THE MOVING BLIND SPOT IN OOD DETECTION: MIXING MAKES UNIVERSAL RATES UNIFORM

Anonymous authors
Paper under double-blind review

ABSTRACT

An out-of-distribution (OOD) detector learns from in-distribution (ID) data and is judged on a distribution it never observes. Problems sharing an ID law are indistinguishable and form an observation fiber, on which every learner is a schedule fixed in advance. The sample size is only a clock, and since a clock can run at any speed, a fiber admits every universal rate or none. We ask which, for errors against a detector class at all ID/OOD proportions. A fiber can be consistent without being uniformly learnable: each hidden law is eventually served, yet at every time some law is missed, a moving blind spot. Mixing hidden laws preserves consistency and the worst-case value of a fiber, yet our main theorem shows it turns universal rates into uniform ones: a learner attains a rate on the countable-mixture completion of a fiber exactly when it attains it uniformly. The key step, that consistency makes the class-relative optimal risk affine under mixing, holds verbatim for AUC and any criterion linear in the hidden law. Once hidden laws can be mixed, a moving blind spot admits no rate: one hidden law, condensing the blind spots of infinitely many times, defeats any candidate rate. Nontrivial rates arise only across fibers: selection costs an exponentially small error over finitely many ID laws, a sub-exponential one over countably many, and can defeat consistency over uncountably many, by Baire's theorem. With unrestricted detectors on measure-separated domains, all slowness is statistical, from an exact $1/n$ rate to none.

> *"... if they review some, the induction will be insecure, since some of the particulars omitted in the induction may contravene the universal; while if they are to review all, they will be toiling at the impossible, since the particulars are infinite and indefinite."* (explained in Appendix J)
> — Sextus Empiricus, *Outlines of Pyrrhonism* II.204 (1933)

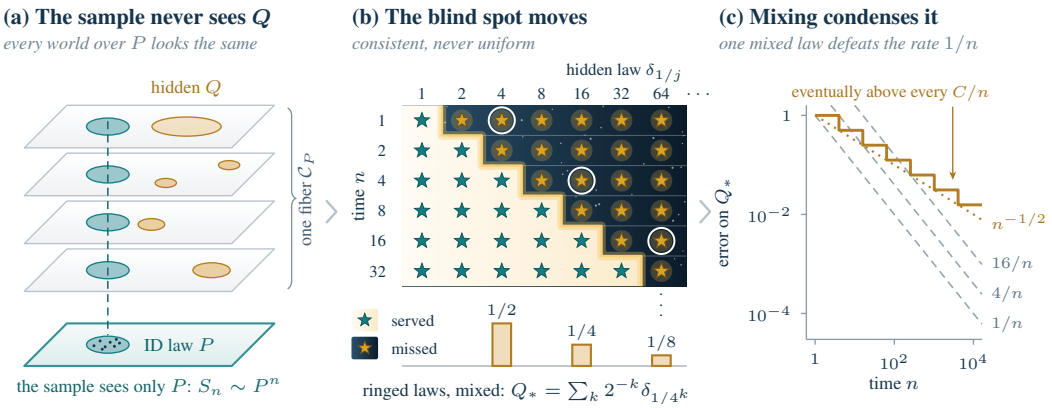

Figure 1: **The paper at a glance.** (a) Worlds (P, Q) sharing the ID law P give the same sample: on this fiber $\mathcal{C}_P$ every learner is a schedule fixed in advance, and n is only a clock (schematic). (b) At time n the prefix schedule serves the points $j \le n$ (Proposition 3.5; powers of 2 shown): every star is lit in time, yet every row keeps stars in the dark. (c) Mixing the ringed blind spots gives one law Q_* with exact error at least $n^{-1/2}$ (dotted), above every C/n once $n > C^2$. By Theorem 3.3, on mixtures only uniform rates survive.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

An out-of-distribution (OOD) detector is trained on in-distribution (ID) data and judged on a distribution it never sees. This asymmetry has a consequence that organizes the whole paper. Two OOD problems with the same ID law generate the same training data; no learner can tell them apart, so whatever it outputs, it outputs for both. We call the problems that share an ID law an *observation fiber*. Within a fiber the sample says nothing about the hidden law on which the detector will be judged, and the sample size serves only as a clock: every learner is equivalent to a schedule of detectors fixed in advance (Lemma 2.1). A clock can run at any speed, so a fiber has every universal rate or none (Proposition 2.2).

What can a clock learn? Two classical answers mark the extremes. Consistency is often possible: Devroye & Wise (1980) showed that accepting only inputs close to the sample eventually rejects any abnormal law that puts no mass on the support of the normal one, the idea behind support estimation (Cuevas & Fraiman, 1997; Schölkopf et al., 2001; Steinwart et al., 2005). Uniform learnability, one sample size for every problem, is often impossible: Fang et al. (2022; 2024) showed that disjoint supports do not suffice, and Garov & Chaudhuri (2025) showed the same for learners that may choose when to stop. In supervised learning the theory of universal rates (Bousquet et al., 2021; Hanneke et al., 2023) maps the ground between such extremes by asking which rates are possible when constants may depend on the distribution. We map it for OOD detection, where the learner never observes the law that decides its error.

The answer is visible already when nothing is left to estimate. Let the ID law be a single known point, and let the detectors reject finite prefixes of a sequence of points outside it. Rejecting the first n points eventually rejects any fixed hidden law on the sequence, since its tail beyond n vanishes; but at every n some hidden law lies beyond the prefix and is missed entirely. In Figure 1(b) every column is lit in time and every row keeps a star in the dark. The blind spot does not shrink. It moves.

Main result. Which fibers have rates, one function of n bounding the error on every hidden law up to constants depending on the law? For point masses, yes: each is served exactly after finite time. For their countable mixtures, no: the error of the prefix schedule on a mixture is its tail, and for every candidate rate some mixture has a heavier tail. We show that this is the general law. Mixing hidden laws, that is, passing from $Q_1, Q_2, \ldots$ to $\sum_k w_k Q_k$, changes neither which schedules are consistent on a fiber nor its best worst-case error, but it turns universal rates into uniform ones: a learner attains a rate on the countable-mixture completion of a fiber if and only if it attains that rate uniformly on the fiber (Theorem 3.3). A consistent fiber that is not uniformly learnable therefore has no rate once its hidden laws can be mixed: for each learner and candidate rate, a single hidden law defeats every constant and every time rescaling infinitely often. That law *condenses* the blind spots of infinitely many times (Figure 1(b,c)). The mechanism is the classical condensation of singularities; what makes it apply to errors measured against a restricted class is that consistency forces the best achievable OOD risk to be affine under mixing (Lemma 3.2).

Two sources of slowness. When the ID law is unknown, the learner must also select a fiber, and its error is at most the sum of the approximation error within the true fiber, governed by the main theorem, and the statistical cost of selection (5). Since a single fiber has every rate or none, nontrivial rates can arise only here. Over finitely many ID laws selection costs an exponentially small error, and the dichotomy becomes a classification into three regimes (Theorem 4.1); over countably many, fiberwise solutions still glue, and when every fiber has $V = 0$ they do so at every sub-exponential rate but not at an exponential one (Theorem 4.2); over uncountably many, gluing can fail, for a reason that is Baire's theorem (Proposition 4.3). With unrestricted detectors the roles reverse: every fiber of measure-separated domains is solved exactly by rejecting outside the ID support, and all slowness is statistical, from an exact $1/n$ rate (Theorem 5.2) to none at all (Theorem 5.3). Table 1 collects the landscape.

Relation to prior work. The learnability theory of OOD detection from ID-only data with class-relative excess is due to Fang et al. (2022; 2024); their same-ID equivalence and finite-mixture linearity conditions are the ancestors of our fibers and completions, and Garov & Chaudhuri (2025) introduced the adaptive-stopping criterion. Consistency of support-based detection is classical (Devroye & Wise, 1980), the rate convention is that of Bousquet et al. (2021), and the diagonal mechanism of our main proof goes back to the condensation of singularities (Banach & Steinhaus, 1927; Deutsch & Hundal, 2010) and to the arbitrarily slow rates of pattern recognition (Devroye,

Table 1: The landscape for expected strong excess. Within a fiber the sample is a clock and rates are all or none (Proposition 2.2); across fibers the sample selects, and nontrivial rates appear; with unrestricted detectors on separated domains only selection remains. "Every rate": for each rate some learner attains it. All lower bounds are fixed-domain, in the sense of (2).

Setting	Consistent	Universal rates	Uniform
Within one fiber, ID law known (Sections 2 and 3)			
countable fiber $\mathcal{C}$	iff windows compatible	every rate, when consistent	iff $V = 0$
its completion $\sigma\mathcal{C}$	iff $\mathcal{C}$ is	every rate if $V = 0$; else none	iff $V = 0$
prefix fiber $\mathcal{C}_{\mathrm{atom}}$ / $\sigma\mathcal{C}_{\mathrm{atom}}$	yes / yes	every / none	no / no
Across fibers (Section 4)			
finitely many ID laws	iff every fiber is	exponential if all $V = 0$; none if a mixture-complete fiber has $V > 0$	iff all $V = 0$
countably many ID laws	iff every fiber is	$e^{-g(n)}$, $g = o(n)$, if all $V = 0$; not e^{-cn} in general	—
uncountably many	can fail, even with trivial fibers	—	—
Unrestricted detectors, measure-separated domains (Section 5)			
interval family	yes	exactly $1/n$	yes
all of $\mathcal{D}^d_{\mathrm{sep}}$	yes	none	no

1982; Devroye et al., 1996). New are the reduction of rates to schedules on fibers, the completion and affinity results that let the classical mechanism act on class-relative OOD excess, the classifications they yield within and across fibers, and fixed-domain versions of all our lower bounds. We also show by a counterexample that a deterministic compatibility condition stated as necessary by Fang et al. (2024, Theorem 13) is not necessary, and locate the step that fails (Section 4).

2 Observation fibers

Domains, detectors, and strong excess. Let $\mathcal{X}$ be a separable metric space, $\mathcal{Y} = \{1, \dots, K\}$, and $K + 1$ the rejection label. A domain $D = (P, Q)$ pairs an ID law P on $\mathcal{X} \times \mathcal{Y}$ with a hidden OOD law Q on $\mathcal{X}$; the learner observes $S_n \sim P^n$ and never observes Q. Detectors are measurable maps $\mathcal{X} \to \mathcal{Y} \cup \{K+1\}$ in a fixed nonempty class $\mathcal{H}$, and a loss $0 \le \ell \le M$, zero exactly on the diagonal, prices their decisions. Write $R^{\mathrm{in}}_P(h) = \mathbb{E}_P \ell(h(X), Y)$ and $R^{\mathrm{out}}_Q(h) = \mathbb{E}_Q \ell(h(X), K+1)$ for the ID and OOD risks, and $E^{\mathrm{in}}_P, E^{\mathrm{out}}_Q$ for their excesses over the best detector in $\mathcal{H}$. The proportion of OOD inputs at test time is unknown, so a detector is judged at all proportions at once: with $R^\alpha_D = (1-\alpha)R^{\mathrm{in}}_P + \alpha R^{\mathrm{out}}_Q$,

$$\mathcal{E}_D(h) = \sup_{\alpha \in [0,1]} \left(R^\alpha_D(h) - \inf_{g \in \mathcal{H}} R^\alpha_D(g) \right) = \max\{E^{\mathrm{in}}_P(h), E^{\mathrm{out}}_Q(h)\}, \tag{1}$$

because the infimum of a sum is at least the sum of the infima and $\alpha \in \{0, 1\}$ attains each endpoint. One detector must serve both endpoints: with one ID point, one OOD point, and the class {accept all, reject all}, each endpoint has a perfect detector, yet every detector has strong excess 1. Excess is relative to $\mathcal{H}$, so a class unable to reject a hidden law is not blamed for missing it. A learner maps the sample and an independent random seed to a detector, formally a Markov kernel from samples to $\mathcal{H}$ (Appendix C.1); it draws a whole detector, and for a probability law μ on $\mathcal{H}$ we write $\widehat{\mathcal{E}}_D(\mu) = \mathbb{E}_{h \sim \mu} \mathcal{E}_D(h)$.

Guarantees and universal rates. A learner A is *uniform* on a family $\mathcal{D}$ of domains if $\sup_{D \in \mathcal{D}} \mathbb{E}\mathcal{E}_D(A(S_n)) \to 0$, and *universal* if $\mathbb{E}\mathcal{E}_D(A(S_n)) \to 0$ for each $D \in \mathcal{D}$. It is *certified* if for all $\varepsilon, \delta \in (0, \frac{1}{2})$ it can stop after an almost surely finite, data-dependent number of observations and return h with $\mathbb{P}_D\{\mathcal{E}_D(h) > \varepsilon\} \le \delta$ for every $D \in \mathcal{D}$, the strong-excess form of the criterion of Garov & Chaudhuri (2025). Following Bousquet et al. (2021), A attains the *universal rate* r, a positive nonincreasing function on $(0, \infty)$ with $r(t) \to 0$ as $t \to \infty$, if each $D \in \mathcal{D}$ has constants $C_D, c_D > 0$ with $\mathbb{E}\mathcal{E}_D(A(S_n)) \le C_D r(c_D n)$ for all n. A family is learnable in one of these senses if some learner is; a universally learnable family is also called *consistent*, and it is *learnable at rate* r if some learner, which may depend on r, attains r. Our no-rate lower bounds take the strongest

individual form,

$$\forall A \ \forall r \ \exists D_* \ \forall C, c > 0 : \qquad \mathbb{E}\mathcal{E}_{D_*}(A(S_n)) > C\, r(cn) \ \text{ for infinitely many } n; \tag{2}$$

one domain, chosen after the learner and the rate, defeats every constant and time rescaling; our other lower bounds are also witnessed by a single domain. Measurability and other conventions are in Appendix A.

The clock. For an ID law P, the *observation fiber* $\mathcal{C}_P = \{(P, Q) \in \mathcal{D}\}$ collects the domains that the learner cannot tell apart.

Lemma 2.1 (Decoupling). *For every learner A and every n, the law μ_n of $A(S_n)$ under P^n is the same for all $D \in \mathcal{C}_P$, and $\mathbb{E}\mathcal{E}_D(A(S_n)) = \widehat{\mathcal{E}}_D(\mu_n)$. Hence a fiber is universally learnable iff some* schedule*, a data-independent sequence (μ_n) of laws on $\mathcal{H}$, has $\widehat{\mathcal{E}}_D(\mu_n) \to 0$ for every D in the fiber; such a schedule is* consistent *on the fiber.*

The hidden law enters neither the sample nor the learner, and that is the proof (Appendix C.1). On a fiber the data carry no information about what the learner does not know: the sample size indexes a schedule fixed in advance, and every question about a fiber is a question about schedules. A clock, moreover, can be run at any speed.

Proposition 2.2 (Rates within a fiber are all or none). *If some learner attains a universal rate on a fiber, then for every universal rate s some learner, depending on s, attains s on the fiber.*

Proof. By decoupling the learner is a schedule with $\widehat{\mathcal{E}}_D(\mu_n) \le C_D r(c_D n)$ for all n and D. Given a rate s, choose integers $\varphi(n)$ with $r(\varphi(n)/n) \le s(n)$ and let the new learner ignore its sample and draw from $\nu_n = \mu_{\varphi(n)}$ (Appendix C.1). For $n \ge 1/c_D$, $\widehat{\mathcal{E}}_D(\nu_n) \le C_D r(c_D \varphi(n)) \le C_D r(\varphi(n)/n) \le C_D s(n)$; for smaller n, $\widehat{\mathcal{E}}_D(\nu_n) \le M \le \frac{M}{s(1/c_D)}\, s(n)$. □

Within a fiber, then, the rate question is qualitative: does some schedule have a common modulus of convergence at all? A nontrivial rate, attainable while a faster one is not, needs several ID laws that the sample must tell apart; Sections 4 and 5 show how such rates arise.

Values and windows. The *value* of a fiber $\mathcal{C}$ is the best worst-case excess available at a single time, $V(\mathcal{C}) = \inf_\mu \sup_{D \in \mathcal{C}} \widehat{\mathcal{E}}_D(\mu)$, the upper value of a game in which the learner commits to a law on detectors and nature picks a domain. By decoupling, a fiber is uniformly learnable iff $V(\mathcal{C}) = 0$. A *window* is a finite $F \subset \mathcal{C}$, with value $V_F = \inf_\mu \max_{D \in F} \widehat{\mathcal{E}}_D(\mu)$; it is *compatible* if $V_F = 0$. On a countable fiber, consistency is decided window by window.

