Tight Stochastic Sequential Covers for General VC Classes with Unknown i.i.d. Features

Changlong Wu[1] Wojciech Szpankowski[2,3]

[1] University of Arizona
[2] Purdue University
[3] Jagiellonian University

[1] clwu@arizona.edu
[2,3] szpan@purdue.edu

Abstract

We study hybrid online learning with stochastic features and adversarial labels when the feature distribution is unknown. Prior work gave stochastic sequential covers of log-size $O(d \log^2 T)$ for a general VC class and logarithmic bounds only under stronger star-type structure. We show that finite VC dimension alone is enough: every binary class of VC dimension d admits a distribution-blind exact cover with log-size $O(d \log T + \log(1/\delta))$, matching the worst-case $\Omega(d \log(T/d))$ dependence on the horizon. The key ingredient is a new local cube weight that counts Boolean subcubes incident to each realizable trace. Unlike the equal splits used in earlier constructions, we split mass unevenly according to these cube weights; a coordinate-deletion inequality shows that VC dimension controls the total cost of these splits. This yields, to our knowledge, the first tight $O((d+1)\log T)$ expected worst-case log-loss regret for general VC partitions with unknown i.i.d. features, and an exact polynomial-time implementation for planar affine halfspaces. We also show that the binary phenomenon *does not* extend from VC dimension to pseudo-dimension: a fixed pseudo-dimension-one sigmoid class requires $\Omega(\log^2 T)$ bits for uniform sequential approximation at accuracy $1/T$.

1 Introduction

Hybrid online learning studies sequential prediction when the features are stochastic but the labels may be adversarial (Lazaric and Munos, 2012). We focus on the distribution-blind setting: the feature distribution is unknown, features arrive only sequentially, and a single prediction strategy must work for every i.i.d. feature law. This differs from two commonly studied settings: known-distribution learning, where the feature law is available to the learner, and transductive prediction, where the full feature design is revealed in advance (Bhatt et al., 2023a; Qian et al., 2026).

A central tool in this setting is the stochastic global sequential cover introduced by Wu et al. (2023b). Such a cover is a finite collection of sequential predictors that, with high probability over the feature sequence, simultaneously reproduces the trace of every hypothesis in the class. For a class of VC dimension d, the general bound of Wu et al. (2023b) is $\log |\mathcal{G}| = O(d \log^2 T)$. They obtained the sharper $O(\log T)$ dependence only under additional structure, first through finite star number and then through Star–Littlestone dimension and special compositions such as intervals and cuboids. These conditions are genuinely restrictive: none of the arguments covered even affine halfspaces in $\mathbb{R}^2$, which were singled out as an open case in Wu et al. (2023b, Remark 4). Thus it

was unclear whether the extra logarithm was an artifact of the proof or a real price for working with a general VC class.

Our main result shows that the extra logarithmic factor is only an artifact. Every binary class of finite VC dimension admits a polynomial stochastic sequential cover. More precisely, for every $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}) = d$,

$$\log |\mathcal{G}| \le 6d\mathsf{H}_T + \log \Pi_{\mathcal{H}}(T) + \log(1/\delta) + \log 3, \tag{1}$$

where $\mathsf{H}_T = \sum_{t=1}^{T} 1/t$ and $\Pi_{\mathcal{H}}(T)$ is the growth function. By Sauer–Shelah, this is $O(d \log T + \log(1/\delta))$. We further exhibit a sparse-set class that requires cover log-cardinality $\Omega(d \log(T/d))$. Thus, the logarithmic dependence on T is tight in the worst case.

Our central conceptual contribution is a new combinatorial quantity that controls how probability mass should be split at a branch. Earlier logarithmic-cover constructions (Wu et al., 2023b) build a realization tree and, whenever both labels remain feasible, split the available covering functions equally between the two children. Each such split costs one bit, so obtaining an $O(\log T)$ cover requires a structural argument showing that only $O(\log T)$ relevant splits can occur; this is precisely where star-type assumptions enter. We keep the same basic splitting viewpoint but abandon equal splitting. Each realizable trace is assigned a local weight given by the number of Boolean subcubes incident to it, and the two children receive mass in proportion to these weights. A new coordinate-deletion inequality shows that VC dimension alone controls the total cost of these unequal splits. Under a random ordering, the expected cost is $O(d/t)$ when t coordinates remain, which sums to $O(d \log T)$. This provides the key argument that extends the result from the previously star-type classes to every finite-VC class.

The covering bound (1) has two immediate consequences. First, it essentially resolves the open problem posed by Bhatt and Kim (2021) for the two-level class $\mathcal{F}_{\mathcal{H}} = \{x \mapsto a + (b-a)h(x) : h \in \mathcal{H},\ a, b \in [0,1]\}$. For an unknown feature distribution, Bhatt and Kim (2021) established a general VC-class regret bound of square-root order. Using the stochastic sequential covers, Wu et al. (2023b) subsequently improved this to $O((d+1)\log^2 T)$. Our result further sharpens the rate to $O((d+1)\log T)$, which has the optimal logarithmic dependence on T for worst-case VC classes. To our knowledge, an $O(d \log T)$ guarantee was not previously known even under the weaker i.i.d.-label setting; for example, the general VC-class bounds of Bilodeau et al. (2023) still lose logarithmic factors. Here, our tight logarithmic rate holds in the stronger setting where the feature distribution is unknown and the labels may be chosen adversarially after observing the entire feature sequence.

Second, the improvement changes the computational picture. The earlier general bound (Wu et al., 2023b) gives a cover of size $\exp(O(d \log^2 T)) = T^{O(d \log T)}$, which is superpolynomial in T even when the dimension is fixed; in particular, naive aggregation over that cover is already superpolynomial for planar linear separators. Our bound reduces the cover size to $T^{O(d)}$. Moreover, for affine halfspaces in $\mathbb{R}^2$, we give a more direct exact polynomial-time implementation that avoids explicitly enumerating the cover.

Finally, we show that the result *does not* extend to arbitrary real-valued classes by simply replacing VC dimension with pseudo-dimension. We construct a fixed sigmoid family of pseudo-dimension one for which every distribution-blind uniform sequential $1/T$-cover has log-cardinality $\Omega(\log^2 T)$. Thus, the improvement from $\log^2 T$ to $\log T$ is a genuinely binary phenomenon rather than a generic consequence of finite dimension.

1.1 Related work

Table 1 summarizes the closest regret guarantees. Bhatt and Kim (2021) study the same two-level comparator under conditionally Bernoulli labels, while Bhatt et al. (2023b, Theorem 11) allow adversarial outcomes but retain a square-root term for general VC classes. Logarithmic rates are

Work	Feature law	Access	Rate
Bhatt and Kim (2021)	unknown	causal	$\sqrt{dT \log T}$
Bhatt et al. (2023b)	unknown	causal	square-root
Bhatt et al. (2023a)	known	causal	$\log T$
Qian et al. (2026)	any	design	$\log T$
Wu et al. (2023b)	unknown	causal	$(d{+}1) \log^2 T$
This paper	unknown	causal	$(d{+}1) \log T$

Table 1: Feature access and regret rates in closely related results. "Causal" means that features are revealed one at a time; "design" means that the full feature sequence is available in advance. The labels are adversarial except in Bhatt and Kim (2021), which assumes conditionally Bernoulli labels. The first, second, fifth, and sixth rows concern two-level probability models, whereas Qian et al. (2026) study logistic regression. The rates summarize the dependence relevant to this comparison; the protocols and comparator classes are discussed in Appendix I.

known under stronger information assumptions, such as a known feature distribution or advance access to the full feature design (Bhatt et al., 2023a; Qian et al., 2026). The difficulty of removing distributional knowledge also appears in smoothed online learning: most earlier results assume that the reference measure controlling the adversary is known, whereas Blanchard (2025) develop a new recursive-cover method to obtain agnostic sublinear regret when that reference measure is unknown. Our result establishes the logarithmic rate for general finite-VC partition classes while retaining both an unknown feature distribution and sequential feature access.

Our proof builds on the random-order viewpoint of one-inclusion prediction and the classical Sauer–Shelah trace bounds (Haussler et al., 1994; Sauer, 1972; Shelah, 1972). Boolean cubes and related cubical structure have also been studied extensively in the geometry of concept classes and sample compression (Rubinstein et al., 2009; Chalopin et al., 2019). We use the local number of cubes incident to each realizable trace as a probability weight and prove a new coordinate-deletion inequality that controls the resulting unequal splits. For planar halfspaces, this construction can be implemented efficiently using classical halfspace enumeration. This setting concerns two-level linear partitions rather than sigmoid probabilities, and is therefore distinct from efficient improper logistic prediction under norm-dependent guarantees (Foster et al., 2018; Jézéquel et al., 2020) or with advance access to the feature design (Qian et al., 2026). Appendix I gives a more detailed comparison.

2 Setup and Main Result

We consider sequential binary prediction over a horizon T. At round t, the learner observes a feature $X_t \in \mathcal{X}$, outputs a probability $\widehat{p}_t \in (0,1)$ for the event $Y_t = 1$, and then observes the binary outcome $Y_t \in \{0,1\}$. The feature sequence is stochastic, whereas the labels may be arbitrary.

A sequential strategy A maps (x^t, y^{t-1}) to $\widehat{p}_t$. Under logarithmic loss,

$$\ell(p, y) = -y \log p - (1-y) \log(1-p),$$

its regret relative to a comparator class $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ is

$$R_T(A; x^T, y^T; \mathcal{F}) = \sum_{t=1}^{T} \ell(\widehat{p}_t, y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t). \tag{2}$$

The quantity of interest is the expected worst-case regret (Wu et al., 2023b):

$$\widetilde{r}_T(\mathcal{F}) = \inf_A \sup_\mu \mathbb{E}_{X^T\sim\mu^T}\Big[\sup_{y^T\in\{0,1\}^T} R_T(A; X^T, y^T; \mathcal{F})\Big]. \tag{3}$$

The order of the quantifiers is important. The strategy is chosen without knowing the feature law μ, while the label sequence may depend on the entire realized feature sequence. Thus the learner must be distribution-blind and the guarantee is uniform over labels.

Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a binary hypothesis class with finite VC dimension $d = \mathrm{VC}(\mathcal{H})$. For a feature prefix $x^t = (x_1, \dots, x_t)$, define the trace class

$$C_t(x^t) = \{(h(x_1), \dots, h(x_t)) : h \in \mathcal{H}\} \subseteq \{0,1\}^t \tag{4}$$

and the growth function

$$\Pi_{\mathcal{H}}(T) = \sup_{x^T\in\mathcal{X}^T} |C_T(x^T)|.$$

We write $\mathsf{H}_T = \sum_{t=1}^T 1/t$ and use natural logarithms throughout. We assume that

$$\{x^t : v \in C_t(x^t)\} \tag{5}$$

is measurable for every t and v; Appendix H verifies this for the standard examples used in the paper.

We will use the following sequential covering notion from (Wu et al., 2023b).

Definition 1 (Stochastic global sequential cover)**.** Fix a class $\mathcal{P}$ of laws on $\mathcal{X}^T$. A finite family

$$\mathcal{G} \subseteq \{g : \cup_{t=1}^T \mathcal{X}^t \to \{0,1\}\}$$

is an (exact) stochastic global sequential cover of $\mathcal{H}$ at failure probability δ if, for every $P \in \mathcal{P}$,

$$P\left\{\forall h \in \mathcal{H}\ \exists g \in \mathcal{G}\ \ \forall t \le T :\ g(X^t) = h(X_t)\right\} \ge 1 - \delta. \tag{6}$$

The family $\mathcal{G}$ may depend on $\mathcal{H}, T$, and δ, but not on the feature law P.

The event in (6) requires simultaneous coverage of all $h \in \mathcal{H}$. Thus the comparator in regret minimization (3) may be selected after the feature sequence is observed. Moreover, each $g \in \mathcal{G}$ is causal: at time t it depends only on X^t.

Theorem 2 (Polynomial stochastic sequential covers)**.** *For every $T \ge 1$ and $\delta \in (0,1)$, there is an exact stochastic global sequential cover of $\mathcal{H}$ for the class of all exchangeable laws on $\mathcal{X}^T$ such that*

$$\log|\mathcal{G}| \le 6d\mathsf{H}_T + \log\Pi_{\mathcal{H}}(T) + \log(1/\delta) + \log 3. \tag{7}$$

Consequently,

$$\log|\mathcal{G}| \le 7d(1 + \log T) + \log(1/\delta) + \log 3. \tag{8}$$

For fixed $d \ge 1$, $T \ge 2$, and $\delta = 1/T$, one has $|\mathcal{G}| = T^{O(d)}$. If $d = 0$, a singleton cover suffices.

