+97] Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. “Scale-sensitive dimensions, uniform convergence, and learnability”. In: *Journal of the ACM* 44.4 (1997), pp. 615–631.
- [Asc+25] Ruben Ascoli, Livia Betti, Justin Cheigh, Alex Iosevich, Ryan Jeong, Xuyan Liu, Brian McDonald, Wyatt Milgrim, Steven J. Miller, Francisco Romero Acosta, and Santiago Velazquez Iannuzzelli. “VC-Dimension of Hyperplanes Over Finite Fields”. In: *Graphs and Combinatorics* 41.2 (2025).
- [BDR21] Adam Block, Yuval Dagan, and Alexander Rakhlin. “Majorizing Measures, Sequential Complexities, and Online Learning”. In: *Proceedings of the 34th Annual Conference on Learning Theory*. 2021, pp. 587–590.
- [BKP97] Peter L Bartlett, Sanjeev R Kulkarni, and S Eli Posner. “Covering numbers for real-valued function classes”. In: *IEEE transactions on information theory* 43.5 (1997), pp. 1721–1724.
- [Blu+89] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. “Learnability and the Vapnik-Chervonenkis dimension”. In: *Journal of the ACM* 36.4 (1989), pp. 929–965.
- [BLW96] Peter L. Bartlett, Philip M. Long, and Robert C. Williamson. “Fat-Shattering and the Learnability of Real-Valued Functions”. In: *Journal of Computer and System Sciences* 52.3 (1996), pp. 434–452.
- [BM02] Peter L Bartlett and Shahar Mendelson. “Rademacher and gaussian complexities: Risk bounds and structural results”. In: *Journal of machine learning research* 3 (2002), pp. 463–482.
- [BPS09] Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. “Agnostic Online Learning”. In: *Proceedings of the 22nd Annual Conference on Learning Theory*. 2009.
- [BY20] Omri Ben-Eliezer and Eylon Yogev. “The Adversarial Robustness of Sampling”. In: *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. 2020, pp. 49–62.
- [CL06] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- [DEG16] Kunal Dutta, Esther Ezra, and Arijit Ghosh. “Two Proofs for Shallow Packings”. In: *Discrete & Computational Geometry* 56.4 (2016), pp. 910–939. ISSN: 0179-5376.
- [Dud67] Richard M Dudley. “The sizes of compact subsets of Hilbert space and continuity of Gaussian processes”. In: *Journal of Functional Analysis* 1.3 (1967), pp. 290–330.

- [GG15] Pierre Gaillard and Sébastien Gerchinovitz. “A Chaining Algorithm for Online Non-parametric Regression”. In: *Proceedings of The 28th Conference on Learning Theory*. Ed. by Peter Grünwald, Elad Hazan, and Satyen Kale. 2015, pp. 764–796.
- [Hau95] David Haussler. “Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension”. In: *Journal of Combinatorial Theory, Series A* 69.2 (1995), pp. 217–232.
- [KS94] Michael J. Kearns and Robert E. Schapire. “Efficient distribution-free learning of probabilistic concepts”. In: *Journal of Computer and System Sciences* 48.3 (1994), pp. 464–497. ISSN: 0022-0000.
- [Lit88] Nick Littlestone. “Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm”. In: *Machine Learning* 2.4 (1988), pp. 285–318.
- [MM24] Maryanthe Malliaris and Shay Moran. “Model theory and agnostic online learning via excellent sets”. In: *Transactions of the American Mathematical Society* 377.11 (2024), pp. 7753–7776.
- [MS14] Maryanthe Malliaris and Saharon Shelah. “Regularity lemmas for stable graphs”. In: *Transactions of the American Mathematical Society* 366.3 (2014), pp. 1551–1585.
- [MV03] Shahr Mendelson and Roman Vershynin. “Entropy and the combinatorial dimension”. In: *Inventiones mathematicae* 152.1 (2003), pp. 37–55.
- [RS14] Alexander Rakhlin and Karthik Sridharan. “Online Non-Parametric Regression”. In: *The 27th Conference on Learning Theory*. 2014, pp. 1232–1264.
- [RST10] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. “Online Learning: Random Averages, Combinatorial Parameters, and Learnability”. In: *Advances in Neural Information Processing Systems*. 2010, pp. 1984–1992.
- [RST11] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. “Online Learning: Beyond Regret”. In: *Proceedings of the 24th Annual Conference on Learning Theory*. Vol. 19. 2011, pp. 559–594.
- [RST15a] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. “Online Learning via Sequential Complexities”. In: *Journal of Machine Learning Research* 16.6 (2015), pp. 155–186.
- [RST15b] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. “Sequential complexities and uniform martingale laws of large numbers”. In: *Probability theory and related fields* 161.1 (2015), pp. 111–153.
- [RST17] Alexander Rakhlin, Karthik Sridharan, and Alexandre B. Tsybakov. “Empirical entropy, minimax regret and minimax risk”. In: *Bernoulli* 23.2 (2017), pp. 789–824.
- [RST24] Vinod Raman, Unique Subedi, and Ambuj Tewari. “Online Learning with Set-valued Feedback”. In: *Proceedings of 37th Conference on Learning Theory*. Ed. by Shipra Agrawal and Aaron Roth. 2024, pp. 4381–4412.
- [Sau72] Norbert Sauer. “On the density of families of sets”. In: *Journal of Combinatorial Theory, Series A* 13.1 (1972), pp. 145–147.
- [Sha+26] Shihao Shao, Cong Fang, Zhouchen Lin, and Dacheng Tao. *Partial Feedback Online Learning*. 2026. URL: <https://arxiv.org/abs/2601.21462>.
- [She72] Saharon Shelah. “A combinatorial problem; stability and order for models and theories in infinitary languages”. In: *Pacific Journal of Mathematics* 41.1 (1972), pp. 247–261.

- [Tal05] Michel Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer, 2005.
- [VC71] V. N. Vapnik and A. Ya. Chervonenkis. “On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities”. In: *Theory of Probability & Its Applications* 16.2 (1971), pp. 264–280.
- [VW96] A. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1996.