Theorem 2.3 (Finite windows). *Let $\mathcal{C} = \{D_1, D_2, \ldots\}$ be countable and $F_m = \{D_1, \ldots, D_m\}$. Then $\inf_{(\mu_n)} \sup_{D \in \mathcal{C}} \limsup_n \widehat{\mathcal{E}}_D(\mu_n) = \sup_m V_{F_m}$. In particular, $\mathcal{C}$ is universally learnable iff every window has value zero, iff every window admits deterministic detectors with arbitrarily small maximal excess.*

Proof. For any schedule and window F,

$$V_F \le \limsup_n \max_{D \in F} \widehat{\mathcal{E}}_D(\mu_n) = \max_{D \in F} \limsup_n \widehat{\mathcal{E}}_D(\mu_n).$$

Conversely, taking μ_m within $1/m$ of optimal on F_m gives $\limsup_n \widehat{\mathcal{E}}_D(\mu_n) \le \sup_m V_{F_m}$ for each D, which lies in F_m for all large m. Finally $V_F \le v_F \le |F| V_F$ for the deterministic value $v_F = \inf_h \max_{D \in F} \mathcal{E}_D(h)$, since a law with small expected sum over F charges some detector with small sum. □

A schedule can serve compatible windows by taking them in turn; what it cannot do is serve every domain at the same time, and V measures that. A countable fiber is therefore in one of three positions: some window is incompatible; all windows are compatible but $V > 0$; or $V = 0$. In the middle position each domain is eventually served, yet at every time some domain is not. This is the *moving blind spot*; in general, a fiber has a moving blind spot if it is consistent but $V > 0$.

Theorem 2.4 (Certificates adapt only to observations). *Uniform learnability implies certified learnability, which implies universal learnability. If $\mathcal{D}$ has finitely many distinct ID laws, certified and uniform learnability coincide; on a fiber, both are equivalent to $V = 0$.*

Proof of the collapse; details in Appendix C.2. A stopping rule sees only ID data, so on a fiber its stopping time T and output H have one joint law for all hidden laws, and a certificate at (ε, δ) gives $\sup_{D\in\mathcal{C}} \mathbb{E}\mathcal{E}_D(H) \le \varepsilon + M\delta$. Over finitely many ID laws, almost sure stopping gives one N with $\max_P \mathbb{P}_P(T > N) \le \delta$; truncating at N adds at most $M\delta$, uniformly, and letting $\varepsilon, \delta \downarrow 0$ yields a uniform learner. Markov's inequality gives the converse. □

A certificate can outdo uniform learning only across infinitely many ID laws (Garov & Chaudhuri, 2025); on a fiber there is nothing observable to adapt to.

3 Mixing makes rates uniform

A universally learnable fiber has every rate or none (Proposition 2.2). On a countable fiber in the middle position the moving blind spot costs nothing: a schedule that takes the windows in turn serves each domain at any prescribed speed (Corollary 3.4). This section shows that it costs every rate as soon as hidden laws can be mixed.

Completion. Novelty seldom arrives pure: a stream of OOD inputs is typically a mixture of sources, and a mixture of plausible OOD laws is again plausible. For a fiber $\mathcal{C} = \{(P, Q) : Q \in \mathcal{Q}\}$ we call

$$\sigma\mathcal{C} = \left\{ \left(P, \textstyle\sum_k w_k Q_k \right) : Q_k \in \mathcal{Q},\ w_k \ge 0,\ \sum_k w_k = 1 \right\}$$

its *mixture completion*, and $\mathcal{C}$ *mixture-complete* if $\sigma\mathcal{C} = \mathcal{C}$. Mixing leaves the ID law unchanged, so $\sigma\mathcal{C}$ is again a fiber, and $\sigma\sigma\mathcal{C} = \sigma\mathcal{C}$. Write $a(h) = E^{\text{in}}_P(h)$ and $b(Q) = \inf_{h\in\mathcal{H}} R^{\text{out}}_Q(h)$, so that $\mathcal{E}_{(P,Q)}(h) = \max\{a(h), R^{\text{out}}_Q(h) - b(Q)\}$.

Proposition 3.1 (Mixing preserves consistency and value). *For every fiber $\mathcal{C}$ and every law μ on $\mathcal{H}$, $\sup_{D\in\sigma\mathcal{C}} \widehat{\mathcal{E}}_D(\mu) = \sup_{D\in\mathcal{C}} \widehat{\mathcal{E}}_D(\mu)$, and a schedule is consistent on $\sigma\mathcal{C}$ iff it is consistent on $\mathcal{C}$. Hence $V(\sigma\mathcal{C}) = V(\mathcal{C})$, and $\sigma\mathcal{C}$ is universally, uniformly, or certifiably learnable exactly when $\mathcal{C}$ is.*

Proof. Let $Q_* = \sum_k w_k Q_k$ and $D_k = (P, Q_k)$. Raw risk is linear, $R^{\text{out}}_{Q_*} = \sum_k w_k R^{\text{out}}_{Q_k}$, while $b(Q_*) \ge \sum_k w_k b(Q_k)$ because the infimum of a sum is at least the sum of the infima. Hence $\mathcal{E}_{(P,Q_*)} \le \sum_k w_k \mathcal{E}_{D_k}$ on $\mathcal{H}$, and by Tonelli $\widehat{\mathcal{E}}_{(P,Q_*)}(\mu) \le \sum_k w_k \widehat{\mathcal{E}}_{D_k}(\mu) \le \sup_k \widehat{\mathcal{E}}_{D_k}(\mu)$. This gives the first claim, and dominated convergence sends the weighted sum to zero along any schedule consistent on $\mathcal{C}$. Certification follows by Theorem 2.4. □

Completion is thus invisible to consistency and to the value. Rates live in the reverse inequality, and that is where consistency enters.

Lemma 3.2 (Consistency makes the optimal OOD risk affine). *If one schedule is consistent on every (P, Q_k), then for all probability weights (w_k) and $Q_* = \sum_k w_k Q_k$,*

$$b(Q_*) = \textstyle\sum_k w_k b(Q_k), \quad \textit{and hence} \quad \mathcal{E}_{(P,Q_*)}(h) \ge w_k\, \mathcal{E}_{(P,Q_k)}(h) \quad \textit{for all } h \in \mathcal{H} \textit{ and all } k. \tag{3}$$

Proof. Along the schedule (ν_n), $0 \le \mathbb{E}_{\nu_n} R^{\text{out}}_{Q_k} - b(Q_k) \le \widehat{\mathcal{E}}_{(P,Q_k)}(\nu_n) \to 0$ for each k. Risks lie in $[0, M]$, so Tonelli and dominated convergence give $\mathbb{E}_{\nu_n} R^{\text{out}}_{Q_*} \to \sum_k w_k b(Q_k)$, and $b(Q_*)$, which is at most every such expectation, is at most the limit. With the reverse inequality, $E^{\text{out}}_{Q_*} = \sum_j w_j E^{\text{out}}_{Q_j} \ge w_k E^{\text{out}}_{Q_k}$; since also $a \ge w_k a$, the bound on $\mathcal{E}$ follows. □

Without consistency b is only concave: if no detector rejects two alien points together, each point mass has optimal risk 0 and their fair mixture has optimal risk $\frac{1}{2}$ (Appendix G.3).

Theorem 3.3 (Mixing turns universal rates into uniform rates)**.** *Let $\mathcal{C}$ be a fiber, A a learner, and r a rate. Exactly one of the following holds.*

1. *A attains r uniformly on $\mathcal{C}$: for some $C_0, c_0 > 0$, $\sup_{D\in\mathcal{C}} \mathbb{E}\mathcal{E}_D(A(S_n)) \le C_0\, r(c_0 n)$ for all n. The same bound then holds on $\sigma\mathcal{C}$.*
2. *A single $D_* \in \sigma\mathcal{C}$ has $\mathbb{E}\mathcal{E}_{D_*}(A(S_n)) > C\, r(cn)$ for infinitely many n, for every $C, c > 0$.*

Thus A attains r on $\sigma\mathcal{C}$ iff it attains r uniformly on $\mathcal{C}$. If $V(\mathcal{C}) > 0$, the second alternative holds for every learner and every rate, which is (2); if $V(\mathcal{C}) = 0$, then for every positive $\varepsilon_n \to 0$ some data-independent learner has worst-case excess at most ε_n on $\sigma\mathcal{C}$ for all n. In particular, a mixture-complete fiber is learnable at some universal rate iff it is uniformly learnable.

Proof. Let μ_n be the output laws of A on the fiber (Lemma 2.1) and $s_n = \sup_{D\in\mathcal{C}} \widehat{\mathcal{E}}_D(\mu_n)$, which is also the supremum over $\sigma\mathcal{C}$ by Proposition 3.1; this proves that alternative 1 transfers to $\sigma\mathcal{C}$. Suppose it fails. Then for every k the set $\{n : s_n > 4^k r(n/k)\}$ is infinite, for otherwise $s_n \le 4^k r(n/k)$ for $n \ge N$ and $s_n \le M \le \frac{M}{r(N/k)} r(n/k)$ for $n < N$, a uniform bound with $c_0 = 1/k$. If (μ_n) is not consistent on $\sigma\mathcal{C}$, some $D \in \sigma\mathcal{C}$ has $\limsup_n \widehat{\mathcal{E}}_D(\mu_n) > 0$, which exceeds $C\, r(cn) \to 0$ infinitely often. Otherwise choose times $n_1 < n_2 < \cdots$ with $s_{n_k} > 4^k r(n_k/k) > 0$, and at each time a domain $(P, Q_k) \in \mathcal{C}$ in its blind spot, $\widehat{\mathcal{E}}_{(P,Q_k)}(\mu_{n_k}) > s_{n_k}/2$. Condense them into one hidden law, $Q_* = \sum_k 2^{-k} Q_k$, fixed before any constant is chosen. By Lemma 3.2, $\widehat{\mathcal{E}}_{(P,Q_*)}(\mu_{n_k}) > 2^{k-1} r(n_k/k)$. Given $C, c > 0$, every large k has $1/k \le c$ and $2^{k-1} \ge C$, whence $C\, r(cn_k) \le C\, r(n_k/k) < \widehat{\mathcal{E}}_{(P,Q_*)}(\mu_{n_k})$. The alternatives exclude each other, since the bound of the first applies to D_* with $C = C_0$ and $c = c_0$. Finally, $V(\mathcal{C}) > 0$ gives $s_n \ge V(\mathcal{C})$ for all n, which excludes the first alternative; and $V(\mathcal{C}) = 0$ gives $V(\sigma\mathcal{C}) = 0$ by Proposition 3.1, so each μ_n can be chosen with worst-case excess at most ε_n, ignoring the data. □

The theorem has the shape of the uniform boundedness principle, which Banach & Steinhaus (1927) published as a principle of condensation of singularities: a sequence of operators is either uniformly controlled at a given rate or slower at a single point; Deutsch & Hundal (2010) sharpened this to uniform versus almost arbitrarily slow convergence. In learning theory the same diagonal pattern yields arbitrarily slow rates for consistent classification with one fixed distribution on countably many atoms (Devroye, 1982; Devroye et al., 1996) and the individual lower bounds of Antos & Lugosi (1998); Bousquet et al. (2021). There the slowness is statistical: unseen atoms hide their labels from the sample. Here it is not. The ID law may be known and the sample irrelevant; the slowness belongs to the hidden law alone, and what makes the diagonal pattern apply is structure. Decoupling turns every learner into one schedule for a whole fiber, and consistency makes the class-relative excess, whose benchmark b may be positive and need not be attained, retain on a mixture the fraction w_k of each component's excess. Countable mixtures are needed: the finite mixtures of point masses in Proposition 3.5 are all served exactly after finite time, although the value is one. The theorem holds verbatim at a fixed OOD proportion, for AUC, and for every class-relative criterion linear in the hidden law (Proposition D.1).

Corollary 3.4 (Countable fibers and their completions)**.** *Let $\mathcal{C}$ be a countable fiber. Exactly one of the following holds.*

1. *Some window has positive value: neither $\mathcal{C}$ nor $\sigma\mathcal{C}$ is universally learnable.*
2. *Every window has value zero but $V(\mathcal{C}) > 0$: $\mathcal{C}$ is universally learnable at every rate, whereas $\sigma\mathcal{C}$ is universally learnable at no rate; neither is uniformly or certifiably learnable.*
3. *$V(\mathcal{C}) = 0$: both $\mathcal{C}$ and $\sigma\mathcal{C}$ are uniformly learnable at every rate.*

Proof. Everything except the rates on $\mathcal{C}$ in part 2 follows from Theorems 2.3, 2.4 and 3.3 and Proposition 3.1. For a rate r, choose μ_n with $\max_{j\le n} \widehat{\mathcal{E}}_{D_j}(\mu_n) \le r(n)$; as r is nonincreasing, $\widehat{\mathcal{E}}_{D_j}(\mu_n) \le \max\{1, M/r(j)\}\, r(n)$ for all n. □

The prefix fiber. The moving blind spot of the introduction is the middle case, and it is now a corollary.

Proposition 3.5 (The prefix fiber). *Let $\mathcal{X} = \{0\} \cup \{1/j : j \in \mathbb{N}\}$, $K = 1$, $P = \delta_0 \otimes \delta_1$, zero–one loss, and $\mathcal{H}_{\text{seq}} = \{h_m : m \in \mathbb{N}\}$, where h_m accepts 0 and rejects exactly $1, \frac{1}{2}, \ldots, \frac{1}{m-1}$. Let $\mathcal{C}_{\text{atom}} = \{(P, \delta_{1/j}) : j \in \mathbb{N}\}$, whose completion $\sigma\mathcal{C}_{\text{atom}}$ contains every hidden law on $\{1/j : j \in \mathbb{N}\}$. For every such Q,*

$$\mathcal{E}_{(P,Q)}(h_m) = Q\{1/j : j \geq m\}; \tag{4}$$

every window of $\mathcal{C}_{\text{atom}}$ has value zero, and $V(\mathcal{C}_{\text{atom}}) = 1$. Hence the schedule h_{n+1} attains every rate on $\mathcal{C}_{\text{atom}}$ and is consistent on $\sigma\mathcal{C}_{\text{atom}}$, on which no learner attains any rate, although every domain has $Q(\operatorname{supp} P_X) = 0$.

Proof. ID risks vanish and $\inf_m Q\{1/j : j \geq m\} = 0$, which gives (4). A window is served exactly by rejecting a long enough prefix, and h_{n+1} is exact on $(P, \delta_{1/j})$ for $n \geq j$. A law on $\mathcal{H}_{\text{seq}}$ is a random index L, with excess $\mathbb{P}(L \leq j) \to 1$ on $(P, \delta_{1/j})$ as $j \to \infty$, so $V = 1$. Apply Corollary 3.4. □

The error of h_{n+1} on a hidden law is that law's tail, and tails decay at any speed: geometrically for $Q\{1/j\} = 2^{-j}$, like $1/n$ for $1/(j(j+1))$, like $1/\log n$ for heavier weights (Appendix D). For every candidate rate some countable mixture has a heavier tail; Figure 1(c) shows one built against $r(n) = 1/n$. Theorem 3.3 says that every fiber with a moving blind spot behaves in this way once its hidden laws can be mixed.

The blind spot chases a missing limit. The acceptance sets of $\mathcal{H}_{\text{seq}}$ form a chain, so the class has VC dimension one. Adjoining the detector h_∞ that rejects every $1/j$ keeps the VC dimension at one but makes $V(\sigma\mathcal{C}_{\text{atom}}) = 0$: the consistent schedule was converging to a detector that the class did not contain. Compactness always has this effect: if the laws on detectors form a compact space on which every excess is lower semicontinuous, compatible windows force one law with zero excess everywhere, and the middle position cannot occur (Proposition C.8). A moving blind spot requires a noncompact class, a property that combinatorial capacity does not see.