The theorem is stated for exchangeable feature laws, which is slightly stronger than the i.i.d. setting of interest. Exchangeability is sufficient because the proof averages over a random permutation of the observed coordinates, and such a permutation leaves the law unchanged. Compared with Wu et al. (2023b, Theorem 4), our bound removes one factor of $\log T$ and improves the confidence

dependence from $\log T \log(1/\delta)$ to $\log(1/\delta)$, thereby resolving the open questions posed in Wu et al. (2023b, Remark 4).

At failure probability $1/T$, Theorem 2 gives $\log|\mathcal{G}| = O(d \log T)$. Standard aggregation over the resulting finite family then gives logarithmic regret for binary prediction (Wu et al., 2023b). In Section 4, we use the underlying trace distribution directly to obtain the sharper bound for the two-level probability class.

The logarithmic dependence on T in (8) cannot be improved in general.

Proposition 3 (Worst-case lower bound)**.** *For every pair of integers $1 \le d \le T$, there is a binary class $\mathcal{H}_d$ of VC dimension d and a nonatomic i.i.d. feature law such that every exact stochastic sequential cover with failure probability $\delta < 1$ satisfies*

$$|\mathcal{G}| \ge \sum_{k=0}^{d} \binom{T}{k}. \tag{9}$$

Consequently, if $T \ge 2d$, then

$$\log|\mathcal{G}| \ge d \log(T/d).$$

For the lower bound, take the class of indicators of subsets of $[0,1]$ of cardinality at most d. Under any nonatomic feature law, the T sampled points are distinct almost surely, and the induced trace class contains exactly the binary vectors with at most d ones. A deterministic sequential rule produces only one trace on a fixed sample, so an exact simultaneous cover must contain at least as many rules as there are realizable traces. The proof is given in Appendix F. For fixed d, this matches the $O(d \log T)$ upper bound in its dependence on T.

3 Proof of The Covering Theorem

We now prove Theorem 2. We assign each realizable trace a local cube weight, use these weights to define sequential transition probabilities, and control the resulting path loss by exposing the coordinates in random reverse order. Arithmetic coding then converts the trace probabilities into a finite deterministic cover. The main argument is given here; technical details are deferred to the appendix.

3.1 Local cube potential

We begin with the local weight that makes the construction sensitive to branch asymmetry. Let I be a finite coordinate set, let $C \subseteq \{0,1\}^I$ be nonempty, and for $v \in C$ and $A \subseteq I$ write $v \oplus \mathbf{1}_A$ for the vector obtained by flipping the coordinates in A.

Definition 4 (Centered-cube weight)**.** For $v \in C$, define

$$\mathcal{Q}_C(v) = \{S \subseteq I : v \oplus \mathbf{1}_A \in C \text{ for every } A \subseteq S\}, \tag{10}$$

$$W_C(v) = |\mathcal{Q}_C(v)|. \tag{11}$$

A set $S \in \mathcal{Q}_C(v)$ specifies a Boolean cube contained in C with every coordinate outside S fixed according to v; such cubical structure of concept classes is classical (Rubinstein et al., 2009; Chalopin et al., 2019). In particular, S is shattered by C, so $|S| \le d$ whenever $\mathrm{VC}(C) \le d$. Informally, $W_C(v)$ measures the local combinatorial mass of the class around v: a trace incident to many subcubes retains more local freedom and is therefore safer to bet on. The empty cube always belongs to $\mathcal{Q}_C(v)$, so $W_C(v) \ge 1$.

For $j \in I$, let the map π_j denote deletion of coordinate j, define $C^{-j} = \pi_j(C)$, and write $p = \pi_j(v)$. Define

$$F_j(p) = \{u \in C : \pi_j(u) = p\}, \tag{12}$$

$$R_j(C, v) = \frac{\sum_{u \in F_j(p)} W_C(u)}{W_{C^{-j}}(p)}, \tag{13}$$

Let $v^{(j)} = v \oplus \mathbf{1}_j$ denote the vector obtained by flipping coordinate j. Then the fiber $F_j(p)$ is either $\{v\}$ or $\{v, v^{(j)}\}$, according to whether $v^{(j)} \in C$.

The following lemma is the main combinatorial ingredient of our argument.

Lemma 5 (Coordinate-deletion inequality)**.** *If* $\mathrm{VC}(C) \le d$*, then, for every* $v \in C$*,*

$$0 < R_j(C, v) \le 4 \quad (j \in I), \qquad \sum_{j \in I} (R_j(C, v) - 1)_+ \le 6d. \tag{14}$$

Consequently,

$$\sum_{j \in I} (\log R_j(C, v))_+ \le 6d. \tag{15}$$

The important feature of (14) is that the budget $6d$ is shared across all candidate deletions. This yields a cost of order d/t when t coordinates remain. We give the main calculation here and defer the full proof to Appendix A. If the fiber $F_j(p)$ contains only v, then no centered cube at v contains j and $R_j(C, v) \le 1$. Otherwise the fiber is $\{v, v^{(j)}\}$; writing $B_j(C, v)$ for the number of centered cubes at v that use j and splitting cubes according to whether they use j gives

$$W_C(v) + W_C(v^{(j)}) \le W_{C^{-j}}(p) + 3B_j(C, v), \tag{16}$$

while projection of cube gives $W_C(v) \le 2W_{C^{-j}}(p)$, thus $(R_j(C, v) - 1)_+ \le 6B_j(C, v)/W_C(v)$. Summing the incidence count

$$\sum_{j \in I} B_j(C, v) = \sum_{S \in \mathcal{Q}_C(v)} |S| \le dW_C(v) \tag{17}$$

then yields (14).

The VC dimension enters only through (17): each centered cube S contributes once for every coordinate $j \in S$, and $|S| \le d$ because S is shattered by C.

3.2 Sequential trace probabilities

The deletion inequality is a static statement about a finite trace class. We now turn it into a sequential probability assignment for *realizable* label sequences. The key point is that a branch receives probability in proportion to the local cube mass that survives after taking that branch. Suppose $y^{t-1} \in C_{t-1}(x^{t-1})$, with feasible next labels

$$A_t(x^t, y^{t-1}) = \{b \in \{0, 1\} : y^{t-1}b \in C_t(x^t)\}.$$

For $b \in A_t(x^t, y^{t-1})$, set

$$q_t(b \mid x^t, y^{t-1}) = \frac{W_{C_t(x^t)}(y^{t-1}b)}{\sum_{c \in A_t(x^t, y^{t-1})} W_{C_t(x^t)}(y^{t-1}c)}. \tag{18}$$

For $b \notin A_t(x^t, y^{t-1})$, set $q_t(b \mid x^t, y^{t-1}) = 0$. The probability assignment is causal and does not involve the feature law.

Note that the probability assignment in (18) is used only as a splitting weight in the cover construction of Section 3.4; it is *not* itself the predictor used for the hybrid learning problem.

For $h \in \mathcal{H}$ write $v_t^h = (h(x_1), \dots, h(x_t))$ and define the cumulative trace log-loss

$$L_T(h; x^T) = -\sum_{t=1}^{T} \log q_t(h(x_t) \mid x^t, v_{t-1}^h). \tag{19}$$

The following lemma bounds the cumulative trace log-loss using the cube-weight ratios (13).

Lemma 6. *Let $w_t = W_{C_t(x^t)}(v_t^h)$, with $w_0 = 1$, and let R_t be the ratio* (13) *obtained by deleting the last coordinate of $C_t(x^t)$ at v_t^h. Then*

$$L_T(h; x^T) = \sum_{t=1}^{T} \log R_t - \log w_T \le \sum_{t=1}^{T} (\log R_t)_+, \tag{20}$$

Proof. Along the trace v_t^h, the denominator in (18) equals $R_t w_{t-1}$, so $q_t = w_t/(R_t w_{t-1})$. Thus

$$-\log q_t = \log R_t + \log w_{t-1} - \log w_t.$$

Summing over t telescopes, and the desired inequality follows from $w_T \ge 1$. □

Example 7. Take the class consisting of the zero function and all singleton indicators, whose traces on distinct features are $C_t = \{0^t, e_1, \dots, e_t\}$. Its VC dimension is one. The zero trace has t single-coordinate flips that remain in C_t, so $W_{C_t}(0^t) = t + 1$, whereas every other trace has weight 2. Along the all-zero path, both extensions remain feasible at every round, but the cube-weighted split is highly asymmetric:

$$q_t(0 \mid x^t, 0^{t-1}) = \frac{t+1}{t+3}.$$

Hence the probability assigned to the all-zero trace over T rounds is $6/((T+2)(T+3))$, and its cumulative trace log-loss is only $O(\log T)$. In contrast, splitting equally whenever both extensions are feasible would assign probability 2^{-T} and incur log-loss $T \log 2$. Figure 1 illustrates the weights and the split at $t = 3$.

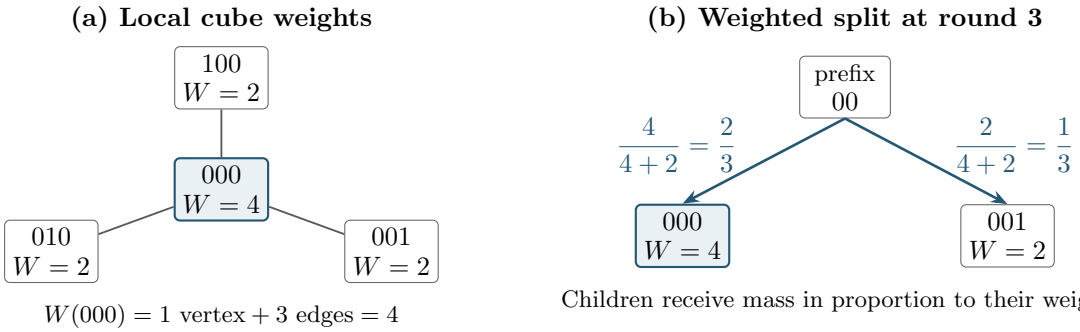

Figure 1: Cube weights and asymmetric branching for $C_3 = \{000, 100, 010, 001\}$. (a) Each weight counts the vertex and its incident edges; no higher-dimensional cubes occur. (b) The extensions of prefix 00 receive probabilities 2/3 and 1/3 under (18).

3.3 Random-order control of realizable traces

The bound in Lemma 6 holds pointwise for every realizable trace. We now convert it into a uniform high-probability bound under exchangeable feature sequences. This uniform trace guarantee is the probabilistic heart of our arguments.