4 Selecting the Fiber

When the ID law is unknown, the learner must also decide which fiber it is on. Suppose it estimates the index $\widehat{i}_n$ of the observed law among candidates $P_1, P_2, \ldots$ and then draws a detector from a schedule $(\nu_n^{(i)})$ for the selected fiber, independently of the sample. For the true index i and every $D \in \mathcal{C}_{P_i}$,

$$\mathbb{E}\mathcal{E}_D(A(S_n)) \;\leq\; \underbrace{\widehat{\mathcal{E}}_D\big(\nu_n^{(i)}\big)}_{\text{within the fiber}} \;+\; \underbrace{M\,\mathbb{P}_{P_i^n}\big(\widehat{i}_n \neq i\big)}_{\text{selection across fibers}}. \tag{5}$$

Theorem 4.1 (Finitely many ID laws). *Let $\mathcal{D}$ have finitely many distinct ID laws. Then $\mathcal{D}$ is universally learnable iff each fiber is, and certifiably iff uniformly iff every fiber has $V = 0$; the selection term in (5) can be made at most $C_0 e^{-c_0 n}$. Consequently, if the fibers are mixture-complete, exactly one regime occurs: (a) every fiber has $V = 0$, and some learner has $\sup_{D \in \mathcal{D}} \mathbb{E}\mathcal{E}_D(A(S_n)) \leq Ce^{-cn}$; (b) every fiber is universally learnable but some has $V > 0$, and $\mathcal{D}$ is universally learnable at no rate; (c) some fiber is not universally learnable, and neither is $\mathcal{D}$.*

Pairwise tests of empirical frequencies identify the true law with error $C_0 e^{-c_0 n}$, and (5) adds this to the fiberwise term; in (b), every learner restricts to one schedule on a fiber with $V > 0$, where Theorem 3.3 applies (Appendices D.1 and E.1). By Proposition 3.1, the three regimes classify the completion of any finite-ID family. The only exponential term in (a) is the selection error: within one known fiber rates are all or none (Proposition 2.2), and with $V = 0$ every prescribed rate is attained uniformly (Theorem 3.3). The exponential is sharp when fibers conflict: for two Bernoulli ID laws whose fibers need different detectors, the Bretagnolle–Huber inequality (Bretagnolle & Huber, 1979) forces expected excess at least $\frac{1}{4} e^{-n\,\mathrm{KL}(P_1 \| P_2)}$ on one of them, so no faster rate is attainable (Proposition E.2).

Deterministic compatibility. Uniform learnability over finitely many ID laws asks for $V = 0$ on every fiber, a condition on laws over detectors, not on single detectors. A fiber with the diagonal profile $\mathcal{E}_{(P,\delta_{x_j})}(h_k) = \mathbf{1}\{j = k\}$ has $V = 0$, since uniform laws on m detectors give $1/m$, yet

every deterministic detector has excess one on some domain; when the ID law is a fair coin, a deterministic learner that reads fair bits from its sample attains 2^{-n}. This contradicts the necessity of the deterministic Condition 4 in Fang et al. (2024, Theorem 13) as stated; the failing step exchanges "for every domain, in expectation over samples" with "for some sample, for every domain", and the repair is local (Appendix G.2).

Theorem 4.2 (Countably many ID laws). *Let $\mathcal{D}$ have countably many distinct ID laws. Then $\mathcal{D}$ is universally learnable iff each fiber is, whatever the cardinality of the fibers. If every fiber has $V = 0$, then for every nondecreasing $g \to \infty$ with $g(n) = o(n)$ some learner has $\sup_{D \in \mathcal{C}_{P_i}} \mathbb{E}\mathcal{E}_D(A(S_n)) \le C_i e^{-g(n)}$ for all n and i; the condition $g(n) = o(n)$ cannot be dropped.*

The learner selects the first enumerated law within $\eta_n \to 0$ of the empirical law, in a metric built from a countable algebra of events. Only finitely many earlier candidates must be excluded, each at positive distance, so selection is eventually correct, and bounded differences make its error $e^{-g(n)}$. Bernoulli laws approaching one of them in relative entropy show that no consistent learner is exponentially fast on every domain, and $\sigma\mathcal{C}_{\mathrm{atom}}$ shows that fiberwise $V = 0$ cannot be dropped (Appendix E.2).

Uncountably many ID laws. Selection is continuous. Every learner's output law is the image of P^n under one kernel, so $\mathrm{TV}(\mu_n^P, \mu_n^{P'}) \le \mathrm{TV}(P^n, P'^n) \le n\,\mathrm{TV}(P, P')$, and a universal learner makes its fiberwise choice as a pointwise limit of continuous functions of the ID law. Over a countable family this is no restriction; over an uncountable one it is, because such limits are continuous on a comeager set.

Proposition 4.3 (Consistent fibers need not glue). *Let $K = 1$ and zero–one loss. For $t \in (0, 1)$ let P_t be the Bernoulli(t) law on $\{0, 1\}$ with label 1, let $Q_t = \delta_2$ for rational t and $Q_t = \delta_3$ otherwise, and let $\mathcal{H} = \{h_2, h_3\}$, where h_x rejects only x. Each fiber of $\mathcal{D} = \{(P_t, Q_t) : t \in (0, 1)\}$ is a single domain with an exact detector, yet every learner whose expected excess vanishes at all rational t has $\limsup_n \mathbb{E}\mathcal{E}_{(P_t,Q_t)}(A(S_n)) = 1$ for a comeager set of irrational t. In particular $\mathcal{D}$ is not universally learnable.*

Proof. The probability $\phi_n(t)$ of outputting h_2 under P_t^n is a polynomial in t, and the expected excess is $1 - \phi_n(t)$ at rational and $\phi_n(t)$ at irrational t. If $\phi_n \to 1$ on the rationals, each set $\bigcup_{n \ge N}\{\phi_n > 1 - 1/m\}$ is open and contains the rationals, hence is dense in $(0, 1)$. Intersecting over $N, m \in \mathbb{N}$ and with the irrationals gives a comeager set, on which $\limsup_n \phi_n = 1$; it is nonempty, indeed dense, by Baire's theorem (Oxtoby, 1980). □

The failure set is comeager but may be Lebesgue-null: Cover (1973) gives tests that err only finitely often, almost surely, at every rational and at almost every irrational. The obstruction is topological, not measure-theoretic, in the spirit of the discernibility criterion of Dembo & Peres (1994). Unrestricted detectors solve $\mathcal{D}$ at once by rejecting $\{2, 3\}$; what fails is selection inside a class.

5 Unrestricted detectors on separated domains

Are the obstructions above artifacts of restricting the detectors? With the class $\mathcal{H}_{\mathrm{all}}$ of all measurable detectors, the fibers of measure-separated domains become trivial. On

$$\mathcal{D}_{\mathrm{sep}}^d = \{(P, Q) : P \in \mathcal{P}(\mathbb{R}^d \times \mathcal{Y}),\ Q \in \mathcal{P}(\mathbb{R}^d),\ Q(\operatorname{supp} P_X) = 0\},$$

the detector that rejects outside $\operatorname{supp} P_X$ and applies a Bayes rule on it has zero strong excess on every domain of its fiber. Every fiber has $V = 0$, attained by one detector, so the first term of (5) vanishes, and learning becomes the statistical problem of approximating that detector from the sample: support estimation (Devroye & Wise, 1980; Cuevas & Fraiman, 1997). All slowness is now statistical, and it shows both faces of classical estimation.

Theorem 5.1 (Consistency). *For all $d, K \ge 1$ and bounded losses, $(\mathcal{D}_{\mathrm{sep}}^d, \mathcal{H}_{\mathrm{all}})$ is universally learnable by the rule that rejects empty dyadic cubes of side $s_n = 2^{-\lceil (\log_2 n)/(2d) \rceil}$ and minimizes the empirical conditional loss on occupied ones.*

Theorem 5.2 (An exact linear rate). *With $K = 1$, zero–one loss, and $\mathcal{H}_{\text{all}}$, let $\mathcal{D}_{\text{int}} = \{(\text{Unif}[0,\theta] \otimes \delta_1, Q) : \theta \in [\frac{1}{4}, \frac{3}{4}],\ Q((\theta, 1]) = 1\}$. Accepting $[0, \max_i X_i]$ has expected strong excess exactly $1/(n+1)$ on every domain, and for every learner a single $D_* \in \mathcal{D}_{\text{int}}$ has $\mathbb{E}\mathcal{E}_{D_*}(A(S_n)) \geq 1/(25n)$ for infinitely many n.*

Theorem 5.3 (No rate). *For $K = 1$ and zero–one loss, $(\mathcal{D}^d_{\text{sep}}, \mathcal{H}_{\text{all}})$ is universally learnable at no rate, for every $d \geq 1$: for every learner and rate, a single domain supported on a segment of a coordinate axis satisfies (2).*

Accepted cubes lie within $\sqrt{d}\, s_n$ of the ID support, which carries no OOD mass, and martingale convergence controls the ID excess (Appendix F.1); for binary detection this is Devroye & Wise (1980, Theorem 1). The two lower bounds again build one domain that carries infinitely many finite-time difficulties, now statistical ones. In Theorem 5.2, endpoints $a < b$ at distance $O(1/n)$ give sample laws within total variation $\frac{1}{8}$ yet demand opposite decisions on the sliver $(a, b]$; nesting such dilemmas, with OOD mass committed above all later endpoints, yields one limiting endpoint and one OOD law. In Theorem 5.3, level k hides m_k pairs of small cells, and an unseen switch decides which cell of each pair carries ID mass and which OOD mass. At time n_k most pairs receive no sample, so every rule errs on half their mass, and a limsup argument over random switches selects one domain that is hard at infinitely many times (Appendices F.2 and F.3). Together with Garov & Chaudhuri (2025, Theorems 2 and 8), which rule out certification on this family, Theorem 5.1 separates consistency from certification. For rankings, the nearest-neighbour score is consistent in AUC on every measure-separated domain, almost surely along one training sequence (Theorem F.1).

6 Discussion

What the sample buys. Within a fiber, ID data buy nothing but time, and time can be bought at any speed: a schedule fixed in advance does as well as any learner, and a fiber has every rate or none (Lemma 2.1 and Proposition 2.2). A nontrivial rate therefore needs several ID laws that the sample must tell apart, as with the exponential of Theorem 4.1, sharp for conflicting fibers, or the $1/n$ of Theorem 5.2. The main theorem decides the fiberwise alternative for hidden laws that can be mixed: all rates when $V = 0$, none when $V > 0$. When $V > 0$, a quantitative guarantee needs an assumption that shrinks the hidden family: on the prefix fiber, a uniform tail bound $\sup_Q Q\{1/j : j > n\} \leq \tau_n$ gives the uniform rate τ_n by (4). ID data cannot verify such an assumption.

Boundaries. Two examples mark where the framework ends (Appendices G.1 and G.3). *Overlap*: with $\mathcal{H}_{\text{all}}$ and zero–one loss, $\widehat{\mathcal{E}}_D(\mu) \geq \frac{1}{2}(1 - \text{TV}(P_X, Q))$, so a single domain is universally learnable iff $P_X \perp Q$. *Uncountable fibers*: if the detectors accept a known ID point and reject finite subsets of $[0, 1]$, every countable subfamily of the fiber of all OOD laws on $[0, 1]$ has a consistent schedule while the fiber has none, because any schedule leaves Lebesgue-almost every point untouched forever; Theorem 2.3 needs countability.

Open problems. Two questions are sharpened by this picture. When do fiberwise solutions glue over an uncountable family of ID laws? Proposition 4.3 shows that the continuity of selection is a constraint, and a criterion in the spirit of Dembo & Peres (1994) would complete Section 4. And which assumptions on the hidden family give uniform rates for natural classes? For the prefix and near-OOD fibers the answer is tightness at the blind spot (Appendices D.4 and D.6), which no ID sample can check.

Other observation structures. Within a fiber we used only that the sample law depends on the domain through the ID law. Observations that depend on the hidden law, such as the contaminated unlabeled sample of semi-supervised novelty detection (Blanchard et al., 2010), refine the fibers; decoupling and Theorems 2.3 and 2.4 hold on the refined fibers verbatim, although mixtures of hidden laws need no longer stay in one fiber. When the hidden law is fully observed, the fibers are points and only statistical difficulty remains, as in supervised learning (Bousquet et al., 2021). OOD detection sits at the other extreme, and the moving blind spot is its signature.

Takeaway. In OOD detection the sample buys selection and time buys consistency; a rate against hidden laws that can be mixed can only be bought with an assumption about them.

REPRODUCIBILITY STATEMENT

The results are mathematical. The main text gives complete proofs of Propositions 3.1–4.3 and 3.5, Theorems 2.3 and 3.3, Lemma 3.2, and Corollary 3.4, and proof sketches of Theorems 2.4, 4.1 and 4.2, completed in Appendices C.2, D.1, E.1 and E.2. The appendices prove the statistical results of Section 5, the general theory of fiber games, and every boundary example. The exact analytic script for the tail computations is included with the source. No empirical benchmark results are claimed.

AI USE STATEMENT

The original research ideas and initial manuscript draft were developed by the authors. Large language models were used in this revision to help reorganize the paper around observation fibers; to check the proofs of mathematical claims and fill in some routine steps; to formulate the open problems and short observations of Appendix I; to polish the text; to help produce the figures. The authors reviewed all AI-assisted content and take full responsibility for the correctness of the manuscript.

ETHICS STATEMENT

This work develops mathematical guarantees and counterexamples; it involves no human participants, personal data, or empirical deployment. Its intended benefit is to clarify what ID-only observations can support and where further assumptions are needed. OOD rejection can exclude valid but underrepresented inputs, while missed OOD cases can create misplaced confidence. Distributional novelty should therefore not be equated with harmfulness or lack of value. The asymptotic guarantees do not certify a deployed detector, validate a learned representation, or establish fairness across groups. Applications require separate validation of the domain assumptions, error costs, and effects on affected users.

REFERENCES

András Antos and Gábor Lugosi. Strong minimax lower bounds for learning. *Machine Learning*, 30(1):31–56, 1998. doi: 10.1023/A:1007454427662. URL https://www.szit.bme.hu/~antos/publ.html.

Stefan Banach and Hugo Steinhaus. Sur le principe de la condensation de singularités. *Fundamenta Mathematicae*, 9:50–61, 1927. doi: 10.4064/fm-9-1-50-61.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023. doi: 10.1214/22-AOS2244. URL https://arxiv.org/abs/2104.08279.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 129–136, 2010. URL https://proceedings.mlr.press/v9/david10a.html.

Julian Bitterwolf, Alexander Meinke, and Matthias Hein. Certifiably adversarially robust detection of out-of-distribution data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16085–16095, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/b90c46963248e6d7aab1e0f429743ca0-Abstract.html.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11(99):2973–3009, 2010. URL https://www.jmlr.org/papers/v11/blanchard10a.html.

Olivier Bousquet, Steve Hanneke, Shay Moran, Ramon van Handel, and Amir Yehudayoff. A theory of universal learning. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 532–541, 2021. doi: 10.1145/3406325.3451087. URL https://arxiv.org/abs/2011.04483.

J. Bretagnolle and C. Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47(2):119–137, 1979. doi: 10.1007/BF00535278.

Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. SegmentMeIfYouCan: A benchmark for anomaly segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/d67d8ab4f4c10bf22aa353e27879133c-Abstract-round2.html.

Thomas M. Cover. On determining the irrationality of the mean of a random variable. *The Annals of Statistics*, 1(5):862–871, 1973. doi: 10.1214/aos/1176342507.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006. doi: 10.1002/047174882X.

Antonio Cuevas and Ricardo Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, 25(6):2300–2312, 1997. doi: 10.1214/aos/1030741073. URL https://www.jstor.org/stable/2959033.

Amir Dembo and Yuval Peres. A topological criterion for hypothesis testing. *The Annals of Statistics*, 22(1):106–117, 1994. doi: 10.1214/aos/1176325360. URL https://statistics.stanford.edu/technical-reports/topological-criterion-hypothesis-testing.

Frank Deutsch and Hein Hundal. Slow convergence of sequences of linear operators I: Almost arbitrarily slow convergence. *Journal of Approximation Theory*, 162(9):1701–1716, 2010. doi: 10.1016/j.jat.2010.05.001. URL https://www.sciencedirect.com/science/article/pii/S0021904510000924.

Luc Devroye. Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-4(2):154–157, 1982. doi: 10.1109/TPAMI.1982.4767222.

Luc Devroye and Gary L. Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488, 1980. doi: 10.1137/0138038. URL https://epubs.siam.org/doi/10.1137/0138038.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996. doi: 10.1007/978-1-4612-0711-5. URL https://www.szit.bme.hu/~gyorfi/pbook.pdf.

Xuefeng Du, Zhen Fang, Ilias Diakonikolas, and Yixuan Li. How does unlabeled data provably help out-of-distribution detection? In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jlEjB8MVGa.

Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f0e91b1314fa5eabf1d7ef6d1561ecec-Abstract-Conference.html.

Zhen Fang, Yixuan Li, Feng Liu, Bo Han, and Jie Lu. On the learnability of out-of-distribution detection. *Journal of Machine Learning Research*, 25(84):1–83, 2024. URL https://jmlr.org/papers/v25/23-1257.html.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7068–7081, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/3941c4358616274ac2436eacf67fae05-Abstract.html.

Konstantin Garov and Kamalika Chaudhuri. A closer look at the learnability of out-of-distribution (OOD) detection, 2025. URL https://arxiv.org/abs/2501.08821.

Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, et al. Does your dermatology classifier know what it doesn't know? Detecting the long-tail of unseen conditions. *Medical Image Analysis*, 75:102274, 2022. doi: 10.1016/j.media.2021.102274. URL https://arxiv.org/abs/2104.03829.