Proposition 8 (Uniform trace log-loss)**.** *Let P be any exchangeable law on $\mathcal{X}^T$. Then*

$$\mathbb{E}_P\left[\sup_{h\in\mathcal{H}} L_T(h;X^T)\right] \le 6d\mathsf{H}_T + \log \Pi_{\mathcal{H}}(T), \tag{21}$$

and for every $\delta \in (0,1)$,

$$P\Big\{\sup_{h\in\mathcal{H}} L_T(h;X^T) > 6d\mathsf{H}_T + \log \Pi_{\mathcal{H}}(T) + \log(1/\delta)\Big\} \le \delta. \tag{22}$$

Proof. Fix a deterministic sample x^T and a full trace $v \in C_T(x^T)$. Let σ be a uniformly random permutation of $[T]$ and put $S_t = \{\sigma(1),\dots,\sigma(t)\}$. Write C_{S_t} for the projection of $C_T(x^T)$ onto S_t and v_{S_t} for the corresponding projection of v, and define

$$Y_t = \big(\log R_{\sigma(t)}(C_{S_t}, v_{S_t})\big)_+.$$

Let $\mathcal{F}_t$ be the sigma-field generated by the suffix $(\sigma(t+1),\dots,\sigma(T))$, with $\mathcal{F}_T$ trivial. Here, we may view the permutation as being revealed *backward*, from $\sigma(T)$ to $\sigma(1)$. Conditional on $\mathcal{F}_t$, the set S_t is fixed and $\sigma(t)$ is uniform on S_t. Therefore Lemma 5 gives

$$\begin{aligned}\mathbb{E}[e^{Y_t} \mid \mathcal{F}_t] &= \frac{1}{t}\sum_{j\in S_t} \exp((\log R_j(C_{S_t}, v_{S_t}))_+) \\ &= 1 + \frac{1}{t}\sum_{j\in S_t} (R_j(C_{S_t}, v_{S_t}) - 1)_+ \le 1 + \frac{6d}{t}.\end{aligned} \tag{23}$$

Iterating conditional expectations,

$$\mathbb{E}_\sigma \exp\left(\sum_{t=1}^T Y_t\right) \le \prod_{t=1}^T \left(1 + \frac{6d}{t}\right) \le e^{6d\mathsf{H}_T}. \tag{24}$$

By Lemma 6, the cumulative trace log-loss in the permuted order is at most $\sum_t Y_t$. Hence, for every fixed terminal trace v,

$$\mathbb{E}_\sigma e^{L_T(v;x^T_\sigma)} \le e^{6d\mathsf{H}_T}.$$

There are at most $|C_T(x^T)| \le \Pi_{\mathcal{H}}(T)$ realizable terminal traces. Since

$$e^{\max_v L_T(v;x^T_\sigma)} \le \sum_{v\in C_T(x^T)} e^{L_T(v;x^T_\sigma)},$$

we obtain

$$\mathbb{E}_\sigma e^{\max_v L_T(v;x^T_\sigma)} \le \Pi_{\mathcal{H}}(T) e^{6d\mathsf{H}_T}. \tag{25}$$

Taking logarithms and Jensen's inequality gives

$$\mathbb{E}_\sigma \max_v L_T(v;x^T_\sigma) \le 6d\mathsf{H}_T + \log \Pi_{\mathcal{H}}(T),$$

while Markov's inequality applied to (25) gives, for every $\delta \in (0,1)$,

$$\Pr_{\sigma}\Big\{\max_{v} L_T(v; x_\sigma^T) > 6d\mathsf{H}_T + \log \Pi_{\mathcal{H}}(T) + \log(1/\delta)\Big\} \le \delta.$$

Finally let X^T have an exchangeable law and let σ be independent and uniform. Then $X_\sigma^T \overset{d}{=} X^T$. Averaging the two preceding bounds over X^T therefore proves (21) and (22), hence Proposition 8. □

The reverse exposure of the permutation is used only for the analysis; the learner still observes the features in their original order. Its role is to make the next deleted coordinate uniform among those that remain. The exponential-moment bound then allows us to control all realizable traces simultaneously at an additional cost of only $\log \Pi_{\mathcal{H}}(T)$. Appendix B provides the full permutation bookkeeping.

3.4 Constructing the sequential cover via arithmetic coding

Proposition 8 gives a uniform lower bound on the probability assigned to every realizable trace. It remains to convert this probabilistic guarantee into a fixed finite family of deterministic sequential rules. For any fixed feature sequence, the probabilities in (18) define a binary probability tree over the realizable traces. We construct a finite family of sequential rules whose paths cover this tree using an interval construction reminiscent of arithmetic coding (Witten et al., 1987).

Assign the root of the probability tree the interval $I_\varnothing = [0,1)$. Recursively, suppose a realizable history y^{t-1} has been assigned an interval $I_{y^{t-1}}$. Partition this interval into two consecutive subintervals $I_{y^{t-1}0}$ and $I_{y^{t-1}1}$ whose lengths are, respectively, $q_t(0 \mid x^t, y^{t-1})$ and $q_t(1 \mid x^t, y^{t-1})$ times the length of the parent interval. For any realizable trace $v_T^h = (h(x_1), \dots, h(x_T))$, the length of its terminal interval is the product of the transition probabilities along the path:

$$|I_{v_T^h}| = \prod_{t=1}^{T} q_t(h(x_t) \mid x^t, v_{t-1}^h) = e^{-L_T(h; x^T)}. \tag{26}$$

Writing

$$L_* = 6d\mathsf{H}_T + \log \Pi_{\mathcal{H}}(T) + \log(1/\delta),$$

take the fixed grid

$$U = \left\{ \frac{k}{M} : k = 0, \dots, M-1 \right\}, \qquad M = \lceil 2e^{L_*} \rceil.$$

Each grid point $u \in U$ defines a deterministic sequential rule g_u: starting from the root, at each round g_u follows the unique child interval containing u and predicts the label associated with that child. Note that g_u is sequential: at round t, its prediction depends only on the observed feature prefix X^t. Figure 2 illustrates the construction for the same singleton class as Figure 1.

Proof of Theorem 2. Let E denote the event in Proposition 8 on which

$$\sup_{h \in \mathcal{H}} L_T(h; X^T) \le L_*.$$

By (22), every exchangeable feature law satisfies $P(E) \ge 1 - \delta$.

Fix a realization $x^T \in E$ and any $h \in \mathcal{H}$. The nested interval associated with the trace

$$(h(x_1), \dots, h(x_T))$$

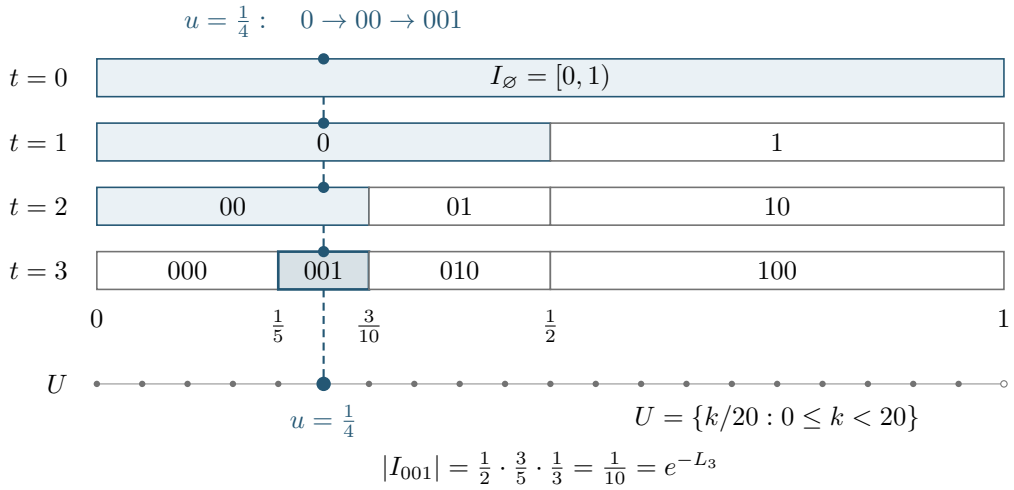

Figure 2: Arithmetic intervals for $C_3 = \{000, 100, 010, 001\}$. Each row partitions $[0, 1)$ into the intervals of feasible prefixes. Shading marks the nested intervals containing the fixed seed $u = 1/4$, whose rule predicts $0, 0, 1$. The terminal interval has length equal to the product of its path probabilities. The illustrative grid has $M = 20$; all terminal intervals have length at least $1/10 = 2/M$ and therefore contain grid points. In the general construction, $M = \lceil 2e^{L_*} \rceil$. All intervals are left-closed and right-open.

has terminal length, by (26),

$$e^{-L_T(h;x^T)} \ge e^{-L_*}.$$

Since $M \ge 2e^{L_*}$, every interval in $[0, 1)$ of length at least e^{-L_*} contains a point of U; Appendix C gives the elementary boundary check. Choose such a point u. Because the intervals are nested, the rule g_u follows the target branch at every prefix, and therefore

$$g_u(x^t) = h(x_t), \qquad t = 1, \dots, T.$$

The same grid U is fixed before the feature law is chosen, and the argument holds simultaneously for every h. Thus

$$\mathcal{G} = \{g_u : u \in U\}$$

is an exact distribution-blind stochastic global sequential cover on E.

It remains only to bound its size. Since $L_* \ge 0$,

$$|\mathcal{G}| \le M \le 2e^{L_*} + 1 \le 3e^{L_*},$$

and hence

$$\log |\mathcal{G}| \le 6d\mathsf{H}_T + \log \Pi_{\mathcal{H}}(T) + \log(1/\delta) + \log 3,$$

which is (7). Sauer–Shelah gives

$$\Pi_{\mathcal{H}}(T) \le \sum_{k=0}^{d} \binom{T}{k} \le (eT/d)^d$$

when $1 \le d \le T$. If $d > T$, then $\log \Pi_{\mathcal{H}}(T) \le T \log 2 \le d(1 + \log T)$. Together with $\mathsf{H}_T \le 1 + \log T$, these bounds give (8) for every $d \ge 1$. If $d = 0$, all hypotheses agree pointwise, and a singleton cover suffices. This proves Theorem 2. □

4 Application to Log-loss

We now give an immediate consequence of Theorem 2 for log-loss regrets; see Wu et al. (2023b) for further applications of stochastic sequential covers.

Consider the two-level probability class

$$\mathcal{F}_{\mathcal{H}} = \{x \mapsto a + (b-a)h(x) : h \in \mathcal{H},\ a, b \in [0,1]\}, \tag{27}$$

where $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ is a binary-valued class. Here, h partitions the feature space into two regions, while a and b specify the probability of label 1 on the two regions. This class has been studied in Bhatt and Kim (2021); Bilodeau et al. (2023); Wu et al. (2023b).

Theorem 9 (Logarithmic regret)**.** *For $T \geq 2$,*

$$\widetilde{r}_T(\mathcal{F}_{\mathcal{H}}) \leq \left(1 + \frac{1}{T}\right)\left(6d\mathsf{H}_T + \log \Pi_{\mathcal{H}}(T) + 3\log T + \log 6\right) + 3. \tag{28}$$

In particular,

$$\widetilde{r}_T(\mathcal{F}_{\mathcal{H}}) = O((d+1)\log T).$$

Proof. Apply Theorem 2 with $\delta = 1/T$, and let $\mathcal{G}$ be the resulting exact cover of $\mathcal{H}$. Let

$$\Theta_T = \left\{0, \frac{1}{T-1}, \frac{2}{T-1}, \ldots, 1\right\},$$

so $|\Theta_T| = T$ and every point of $[0,1]$ is within $1/T$ of some element of Θ_T. For $g \in \mathcal{G}$ and $\bar{a}, \bar{b} \in \Theta_T$, define

$$g_{\bar{a},\bar{b}}(x^t) = \bar{a} + (\bar{b} - \bar{a})g(x^t),$$

and let

$$\mathcal{G}' = \{g_{\bar{a},\bar{b}} : g \in \mathcal{G},\ \bar{a}, \bar{b} \in \Theta_T\}.$$

Then $|\mathcal{G}'| \leq T^2|\mathcal{G}|$. On the success event of $\mathcal{G}$, every $f(x) = a + (b-a)h(x) \in \mathcal{F}_{\mathcal{H}}$ is uniformly approximated along the realized sequence to accuracy $1/T$ by some member of $\mathcal{G}'$. Hence $\mathcal{G}'$ is a stochastic sequential $1/T$-cover of $\mathcal{F}_{\mathcal{H}}$ with failure probability $1/T$. By Wu et al. (2023b, Theorem 2),

$$\widetilde{r}_T(\mathcal{F}_{\mathcal{H}}) \leq 3 + \left(1 + \frac{1}{T}\right)\log(|\mathcal{G}'| + 1).$$

Since $|\mathcal{G}'| \leq T^2|\mathcal{G}|$ and $|\mathcal{G}| \geq 1$,

$$\log(|\mathcal{G}'| + 1) \leq 2\log T + \log|\mathcal{G}| + \log 2.$$

Applying Theorem 2 with $\delta = 1/T$ yields (28). □

Theorem 9 achieves the *optimal* logarithmic dependence on T for general VC classes, essentially closing the line of work initiated in Bhatt and Kim (2021) and further developed in Bilodeau et al. (2023); Wu et al. (2023b); Bhatt et al. (2023b).

A naive implementation of the predictor achieving the regret bound in Theorem 9 aggregates over the cover $\mathcal{G}'$, whose size is $T^{O(d+1)}$. For fixed d, this gives polynomially many experts; a polynomial-time implementation additionally requires efficient construction and evaluation of those experts. Finite VC dimension alone does not supply this computational access. For planar linear classifiers, using the probability assignment in (18), we construct the predictor directly without

explicitly aggregating over the cover. The full construction and analysis are given in Appendix D. For any $\mathcal{H}$ and every exchangeable feature law, the direct predictor satisfies the sharper bound

$$\mathbb{E}\left[\sup_{y^T} R_T(A; X^T, y^T; \mathcal{F}_\mathcal{H})\right] \le 6d\mathsf{H}_T + \log \Pi_\mathcal{H}(T) + 2\log(T+1). \tag{29}$$

Theorem 10 (Efficient prediction for planar halfspaces)**.** *Let $\mathcal{H}$ be the class of affine halfspaces in $\mathbb{R}^2$. There exists a distribution-blind sequential predictor for $\mathcal{F}_\mathcal{H}$ with $O(\log T)$ expected worst-case log-loss regret under unknown i.i.d. features. At round t, the predictor maintains at most $t^2 - t + 2$ realizable labelings of the observed feature prefix. For rational features, it admits an exact implementation with polynomial bit complexity, using $O(T^6)$ combinatorial and rational-arithmetic operations over T rounds.*

Indeed, affine halfspaces in $\mathbb{R}^2$ realize at most $t^2 - t + 2 = O(t^2)$ labelings on t points. Since their VC dimension is three, each cube weight can be computed by examining $O(t^3)$ candidate coordinate sets. Thus one round requires $O(t^5)$ operations, giving $O(T^6)$ over the full horizon. Appendix E gives the implementation and bit-complexity details.

5 Limitation of Pseudo-Dimension

The binary result naturally raises the question of whether the same phenomenon extends to real-valued classes, with VC dimension replaced by pseudo-dimension. For $\mathcal{F} \subseteq [0,1]^\mathcal{X}$, the pseudo-dimension is the VC dimension of the subgraph class

$$\{(x, r) \mapsto \mathbf{1}\{f(x) \ge r\} : f \in \mathcal{F}\}.$$

The corresponding sequential covering notion extends to $[0,1]$-valued predictors and requires

$$|g(X^t) - f(X_t)| \le \alpha$$

on every round (Wu et al., 2023b).

It was shown in Wu et al. (2023b, Appendix E) that a real-valued class $\mathcal{F}$ of pseudo-dimension d admits a stochastic sequential α-cover at failure probability δ with log-cardinality

$$O\big(d\log^2(T/\alpha) + \log(T/\alpha)\log(1/\delta)\big).$$

For log loss, taking $\alpha = \delta = 1/T$ yields an $O(d\log^2 T)$ regret bound for every such class with $d \ge 1$. Quite surprisingly, our next theorem shows that, unlike in the binary case, this extra logarithmic factor *cannot* in general be removed: even for a pseudo-dimension-one class, a stochastic sequential $1/T$-cover with log-cardinality $O(\log T)$ is *impossible*.

Theorem 11 (Pseudo-dimension does not imply polynomial covers)**.** *There is a fixed class $\mathcal{F}_* \subseteq (0,1)^\mathcal{X}$ with $\mathrm{Pdim}(\mathcal{F}_*) = 1$ and universal constants $c, T_0 > 0$ such that, for every $T \ge T_0$, every distribution-blind stochastic sequential $1/T$-cover of $\mathcal{F}_*$ at failure probability $1/4$ satisfies*

$$\log|\mathcal{G}| \ge c\log^2 T. \tag{30}$$

The class can be taken to be the sigmoid family

$$\mathcal{F}_* = \{(a,b) \mapsto \sigma(a\theta + b) : \theta \in (0,1)\}, \qquad \sigma(s) = \frac{1}{1+e^{-s}}, \tag{31}$$

on the domain $(a,b) \in (0,\infty) \times \mathbb{R}$.

We sketch the key idea below and refer to Appendix G for the complete proof.

Let $K \asymp 1/\alpha$. We construct a full K-ary tree in which each node u is associated with a parameter interval $I_u \subset (0,1)$ and a feature x_u. Its children $I_{u1}, \ldots, I_{uK}$ are disjoint subintervals of I_u chosen so that, whenever $\theta \in I_{uk}$, the value $f_\theta(x_u)$ lies in one of K output bands separated by more than 2α. Therefore, any sequential predictor that approximates $f_\theta(x_u)$ within accuracy α uniquely identifies which child interval I_{uk} contains θ. The key idea is thus to reduce the covering problem to identifying a hidden root-to-leaf path in this tree.

We randomize the hard instance in two stages. First choose a hidden path of depth L

$$B = (B_1, \ldots, B_L), \qquad B_1, \ldots, B_L \overset{\text{iid}}{\sim} \text{Unif}[K],$$

and choose θ_B in the corresponding terminal interval $I_{B_1 \cdots B_L}$. The node at depth r on this path has feature

$$x_r(B) = x_{B_1 \cdots B_{r-1}}.$$

For this fixed path B, let μ_B be the uniform distribution on $x_1(B), \ldots, x_L(B)$, and draw features

$$X_1, \ldots, X_T \overset{\text{iid}}{\sim} \mu_B.$$

Equivalently, we can draw independent depths

$$D_1, \ldots, D_T \overset{\text{iid}}{\sim} \text{Unif}[L]$$

and set $X_t = x_{D_t}(B)$. Thus B determines the hidden path and target function, while D_t determines which node on that fixed path is sampled at round t. Conditional on B, the features are i.i.d.

Call round t a *record* if

$$D_t > \max_{s<t} D_s,$$

with the first round counted as a record. Suppose $D_t = r$ is a record. Since all previously sampled depths are smaller than r, the feature prefix X^t depends only on $B_1, \ldots, B_{r-1}$. Any sequential rule must therefore choose its prediction before learning the fresh digit B_r. On the other hand, $f_{\theta_B}(X_t)$ lies in the output band indexed by B_r. Because the bands are separated by more than 2α, predicting within α uniquely identifies B_r. Hence any fixed sequential rule succeeds at this record with conditional probability at most $1/K$.

Different records expose fresh path digits. Therefore, if R records occur, any fixed sequential rule tracks the target with probability at most K^{-R}. For a cover of size M, a union bound gives

$$\frac{3}{4} \le \mathbb{E} \min\{1, MK^{-R}\}. \tag{32}$$

As $L \to \infty$, collisions among the sampled depths vanish, and the relative order of $D_1, \ldots, D_T$ converges to that of a uniformly random permutation of $[T]$. Thus R converges to the corresponding record count, which is of order $\log T$. Combining (32) with the generating function of the record count yields

$$\log M = \Omega(\log T \log K) = \Omega\left(\log T \log \frac{1}{\alpha}\right).$$

Taking $\alpha = 1/T$ gives the claimed $\Omega(\log^2 T)$ lower bound. Appendix G gives the full construction and proof, including the conditioning argument and the record-count calculation.

6 Discussion

We established tight stochastic sequential covering bounds for general finite-VC classes, a key ingredient for sharp regret guarantees in distribution-blind hybrid online learning. For the two-level probability class under log loss, the new covering bound yields the optimal logarithmic dependence on the horizon, and similar reductions may improve other prediction problems, including mixable losses and contextual bandits. Our pseudo-dimension lower bound shows that the binary theory does not extend automatically to real-valued classes. For structured models such as logistic regression with unknown i.i.d. features, it remains open whether dimension alone suffices for $O(d \log T)$ distribution-blind regret.

AI Use Statement

Generative AI tools were used in the development of this work. In particular, OpenAI GPT-5.6 Sol provided the key cube-potential argument underlying the main covering theorem, including the idea of weighting traces by their incident Boolean subcubes and using these weights for unequal probability splitting. The authors formulated the research problem, supplied the relevant prior results and partial progress, directed and iteratively steered the exploration, and independently checked the resulting mathematical arguments and proofs. Anthropic Claude and OpenAI Codex were also used to help investigate extensions, search and organize relevant literature, revise exposition, implement verification code and predictors, and check the compiled manuscript.

References

Agarwal, N., Kale, S., and Zimmert, J. (2022). Efficient methods for online multiclass logistic regression. In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pages 3–33.

Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. (1997). Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631.

Ben-David, S., Pál, D., and Shalev-Shwartz, S. (2009). Agnostic online learning. In *Proceedings of the 22nd Annual Conference on Learning Theory*.

Bhatt, A., Haghtalab, N., and Shetty, A. (2023a). Smoothed analysis of sequential probability assignment. In *Advances in Neural Information Processing Systems*, volume 36.

Bhatt, A. and Kim, Y.-H. (2021). Sequential prediction under log-loss with side information. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 340–344.

Bhatt, A., Ryu, J. J., and Kim, Y.-H. (2023b). On universal portfolios with continuous side information. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 4147–4163.

Bilodeau, B., Foster, D. J., and Roy, D. M. (2020). Tight bounds on minimax regret under logarithmic loss via self-concordance. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 919–929.

Bilodeau, B., Foster, D. J., and Roy, D. M. (2023). Minimax rates for conditional density estimation via empirical entropy. *The Annals of Statistics*, 51(2):762–790.

Blanchard, M. (2025). Agnostic smoothed online learning. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 1997–2006. ACM.

Bousquet, O., Hanneke, S., Moran, S., Shafer, J., and Tolstikhin, I. (2023). Fine-grained distribution-dependent learning curves. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5890–5924.

Bousquet, O., Hanneke, S., Moran, S., van Handel, R., and Yehudayoff, A. (2021). A theory of universal learning. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 532–541.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Chalopin, J., Chepoi, V., Moran, S., and Warmuth, M. K. (2019). Unlabeled sample compression schemes and corner peelings for ample and maximum classes. In *46th International Colloquium on Automata, Languages, and Programming*, volume 132 of *Leibniz International Proceedings in Informatics*, pages 34:1–34:15.

Chardon, H., Pathak, R., and Zhivotovskiy, N. (2026). Beyond modern asymptotics for log-likelihood ratios in logistic regression. arXiv preprint arXiv:2608.02507.

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334.

Di Gennaro, F., Chakraborty, S., and Zhivotovskiy, N. (2026). Efficient logistic regression with mixture of sigmoids. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 2053–2061.

Drmota, M., Jacquet, P., Wu, C., and Szpankowski, W. (2024). Minimax regret with unbounded weights. In *IEEE International Symposium on Information Theory*, pages 2305–2310.

Drmota, M., Jacquet, P., Wu, C., and Szpankowski, W. (2025). Precise regularized minimax regret with unbounded weights. *IEEE Transactions on Information Theory*, 71(11):8753–8776.

Drmota, M., Jacquet, P., Wu, C., and Szpankowski, W. (2026). Phase transition of regret for logistic regression with large weights. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pages 1–28.

Edelsbrunner, H., O'Rourke, J., and Seidel, R. (1986). Constructing arrangements of lines and hyperplanes with applications. *SIAM Journal on Computing*, 15(2):341–363.

Foster, D. J., Kale, S., Luo, H., Mohri, M., and Sridharan, K. (2018). Logistic regression: The importance of being improper. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 167–208.

Haussler, D., Littlestone, N., and Warmuth, M. K. (1994). Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292.

Hazan, E., Koren, T., and Levy, K. Y. (2014). Logistic regression: Tight bounds for stochastic and online optimization. In *Proceedings of the 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 197–209.

Jacquet, P., Shamir, G., and Szpankowski, W. (2021). Precise minimax regret for logistic regression with categorical feature values. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 755–771.

Jacquet, P., Shamir, G. I., and Szpankowski, W. (2022). Precise minimax regret for logistic regression. In *IEEE International Symposium on Information Theory*, pages 444–449.

Jézéquel, R., Gaillard, P., and Rudi, A. (2020). Efficient improper learning for online logistic regression. In *Proceedings of the Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2085–2108.

Jézéquel, R., Gaillard, P., and Rudi, A. (2021). Mixability made efficient: Fast online multiclass logistic regression. In *Advances in Neural Information Processing Systems*, volume 34.

Jia, Z., Polyanskiy, Y., and Rakhlin, A. (2025a). A gapped scale-sensitive dimension and lower bounds for offset rademacher complexity. arXiv:2509.20618.

Jia, Z., Rakhlin, A., and Polyanskiy, Y. (2025b). On the minimax regret of sequential probability assignment via square-root entropy. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 2958–3016.

Kakade, S. M. and Ng, A. Y. (2004). Online bounds for Bayesian algorithms. In *Advances in Neural Information Processing Systems*, volume 17, pages 641–648.

Lazaric, A. and Munos, R. (2012). Learning with stochastic inputs and adversarial outputs. *Journal of Computer and System Sciences*, 78(5):1516–1537.

Mathiasen, M. E., Qian, J., and Zhivotovskiy, N. (2026). An optimal agnostic PAC algorithm. arXiv:2608.06363.

Mayo, J. J., Hadiji, H., and van Erven, T. (2022). Scale-free unconstrained online learning for curved losses. In *Proceedings of the Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4464–4497.

Mourtada, J. and Gaïffas, S. (2022). An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *Journal of Machine Learning Research*, 23(31):1–49.