Steve Hanneke and Shay Moran. A theory of universal agnostic learning, 2026. URL https://arxiv.org/abs/2601.20961.

Steve Hanneke, Shay Moran, and Qian Zhang. Universal rates for multiclass learning. In *Proceedings of the Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5615–5681, 2023. URL https://proceedings.mlr.press/v195/hanneke23a.html.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1610.02136.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyxCxhRcY7.

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

Gregor Kasieczka, Benjamin Nachman, David Shih, Oz Amram, Anders Andreassen, et al. The LHC Olympics 2020: A community challenge for anomaly detection in high energy physics. *Reports on Progress in Physics*, 84(12):124201, 2021. doi: 10.1088/1361-6633/ac36b9. URL https://arxiv.org/abs/2101.08320.

Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training OOD detectors in their natural habitats. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10848–10865, 2022. URL https://proceedings.mlr.press/v162/katz-samuels22a.html.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

Bo Liu, Li-Ming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. How good are LLMs at out-of-distribution detection? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8211–8222, 2024. URL https://aclanthology.org/2024.lrec-main.720/.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html.

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

Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. Classification without labels: learning from mixed samples in high energy physics. *Journal of High Energy Physics*, 2017(10):174, 2017. doi: 10.1007/JHEP10(2017)174. URL https://arxiv.org/abs/1708.02949.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35087–35102, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/e43a33994a28f746dcfd53eb51ed3c2d-Abstract-Conference.html.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.09136.

John C. Oxtoby. *Measure and Category*, volume 2 of *Graduate Texts in Mathematics*. Springer, second edition, 1980. doi: 10.1007/978-1-4684-9339-9.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, 2013. doi: 10.1109/TPAMI.2012.256. URL https://doi.org/10.1109/TPAMI.2012.256.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001. doi: 10.1162/089976601750264965.

Sextus Empiricus. *Outlines of Pyrrhonism*. William Heinemann, London, 1933. Translated by R. G. Bury. Loeb Classical Library 273.

Ingo Steinwart, Don Hush, and Clint Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6(8):211–232, 2005. URL https://jmlr.org/papers/v6/steinwart05a.html.

Charles J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 1977. doi: 10.1214/aos/1176343886.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20827–20840, 2022. URL https://proceedings.mlr.press/v162/sun22d.html.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. doi: 10.1007/b13794.

David Williams. *Probability with Martingales*. Cambridge University Press, 1991. doi: 10.1017/CBO9780511813658.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey, 2021. URL https://arxiv.org/abs/2110.11334. Revised 2024.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/d201587e3a84fc4761eadc743e9b3f35-Abstract-Datasets_and_Benchmarks.html.

Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12427–12436, 2021. URL https://proceedings.mlr.press/v139/zhang21g.html.

David Zimmerer, Peter M. Full, Fabian Isensee, Paul Jäger, Tim Adler, et al. MOOD 2020: A public benchmark for out-of-distribution detection and localization on medical images. *IEEE Transactions on Medical Imaging*, 41(10):2728–2738, 2022. doi: 10.1109/TMI.2022.3170077. URL https://doi.org/10.1109/TMI.2022.3170077.

TABLE OF CONTENTS

1 Introduction **2**

2 Observation fibers **3**

3 Mixing makes rates uniform **5**

4 Selecting the fiber **7**

5 Unrestricted detectors on separated domains **8**

6 Discussion **9**

A Notation, conventions, and the map of results **16**

A.1 Standing assumptions and conventions 17

A.2 The guarantees as quantifier strings 18

A.3 A dictionary: learning on a fiber as geometry of profiles 18

B Classical facts used in the proofs **18**

C Observation fibers: proofs and the general game **21**

C.1 Proof of Lemma 2.1 21

C.2 Proof of Theorem 2.4 21

C.3 The game on a fiber 22

C.4 Rankings use the same game 25

D Mixing: the finite-ID classification and the role of each hypothesis **26**

D.1 Proof of Theorem 4.1 26

D.2 The zero-optimum case 26

D.3 The role of each hypothesis 26

D.4 Tails of the prefix schedule 27

D.5 One theorem for every linear criterion 27

D.6 Near and far OOD: a continuous blind spot 28

D.7 Relation to the condensation of singularities 29

E Selecting the fiber: proofs **30**

E.1 Finitely many ID laws 30

E.2 Countably many ID laws 31

E.3 Uncountably many ID laws: the Baire obstruction 32

F Unrestricted detectors: proofs **33**

F.1 Consistency: proof of Theorem 5.1 33

F.2 An exact linear rate: proof of Theorem 5.2 34
F.3 No rate: proof of Theorem 5.3 35
F.4 Nearest-neighbour ranking 36

G Where the framework ends **36**
G.1 An uncountable fiber without a consistent schedule 36
G.2 Deterministic compatibility and a published necessity claim 37
G.3 Overlap and genuine conflicts 37

H Related work in detail **38**

I Open problems and directions from applications **39**
I.1 Deployments are mixtures 39
I.2 Near OOD in safety-critical perception 40
I.3 Foundation models and observations that see the hidden law 41
I.4 Adaptive adversaries 42
I.5 Scientific discovery 42
I.6 Continuous families of environments 43
I.7 From theorems to practice 43

J On the epigraph **43**

GUIDE TO THE APPENDICES

The appendices follow the architecture of the main text. Appendix A fixes notation and conventions, writes every guarantee as a string of quantifiers, translates learning on a fiber into the geometry of its profile set, and draws the map of results; Appendix B restates the classical facts used in proofs. Appendix C proves decoupling and the certification hierarchy and develops the general theory of fiber games: pure and mixed values, finite windows, point reductions, compactness, and rankings. Appendix D completes the finite-ID classification, tests every hypothesis of the mixing theorem, extends it to every class-relative criterion that is linear in the hidden law, including AUC, and gives a continuous example that separates near from far OOD. Appendix E proves the selection results: finite-ID identification and the sharpness of its exponential rate, countable gluing with its sub-exponential boundary, and the Baire obstruction. Appendix F proves the statistical results for unrestricted detectors and the nearest-neighbour ranking theorem. Figures in Appendices C, D and F draw the three positions of a fiber, the near/far example, and the two lower-bound constructions. Appendix G gives the uncountable fiber, the compatibility correction, and the overlap conflicts, and Appendix H places the work in the literature. Appendix I reads applications, from medical imaging and driving to particle physics and foundation models, through fibers and states the open problems they raise, and Appendix J reads the epigraph in the terms of the paper.

A NOTATION, CONVENTIONS, AND THE MAP OF RESULTS

Table 2 collects the notation and Appendix A.1 the standing assumptions and conventions. Appendix A.2 writes every guarantee of the paper as a string of quantifiers, Appendix A.3 translates learning on a fiber into the geometry of its profile set, and Figure 2 shows which results are used to prove which.

Table 2: Notation, with the place where each object is introduced.

Symbol	Meaning	Where
$\mathcal{X}$; $\mathcal{Y} = \{1, \dots, K\}$; $K + 1$	instance space; label set; rejection label	Section 2
$D = (P, Q)$; P_X	domain: ID law on $\mathcal{X} \times \mathcal{Y}$ and hidden OOD law on $\mathcal{X}$; ID marginal	Section 2
$S_n \sim P^n$; A, $A(S_n)$	ID sample; learner and its output detector	Section 2
$\mathcal{H}$; $\mathcal{H}_{\mathrm{all}}$	detector class; all measurable detectors	Section 2
ℓ, M	loss and its upper bound	Section 2
R_P^{in}, R_Q^{out}; E_P^{in}, E_Q^{out}	ID and OOD risks; excesses over the best detector in $\mathcal{H}$	Section 2
R_D^{α}; $\mathcal{E}_D(h)$	risk at OOD proportion α; strong excess (1)	Section 2
μ, ν; $\mathcal{P}(\mathcal{H})$; $\widehat{\mathcal{E}}_D(\mu)$	laws on detectors; all such laws; expected strong excess	Section 2
r, s; C_D, c_D; φ	rates (positive, nonincreasing on $(0, \infty)$, tending to 0); domain-dependent constants, (2); time change	Section 2
$\mathcal{C}_P$; $\mathcal{C}$; $\mathcal{Q}$	observation fiber of P; a generic fiber; its family of hidden laws	Sections 2 and 3
schedule; consistent	data-independent sequence of laws on $\mathcal{H}$; with excess $\to 0$ on every domain	Lemma 2.1
$V(\mathcal{C})$; V_F; window	value; window value; finite subset of a fiber	Section 2
$\sigma\mathcal{C}$; mixture-complete	countable-mixture completion; $\sigma\mathcal{C} = \mathcal{C}$	Section 3
$a(h)$; $b(Q)$	ID excess; optimal OOD risk in the class	Section 3
$\mathcal{H}_{\mathrm{seq}}$, h_m; $\overline{\mathcal{H}}$, h_∞	prefix-rejection class; with the limit detector adjoined	Proposition 3.5
$\mathcal{C}_{\mathrm{atom}}$; $\sigma\mathcal{C}_{\mathrm{atom}}$	point-mass fiber; all hidden laws on $\{1/j\}$	Proposition 3.5
$\widehat{P}_n$; $\widehat{\imath}_n$; d_0; η_n; g	empirical law; selected index; law metric; threshold; exponent	Section 4, Appendix E.2
$\mathcal{D}_{\mathrm{int}}$; $\mathcal{D}_{\mathrm{sep}}^d$	interval family; measure-separated domains on $\mathbb{R}^d$	Section 5
L_Q; e_Q	linear criterion; its class-relative excess	Appendix D.5
TV; dist; supp; AUC$(r; D)$	total variation; distance; support; area under the ranking curve	Section 5, Appendix F.4
V_T^{p}, V_T^{m}, W_T^{p}, W_T^{m}	pure and mixed static and sequential values of a game	Appendix C.3
$\mathbf{1}$; δ_x; Ber(p); Unif	indicator; point mass at x; Bernoulli and uniform laws	throughout

A.1 Standing assumptions and conventions

Throughout, $\mathcal{X}$ is a separable metric space with its Borel sigma field, the label set $\mathcal{Y} = \{1, \dots, K\}$ is finite, the loss ℓ takes values in $[0, M]$ and vanishes exactly on the diagonal, the detector class $\mathcal{H}$ is nonempty, and learners may randomize.

Measurability. The class $\mathcal{H}$ carries a sigma field on which laws μ are defined, and evaluation $(h, x) \mapsto \ell(h(x), y)$ is jointly measurable for every $y \in \mathcal{Y} \cup \{K + 1\}$. By Tonelli's theorem, $h \mapsto R_P^{\mathrm{in}}(h)$, $h \mapsto R_Q^{\mathrm{out}}(h)$ and $h \mapsto \mathcal{E}_D(h)$ are then measurable for every Borel P and Q. In particular a countable mixture $Q_* = \sum_k w_k Q_k$ is a Borel probability law with $R_{Q_*}^{\mathrm{out}}(h) = \sum_k w_k R_{Q_k}^{\mathrm{out}}(h)$ by monotone convergence, so completions never leave the measurable setting. Finite and countable classes carry their discrete sigma fields. For the class $\mathcal{H}_{\mathrm{all}}$ of all measurable detectors no standard Borel parameterization is assumed: a randomized learner over $\mathcal{H}_{\mathrm{all}}$ is a map $(s, u, x) \mapsto h_{s,u}(x)$, jointly measurable in the sample s, an independent seed u, and the input x, and all expected risks are computed by Tonelli's theorem. A possibly randomized learner at sample size n is a Markov kernel from samples to $\mathcal{H}$ (Appendix C.1). All learnability notions are statistical existence notions; no statement concerns computation.

Two ways to charge a randomized learner. A randomized learner draws a whole detector and is charged $\widehat{\mathcal{E}}_D(\mu) = \mathbb{E}_\mu \max\{E_P^{\mathrm{in}}, E_Q^{\mathrm{out}}\}$. Charging $\max\{\mathbb{E}_\mu E_P^{\mathrm{in}}, \mathbb{E}_\mu E_Q^{\mathrm{out}}\}$ instead changes every value by at most a factor two, since

$$\max\{\mathbb{E}_\mu E_P^{\mathrm{in}}, \mathbb{E}_\mu E_Q^{\mathrm{out}}\} \le \widehat{\mathcal{E}}_D(\mu) \le \mathbb{E}_\mu E_P^{\mathrm{in}} + \mathbb{E}_\mu E_Q^{\mathrm{out}} \le 2 \max\{\mathbb{E}_\mu E_P^{\mathrm{in}}, \mathbb{E}_\mu E_Q^{\mathrm{out}}\}.$$

Every learnability notion, every rate class, and every statement that a value is zero is therefore the same under both conventions; numerical constants, such as $1/(25n)$ in Theorem 5.2 and $\frac{1}{4}$ in

Table 3: The guarantees of the paper, written as strings of quantifiers.

Notion	Quantifier string	Where
uniform	$\forall \varepsilon > 0\ \exists N\ \forall n \geq N\ \forall D \in \mathcal{D} :\ e_D(n) \leq \varepsilon$	Section 2
universal (consistent)	$\forall D \in \mathcal{D}\ \forall \varepsilon > 0\ \exists N\ \forall n \geq N :\ e_D(n) \leq \varepsilon$	Section 2
certified	$\forall \varepsilon, \delta \in (0, \frac{1}{2})\ \forall D \in \mathcal{D} :\ \mathbb{P}_D(T < \infty) = 1$ and $\mathbb{P}_D\{\mathcal{E}_D(H) > \varepsilon\} \leq \delta$	Section 2
universal rate r	$\forall D \in \mathcal{D}\ \exists C, c > 0\ \forall n :\ e_D(n) \leq C\, r(cn)$	Section 2
uniform rate r	$\exists C, c > 0\ \forall n\ \forall D \in \mathcal{D} :\ e_D(n) \leq C\, r(cn)$	Theorem 3.3
no rate, fixed domain	$\forall A\ \forall r\ \exists D \in \mathcal{D}\ \forall C, c > 0\ \forall N\ \exists n \geq N :\ e_D(n) > C\, r(cn)$	(2)

Proposition E.2, refer to $\widehat{\mathcal{E}}_D$. Certificates concern a single returned detector and do not depend on the convention.

Strong and fixed-proportion excess. By (1), $R_D^\alpha(h) - \inf_{g \in \mathcal{H}} R_D^\alpha(g) \leq \mathcal{E}_D(h)$ for every $\alpha \in [0, 1]$. Guarantees for the strong excess therefore hold at every fixed proportion of OOD inputs, and impossibility results at a fixed, even known, proportion transfer to the strong excess; this is how Fact B.9 enters. The strong excess asks for more than all fixed proportions together, since one detector must serve both endpoints (Section 2).

Rates. A rate is a positive nonincreasing function on $(0, \infty)$ that tends to zero at infinity, and only its behaviour at infinity matters: excesses are at most M, so changing r on a bounded interval can be absorbed into the constant C_D of an upper bound and affects finitely many n in a lower bound. The constant c_D rescales time, as in Bousquet et al. (2021), so that $r(t)$ and $r(t/2)$ define the same rate class. "Learnable at every rate" means that for each rate r some learner, which may depend on r, attains r; no single learner is claimed to attain every rate at once.

A.2 The guarantees as quantifier strings

For a learner A on a family $\mathcal{D}$ write $e_D(n) = \mathbb{E}\mathcal{E}_D(A(S_n))$, and for a sequential learner write (T, H) for its stopping time and returned detector. Table 3 spells out each notion. Every positive notion asks for some learner, so its string is preceded by $\exists A$; the fixed-domain lower bound (2) is asked of every learner. Uniform and universal learnability differ only in the position of $\forall D$, and so do uniform and universal rates. On a fiber, exchanging $\forall D$ with the quantifiers that follow it is exactly what a moving blind spot prevents. Theorem 3.3 is a theorem about this exchange for rates: on a mixture completion, a learner satisfies the string "$\forall D\ \exists C, c$" with rate r iff it satisfies "$\exists C, c\ \forall D$" with the same rate on the original fiber; otherwise the negation of the first string holds with a single domain, which is the lower bound in the last row.

A.3 A dictionary: learning on a fiber as geometry of profiles

For a fiber $\mathcal{C}$, let $\mathcal{R}_\mathcal{C} = \{(\widehat{\mathcal{E}}_D(\mu))_{D \in \mathcal{C}} : \mu \in \mathcal{P}(\mathcal{H})\} \subset [0, M]^\mathcal{C}$ be its profile set, a convex set (Appendix C.3). By decoupling, a learner on $\mathcal{C}$ is a sequence in $\mathcal{R}_\mathcal{C}$, one profile per sample size, and every notion of Sections 2 and 3 is a statement about how such sequences approach the origin. Table 4 collects the translation. Read top to bottom, the table is the paper in miniature: consistency is convergence in the product topology, uniform learning is convergence in the supremum norm, the moving blind spot is the gap between them, and mixing hidden laws closes the gap for rates.