Okoroafor, P., Kleinberg, R., and Kim, M. P. (2025). Near-optimal algorithms for omniprediction. In *Proceedings of the 66th Annual Symposium on Foundations of Computer Science*.

Okoroafor, P., Kleinberg, R., and Kim, M. P. (2026). Oracle-efficient hybrid learning with constrained adversaries. In *International Conference on Learning Representations*.

Qian, J., Rakhlin, A., and Zhivotovskiy, N. (2026). Refined risk bounds for unbounded losses via transductive priors. *Journal of Machine Learning Research*, 27(26):1–64.

Rakhlin, A. and Sridharan, K. (2015). Sequential probability assignment with binary alphabets and large classes of experts. arXiv:1501.07340.

Rakhlin, A., Sridharan, K., and Tewari, A. (2011). Online learning: Stochastic and constrained adversaries. arXiv:1104.5070.

Rakhlin, A., Sridharan, K., and Tewari, A. (2015). Online learning via sequential complexities. *Journal of Machine Learning Research*, 16(6):155–186.

Rubinstein, B. I. P., Bartlett, P. L., and Rubinstein, J. H. (2009). Shifting: One-inclusion mistake bounds and sample compression. *Journal of Computer and System Sciences*, 75(1):37–59.

Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147.

Shamir, G. I. (2020). Logistic regression regret: What's the catch? In *Proceedings of the Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3296–3319.

Shamir, G. I. and Szpankowski, W. (2021). A lower bound for regret in logistic regression. In *IEEE International Symposium on Information Theory*, pages 2507–2512.

Shelah, S. (1972). A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261.

Simon, H. U. and Szörényi, B. (2010). One-inclusion hypergraph density revisited. *Information Processing Letters*, 110(8–9):341–344.

Voitovych, S., Shetty, A., Golowich, N., and Rakhlin, A. (2026). Learning with simulators: No regret in a computationally bounded world. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 6533–6591.

Vovk, V. (1990). Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–383. Morgan Kaufmann.

Witten, I. H., Neal, R. M., and Cleary, J. G. (1987). Arithmetic coding for data compression. *Communications of the ACM*, 30(6):520–540.

Wu, C., Grama, A., and Szpankowski, W. (2023a). Online learning in dynamically changing environments. In *Proceedings of the Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 325–358.

Wu, C., Grama, A., and Szpankowski, W. (2026). Online universal learning from information-theoretic perspective. *Foundations and Trends in Communications and Information Theory*, 23(3–4):225–443.

Wu, C., Heidari, M., Grama, A., and Szpankowski, W. (2023b). Expected worst case regret via stochastic sequential covering. *Transactions on Machine Learning Research*.

Wu, C., Heidari, M., Grama, A., and Szpankowski, W. (2023c). Regret bounds for log-loss via bayesian algorithms. *IEEE Transactions on Information Theory*, 69(9):5971–5989.

Wu, C., Sima, J., and Szpankowski, W. (2024). Oracle-efficient hybrid online learning with unknown distribution. In *Proceedings of the Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4992–5018.

Wu, C., Wang, Y., Grama, A., and Szpankowski, W. (2023d). Learning functional distributions with private labels. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 37728–37744.

A Proof of the Coordinate-Deletion Inequality

We prove Lemma 5. Fix a finite coordinate set I, a nonempty class $C \subseteq \{0,1\}^I$, a vertex $v \in C$, and a coordinate $j \in I$. Let

$$D = W_{C^{-j}}(p), \qquad p = \pi_j(v),$$

and let $B = B_j(C, v)$ denote the number of members of $\mathcal{Q}_C(v)$ containing j.

If $F_j(p) = \{v\}$, then $v \oplus \mathbf{1}_{\{j\}} \notin C$. Hence no centered cube at v contains j. Deleting coordinate j maps every centered cube at v injectively to a centered cube at p in C^{-j}. Consequently,

$$W_C(v) \le D, \qquad R_j(C, v) \le 1, \qquad B = 0.$$

Suppose now that

$$F_j(p) = \{v, v^{(j)}\}, \qquad v^{(j)} = v \oplus \mathbf{1}_{\{j\}}.$$

For $S \subseteq I \setminus \{j\}$, define

$$a_S = \mathbf{1}\{S \in \mathcal{Q}_C(v)\}, \qquad b_S = \mathbf{1}\{S \in \mathcal{Q}_C(v^{(j)})\}, \qquad c_S = \mathbf{1}\{S \in \mathcal{Q}_{C^{-j}}(p)\}.$$

A cube on $S \cup \{j\}$ is present at v if and only if the complete S-cube is present in both j-layers. The same equivalence holds at $v^{(j)}$. Therefore,

$$W_C(v) = \sum_{S \subseteq I \setminus \{j\}} (a_S + a_S b_S), \tag{33}$$

$$W_C(v^{(j)}) = \sum_{S \subseteq I \setminus \{j\}} (b_S + a_S b_S), \tag{34}$$

$$B = \sum_{S \subseteq I \setminus \{j\}} a_S b_S. \tag{35}$$

Whenever $a_S = 1$ or $b_S = 1$, the projected S-cube is present in C^{-j}, and hence $a_S, b_S \le c_S$. For the four possible pairs (a_S, b_S),

$$a_S + b_S + 2a_S b_S \le c_S + 3a_S b_S.$$

Summing and using (33)–(35) gives

$$W_C(v) + W_C(v^{(j)}) \le D + 3B. \tag{36}$$

Every cube counted by B projects to a centered cube at p, so $B \le D$. It follows from (36) that

$$R_j(C, v) \le 1 + 3B/D \le 4.$$

Moreover, (33) and $a_S, a_S b_S \le c_S$ imply

$$W_C(v) \le 2D.$$

Thus, in both fiber cases,

$$(R_j(C, v) - 1)_+ \le \frac{3B_j(C, v)}{W_{C^{-j}}(p)} \le \frac{6B_j(C, v)}{W_C(v)}. \tag{37}$$

Finally, double counting cube–coordinate incidences gives

$$\sum_{j \in I} B_j(C, v) = \sum_{j \in I} \sum_{S \in \mathcal{Q}_C(v)} \mathbf{1}\{j \in S\} = \sum_{S \in \mathcal{Q}_C(v)} |S| \le d|\mathcal{Q}_C(v)| = dW_C(v),$$

because every $S \in \mathcal{Q}_C(v)$ is shattered and therefore has cardinality at most d. Summing (37) proves

$$\sum_{j \in I} (R_j(C, v) - 1)_+ \leq 6d.$$

The logarithmic statement follows from

$$(\log r)_+ \leq (r-1)_+, \qquad r > 0.$$

B Random-Order Analysis

We prove Proposition 8. Fix a deterministic feature sequence $x^T = (x_1, \ldots, x_T)$ and a full trace

$$v = (h(x_1), \ldots, h(x_T)) \in C_T(x^T).$$

For $S \subseteq [T]$, define

$$C_S = \{(h(x_i))_{i \in S} : h \in \mathcal{H}\}, \qquad v_S = v|_S.$$

If $j \in S$, let

$$R(S, j; v_S) = \frac{\sum_{u \in C_S : u|_{S \setminus \{j\}} = v_{S \setminus \{j\}}} W_{C_S}(u)}{W_{C_{S \setminus \{j\}}}(v_{S \setminus \{j\}})}.$$

This is precisely the ratio in (13) for the class C_S and coordinate j.

Let σ be a uniformly random permutation of $[T]$, and set

$$S_t = \{\sigma(1), \ldots, \sigma(t)\}.$$

When the features are presented in the order $x_{\sigma(1)}, \ldots, x_{\sigma(T)}$, the forward ratio at time t is

$$R_t^\sigma = R(S_t, \sigma(t); v_{S_t}).$$

Define

$$Y_t = (\log R_t^\sigma)_+.$$

Expose the permutation in the reverse order

$$\sigma(T), \sigma(T-1), \ldots, \sigma(1).$$

Let $\mathcal{F}_t$ be the sigma-algebra generated by the already exposed suffix $\sigma(T), \ldots, \sigma(t+1)$. Before $\sigma(t)$ is exposed, the set S_t is known and $\sigma(t)$ is uniform on S_t, even conditional on $\mathcal{F}_t$. Lemma 5 therefore gives

$$\mathbb{E}_\sigma[e^{Y_t} \mid \mathcal{F}_t] = \frac{1}{t} \sum_{j \in S_t} \exp\big((\log R(S_t, j; v_{S_t}))_+\big) \tag{38}$$

$$= \frac{1}{t} \sum_{j \in S_t} \max\{1, R(S_t, j; v_{S_t})\} \tag{39}$$

$$= 1 + \frac{1}{t} \sum_{j \in S_t} (R(S_t, j; v_{S_t}) - 1)_+ \tag{40}$$

$$\leq 1 + \frac{6d}{t}. \tag{41}$$

In the reverse exposure order, the factors corresponding to already exposed coordinates are measurable. Iterated conditional expectation and (41) yield

$$\mathbb{E}_\sigma \exp\left(\sum_{t=1}^{T} Y_t\right) \le \prod_{t=1}^{T}\left(1+\frac{6d}{t}\right) \le e^{6d\mathsf{H}_T}. \tag{42}$$

By Lemma 6,

$$L_T(v;\sigma) \le \sum_{t=1}^{T} Y_t.$$

Consequently,

$$\mathbb{E}_\sigma e^{L_T(v;\sigma)} \le e^{6d\mathsf{H}_T}. \tag{43}$$

Markov's inequality implies that, for every $\varepsilon \in (0,1)$,

$$\mathbb{P}_\sigma\left\{L_T(v;\sigma) > 6d\mathsf{H}_T + \log(1/\varepsilon)\right\} \le \varepsilon. \tag{44}$$

Let

$$N(x^T) = |C_T(x^T)|.$$

There are only $N(x^T)$ distinct full traces, even if $\mathcal{H}$ is infinite. Applying (44) with $\varepsilon = \delta/N(x^T)$ and taking a union bound gives

$$\mathbb{P}_\sigma\left\{\max_{v\in C_T(x^T)} L_T(v;\sigma) > 6d\mathsf{H}_T + \log N(x^T) + \log(1/\delta)\right\} \le \delta. \tag{45}$$

For the expectation bound, use log-sum-exp, Jensen's inequality, and (43):

$$\begin{aligned}
\mathbb{E}_\sigma \max_{v\in C_T(x^T)} L_T(v;\sigma) &\le \mathbb{E}_\sigma \log \sum_{v\in C_T(x^T)} e^{L_T(v;\sigma)} && (46)\\
&\le \log \sum_{v\in C_T(x^T)} \mathbb{E}_\sigma e^{L_T(v;\sigma)} && (47)\\
&\le 6d\mathsf{H}_T + \log N(x^T). && (48)
\end{aligned}$$

Now let X^T have an exchangeable law P, and let σ be an independent uniform permutation. For the bad-event indicator in (45), exchangeability and Fubini's theorem give

$$\mathbb{E}_X B(X^T) = \mathbb{E}_X \mathbb{E}_\sigma B(X_{\sigma(1)},\dots,X_{\sigma(T)}) \le \delta.$$

Similarly, applying (48) conditionally on X^T and using $N(X^T) \le \Pi_{\mathcal{H}}(T)$ gives

$$\mathbb{E}_X \sup_{h\in\mathcal{H}} L_T(h;X^T) \le 6d\mathsf{H}_T + \log \Pi_{\mathcal{H}}(T).$$

This proves Proposition 8.

C Arithmetic-Interval Conversion

We give the complete conversion from the trace probabilities to a finite deterministic cover.

For a fixed feature prefix, assign the empty history the interval

$$I_\varnothing = [0, 1).$$

Suppose a realizable history y^{t-1} has been assigned the interval $I_{y^{t-1}} = [a, b)$. Put

$$c = a + q_t(0 \mid x^t, y^{t-1})(b - a)$$

and define

$$I_{y^{t-1}0} = [a, c), \qquad I_{y^{t-1}1} = [c, b).$$

A zero-probability child receives the empty interval. For a realizable trace v_T^h, induction gives

$$|I_{v_T^h}| = \prod_{t=1}^{T} q_t(h(x_t) \mid x^t, v_{t-1}^h) = e^{-L_T(h;x^T)}.$$

Let

$$L_* = 6d\mathsf{H}_T + \log \Pi_{\mathcal{H}}(T) + \log(1/\delta), \qquad M = \lceil 2e^{L_*} \rceil,$$

and define the fixed grid

$$U = \left\{ \frac{k}{M} : k = 0, \ldots, M - 1 \right\}.$$

Every half-open interval $J = [a, b) \subseteq [0, 1)$ of length at least e^{-L_*} contains an element of U. Indeed, $|J| \geq 2/M$. If $k = \lceil Ma \rceil$, then

$$a \leq k/M < a + 1/M < b.$$

The case $k = M$ would imply $1 - a < 1/M$, contradicting $|J| \geq 2/M$.

For each $u \in U$, define a deterministic sequential function g_u recursively. Given x^t, first reconstruct

$$g_u(x^1), \ldots, g_u(x^{t-1})$$

and the corresponding current interval. Split that interval using (18) and output the bit whose child interval contains u. The half-open children are disjoint and partition the parent, so the output is unique. Starting from the empty history, induction shows that the generated history remains realizable: an infeasible child has probability zero and hence an empty interval.

On the event

$$\sup_{h \in \mathcal{H}} L_T(h; X^T) \leq L_*,$$

every target terminal interval has length at least e^{-L_*} and contains a grid point u. Nestedness then implies

$$g_u(X^t) = h(X_t) \qquad \text{for all } t \leq T.$$

Thus $\mathcal{G} = \{g_u : u \in U\}$ is an exact stochastic global sequential cover. Finally,

$$|\mathcal{G}| \leq M \leq 2e^{L_*} + 1 \leq 3e^{L_*},$$

which proves (7). The simpler bound (8) follows from Sauer–Shelah and $\mathsf{H}_T \leq 1 + \log T$.

D Direct Mixture Implementation

We now give the direct implementation used in Theorem 10. Rather than enumerating the sequential cover, we maintain a mixture over the realizable traces of the observed feature prefix.

For $v = (v_1, \dots, v_t) \in C_t(x^t)$, define its sequential probability under (18) by

$$\pi_t(v \mid x^t) = \prod_{s=1}^{t} q_s(v_s \mid x^s, v_{1:s-1}).$$

Conditional on a trace v, we use independent beta–Bernoulli mixtures for the labels observed on its two sides. For $c \in \{0,1\}$, let

$$n_c(v) = |\{s \le t : v_s = c\}|, \qquad s_c(v; y^t) = \sum_{s \le t:\, v_s = c} y_s,$$

and define

$$B_v(y^t) = \prod_{c=0}^{1} \int_0^1 p^{s_c(v;y^t)} (1-p)^{n_c(v) - s_c(v;y^t)}\, dp.$$

Using the beta integral,

$$B_v(y^t) = \prod_{c=0}^{1} \frac{1}{(n_c(v)+1)\binom{n_c(v)}{s_c(v;y^t)}}.$$

We combine the trace and label probabilities through

$$Q(y^t \| x^t) = \sum_{v \in C_t(x^t)} \pi_t(v \mid x^t) B_v(y^t). \tag{49}$$

We first verify that Q is a sequential probability assignment and derive an explicit recursion. Let $v = ub \in C_t(x^t)$, where $u \in C_{t-1}(x^{t-1})$ and $b \in \{0,1\}$. By construction,

$$\pi_t(ub \mid x^t) = \pi_{t-1}(u \mid x^{t-1}) q_t(b \mid x^t, u).$$

Moreover, conditional on the child trace $v = ub$, the beta–Bernoulli predictive probability of $Y_t = 1$ is

$$p_t(v) = \frac{1 + \sum_{s<t:\, u_s = b} y_s}{2 + |\{s < t : u_s = b\}|}. \tag{50}$$

Consequently,

$$B_v(y^t) = B_u(y^{t-1}) p_t(v)^{y_t} (1 - p_t(v))^{1-y_t}.$$

In particular,

$$\sum_{y_t \in \{0,1\}} B_v(y^t) = B_u(y^{t-1}). \tag{51}$$

Summing first over the two possible outcomes and then over the feasible children of each parent gives

$$\sum_{y_t \in \{0,1\}} Q(y^t \| x^t) = Q(y^{t-1} \| x^{t-1}).$$

Thus Q defines a sequential probability assignment, and the corresponding prediction is

$$\widehat{p}_t = \frac{Q(y^{t-1}1 \| x^t)}{Q(y^{t-1} \| x^{t-1})}.$$

We next write this prediction as an explicit online recursion. After observing (x^{t-1}, y^{t-1}), let

$$\rho_{t-1}(u) = \frac{\pi_{t-1}(u \mid x^{t-1}) B_u(y^{t-1})}{Q(y^{t-1} \| x^{t-1})}$$

be the posterior mass of $u \in C_{t-1}(x^{t-1})$, with $\rho_0(\varnothing) = 1$. Once x_t is observed, propagate this mass to each feasible child $v = ub$ according to

$$r_t(v) = \rho_{t-1}(u) q_t(b \mid x^t, u).$$

Using the local prediction $p_t(v)$ from (50), the learner predicts

$$\widehat{p}_t = \sum_{v \in C_t(x^t)} r_t(v) p_t(v).$$

After observing y_t, the posterior is updated by

$$\rho_t(v) = \frac{r_t(v) p_t(v)^{y_t} (1 - p_t(v))^{1-y_t}}{\widehat{p}_t^{\,y_t} (1 - \widehat{p}_t)^{1-y_t}}.$$

The beta–Bernoulli predictions lie strictly between zero and one, so all denominators are positive. The identities above show by induction that this recursion implements exactly the mixture (49).

Regret of the direct mixture. Sequential consistency implies that the cumulative log loss of the predictor is $-\log Q(y^T \| x^T)$. Fix a comparator $f(x) = a + (b-a)h(x) \in \mathcal{F}_{\mathcal{H}}$ and let $v = h(x^T)$. Retaining only the term corresponding to v in (49) gives

$$Q(y^T \| x^T) \geq \pi_T(v \mid x^T) B_v(y^T).$$

Let n_c be the number of rounds with $v_t = c$ and s_c the number of ones observed on those rounds. Writing $\theta_0 = a$ and $\theta_1 = b$, the elementary inequality

$$\binom{n_c}{s_c} \theta_c^{s_c} (1 - \theta_c)^{n_c - s_c} \leq 1$$

implies

$$\frac{1}{(n_c + 1)\binom{n_c}{s_c}} \geq \frac{\theta_c^{s_c} (1 - \theta_c)^{n_c - s_c}}{n_c + 1}.$$

Since $(n_0 + 1)(n_1 + 1) \leq (T + 1)^2$ and $-\log \pi_T(v \mid x^T) = L_T(h; x^T)$, we obtain

$$-\log Q(y^T \| x^T) \leq \sum_{t=1}^{T} \ell(f(x_t), y_t) + L_T(h; x^T) + 2 \log(T + 1).$$

Taking the infimum over $f \in \mathcal{F}_{\mathcal{H}}$ therefore gives, for every x^T and y^T,

$$R_T(A; x^T, y^T; \mathcal{F}_{\mathcal{H}}) \leq \sup_{h \in \mathcal{H}} L_T(h; x^T) + 2 \log(T + 1).$$

Taking the supremum over y^T, followed by expectation over the features, and applying Proposition 8 proves (29) for every exchangeable feature law.

E Efficient Implementation for Planar Halfspaces

We now prove Theorem 10. The implementation has two components: enumerating all realizable labelings of the observed feature prefix and computing their cube weights.

Let $x_i = (a_i, b_i) \in \mathbb{R}^2$. We first show that the trace class $C_t(x^t)$ can be enumerated exactly. Under the convention

$$h(x) = \mathbf{1}\{w^\top x + c \geq 0\},$$

every nonconstant feasible labeling of a finite sample can be realized with a strict margin. Indeed, the negatively labeled points already satisfy strict inequalities, while increasing c by a sufficiently small amount makes all positive inequalities strict without changing any negative label.

A strict labeling is stable under a sufficiently small perturbation of the normal vector. Hence it suffices to consider normals with nonzero first coordinate. Up to scaling and complementing the labeling, such a normal can be written as $(1, s)$. For a fixed s, the ordering of the projected values

$$a_i + sb_i$$

changes only when two projections coincide, namely at one of the critical values

$$s = -\frac{a_i - a_j}{b_i - b_j}, \qquad b_i \neq b_j.$$

There are $O(t^2)$ such values. Choose one rational representative from each interval between consecutive distinct critical values and from the two unbounded intervals. For each representative direction, sort the projected values and enumerate all threshold cuts and their complements. Equal projected values are kept together, and cuts are made only between consecutive distinct values. Since every strictly feasible labeling persists on an open set of directions, this procedure finds every nonconstant realizable labeling. Constant labelings are added separately, and duplicate traces are removed. The $O(t^3)$ candidate traces can be generated and deduplicated in $O(t^4)$ operations by storing their t bits in a binary trie.

The number of distinct labelings is at most the classical affine halfspace growth bound

$$|C_t(x^t)| \leq 2\sum_{k=0}^{2}\binom{t-1}{k} = t^2 - t + 2.$$

Degeneracies and repeated feature points can only reduce this number.

It remains to compute the cube weights required by (18). Planar affine halfspaces have VC dimension three, so every centered cube contains at most three coordinates. For a fixed trace $v \in C_t(x^t)$, it therefore suffices to enumerate the

$$\sum_{k=0}^{3}\binom{t}{k} = O(t^3)$$

coordinate sets of size at most three. To test the candidate cubes, assign an integer identifier to each trace and precompute a table of one-coordinate-flip neighbors. Comparing every pair of the $O(t^2)$ traces in $O(t)$ operations identifies every such neighbor and takes $O(t^5)$ operations in total. A table entry records the neighboring trace's identifier, or records that the neighbor is absent. For each subset $A \subseteq S$ of a candidate set S, start at v and follow the neighbor-table entries for the coordinates of A in a fixed order. This checks all $2^{|S|} \leq 8$ corners using at most $\sum_{A\subseteq S}|A| \leq 12$

table accesses; a missing neighbor makes the candidate infeasible. Thus all cube weights can be computed in

$$O(t^2) \cdot O(t^3) = O(t^5)$$

operations at round t. Once these weights are known, the transition probabilities in (18) and the mixture recursion of Appendix D require only lower-order work. Summing over $t \le T$ gives

$$\sum_{t=1}^{T} O(t^5) = O(T^6)$$

combinatorial and rational-arithmetic operations.

Finally, suppose the input coordinates are rational. The critical slopes are rational, and rational representatives between consecutive critical slopes can be chosen with bit length polynomial in the input bit length. The cube weights are integers of size $O(t^3)$, so each transition probability in (18) has numerator and denominator of bit length $O(\log(t+1))$. The product defining $\pi_t(v \mid x^t)$ therefore has numerator and denominator of bit length $O(t \log(t+1))$. Since $\binom{n}{s} \le 2^n$, the beta–Bernoulli factor $B_v(y^t)$ has numerator and denominator of bit length $O(t)$. Each unnormalized mass $\pi_t(v \mid x^t)B_v(y^t)$ consequently has $O(t \log(t+1))$-bit components. Summing at most $O(t^2)$ such masses gives $Q(y^t \| x^t)$ with numerator and denominator of bit length $O(t^3 \log(t+1))$, even using the product of the individual denominators as a common denominator. The posterior masses and prediction probabilities are ratios of these quantities, so their reduced fractions also have polynomial bit length. Performing the rational arithmetic with reduction after each operation therefore yields an exact implementation with bit complexity polynomial in T and the input bit length.

F A Matching Lower Bound for Exact Binary Covers

Fix $d \ge 1$ and let $\mathcal{X} = [0,1]$. Consider

$$\mathcal{H}_d = \{x \mapsto \mathbf{1}\{x \in B\} : B \subseteq [0,1],\ |B| \le d\}.$$

This class has VC dimension d. Let μ be any nonatomic distribution on $[0,1]$. With probability one, $X_1, \dots, X_T$ are distinct. On every such sample, the trace class consists of all binary vectors with at most d ones and hence has cardinality

$$\sum_{k=0}^{\min\{d,T\}} \binom{T}{k}.$$

On a fixed sample, one deterministic sequential function produces one binary trace. Therefore an exact cover that simultaneously covers all hypotheses must contain at least as many functions as there are distinct traces. If its success probability is positive, this lower bound holds on at least one sample in the success event, and consequently

$$|\mathcal{G}| \ge \sum_{k=0}^{\min\{d,T\}} \binom{T}{k}.$$

For $T \ge 2d$,

$$\log |\mathcal{G}| \ge \log \binom{T}{d} \ge d \log(T/d).$$

For fixed positive d, this matches the $\log T$ order of Theorem 2. It does not establish optimal dependence on jointly growing d and T.