B Classical facts used in the proofs

We restate, in the form used, the classical results that the proofs invoke. Results cited only for context, such as the consistency theorem of Devroye & Wise (1980), the learnability theory of Fang et al. (2022; 2024), the universal-rate theory of Bousquet et al. (2021), and the slow-convergence theorems of Banach & Steinhaus (1927); Deutsch & Hundal (2010), are not invoked in any proof. One external impossibility result enters Table 1 and is restated last.

Table 4: Learning on a fiber $\mathcal{C}$ and the geometry of its profile set $\mathcal{R}_{\mathcal{C}} \subset [0, M]^{\mathcal{C}}$.

On the fiber	In the profile set	Where
a learner	a sequence in $\mathcal{R}_{\mathcal{C}}$, one term per sample size	Lemma 2.1
consistent (universally learnable)	a sequence in $\mathcal{R}_{\mathcal{C}}$ converges to 0 coordinatewise	Lemma 2.1 and Proposition C.1
every window compatible	0 lies in the product closure of $\mathcal{R}_{\mathcal{C}}$	Proposition C.1
uniformly, equivalently certifiably, learnable	0 lies in the supremum-norm closure; $V(\mathcal{C}) = 0$	Proposition C.1, Theorem 2.4
countable fiber: windows decide consistency	$[0, M]^{\mathcal{C}}$ is metrizable, so closure and sequential closure agree	Theorem 2.3
moving blind spot	a sequence tends to 0 coordinatewise while $\mathcal{R}_{\mathcal{C}}$ stays at supremum distance $V(\mathcal{C}) > 0$ from 0	Section 2
uncountable fiber, compatible but not consistent	0 lies in the product closure but is not a sequential limit	Appendix G.1
compact class, lower semicontinuous excesses	the middle position is impossible	Proposition C.8
running the clock faster	reindexing the sequence along a subsequence	Proposition 2.2
mixture completion $\sigma\mathcal{C}$	new coordinates are bounded by weighted averages of old ones, so supremum norms and coordinatewise limits are unchanged	Proposition 3.1
a universal rate on $\sigma\mathcal{C}$	convergence at rate r in the supremum norm on $\mathcal{C}$	Theorem 3.3
any class-relative criterion linear in the hidden law, such as AUC	the same geometry, with the same conclusions	Proposition D.1

Fact B.1 (Hoeffding's inequality; Hoeffding, 1963)**.** Let $X_1, \dots, X_n$ be independent with $a_i \le X_i \le b_i$, and let $S = \sum_i X_i$. Then $\mathbb{P}\{S - \mathbb{E}S \ge t\} \le \exp\big(-2t^2/\sum_i (b_i - a_i)^2\big)$ for every $t > 0$. In particular, the empirical frequency $\widehat{P}_n(B)$ of an event B under n independent draws from P satisfies $\mathbb{P}\{|\widehat{P}_n(B) - P(B)| > t\} \le 2e^{-2nt^2}$.

Used in Appendix E.1 for the pairwise identification tests.

Fact B.2 (Bounded-differences inequality; McDiarmid, 1989)**.** Let $X_1, \dots, X_n$ be independent and let f be a measurable function of $(X_1, \dots, X_n)$ such that changing the ith coordinate alone changes f by at most c_i. Then $\mathbb{P}\{f - \mathbb{E}f \ge t\} \le \exp\big(-2t^2/\sum_i c_i^2\big)$ for every $t > 0$.

Used in Appendix E.2 with $c_i = 1/n$, where it gives e^{-2nt^2}.

Fact B.3 (Relative entropy under products and kernels; Cover & Thomas, 2006)**.** For probability laws P, Q on a common space, $\mathrm{KL}(P^{\otimes n} \| Q^{\otimes n}) = n\,\mathrm{KL}(P\|Q)$. If K is a Markov kernel and PK, QK denote the laws of its output when its input follows P or Q, then $\mathrm{KL}(PK\|QK) \le \mathrm{KL}(P\|Q)$. For Bernoulli laws the relative entropy is $\mathrm{kl}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1-p}{1-q}$, with the conventions $0 \log 0 = 0$ and $\mathrm{kl}(p, q) = \infty$ when $q \in \{0, 1\}$ and $p \ne q$.

Used in Appendix E.2 to bound the probability of an output under a null law by its probability under nearby alternatives.

Fact B.4 (Upward martingale convergence; Williams, 1991)**.** Let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$ be sigma fields with $\mathcal{F}_\infty = \sigma(\bigcup_n \mathcal{F}_n)$, and let X be integrable. Then $\mathbb{E}[X \mid \mathcal{F}_n] \to \mathbb{E}[X \mid \mathcal{F}_\infty]$ almost surely and in L^1. If $\mathcal{F}_\infty$ is the Borel sigma field of $\mathbb{R}^d$ and $X = c(X')$ for a bounded Borel function c of a random vector X', the limit is $c(X')$.

Used in Appendix F.1 with the sigma fields generated by nested dyadic partitions.

Fact B.5 (Dominated convergence for series)**.** Let $a_k(n) \to a_k$ for each k and $|a_k(n)| \le b_k$ with $\sum_k b_k < \infty$. Then $\sum_k a_k(n) \to \sum_k a_k$. This is dominated convergence for counting measure; with Tonelli's theorem it justifies exchanging an expectation with a nonnegative series.

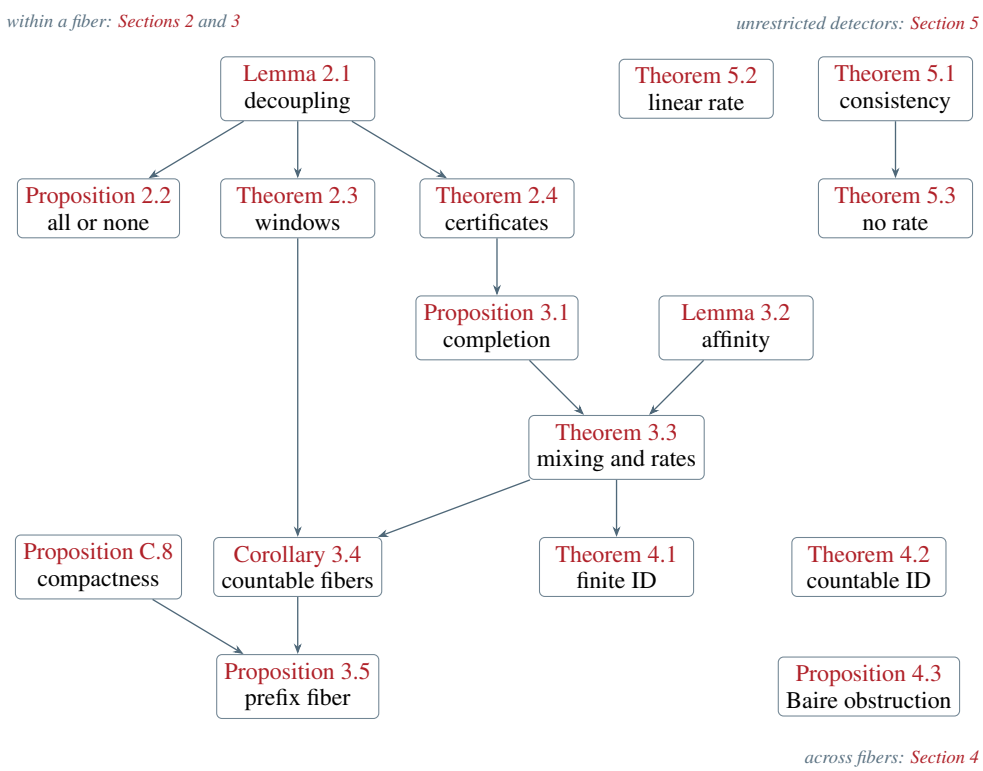

Figure 2: Map of results: an arrow means that the first result is used directly in the proof of the second, and uses implied by transitivity are omitted. Classical facts (Appendix B) are not drawn: dominated convergence enters Proposition 3.1 and Lemma 3.2; Hoeffding's inequality Theorem 4.1; bounded differences and relative entropy Theorem 4.2; Baire category Proposition 4.3; martingale convergence and elementary measure facts Theorems 5.1–5.3. The decomposition (5) enters Theorems 4.1 and 4.2; Proposition C.8 enters only the remark after Proposition 3.5; the nearest-neighbour ranking theorem (Appendix F.4) stands alone.

Used in Proposition 3.1 and Lemma 3.2, where the weights are probability masses and the terms are bounded risks.

Fact B.6 (Elementary measure-theoretic facts)**.** (i) Total variation is subadditive under products, $\mathrm{TV}(P^{\otimes n}, Q^{\otimes n}) \le n\ \mathrm{TV}(P, Q)$, and contracts under Markov kernels, $\mathrm{TV}(PK, QK) \le \mathrm{TV}(P, Q)$. (ii) If $A_1 \supset A_2 \supset \cdots$ and $\mu(A_1) < \infty$, then $\mu(\bigcap_k A_k) = \lim_k \mu(A_k)$. (iii) For events G_k in a probability space, $\mathbb{P}(\limsup_k G_k) \ge \limsup_k \mathbb{P}(G_k)$.

Item (i) is used in Section 4 for the continuity of selection and in Appendix F.2 to compare the sample laws of nearby endpoints; (ii) in Appendix F.1; (iii) in Appendix F.3 to select one hard domain without independence between events.

Fact B.7 (Bretagnolle–Huber inequality; Bretagnolle & Huber, 1979; see Tsybakov, 2009, Chapter 2)**.** For probability laws P, Q on a common space and every measurable ψ with values in $[0, 1]$, $\mathbb{E}_P \psi + \mathbb{E}_Q(1 - \psi) \ge \int \min(dP, dQ) \ge \frac{1}{2} \exp(-\mathrm{KL}(P \| Q))$.

Used in Appendix E.1 to show that the exponential rate over finitely many ID laws can be sharp.

Fact B.8 (Baire category theorem; Oxtoby, 1980)**.** In a completely metrizable space, in particular in any open interval of $\mathbb{R}$, a countable intersection of dense open sets is dense. A set containing such an intersection is called comeager, and its complement meager. In a space without isolated points every countable set, such as $\mathbb{Q} \cap (0, 1)$ in $(0, 1)$, is meager, and a comeager set with a meager set removed is still comeager, hence nonempty.

Used in Proposition 4.3 and Appendix E.3.

Fact B.9 (No certificate on separated supports; Garov & Chaudhuri, 2025, Theorems 2 and 8, in the form used here)**.** In the binary zero–one setting with unrestricted detectors, no learner with an almost surely finite, data-dependent stopping time returns, for every domain with disjoint ID and OOD supports in $\mathbb{R}^d$, a detector whose excess at a fixed known mixture weight is at most ε with probability at least $1-\delta$; the same holds on the subfamily with convex ID support when $d \geq 2$.

Used in Section 5 and Table 1. Since strong excess dominates the excess at any fixed weight, the strong-excess certificate of Section 2 fails on the same families; disjoint supports form a subfamily of $\mathcal{D}^d_{\mathrm{sep}}$. Appendix H discusses the transfer.

C Observation fibers: proofs and the general game

This appendix proves the two structural facts of Section 2, decoupling and the certification hierarchy, and then develops the game on a fiber in the generality in which it is natural: an arbitrary set of nonnegative bounded examinations of an arbitrary strategy space.

C.1 Proof of Lemma 2.1

Fix a sample size n. A possibly randomized learner is a Markov kernel κ_n from $(\mathcal{X} \times \mathcal{Y})^n$ to $\mathcal{H}$: for each sample s, $\kappa_n(s, \cdot)$ is a probability law on $\mathcal{H}$, and $s \mapsto \kappa_n(s, B)$ is measurable for every measurable $B \subset \mathcal{H}$. A deterministic learner is the case $\kappa_n(s, \cdot) = \delta_{A_n(s)}$, and a learner driven by an independent seed $U \sim \rho$ through a measurable map $A_n(s, u)$ is the kernel $\kappa_n(s, B) = \rho\{u : A_n(s, u) \in B\}$. In every case the output law is

$$\mu_n(B) = \int \kappa_n(s, B)\, dP^n(s),$$

a probability law on $\mathcal{H}$ by Tonelli's theorem. It depends on $D = (P, Q)$ only through P, since the sample law P^n and the kernel are its only ingredients; hence μ_n is common to all $D \in \mathcal{C}_P$. For $D \in \mathcal{C}_P$, $\mathcal{E}_D : \mathcal{H} \to [0, M]$ is measurable (Appendix A.1), and Tonelli's theorem gives

$$\mathbb{E}\, \mathcal{E}_D(A(S_n)) = \iint \mathcal{E}_D(h)\, \kappa_n(s, dh)\, dP^n(s) = \int \mathcal{E}_D(h)\, d\mu_n(h) = \widehat{\mathcal{E}}_D(\mu_n).$$

If A is universally consistent on the fiber, the fixed sequence (μ_n) is a consistent schedule. Conversely, a consistent schedule (μ_n) is realized by the constant kernel $\kappa_n(s, \cdot) = \mu_n$, a learner that ignores its sample. Thus universal learnability of $\mathcal{C}_P$ is equivalent to the existence of a consistent schedule.

The law μ_n need not be a point mass: a deterministic learner under a random sample can have a nontrivial output law, as the fair-bit learner of Appendix G.2 shows, and it is this law that the fiber sees. The identity holds at each n separately, with no assumption on the cardinality of the fiber or on the hidden laws; every further statement about a fiber is a statement about sequences of laws on $\mathcal{H}$.

C.2 Proof of Theorem 2.4

Uniform implies certified. A uniform expected bound tending to zero gives a deterministic sample size at which the bound is at most $\varepsilon\delta$, and Markov's inequality proves the certificate condition without any knowledge of the domain.

Certified implies universal. Let $B(\varepsilon, \delta)$ be a certified learner. For $j \geq 1$ put $\varepsilon_j = \delta_j = 2^{-j-1}$ and run $B_j = B(\varepsilon_j, \delta_j)$ on a common infinite ID sample stream, each with its own seed. Let T_j and H_j be its almost surely finite sample count and output; each run obeys $\mathbb{P}_D\{\mathcal{E}_D(H_j) > \varepsilon_j\} \leq \delta_j$. At budget n, simulate the first $\lfloor \log_2 n \rfloor$ runs on the available prefix and select the largest index J_n among those that stopped within the budget, with a fixed default if none did. Fix D and j_0. Once $\lfloor \log_2 n \rfloor \geq j_0$, failure to select an index at least j_0 has probability at most $\mathbb{P}_D(T_{j_0} > n)$. Selection may depend on outputs and stopping times, so we do not condition on it; a union bound over the unconditional failure events gives

$$\mathbb{E}_D \mathcal{E}_D(A(S_n)) \leq \varepsilon_{j_0} + M \sum_{j \geq j_0} \delta_j + M\mathbb{P}_D(T_{j_0} > n) = 2^{-j_0-1} + M2^{-j_0} + M\mathbb{P}_D(T_{j_0} > n).$$

Let $n \to \infty$ and then $j_0 \to \infty$. Dependence through the shared stream does not affect the marginal certificates or the union bound. For computational sequential learners, dovetail the simulations with a growing runtime budget; each terminating run is eventually included.

Collapse on a fiber and over finitely many ID laws. On a fiber, (T, H) has one law for all hidden laws, so a certificate at (ε, δ) gives $\sup_{D \in \mathcal{C}} \mathbb{E}\mathcal{E}_D(H) \le \varepsilon + M\delta$ and hence $V(\mathcal{C}) \le \varepsilon + M\delta$; letting $\varepsilon, \delta \downarrow 0$ gives $V(\mathcal{C}) = 0$, which is uniform learnability by decoupling. Over finitely many ID laws, for each k run a certificate with $\varepsilon_k = \delta_k = 2^{-k-1}$, choose a common truncation level N_k with $\max_P \mathbb{P}_P(T > N_k) \le \delta_k$ over the finitely many ID laws, and make the N_k strictly increasing. At sample size n, use the largest k with $N_k \le n$, stopping at N_k if the certificate has not stopped and ignoring surplus observations. Its worst-domain expected excess is at most $(1 + 2M)2^{-k-1}$, and $k \to \infty$ with n. No countability of the fibers or external randomization beyond that of the certified learner is needed.

C.3 The game on a fiber

By decoupling, a fiber is a game: the learner commits to a law on detectors, nature picks a domain, and the payoff is the expected excess. We develop this game abstractly, since the abstract form also covers rankings (Appendix C.4) and shows which statements depend on countability.

Examinations, profiles, and four values. Let $\mathcal{H}$ be a nonempty measurable strategy space and T a nonempty index set. Each $t \in T$ is an *examination* $e_t : \mathcal{H} \to [0, B]$, measurable, with $B < \infty$. Write $\mathcal{P}(\mathcal{H})$ for the probability laws on $\mathcal{H}$, $\bar{e}_t(\mu) = \int e_t \, d\mu$, and

$$\mathcal{R}_T^{\mathrm{p}} = \{(e_t(h))_{t \in T} : h \in \mathcal{H}\}, \qquad \mathcal{R}_T^{\mathrm{m}} = \{(\bar{e}_t(\mu))_{t \in T} : \mu \in \mathcal{P}(\mathcal{H})\}$$

for the pure and mixed profile sets; the latter is convex and contains the former. On an OOD fiber, $t = D$, $e_D = \mathcal{E}_D$ and $B = M$, and mixed profile sequences are exactly the schedules. For nonempty finite $F \subset T$ and for T itself put

$$V_F^{\mathrm{m}} = \inf_{\mu} \max_{t \in F} \bar{e}_t(\mu), \qquad V_F^{\mathrm{p}} = \inf_{h} \max_{t \in F} e_t(h), \tag{6}$$

$$V_T^{\mathrm{m}} = \inf_{\mu} \sup_{t \in T} \bar{e}_t(\mu), \qquad V_T^{\mathrm{p}} = \inf_{h} \sup_{t \in T} e_t(h), \tag{7}$$

$$W_T^{\mathrm{m}} = \inf_{(\mu_n)} \sup_{t \in T} \limsup_n \bar{e}_t(\mu_n), \qquad W_T^{\mathrm{p}} = \inf_{(h_n)} \sup_{t \in T} \limsup_n e_t(h_n). \tag{8}$$

A *pure sequence* is a data-independent sequence of individual strategies, not an arbitrary deterministic learner, whose output under a random sample can be mixed. A sequence is *resolving* if every coordinate tends to zero; on an OOD fiber, resolving mixed sequences are consistent schedules. All four values lie in $[0, B]$, and constant sequences and point masses give

$$W_T^{\mathrm{m}} \le V_T^{\mathrm{m}} \le V_T^{\mathrm{p}}, \qquad W_T^{\mathrm{m}} \le W_T^{\mathrm{p}} \le V_T^{\mathrm{p}}.$$

The value $V(\mathcal{C})$ of the main text is V_T^{m}, and V_F is V_F^{m}.

Proposition C.1 (Three closures)**.** *For any T:*

1. *$V_T^{\mathrm{m}} = 0$ iff zero lies in the supremum-norm closure of $\mathcal{R}_T^{\mathrm{m}}$, iff some mixed sequence converges to zero uniformly; likewise for pure profiles and V_T^{p}.*
2. *A mixed resolving sequence exists iff zero is a sequential limit of $\mathcal{R}_T^{\mathrm{m}}$ in the product topology, and then $W_T^{\mathrm{m}} = 0$; likewise for pure sequences.*
3. *Zero lies in the product closure of $\mathcal{R}_T^{\mathrm{m}}$ iff $V_F^{\mathrm{m}} = 0$ for every finite $F \subset T$. Moreover*

$$V_F^{\mathrm{m}} \le V_F^{\mathrm{p}} \le |F| \, V_F^{\mathrm{m}}, \tag{9}$$

so the product closures of the two profile sets contain zero simultaneously.

Proof. The supremum norm measures the largest coordinate, which gives the first statement. Convergence in the product topology is coordinatewise, which gives the second. Basic product neighbourhoods constrain finitely many coordinates, which gives the equivalence in the third. For (9), the lower bound uses point masses. Given $\eta > 0$, a mixed strategy with maximum at most $V_F^{\mathrm{m}} + \eta$ has expected sum at most $|F|(V_F^{\mathrm{m}} + \eta)$, so some strategy has sum below that bound plus η; by nonnegativity its maximum is at most its sum. Let $\eta \downarrow 0$. □

Product closure is not sequential closure in general: a sequence indexed by sample size must be constructed, and a net or a family of separately chosen sequences is a different object. Appendix G.1 gives an OOD fiber with zero in the product closure but no resolving sequence.

Theorem C.2 (Master formulas for countable examinations)**.** *Let T be countable and let nonempty finite sets $F_1 \subset F_2 \subset \cdots$ exhaust T. Then*

$$W_T^{\mathrm{m}} = \sup_m V_{F_m}^{\mathrm{m}} = \lim_m V_{F_m}^{\mathrm{m}}, \qquad W_T^{\mathrm{p}} = \sup_m V_{F_m}^{\mathrm{p}} = \lim_m V_{F_m}^{\mathrm{p}}. \tag{10}$$

The following are equivalent: a mixed resolving sequence exists; $W_T^{\mathrm{m}} = 0$; a pure resolving sequence exists; $W_T^{\mathrm{p}} = 0$; all mixed window values vanish; all pure window values vanish.

Proof. For any mixed sequence and finite F, maximum and limsup commute: $V_F^{\mathrm{m}} \le \limsup_n \max_{t\in F} \bar{e}_t(\mu_n) = \max_{t\in F} \limsup_n \bar{e}_t(\mu_n) \le \sup_{t\in T} \limsup_n \bar{e}_t(\mu_n)$. Take the supremum over F_m and the infimum over sequences. Conversely, choose μ_m with $\max_{t\in F_m} \bar{e}_t(\mu_m) \le V_{F_m}^{\mathrm{m}} + 1/m$; every fixed t eventually lies in F_m, so its limsup is at most $\sup_m V_{F_m}^{\mathrm{m}}$. The pure case is identical, and monotonicity makes each supremum a limit. If all window values vanish, the diagonal sequence resolves; conversely a resolving sequence forces every window value to vanish, and (9) equates the pure and mixed zero conditions. □

When T is finite, (10) gives $W_T = V_T$ in both versions: a gap between the sequential and the static value needs infinitely many examinations. Theorem 2.3 is the mixed case on a countable fiber.

Proposition C.3 (Randomization changes positive values)**.** *For every $k \ge 2$ there is a k-domain OOD fiber and a k-element class with $V_T^{\mathrm{m}} = W_T^{\mathrm{m}} = 1/k$ and $V_T^{\mathrm{p}} = W_T^{\mathrm{p}} = 1$.*

Proof. Let $P = \delta_z \otimes \delta_1$, $K = 1$, zero–one loss, and $D_j = (P, \delta_{x_j})$ for k distinct alien points. Let h_i accept z and x_i and reject every other alien point. Both separate infima vanish on each domain because $k \ge 2$, and $\mathcal{E}_{D_j}(h_i) = \mathbf{1}\{i = j\}$. Mixed coordinates are the probabilities of the k strategies, so the best maximum is $1/k$, attained by the uniform mixture; every pure strategy has maximum one. The master formula gives the sequential values; directly, a pure sequence repeats some strategy infinitely often. □

Letting $k \to \infty$ shows that no bound $W_T^{\mathrm{p}} \le f(W_T^{\mathrm{m}})$ with $f(u) \to 0$ as $u \downarrow 0$ holds for all games; for a fixed window, (9) remains valid.

Countable alien sets. Point masses often suffice to decide consistency. The next theorem makes this precise and also covers uncountable fibers.

Theorem C.4 (Point examinations)**.** *Let $K \ge 1$, zero–one loss, an ID law P, and a nonempty countable Borel set $S \subset \mathcal{X}$ with $P_X(S) = 0$ be given, and suppose every $x \in S$ is rejected by some $h \in \mathcal{H}$. Let $\mathcal{C} = \{(P, Q) : Q(S) = 1\}$, $b(h) = E_P^{\mathrm{in}}(h)$, and $a_x(h) = \mathbf{1}\{h(x) \ne K + 1\}$.*

1. *A sequence (μ_n) is consistent on $\mathcal{C}$ iff $\mathbb{E}_{\mu_n} b \to 0$ and $\mathbb{E}_{\mu_n} a_x \to 0$ for every $x \in S$.*
2. *The following are equivalent: $\mathcal{C}$ is universally learnable; every window of $\mathcal{C}$ has value zero; every window admits deterministic detectors with arbitrarily small excess; for every finite $F \subset S$ and $\varepsilon > 0$ some $h \in \mathcal{H}$ has $b(h) \le \varepsilon$ and rejects all of F. When they hold, a deterministic consistent schedule exists.*

Proof. Individual rejectability gives zero optimal OOD risk for each $D_x = (P, \delta_x)$, so

$$\mathcal{E}_{D_x}(h) = \max\{b(h), a_x(h)\}. \tag{11}$$

Since each $D_x \in \mathcal{C}$, consistency implies the two limits. Conversely, for $D = (P, Q) \in \mathcal{C}$, nonnegativity of the optimal risks and countable additivity give

$$\widehat{\mathcal{E}}_D(\mu_n) \le \mathbb{E}_{\mu_n} b + \mathbb{E}_{\mu_n} R_Q^{\mathrm{out}} = \mathbb{E}_{\mu_n} b + \sum_{x\in S} Q\{x\} \mathbb{E}_{\mu_n} a_x, \tag{12}$$

which tends to zero by dominated convergence; this proves part 1. For part 2, consistency makes every window value vanish, and (9) gives the deterministic version. Either version yields the finite-point

condition: given nonempty $F \subset S$ with $q = |F|$ and $t = \min(\varepsilon, 1)/2$, a mixed strategy with all point-domain excesses below $t/(2(q+1))$ has $\mathbb{E}_\mu[b + \sum_{x \in F} a_x] < t/2$ by (11), so some strategy has bracketed sum below t; its $\{0,1\}$-valued point terms vanish and $b \le \varepsilon$. For empty F an approximate ID minimizer suffices. Conversely, given domains $D_i = (P, Q_i)$, $i \le r$, and $0 < \varepsilon < 1$, choose finite subsets of S carrying Q_i-mass at least $1 - \varepsilon$ and let F be their union; a detector with $b \le \varepsilon$ rejecting F has $\mathcal{E}_{D_i} \le \varepsilon$ for all i. Finally, exhaust S by finite F_n and choose h_n rejecting F_n with $b(h_n) \le 1/n$; each point term vanishes eventually, and part 1 shows that (h_n) is consistent on all of $\mathcal{C}$. □

The same tests characterize zero value, though not its size: with $L(\mu) = \max\{\mathbb{E}_\mu b, \sup_{x \in S} \mathbb{E}_\mu a_x\}$, (11) and (12) give

$$L(\mu) \le \sup_{D \in \mathcal{C}} \widehat{\mathcal{E}}_D(\mu) \le 2L(\mu). \tag{13}$$

In the language of Section 3, $\mathcal{C}$ is the completion of its point-mass subfiber, and (12) is Proposition 3.1 for this fiber.

Proposition C.5 (Reductions preserve zero, not values)**.** *Even with one alien point, the value of the actual domains can differ from the value of the ID and point tests.*

Proof. Take $\mathcal{X} = \{0, x\}$, $P = \delta_0 \otimes \delta_1$, $S = \{x\}$, $K = 1$, zero–one loss, and two hypotheses: h_a accepts both points and h_b rejects both. The test profiles (b, a_x) are $(0, 1)$ and $(1, 0)$, with mixed value $1/2$. The only domain D_x has $\mathcal{E}_{D_x} = \max(b, a_x) = 1$ for either hypothesis, so its value is 1. Both formulations agree that no consistent schedule exists. □

Exact formulas such as (10) refer to the examinations in their own definition; substituting convergence-equivalent examinations preserves the zero conditions but not the values, and it does not preserve rates, as Proposition 3.5 shows.

Three positions and compactness.

Corollary C.6 (Compatibility ladder)**.** *For countable examinations, $V_T^{\mathrm{p}} = 0 \Rightarrow V_T^{\mathrm{m}} = 0 \Rightarrow W_T^{\mathrm{m}} = 0 \Leftrightarrow W_T^{\mathrm{p}} = 0 \Leftrightarrow$ every window has value zero, and both one-way implications can be strict on OOD fibers.*

Proof. The implications follow from the definitions and Theorem C.2. For the first strictness, the countable diagonal profiles $\mathcal{E}_{D_j}(h_k) = \mathbf{1}\{j = k\}$ of Appendix G.2 have pure value one and mixed value zero, since uniform mixtures on m strategies give $1/m$. For the second, the point-mass fiber of Proposition 3.5 has mixed value one and a pure consistent schedule. □

Theorem C.7 (Three positions)**.** *For countable examinations exactly one of the following holds. (1) Zero lies outside the product closure of $\mathcal{R}_T^{\mathrm{m}}$: some window has $V_F^{\mathrm{m}} = c > 0$, and every mixed sequence has $\max_{t \in F} \limsup_n \bar{e}_t(\mu_n) \ge c$. (2) Zero lies in the product closure but not in the supremum-norm closure: resolving sequences exist, $V_T^{\mathrm{m}} > 0$, and at every time every mixed strategy has some examination above $V_T^{\mathrm{m}}/2$. (3) Zero lies in the supremum-norm closure, that is, $V_T^{\mathrm{m}} = 0$. On an OOD fiber, case 1 excludes all three guarantees, case 2 is universally but neither uniformly nor certifiably learnable, and case 3 is uniformly and certifiably learnable.*

Proof. The supremum-norm closure lies inside the product closure, so the cases are exhaustive and disjoint. Proposition C.1 identifies case 1 and the zero value, and the max–limsup argument gives the displayed bound. Theorem C.2 constructs resolving sequences in case 2, and the definition of V_T^{m} gives the examination above $V_T^{\mathrm{m}}/2$. Decoupling and Theorem 2.4 translate the cases into learning criteria. □

The same classification applies to the possibly uncountable fiber of Theorem C.4, whose window compatibility and consistency are equivalent by that theorem and whose zero value is equivalent to that of its countable tests by (13).

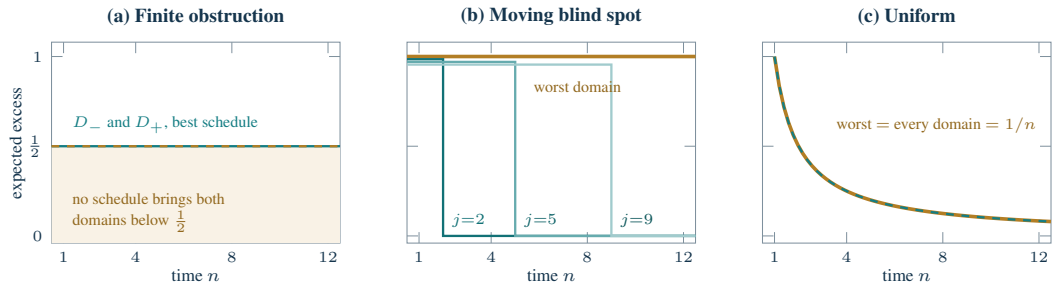

Figure 3: The three positions of Theorem C.7, seen through the expected excess of a schedule over time: a thick line for the worst domain and thin lines for fixed domains. (a) The two-detector conflict of Appendix G.3: a law with $\mathbb{P}(h_-) = p$ has profile $(1-p, p)$, so no schedule brings both domains below $\frac{1}{2}$; the best one keeps both at $\frac{1}{2}$. (b) The prefix fiber with the schedule h_{n+1} (Proposition 3.5): the domain $(P, \delta_{1/j})$ is served from time j on, while the worst domain stays at one. Consistency is convergence of every thin line, uniformity is convergence of the thick one, and the moving blind spot is the gap between them. (c) With the limit detector adjoined, the schedule $(1-\frac{1}{n})\delta_{h_\infty} + \frac{1}{n}\delta_{h_1}$ has excess $1/n$ on every domain.

Proposition C.8 (Compactness removes the middle position)**.** *Let T be arbitrary, and suppose the available laws on strategies form a space that is compact in a topology making every $\bar{e}_t$ lower semicontinuous. If every window has infimum zero over that space, a single available law has all examinations zero; in particular $V_T^{\mathrm{m}} = 0$.*

Proof. The closed sets $\{\mu : \bar{e}_t(\mu) \le 1/k\}$, $t \in T$, $k \ge 1$, have the finite intersection property, since a finite window has a near-optimizer within its smallest required tolerance. Compactness gives a law in all of them, and nonnegativity makes each of its examinations zero. □

The hypothesis concerns the strategies actually available. Adding limit strategies changes the class and hence the problem: with $\overline{\mathcal{H}} = \mathcal{H}_{\mathrm{seq}} \cup \{h_\infty\}$ indexed by the one-point compactification $\mathbb{N} \cup \{\infty\}$ of $\mathbb{N}$, each $m \mapsto \mathcal{E}_{(P,Q)}(h_m) = Q\{1/j : j \ge m\}$ is continuous with limit $0 = \mathcal{E}_{(P,Q)}(h_\infty)$, laws on $\mathbb{N} \cup \{\infty\}$ form a weakly compact set on which $\mu \mapsto \int \mathcal{E}_{(P,Q)}\, d\mu$ is continuous, and the proposition gives $V(\sigma\mathcal{C}_{\mathrm{atom}}; \overline{\mathcal{H}}) = 0$; the acceptance sets of $\overline{\mathcal{H}}$ still form a chain, of VC dimension one. Without compactness, Appendix G.1 shows that uncountably many examinations admit a fourth position relative to sequences: product closure at zero without a resolving sequence.