G Proof of the Pseudo-Dimension Separation

We prove Theorem 11 in the following stronger form.

Theorem 12. *There are universal constants $c > 0$ and $T_0 \geq 1$ such that the class $\mathcal{F}_*$ in (31) has pseudo-dimension one and satisfies the following. For every $T \geq T_0$ and every $\alpha \in (0, 1/64]$, every distribution-blind uniform stochastic sequential α-cover $\mathcal{G}$ of $\mathcal{F}_*$ with failure probability $1/4$ satisfies*

$$\log |\mathcal{G}| \geq c \log T \log(1/\alpha).$$

The proof has three ingredients. First, although $\mathcal{F}_*$ has pseudo-dimension one, a suitable feature can force the value of f_θ to reveal one of $K \asymp 1/\alpha$ well-separated possibilities. Second, we arrange these K-way choices along a hidden tree and choose an i.i.d. feature distribution supported on one hidden root-to-leaf path. Whenever the sampled depth reaches a new record, a sequential predictor must guess a fresh hidden digit. Third, a random permutation has order $\log T$ records, which yields the additional $\log T$ factor.

G.1 Pseudo-dimension one

Recall that

$$\mathcal{F}_* = \{(a, b) \mapsto \sigma(a\theta + b) : \theta \in (0, 1)\}, \qquad a > 0.$$

Fix a feature (a, b) and a threshold $r \in (0, 1)$. Since σ is strictly increasing,

$$\mathbf{1}\{\sigma(a\theta + b) \geq r\} = \mathbf{1}\left\{\theta \geq \frac{\operatorname{logit}(r) - b}{a}\right\}.$$

Thus, for any finite collection of features and thresholds, the induced binary label vector is obtained by varying a single threshold parameter θ on the real line. These binary patterns form a chain and cannot shatter two coordinates. Hence

$$\operatorname{Pdim}(\mathcal{F}_*) \leq 1.$$

Conversely, for example, the single feature $(1, 0)$ with threshold $\sigma(1/2)$ realizes both binary labels as θ varies in $(0, 1)$. Therefore

$$\operatorname{Pdim}(\mathcal{F}_*) = 1.$$

G.2 Encoding a hidden K-ary path

Fix $\alpha \leq 1/64$ and let

$$K = \left\lfloor \frac{1}{16\alpha} \right\rfloor.$$

Then $K \geq 4$. Define K closed intervals

$$J_k = \left[\frac{1}{4} + 4\alpha(k-1), \frac{1}{4} + 4\alpha(k-1) + \alpha\right], \qquad k = 1, \ldots, K.$$

They all lie in $(0, 1)$, and two distinct bands are separated by more than 2α. Consequently,

$$\text{for every } z \in \mathbb{R}, \qquad \operatorname{dist}(z, J_k) \leq \alpha \text{ for at most one } k. \tag{52}$$

We now recursively construct a full K-ary tree of nested intervals in the parameter space $(0,1)$. Let

$$I_\varnothing = (0,1).$$

Suppose the interval

$$I_u = (\ell_u, r_u)$$

has already been associated with a node u. Choose numbers

$$z_- < \operatorname{logit}(\min J_1), \qquad z_+ > \operatorname{logit}(\max J_K),$$

and choose $a_u > 0$ and $b_u \in \mathbb{R}$ so that the affine map

$$\phi_u(\theta) = a_u\theta + b_u$$

maps I_u bijectively onto (z_-, z_+). For $k = 1, \dots, K$, define

$$I_{uk} = \phi_u^{-1}\big(\operatorname{logit}(\operatorname{int} J_k)\big).$$

The intervals $I_{u1}, \dots, I_{uK}$ are nonempty and pairwise disjoint, and their closures are contained in I_u.

Associate with node u the feature

$$x_u = (a_u, b_u).$$

By construction,

$$\theta \in I_{uk} \quad \Longrightarrow \quad f_\theta(x_u) = \sigma(a_u\theta + b_u) \in \operatorname{int} J_k. \tag{53}$$

Thus, evaluating at x_u reveals which one of the K child intervals contains θ, up to accuracy α.

Fix a depth L and let

$$B_1, \dots, B_L \overset{\text{iid}}{\sim} \operatorname{Unif}\{1, \dots, K\}.$$

The vector

$$B = (B_1, \dots, B_L)$$

specifies a random root-to-leaf path. Choose an arbitrary point

$$\theta_B \in I_{B_1 \cdots B_L}.$$

For $r = 1, \dots, L$, define

$$x_r(B) = x_{B_1 \cdots B_{r-1}}.$$

Since the terminal interval is contained in all of its ancestors, (53) gives

$$f_{\theta_B}(x_r(B)) \in J_{B_r}, \qquad r = 1, \dots, L. \tag{54}$$

For each fixed path B, define the feature distribution

$$\mu_B = \frac{1}{L}\sum_{r=1}^{L} \delta_{x_r(B)}.$$

Equivalently, let

$$D_1, \dots, D_T \overset{\text{iid}}{\sim} \operatorname{Unif}\{1, \dots, L\}$$

and set

$$X_t = x_{D_t}(B).$$

Conditional on B, the variables $X_1, \dots, X_T$ are therefore i.i.d. from the fixed distribution μ_B.

The role of distribution-blindness is important here. The cover $\mathcal{G}$ must be fixed before the feature distribution is chosen, so the *same* family $\mathcal{G}$ must work simultaneously for every μ_B. This allows us to randomize the hidden path B after fixing the cover.

G.3 Every record forces a fresh K-way guess

Call round t a *record time* if

$$D_t > \max_{s<t} D_s,$$

with the first round counted as a record. Let

$$R(D^T)$$

denote the total number of record times. The next lemma is the core of the argument.

Lemma 13. *Fix the depth sequence D^T and a deterministic sequential rule*

$$g : \bigcup_{t=1}^{T} \mathcal{X}^t \to [0,1].$$

Then

$$\mathbb{P}_B \left\{ \max_{t \le T} \left| g(X^t) - f_{\theta_B}(X_t) \right| \le \alpha \;\middle|\; D^T \right\} \le K^{-R(D^T)}. \tag{55}$$

Proof. Let

$$t_1 < \cdots < t_R$$

be the record times and write

$$r_j = D_{t_j}.$$

Then

$$r_1 < r_2 < \cdots < r_R.$$

Consider a record time t_j with depth $r = r_j$. Since every previously sampled depth is strictly smaller than r, each feature in the prefix

$$X^{t_j}$$

depends only on

$$B_1, \ldots, B_{r-1}.$$

Indeed,

$$x_q(B) = x_{B_1 \cdots B_{q-1}}$$

depends only on the first $q-1$ digits, and every depth observed up to time t_j is at most r. Therefore

$$g(X^{t_j})$$

is a function of D^T and $B_1, \ldots, B_{r-1}$, and in particular does not depend on the fresh digit B_r.

Define a coarse guess $\widehat{B}_r$ from $g(X^{t_j})$ as follows: if there is a unique k such that

$$\operatorname{dist}(g(X^{t_j}), J_k) \le \alpha,$$

set $\widehat{B}_r = k$; otherwise set $\widehat{B}_r = 0$. Uniqueness follows from (52). The random variable $\widehat{B}_r$ is measurable with respect to

$$D^T, B_1, \ldots, B_{r-1}.$$

If g approximates the target at round t_j, then by (54),

$$f_{\theta_B}(X_{t_j}) = f_{\theta_B}(x_r(B)) \in J_{B_r}.$$

Hence

$$\left|g(X^{t_j}) - f_{\theta_B}(X_{t_j})\right| \le \alpha \quad \Longrightarrow \quad \widehat{B}_r = B_r.$$

Thus the event that g approximates the target on *all* rounds is contained in

$$\bigcap_{j=1}^{R} \{\widehat{B}_{r_j} = B_{r_j}\}.$$

Now reveal the digits in increasing order of the record depths. At depth r_j, the guess $\widehat{B}_{r_j}$ is determined by digits with index strictly smaller than r_j, whereas B_{r_j} is still independent and uniform on $\{1, \dots, K\}$. Consequently,

$$\mathbb{P}\left(\widehat{B}_{r_j} = B_{r_j} \mid D^T, B_1, \dots, B_{r_j-1}\right) \le \frac{1}{K}.$$

Applying the chain rule over the R record depths gives

$$\mathbb{P}_B\left(\bigcap_{j=1}^{R} \{\widehat{B}_{r_j} = B_{r_j}\} \;\middle|\; D^T\right) \le K^{-R}.$$

This proves (55). □

The proof deliberately works with the coarse events $\{\widehat{B}_r = B_r\}$ rather than conditioning on the exact approximation events themselves. An exact approximation event can depend on the precise value of f_{θ_B} and hence on deeper path digits, whereas the coarse guess at a record depth depends only on the already exposed prefix.

G.4 From one rule to a cover

Let

$$M = |\mathcal{G}|.$$

Fix D^T. By Lemma 13 and a union bound over the M sequential rules,

$$\mathbb{P}_B\left\{\exists g \in \mathcal{G} : \max_{t \le T} |g(X^t) - f_{\theta_B}(X_t)| \le \alpha \;\middle|\; D^T\right\} \le \min\{1, MK^{-R(D^T)}\}. \tag{56}$$

We now use the defining quantifiers of a distribution-blind cover. For every fixed path B, the same family $\mathcal{G}$ must be an α-cover under the distribution μ_B. Therefore, for the particular target f_{θ_B},

$$\mathbb{P}_{D^T}\left\{\exists g \in \mathcal{G} : \max_{t \le T} |g(X^t) - f_{\theta_B}(X_t)| \le \alpha\right\} \ge \frac{3}{4}.$$

Averaging this inequality over the random path B and exchanging the order of expectation gives

$$\frac{3}{4} \le \mathbb{E}_{D^T} \min\{1, MK^{-R(D^T)}\}. \tag{57}$$

Thus all that remains is to understand the number of records in an i.i.d. sequence of random depths.

G.5 The record count

Let

$$\mathcal{C}_L = \{D_1, \ldots, D_T \text{ are all distinct}\}.$$

A union bound gives

$$\mathbb{P}(\mathcal{C}_L^c) \le \frac{T(T-1)}{2L}.$$

Conditional on $\mathcal{C}_L$, the relative ordering of $D_1, \ldots, D_T$ is a uniformly random permutation of $[T]$. The record count depends only on this relative ordering. Hence, if R_T denotes the number of records in a uniformly random permutation of $[T]$, then for every function $\psi : \{1, \ldots, T\} \to [0, 1]$,

$$\left|\mathbb{E}\psi(R(D^T)) - \mathbb{E}\psi(R_T)\right| \le \mathbb{P}(\mathcal{C}_L^c).$$

Since (57) holds for every L, letting $L \to \infty$ with

$$\psi(r) = \min\{1, MK^{-r}\}$$

gives

$$\frac{3}{4} \le \mathbb{E}\min\{1, MK^{-R_T}\}. \tag{58}$$

For completeness, recall the probability generating function of R_T. If ρ_t is the rank of the tth element among the first t elements of a uniformly random permutation, then

$$\rho_1, \ldots, \rho_T$$

are independent and ρ_t is uniform on $\{1, \ldots, t\}$. A record occurs exactly when $\rho_t = t$. Therefore

$$\mathbb{E}z^{R_T} = \prod_{t=1}^{T} \left(1 - \frac{1}{t} + \frac{z}{t}\right). \tag{59}$$

Set

$$\lambda = \frac{\log 2}{\log K}.$$

Since $K \ge 4$, we have $0 < \lambda \le 1$. For every $u \ge 0$,

$$\min\{1, u\} \le u^\lambda.$$

Using $K^{-\lambda} = 1/2$ in (58), followed by (59), yields

$$\frac{3}{4} \le M^\lambda \mathbb{E}K^{-\lambda R_T} = M^\lambda \mathbb{E}2^{-R_T} = M^\lambda \prod_{t=1}^{T} \left(1 - \frac{1}{2t}\right) \le M^\lambda e^{-\mathsf{H}_T/2}.$$

Hence

$$\log M \ge \frac{\log K}{\log 2} \left(\frac{\mathsf{H}_T}{2} + \log \frac{3}{4}\right). \tag{60}$$

For $T \ge 2$,

$$\frac{\mathsf{H}_T}{2} + \log \frac{3}{4} \ge \frac{\mathsf{H}_T}{4},$$

and $\mathsf{H}_T \geq \log T$. Moreover, since $\alpha \leq 1/64$,

$$K = \left\lfloor \frac{1}{16\alpha} \right\rfloor \geq \frac{1}{32\alpha}.$$

Because $\log(1/\alpha) \geq \log 64$,

$$\log K \geq \log \frac{1}{\alpha} - \log 32 \geq \frac{1}{6} \log \frac{1}{\alpha}.$$

Substituting these two bounds into (60) gives

$$\log M \geq \frac{1}{24 \log 2} \log T \log(1/\alpha)$$

for all $T \geq 2$. This proves Theorem 12, with a universal constant c. Taking $\alpha = 1/T$ proves Theorem 11.

H Measurability and Standard Examples

For a fixed horizon t, there are finitely many binary vectors $v \in \{0,1\}^t$. Assumption (5) says that the indicator of the event $v \in C_t(x^t)$ is measurable for each such vector. Therefore the finite-valued mapping $x^t \mapsto C_t(x^t)$ is measurable when the set of trace classes carries its discrete sigma-algebra. Cube counts, feasibility decisions, and the ratios in (18) are functions of this finite object and the finite label history. Thus q_t is measurable.

For a fixed seed u, the endpoints of its arithmetic interval are measurable functions of the feature prefix, by induction on t. Comparing u to the split point gives a measurable bit, so g_u is measurable. To see that the uniform log-loss event is measurable without taking an uncountable union, enumerate all 2^T binary vectors. For each v, its realizability and its path probability are measurable. The event is the finite union over v of realizability together with the corresponding loss exceeding the threshold. The cover failure event can similarly be written as a finite union over binary traces, followed by a finite intersection over covering predictors. Supremum regret over binary label sequences is also a finite maximum for the constructed strategies.

If $\mathcal{H}$ is countable and each hypothesis is measurable, (5) is a countable union of finite intersections of measurable sets. This is a sufficient condition, not a restriction to countable hypothesis classes. For the sparse-set class in Appendix F, a labeling is feasible exactly when equal feature values receive equal labels and at most d distinct feature values receive label one. These conditions are Borel on $[0,1]^t$.

For affine halfspaces on $\mathbb{R}^m$, every feasible labeling of a finite sample can also be realized with rational coefficients. To prove this, start from a realizing pair (w, a) for the rule $\mathbf{1}\{\langle w, x\rangle \geq a\}$. Every negatively labeled sample point has a strictly negative value of $\langle w, x_i\rangle - a$. Decrease a by a sufficiently small positive amount. All positive points now have a strictly positive margin, and all negative points retain a strictly negative margin. There are finitely many inequalities, so a sufficiently close rational pair preserves every sign. Thus each finite-feasibility set is a countable union over rational coefficients of Borel sets.

For completeness, affine halfspaces in $\mathbb{R}^m$ have VC dimension $m+1$. An affinely independent set of $m+1$ points can receive any prescribed $\{-1,1\}$ values from an affine function, and is therefore shattered. Any $m+2$ points have a nonzero affine dependence $\sum_i \lambda_i x_i = 0$, $\sum_i \lambda_i = 0$. There is at least one positive and one negative coefficient. Label the points with positive coefficients one and those with negative coefficients zero. A realizing affine function f would satisfy $\lambda_i f(x_i) \geq 0$ for positive coefficients and $\lambda_i f(x_i) > 0$ for negative coefficients. This contradicts $\sum_i \lambda_i f(x_i) = 0$. Coefficients equal to zero do not affect the argument.

I Detailed Related Work and Comparison of Protocols

I.1 Unknown stochastic features and adversarial labels

The hybrid learning framework dates to Lazaric and Munos (2012), whose conference precursor appeared in 2009. Their learners use empirical covers with stochastic inputs and adversarial outputs. Rakhlin et al. (2011) place stochastic and constrained adversaries in a general minimax framework. Distribution knowledge matters: in our game, one sequential strategy is fixed before the choice of an arbitrary feature law. Moreover, the supremum over the entire label sequence is inside the feature expectation. Results for jointly i.i.d. labeled samples or for a known reference measure do not directly imply this guarantee.

The two-level comparator $\mathcal{F}_{\mathcal{H}}$ is established prior work. Bhatt and Kim (2021) study precisely this model under conditionally Bernoulli labels and ask for a constructive logarithmic bound for general finite-VC partitions. Bhatt et al. (2023b, Theorem 11) analyze statewise categorical probability mixtures with adversarial outcomes; their binary two-state case is especially close to our application. Their general i.i.d. bound retains an empirical-process term of square-root order. Thus neither the partition model nor the use of independent Bernoulli mixtures is introduced here. The new ingredient is the cube-weighted sequential distribution over latent traces and its simultaneous random-order guarantee.

Wu et al. (2023b, Example 2, Proposition 2, and Corollary 2) give the direct stochastic-cover predecessor for this same expected-worst-case log-loss problem. Their general VC rate is $O((d+1)\log^2 T)$. They already obtain logarithmic covers for subclasses controlled by star-type dimensions and for certain interval and cuboid compositions. Our binary result applies without these additional structural conditions. The finite-VC polynomial-cover question in their Remark 4, including planar halfspaces, also appears in Wu et al. (2026, Remark 7.1).

Bhatt et al. (2023a) obtain logarithmic regret for two-level VC models under smoothed features relative to a *known* reference distribution; at smoothing parameter one, this includes known-law i.i.d. features. Their transductive reduction does not provide the single distribution-blind strategy required here. Extensions of stochastic covering to unknown changing feature distributions appear in Wu et al. (2023a). Further applications include private-label functional distributions (Wu et al., 2023d) and hybrid-learning reductions for omniprediction (Okoroafor et al., 2025). These results establish the broader relevance of the cover notion, while using different loss or observation models.

I.2 One-inclusion geometry, compression, and computation

The random-order viewpoint originates in one-inclusion prediction (Haussler et al., 1994). The trace growth estimates are the classical Sauer–Shelah inequalities (Sauer, 1972; Shelah, 1972). Cubical structure, shifting, and compression have been studied extensively (Rubinstein et al., 2009; Chalopin et al., 2019). In that language, $W_C(v)$ counts all cubes of the trace class incident to v, including its zero-dimensional cube. These objects are classical; the contribution here is their use as probability weights and the pointwise coordinate-deletion inequality of Lemma 5. It does not require the trace class to be maximum, ample, or to admit a corner peeling. Recent agnostic PAC work also uses Boolean-cube orientations and leave-one-out analysis (Mathiasen et al., 2026); its target is classification risk under jointly i.i.d. data rather than our uniform cumulative log-loss bound.

The multiclass one-inclusion literature deserves particular care. Rubinstein et al. (2009) bound one-inclusion hypergraph density by Pollard pseudo-dimension. Simon and Szörényi (2010) repair the proof, showing that the original assertion that shifting preserves pseudo-dimension is false. The corrected density result concerns expected categorical prediction errors. It does not by itself bound the uniform pathwise probability cost needed by our covering argument. Our lower bound is

consistent with that result: a fresh probability value can require choosing among many separated bands.

Arithmetic intervals (Witten et al., 1987) and mistake-location expert reductions (Ben-David et al., 2009) are established coding devices. Our conversion uses a common fixed seed grid with the cube predictor's uniform log-loss bound. Similarly, planar trace counts and arrangement enumeration are classical (Cover, 1965; Edelsbrunner et al., 1986). The computational contribution is their combination with exact cube weights and the posterior recursion, including polynomial bit complexity for rational planar features. The method is polynomial for fixed dimension; its degree grows with dimension. Oracle-efficient unknown-law hybrid learning (Wu et al., 2024) and the more recent constrained-label setting of Okoroafor et al. (2026) address a different computational model. Simulator-based classification (Voitovych et al., 2026) also gives the learner additional access to the feature process. These results do not justify a lower bound on the running time of earlier log-loss algorithms, and we make no such claim.

I.3 Pseudo-dimension, covering entropy, and log-loss

Classical scale-sensitive dimensions control empirical covers and learnability for real-valued classes (Alon et al., 1997). Gapped variants and their relation to empirical and sequential entropy are studied by Jia et al. (2025a). The distinctions between a cover constructed after a fixed design is known, a cover on a specified feature tree (Rakhlin et al., 2015), and one fixed family of sequential rules valid for every unknown i.i.d. law are essential. The lower bound in Theorem 11 concerns the last notion. It therefore does not contradict small ordinary empirical covers for finite pseudo-dimension classes.

Random-branch lower bounds in distribution-dependent learning use an unseen deeper tree node to force a fresh prediction (Bousquet et al., 2021, 2023). Our construction uses this general principle with nested parameter intervals realizing K separated sigmoid bands. It counts repeated record depths and bounds the success of an entire finite sequential cover. This produces the additional $\log K$ cost within one fixed, ordered sigmoid family. The feature distribution may depend on the horizon and chosen branch; the function class does not. We do not assert a single hard distribution simultaneously for all horizons.

Covering entropy provides useful upper bounds for log-loss but need not characterize its exact minimax regret. Sequential probability assignment and refined entropy bounds are developed by Rakhlin and Sridharan (2015); Bilodeau et al. (2020); smooth truncation and Bayesian covers are analyzed by Wu et al. (2023c). More recent square-root entropy bounds are tailored to the geometry of probability assignment (Jia et al., 2025b). These results motivate retaining the distinction between our uniform approximation lower bound and a binary log-loss lower bound. We have not transferred any adversarial-tree entropy characterization to the distribution-blind i.i.d. game without proof. Bayesian aggregation and mixability themselves are classical (Vovk, 1990; Cesa-Bianchi and Lugosi, 2006).

I.4 Logistic regret and efficient improper learning

Logarithmic logistic regret from Bayesian prediction predates the present work (Kakade and Ng, 2004). Lower bounds for proper bounded-domain learning (Hazan et al., 2014) motivate the improper strategies of Foster et al. (2018). The computational line includes binary logistic prediction (Jézéquel et al., 2020), efficient multiclass aggregation (Jézéquel et al., 2021; Agarwal et al., 2022), and the recent mixture-of-sigmoids implementation (Di Gennaro et al., 2026). These works provide increasingly efficient algorithms under norm-dependent guarantees. Shamir (2020) analyze how the norm constraint and dimension affect regret. Consequently our planar theorem should be read as a

guarantee for two-level partitions; it does not replace these algorithms for sigmoid probabilities.

Mayo et al. (2022) adapt to unknown comparator and feature scales. Their scale-free bounds still depend on the realized scale, including the comparator–feature norm product. The absence of an advance norm bound is therefore distinct from a regret guarantee depending only on dimension and horizon. A dimension-only logarithmic bound for unrestricted logistic regression under unknown, causally revealed feature laws remains open.

I.5 Full-design logistic regret and statistical risk

In the transductive model, the entire feature multiset is available before prediction. Qian et al. (2026, Theorem 13) obtain unrestricted logistic regret at most $6q\log(eT)$ in this model, with adversarial labels and no feature or parameter norm constraint. Their strategy can be approximated in polynomial time at fixed dimension using slab enumeration and log-concave sampling. It is an especially important comparison to both our sequential planar implementation and our unresolved unrestricted-logit question: the missing assumption here is advance access to the feature multiset, rather than simply computational power.

Precise logistic regret through Shtarkov sums has a substantial fixed-design literature, including lower bounds (Shamir and Szpankowski, 2021), categorical and general feature asymptotics (Jacquet et al., 2021, 2022), unbounded weights (Drmota et al., 2024), regularized unbounded weights (Drmota et al., 2025), and large-weight phase transitions (Drmota et al., 2026). The latter also averages fixed-design regret over certain random designs; that ordering still gives the learner the full design in each game. The recent preprint of Chardon et al. (2026, Theorem 4.1) bounds the unrestricted logistic Shtarkov sum for arbitrary fixed designs by $\sum_{j=0}^{q}\binom{T}{j}$, where q is the parameter dimension. Such design-dependent probability assignments do not directly give a sequential strategy for unknown i.i.d. features.

Statistical excess risk is a further distinct target. Bilodeau et al. (2023) analyze well-specified conditional density estimation through empirical entropy. Mourtada and Gaïffas (2022) give an improper estimator for misspecified density estimation and logistic regression, and Qian et al. (2026) derive norm-independent batch risk bounds through transductive priors. Jointly i.i.d. labeled data permit statistical reductions unavailable when the labels are chosen adversarially after inspecting the features. These batch results therefore concern a distinct protocol. In addition, the efficient specialization of Qian et al. (2026) assumes a well-specified logistic response model and regular covariate distributions.