A threshold example. For $P = \delta_0 \otimes \delta_1$, $D_- = (P, \delta_{-1})$, $D_+ = (P, \delta_1)$, and binary zero–one loss, let h_t accept $(-\infty, t]$, including the two constant thresholds. Every h_t either rejects 0 or accepts -1, so $\mathcal{E}_{D_-}(h_t) = 1$; a threshold $0 \le t < 1$ has profile $(1, 0)$. Both the singleton $\{D_-\}$ and the pair have mixed value one: this is a singleton obstruction. The genuinely two-domain conflict, with value $\frac{1}{2}$ and individually solvable singletons, is in Appendix G.3.

C.4 Rankings use the same game

Let $\mathcal{G}$ be a nonempty class of measurable ranking functions, with a sigma field for which $(r, x) \mapsto r(x)$ is jointly measurable, and fix an OOD fiber. For each domain put $e_D(r) = \sup_{g \in \mathcal{G}} \mathrm{AUC}(g; D) - \mathrm{AUC}(r; D) \in [0, 1]$, with AUC as in Appendix F.4. A randomized ranking draws a whole function, and its expected AUC excess is affine in the ranking law, so Proposition C.1 and Theorems C.2 and C.7 apply to AUC examinations with the same cardinality assumptions, and decoupling is unchanged because training depends only on P. Countable AUC fibers are therefore universally learnable exactly when their windows are compatible, and then a data-independent sequence of nonrandom rankings suffices. Defining an AUC certificate by replacing $\mathcal{E}_D$ with e_D in the certificate of Section 2, the truncation proof of Theorem 2.4 shows that on finite-ID spaces it is equivalent to uniform expected AUC excess. Risk impossibility theorems do not transfer to AUC by themselves: the benchmarks differ, as the conflicts of Appendix G.3 show.

D MIXING: THE FINITE-ID CLASSIFICATION AND THE ROLE OF EACH HYPOTHESIS

Proposition 3.1, Lemma 3.2, and Theorem 3.3 are proved in Section 3. This appendix completes Theorem 4.1, tests each hypothesis of the mixing theorem, and places its proof in the analysis literature. Throughout, the loss is bounded by M, the class is nonempty and fixed, and on a fiber with ID law P we write $a(h) = E_P^{\text{in}}(h)$ and $b(Q) = \inf_{h \in \mathcal{H}} R_Q^{\text{out}}(h)$. Mixtures are always mixtures of hidden laws with the same ID law.

D.1 PROOF OF THEOREM 4.1

Let the distinct ID laws be $P_1, \ldots, P_m$ with fibers $\mathcal{C}_i$. The first assertions are proved in Appendix E.1: universal learnability iff every fiber is universally learnable, the equivalence of certified and uniform learnability by Theorem 2.4, uniform learnability iff every $V(\mathcal{C}_i) = 0$ by Proposition E.1, and a selection error at most $C_0 e^{-c_0 n}$ by pairwise tests, with no test and zero error when $m = 1$. Assume now that the fibers are mixture-complete.

Regime (a). Suppose every $V(\mathcal{C}_i) = 0$. If $M = 0$ there is nothing to prove. Otherwise choose, for each i and n, a law $\nu_n^{(i)}$ with worst-case excess on $\mathcal{C}_i$ at most Me^{-n}, and draw from the law selected by the test, independently of the sample. By (5) the uniform bound is $Me^{-n} + MC_0 e^{-c_0 n}$. Uniform implies certified learning by Theorem 2.4. The exponential is an achievable upper rate; with exact identification, for example when the ID laws have disjoint supports, every prescribed rate is attainable.

Regime (b). If every fiber is universally learnable, playing a consistent schedule after the same test gives a universal learner, with no cardinality restriction on the fibers. If in addition $V(\mathcal{C}_i) > 0$ for some i, every global learner induces one schedule on $\mathcal{C}_i$ by Lemma 2.1, and Theorem 3.3 supplies a fixed domain in $\mathcal{C}_i = \sigma\mathcal{C}_i$ satisfying (2). Uniform learning fails on a fiber with positive value, and certified learning fails because it coincides with uniform learning on finite-ID spaces.

Regime (c). If a fiber is not universally learnable, decoupling rules out a universal learner on that fiber, and the hierarchy of Theorem 2.4 rules out the stronger criteria. Since $V(\mathcal{C}_i) = 0$ implies universal learnability, the three regimes are disjoint and exhaustive. For a finite-ID family $\mathcal{D}$ with arbitrary fibers, Proposition 3.1 shows that the fiberwise completion $\sigma\mathcal{D}$ has the same fiberwise values and consistent schedules, so the classification applies to $\sigma\mathcal{D}$. □

D.2 THE ZERO-OPTIMUM CASE

If $b(Q) = 0$ for every admissible Q, the lower bound in Lemma 3.2 needs no consistency, because linearity of raw OOD risk gives $\max\{a(h), R^{\text{out}}_{\sum_j w_j Q_j}(h)\} \geq w_k \max\{a(h), R^{\text{out}}_{Q_k}(h)\}$ directly. The prefix fiber is of this kind. The general lemma is needed because b is only concave in general, and a consistent schedule is exactly what forces it to be affine on countable mixtures of the laws it serves.

D.3 THE ROLE OF EACH HYPOTHESIS

Countable, not finite, mixtures. Keep the ID law and class of Proposition 3.5, but admit only finitely supported hidden laws on $\{1/j : j \geq 1\}$. This family is the convex hull of $\mathcal{C}_{\text{atom}}$, closed under finite mixtures, and its value is one because it contains every alien point mass. Yet h_{n+1} is exact on each fixed domain once n exceeds its largest atom index, so it attains every rate, the finite prefix being absorbed into a domain-dependent constant. Finite convexity cannot replace countable mixtures.

Zero value need not be attained. For the same class, let the fiber contain the single law with $Q\{1/j\} = 2^{-j}$. A singleton is mixture-complete and has value zero, but every detector and every law on $\mathcal{H}_{\text{seq}}$ has positive OOD excess. Positive tolerances are attainable and exact zero is not, which is why Theorem 3.3 quantifies over positive sequences ε_n.

Positive optimal OOD risk is allowed. Adjoin to the prefix fiber an alien point u that every allowed detector accepts, and replace each hidden law Q by $\widetilde{Q} = \beta\delta_u + (1-\beta)Q$ for a fixed $\beta \in (0,1)$. The family remains mixture-complete, $b(\widetilde{Q}) = \beta$, and the excess is $(1-\beta)$ times the original tail.

The value is $1 - \beta > 0$, the fiber is consistent, and it has no rate: Theorem 3.3 applies with nonzero optimal OOD risk.

Consistency is what makes b affine. In the two-domain conflict of Appendix G.3, both point masses have optimal OOD risk zero while their fair mixture has optimal risk $\frac{1}{2}$, since each allowed detector accepts exactly one of the two alien points. No schedule is consistent on both, and $b((Q_- + Q_+)/2) > [b(Q_-) + b(Q_+)]/2$. Lemma 3.2 excludes precisely this strict concavity along a consistent schedule.

Fiberwise hypotheses do not control selection. The Bernoulli family of Proposition 4.3 has singleton, mixture-complete fibers of value zero, yet it is not universally learnable. Mixing concerns the hidden side of one fiber and says nothing about selecting a fiber among uncountably many; countably many ID laws do glue (Theorem 4.2).

D.4 Tails of the prefix schedule

By (4), the error of the schedule h_{n+1} on a hidden law of $\sigma\mathcal{C}_{\mathrm{atom}}$ is its tail $Q\{1/j : j > n\}$. Figure 4 shows three exact tails: $q_j = 2^{-j}$ gives 2^{-n}, $q_j = 1/(j(j+1))$ gives $1/(n+1)$, and $q_j = \log 2\,[1/\log(j+1) - 1/\log(j+2)]$ gives $\log 2/\log(n+2)$. These are exact errors of one schedule, not lower bounds for all learners, which Theorem 3.3 supplies. Conversely, a subfamily with a uniform tail bound $\sup_Q Q\{1/j : j > n\} \le \tau_n$ is uniformly learnable at rate τ_n by the same schedule, and a subfamily of $\sigma\mathcal{C}_{\mathrm{atom}}$ is uniformly learnable iff its tails vanish uniformly. Indeed, if they do not, there are $\varepsilon > 0$ and, for every m, a law Q_m in the subfamily with $Q_m\{1/j : j \ge m\} > \varepsilon$; a law on $\mathcal{H}_{\mathrm{seq}}$ is a random index L, and choosing m with $\mathbb{P}(L \le m) \ge \frac{1}{2}$ gives it excess at least $\varepsilon/2$ on Q_m, so the value is at least $\varepsilon/2$.

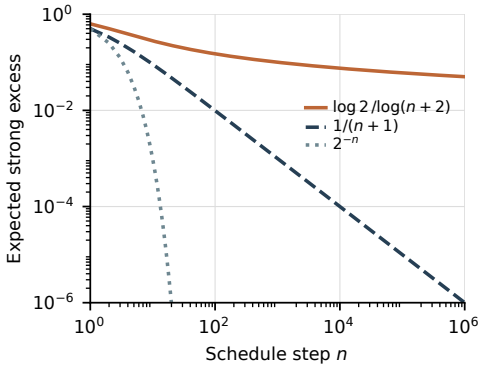

Figure 4: Exact excess of the prefix schedule h_{n+1} on three hidden laws of $\sigma\mathcal{C}_{\mathrm{atom}}$. Changing only the hidden law changes the decay from geometric to reciprocal to logarithmic; a mixture can be slower than any candidate rate (Figure 1(c)).

D.5 One theorem for every linear criterion

The strong excess enters the proofs of Proposition 3.1, Lemma 3.2, and Theorem 3.3 through two properties only: raw losses are linear in the hidden law, and every component of the error is measured against the best strategy in the class. The theorem therefore holds for every criterion with these two properties, including the fixed-proportion excess used by Garov & Chaudhuri (2025) and the AUC excess.

Fix an ID law P and a strategy space $\mathcal{S}$ with a sigma field; strategies may be detectors or rankings. A *linear criterion* assigns to every hidden law Q on $\mathcal{X}$ a measurable function $L_Q : \mathcal{S} \to [0, B]$ such that $L_{\sum_k w_k Q_k}(s) = \sum_k w_k L_{Q_k}(s)$ for every strategy s and every countable mixture; a criterion that does not depend on Q is allowed. Its class-relative excess is $e_Q(s) = L_Q(s) - \inf_{t \in \mathcal{S}} L_Q(t)$, and a *criterion* is a finite maximum $e_Q = \max_{i \le p} e_Q^{(i)}$ of such excesses. Fibers, learners, schedules, values, and rates are defined as in Section 2, with $\widehat{\mathcal{E}}_D(\mu)$ replaced by $\int e_Q \, d\mu$ and M by B.

Proposition D.1 (Mixing makes rates uniform for every linear criterion). *For every criterion of this form, Propositions 3.1 and 2.2 and Theorem 3.3 hold verbatim, and so does Lemma 3.2 in the form: along a schedule consistent on every (P, Q_k), $e_{Q_*} \geq w_k e_{Q_k}$ on $\mathcal{S}$ for every k.*

Proof. Write $b^{(i)}(Q) = \inf_t L_Q^{(i)}(t)$. Linearity and the fact that the infimum of a sum is at least the sum of the infima give $b^{(i)}(Q_*) \geq \sum_k w_k b^{(i)}(Q_k)$, hence $e_{Q_*}^{(i)} \leq \sum_k w_k e_{Q_k}^{(i)}$ for each i. A maximum of weighted averages is at most the weighted average of the maxima, so $e_{Q_*} \leq \sum_k w_k e_{Q_k}$, which is all that the proof of Proposition 3.1 uses. Along a schedule (ν_n) consistent on every (P, Q_k), each component satisfies $0 \leq \int L_{Q_k}^{(i)}\, d\nu_n - b^{(i)}(Q_k) \leq \int e_{Q_k}\, d\nu_n \to 0$, and the argument of Lemma 3.2 gives $b^{(i)}(Q_*) = \sum_k w_k b^{(i)}(Q_k)$. Hence $e_{Q_*}^{(i)} = \sum_j w_j e_{Q_j}^{(i)} \geq w_k e_{Q_k}^{(i)}$ for every component, and taking maxima gives $e_{Q_*} \geq w_k e_{Q_k}$. The proofs of Proposition 2.2 and Theorem 3.3 use only decoupling, which holds because the criterion does not affect the sample, these two inequalities, and the bound B. □

Instances. The strong excess is the case $p = 2$ with $L_Q^{(1)} = R_P^{\text{in}}$, independent of Q, and $L_Q^{(2)} = R_Q^{\text{out}}$. For a fixed proportion $\alpha \in [0, 1]$ of OOD inputs, $L_Q = (1 - \alpha) R_P^{\text{in}} + \alpha R_Q^{\text{out}}$ is linear in Q, and its class-relative excess is the criterion of Garov & Chaudhuri (2025) at a known proportion. For a class $\mathcal{G}$ of rankings carrying a sigma field for which $(r, x) \mapsto r(x)$ is jointly measurable, for instance a countable class, $L_Q(r) = 1 - \text{AUC}(r; (P, Q)) = \int \phi_r\, dQ$ with $\phi_r(z) = 1 - P_X\{x : r(x) > r(z)\} - \frac{1}{2} P_X\{x : r(x) = r(z)\}$ is linear in Q, and its class-relative excess is the AUC excess of Appendix C.4. On a mixture completion, a learner therefore attains a rate for the fixed-proportion excess, or for the AUC excess, iff it attains that rate uniformly. The risk impossibility theorems of the main text do not transfer to AUC by themselves (Appendix G.3), but the mixing mechanism does.

A ranking blind spot. On $\mathcal{X} = \{0\} \cup \{1/j : j \in \mathbb{N}\}$ with $P_X = \delta_0$, let the ranking g_m score 0 by 1, the points $1/j$ with $j < m$ by 0, and the points $1/j$ with $j \geq m$ by 2. Then $\text{AUC}(g_m; (P, \delta_{1/j})) = \mathbf{1}\{j < m\}$, and for every hidden law Q on $\{1/j : j \in \mathbb{N}\}$ the AUC excess of g_m is $Q\{1/j : j \geq m\}$, which is (4). The fiber of point masses has compatible windows and value one, and by Proposition D.1 its completion admits no AUC rate.

D.6 Near and far OOD: a continuous blind spot

The prefix fiber is discrete, but the phenomenon is geometric: a moving blind spot appears where hidden mass can accumulate at a boundary that the class cannot draw exactly. Let $\mathcal{X} = [0, 2]$, $K = 1$, zero–one loss, and $P = \text{Unif}[0, 1] \otimes \delta_1$, and let h_m, $m \in \mathbb{N}$, accept $[0, 1 + \frac{1}{m}]$ with label 1 and reject $(1 + \frac{1}{m}, 2]$. Every h_m accepts the ID support, so all ID risks vanish, and continuity from above gives $\inf_m Q((1, 1 + \frac{1}{m}]) = 0$ for every hidden law Q on $(1, 2]$; hence

$$\mathcal{E}_{(P,Q)}(h_m) = Q\big((1, 1 + \tfrac{1}{m}]\big). \tag{14}$$

The interval $(1, 1 + \frac{1}{m}]$ is the blind spot of h_m: it shrinks toward the ID support without reaching it (Figure 5).

Proposition D.2 (Near and far OOD).
1. *The fiber of all hidden laws on $(1, 2]$ is mixture-complete and consistent, by the schedule h_n, and has value one; no learner attains any rate on it.*
2. *For $\tau \in (0, 1)$, the fiber of hidden laws with $Q((1, 1 + \tau]) = 0$ is mixture-complete with value zero: h_m with $\frac{1}{m} \leq \tau$ is exact on all of it.*
3. *For a nondecreasing $\psi : (0, 1] \to [0, 1]$ with $\psi(1) = 1$ and $\psi(s) \to 0$ as $s \downarrow 0$, the fiber of hidden laws with $Q((1, 1 + s]) \leq \psi(s)$ for all $s \in (0, 1]$ is mixture-complete, and h_n has worst-case excess at most $\psi(1/n)$ on it.*
4. *Adjoining the detector that accepts exactly $[0, 1]$ makes each of these fibers uniformly learnable, by a single exact detector.*

Proof. The three families are defined by constraints that are linear in Q, so they are closed under countable mixtures. For part 1, the schedule h_n has excess $Q((1, 1 + \frac{1}{n}]) \to 0$ by (14); a law on the

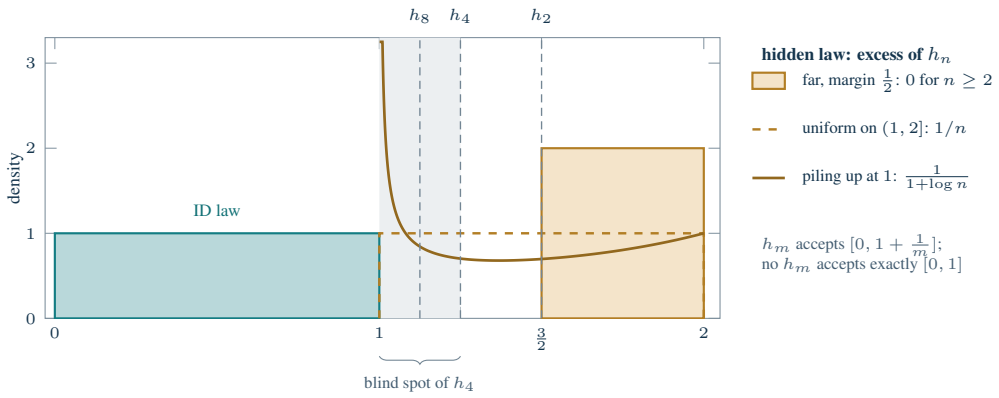

Figure 5: The continuous blind spot of Proposition D.2. The detector h_m accepts $[0, 1 + \frac{1}{m}]$, so the interval $(1, 1 + \frac{1}{m}]$ is accepted although it carries no ID mass. The excess of the schedule h_n on a hidden law is the mass that law puts in the blind spot: zero after finite time for a law with a margin, $1/n$ for the uniform law on $(1, 2]$, and $1/(1 + \log n)$ for the law with distribution function $1/(1 + \log \frac{1}{s})$ at $1 + s$, which piles up at the boundary. Mixtures of laws that pile up at different scales defeat every candidate rate.

class is a random index L, whose excess on the point mass at $1 + \varepsilon$ is $\mathbb{P}(L \leq 1/\varepsilon) \to 1$ as $\varepsilon \downarrow 0$, so the value is one and Theorem 3.3 applies. Parts 2 and 3 follow from (14), and in part 4 the added detector has zero excess on every hidden law on $(1, 2]$. □

The three fibers answer one question, how much hidden mass may sit in the blind spot. A margin removes the blind spot after finite time, as in the far-OOD analyses of Garov & Chaudhuri (2025), which concern unrestricted detectors and unknown ID laws; a density bound near the boundary, $\psi(s) = cs$, gives the rate c/n; and without control, mixtures load the blind spot at every scale and no rate survives. The three fibers share their ID law, so the ID sample cannot tell them apart: which of them describes the deployment is an assumption about the hidden family.

D.7 Relation to the condensation of singularities

The proof of Theorem 3.3 is a gliding-hump argument, the mechanism of the uniform boundedness principle, which Banach & Steinhaus (1927) published as a principle of condensation of singularities, and the theorem has the principle's shape: for a given learner and rate, either uniform control at that rate or a single point that is slower. Deutsch & Hundal (2010) proved the corresponding dichotomy for bounded linear operators on a Banach space: pointwise convergence is either uniform or almost arbitrarily slow, in the sense that for every proposed rate a fixed input is slower infinitely often. In learning theory the diagonal pattern is classical: a single distribution on countably many atoms makes a consistent classification rule arbitrarily slow (Devroye, 1982), (Devroye et al., 1996, Chapter 7), and individual lower bounds keep one distribution along infinitely many sample sizes (Antos & Lugosi, 1998; Bousquet et al., 2021).

What the OOD setting adds is the reason the pattern applies. Decoupling reduces every learner to one sequence of laws on a fiber. The strong excess is not linear in the hidden law, and its benchmark is an optimum over a restricted class that may be positive and unattained; along a consistent schedule, Lemma 3.2 makes the OOD excess affine in the mixture weights, so that the expected OOD excesses act on mixtures of the chosen hidden laws as positive linear functionals of the weights in ℓ^1, and the strong excess of a mixture retains the fraction w_k of each component's excess. Positivity turns the witness into probability weights, and completion keeps it in the fiber. Our proof is self-contained and invokes none of these results.

E SELECTING THE FIBER: PROOFS

This appendix proves the results of Section 4. The within-fiber term of (5) is handled by Sections 2 and 3; here the question is how well a learner can select the fiber from the sample. The decomposition itself holds because the final draw from $\nu_n^{(\widehat{i}_n)}$ is independent of the sample given $\widehat{i}_n$, and the excess is at most M on the event $\widehat{i}_n \neq i$.

E.1 FINITELY MANY ID LAWS

Let the distinct ID laws be $P_1, \dots, P_m$. Necessity of fiberwise universal learnability follows by restricting any learner to one fiber (Lemma 2.1). For sufficiency, let $(\nu_n^{(i)})$ be a consistent schedule on the fiber of P_i.

Identification. For $m \geq 2$, choose for each unordered pair $i \neq j$ a measurable event $B_{ij} = B_{ji}$ with $\delta_{ij} = |P_i(B_{ij}) - P_j(B_{ij})| > 0$, which exists because the laws are distinct, and let $\delta = \min_{i \neq j} \delta_{ij}$. Select the least index i with

$$|\widehat{P}_n(B_{ij}) - P_i(B_{ij})| \leq \delta_{ij}/3 \qquad \text{for every } j \neq i,$$

and index 1 if none qualifies. When P_{i_*} is true, Hoeffding's inequality (Fact B.1) and a union bound show that, outside an event of probability at most $2(m-1)e^{-2n\delta^2/9}$, every comparison involving i_* is within $\delta_{i_* j}/3$ of its mean; then i_* qualifies and every other index fails its comparison with i_*. For $m = 1$ no test is needed. Drawing from $\nu_n^{(\widehat{i})}$ gives, for every D in the true fiber,

$$\mathbb{E}\mathcal{E}_D(A(S_n)) \leq \widehat{\mathcal{E}}_D\big(\nu_n^{(i_*)}\big) + 2(m-1)Me^{-2n\delta^2/9}, \tag{15}$$

which is (5) with an explicit selection error.

Rates. If a fiber is countable, list its domains $D_1, D_2, \dots$ and choose, as in Corollary 3.4, laws $\bar{\nu}_k^{(i)}$ with $\max_{j \leq k} \widehat{\mathcal{E}}_{D_j}(\bar{\nu}_k^{(i)}) \leq M2^{-k}$; then $\widehat{\mathcal{E}}_{D_j}(\bar{\nu}_k^{(i)}) \leq M2^j 2^{-k}$ for all k, and (15) gives a domain-wise exponential rate. On one known countable fiber every rate is attainable, and the exponential comes from the selection term. This acceleration uses an enumeration of the domains themselves; it does not extend to their countable mixtures, whose rates are governed by Theorem 3.3.

Proposition E.1 (Uniform learnability over finitely many ID laws)**.** *For a family with finitely many distinct ID laws, uniform learnability by randomized learners is equivalent to $V(\mathcal{C}_P) = 0$ for every fiber.*

Proof. Necessity: a uniform learner gives, by decoupling, laws $\mu_n^{(P)}$ with $\sup_{D \in \mathcal{C}_P} \widehat{\mathcal{E}}_D(\mu_n^{(P)}) \to 0$. Sufficiency: select $\nu_n^{(i)}$ with worst-case excess at most M/n on the fiber of P_i and draw from the law selected by the test above; the expected excess is at most M/n plus M times the exponentially small selection error, uniformly over domains. If $M = 0$ all risks vanish. □

The randomization is essential, and a deterministic learner can sometimes produce it from its sample (Appendix G.2).

Proposition E.2 (The exponential can be sharp)**.** *Let $\mathcal{X} = \{0,1,2,3\}$, $K = 1$, zero–one loss, and let P_1, P_2 have constant label 1 and marginals $\mathrm{Ber}(\frac{1}{2})$ and $\mathrm{Ber}(\frac{3}{4})$ on $\{0,1\}$. Let $Q_1 = \delta_2$, $Q_2 = \delta_3$, and $\mathcal{H} = \{h_2, h_3\}$, where h_x rejects only x. The family $\{(P_1, Q_1), (P_2, Q_2)\}$ is uniformly learnable at an exponential rate, and every learner has*

$$\max_{i=1,2} \mathbb{E}\mathcal{E}_{(P_i,Q_i)}(A(S_n)) \geq \tfrac{1}{4}e^{-n\kappa}, \qquad \kappa = \mathrm{KL}(P_1 \| P_2) = \tfrac{1}{2}\log\tfrac{4}{3}.$$

Consequently, if a learner attains a rate r on this family, then $r(t) \geq c'e^{-\kappa t/c}$ for some $c, c' > 0$ and all large t; no rate faster than every exponential is attainable.

Proof. Each fiber is a single domain with an exact detector, h_2 for the first and h_3 for the second, so Theorem 4.1 gives the exponential upper bound. All ID risks vanish, $\mathcal{E}_{(P_1,Q_1)}(h_3) =$

$1 = \mathcal{E}_{(P_2,Q_2)}(h_2)$, and the other two excesses are zero. For a learner with kernel κ_n, put $\psi(s) = \kappa_n(s, \{h_3\})$; then the two expected excesses are $\mathbb{E}_{P_1^n}\psi$ and $\mathbb{E}_{P_2^n}(1-\psi)$, and Fact B.7 with $\mathrm{KL}(P_1^n \| P_2^n) = n\kappa$ (Fact B.3) bounds their sum below by $\frac{1}{2}e^{-n\kappa}$. If the learner attains r with constants C_i, c_i, put $C = \max_i C_i$ and $c = \min_i c_i$; then $\frac{1}{4}e^{-n\kappa} \le C\, r(cn)$ for all n, and monotonicity of r gives the stated bound. □

E.2 Countably many ID laws

This subsection proves Theorem 4.2. The qualitative part separates selection from the hidden-domain problem and carries no rate; the quantitative part assumes fiberwise $V = 0$ and a sub-linear exponent, and Proposition E.4 shows that the sub-linear restriction is sharp.

A metric that distinguishes the candidate laws. Write $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. A separable metric space has a countable base, so its Borel sigma field is generated by a countable algebra $\{B_j : j \ge 1\}$, and finite measures that agree on this algebra agree on all Borel sets. Define

$$d_0(P, P') = \sum_{j \ge 1} 2^{-j} |P(B_j) - P'(B_j)|.$$

This is a metric on Borel probability laws: nonnegativity, symmetry and the triangle inequality hold coordinatewise, and the generating-algebra property gives definiteness. No common separation margin between the candidate laws is assumed. For the empirical law $\widehat{P}_n$, each frequency $\widehat{P}_n(B_j)$ has variance at most $1/(4n)$, so Cauchy–Schwarz and Tonelli give

$$\mathbb{E}_{P^n} d_0(\widehat{P}_n, P) \le \sum_{j \ge 1} 2^{-j} \sqrt{\mathrm{Var}(\widehat{P}_n(B_j))} \le \frac{1}{2\sqrt{n}}. \tag{16}$$

Select the first plausible law. Enumerate the candidate ID laws as $P_1, P_2, \ldots$ and fix thresholds $\eta_n \downarrow 0$. Given S_n, choose the least i with $d_0(\widehat{P}_n, P_i) \le \eta_n$, and index 1 if none exists; this is measurable because a least-index event subtracts finitely many earlier events. Fix a true index i_*. If $i_* > 1$ put $\gamma_{i_*} = \min_{j<i_*} d_0(P_j, P_{i_*}) > 0$, a minimum over finitely many earlier candidates. Once $2\eta_n < \gamma_{i_*}$, on the event $d_0(\widehat{P}_n, P_{i_*}) \le \eta_n$ the true index qualifies and every smaller index fails by the triangle inequality, so $\widehat{i}_n = i_*$. Later candidates arbitrarily close to P_{i_*} do not matter to this least-index rule. Hence $\mathbb{P}_{P_{i_*}^n}(\widehat{i}_n \ne i_*) \le \mathbb{P}_{P_{i_*}^n}\{d_0(\widehat{P}_n, P_{i_*}) > \eta_n\}$ after an index-dependent finite prefix, and this tends to zero; with $\eta_n = n^{-1/4}$, Markov's inequality and (16) bound it by $\frac{1}{2}n^{-1/4}$.

Proof of the qualitative part. Necessity is decoupling. For sufficiency, fix a consistent schedule $\nu_n^{(i)}$ on each fiber and, after selecting $\widehat{i}_n$, draw from $\nu_n^{(\widehat{i}_n)}$ independently of the sample; a countable selection among fixed laws is a measurable kernel. For D in the true fiber, (5) gives $\mathbb{E}\mathcal{E}_D(A(S_n)) \le \widehat{\mathcal{E}}_D(\nu_n^{(i_*)}) + M\mathbb{P}_{P_{i_*}^n}(\widehat{i}_n \ne i_*) \to 0$.

Proposition E.3 (Quantitative gluing). *Let the distinct ID laws be at most countable and let every fiber have $V(\mathcal{C}_{P_i}) = 0$. For every nondecreasing $g : \mathbb{N} \to (0, \infty)$ with $g(n) \to \infty$ and $g(n) = o(n)$, some measurable randomized learner has, for each i and a finite constant C_i,*

$$\sup_{D \in \mathcal{C}_{P_i}} \mathbb{E}\mathcal{E}_D(A(S_n)) \le C_i e^{-g(n)} \qquad \text{for every } n \ge 1.$$

Proof. Concentration. Changing one observation changes each $\widehat{P}_n(B_j)$ by at most $1/n$, hence changes $d_0(\widehat{P}_n, P)$ by at most $1/n$. The bounded-differences inequality (Fact B.2) gives $\mathbb{P}_{P^n}\{d_0(\widehat{P}_n, P) > \mathbb{E}d_0(\widehat{P}_n, P) + t\} \le e^{-2nt^2}$, so with (16) and $t = \sqrt{g(n)/(2n)}$ the threshold $\eta_n = \frac{1}{2\sqrt{n}} + \sqrt{\frac{g(n)}{2n}}$ satisfies $\mathbb{P}_{P^n}\{d_0(\widehat{P}_n, P) > \eta_n\} \le e^{-g(n)}$ for every P and n, and $\eta_n \to 0$ because $g(n) = o(n)$.

Identification. Use the first-plausible-law rule with this η_n. For a true index i choose N_i with $2\eta_n < \gamma_i$ for $n \ge N_i$ (with $N_1 = 1$); as above, $\mathbb{P}_{P_i^n}(\widehat{i}_n \ne i) \le e^{-g(n)}$ for $n \ge N_i$.

Gluing. If $M = 0$ all risks vanish. Otherwise $V(\mathcal{C}_{P_i}) = 0$ gives laws $\nu_n^{(i)}$ with $\sup_{D \in \mathcal{C}_{P_i}} \widehat{\mathcal{E}}_D(\nu_n^{(i)}) \le Me^{-g(n)}$. Drawing from $\nu_n^{(\widehat{i}_n)}$, (5) gives $\mathbb{E}\mathcal{E}_D(A(S_n)) \le 2Me^{-g(n)}$ for $n \ge N_i$ and $D \in \mathcal{C}_{P_i}$. For $n < N_i$, boundedness and monotonicity of g give $\mathbb{E}\mathcal{E}_D(A(S_n)) \le M \le Me^{g(N_i)}e^{-g(n)}$, so $C_i = M \max\{2, e^{g(N_i)}\}$ works. □

The bound is uniform within each fiber and pointwise across ID laws: the learner may depend on g, and C_i and N_i depend on the true ID law. Fiberwise $V = 0$ cannot be dropped, since $\sigma\mathcal{C}_{\text{atom}}$ alone is universally learnable at no rate. Whether certified and uniform learning coincide over countably infinite ID families is not settled by this argument.

Proposition E.4 (No exponential rate over countably many ID laws). *Let $\mathcal{X} = \{0,1,2,3\}$, $K = 1$, zero–one loss. For $j \ge 0$ let P_j have constant label 1 and marginal $\mathrm{Ber}(p_j)$ on $\{0,1\}$, with $p_0 = \frac{1}{2}$ and $p_j = \frac{1}{2} + \frac{1}{j+4}$ for $j \ge 1$; let $Q_0 = \delta_2$ and $Q_j = \delta_3$ for $j \ge 1$; and let $\mathcal{H} = \{h_2, h_3\}$, where h_x rejects only x. The family $\mathcal{D} = \{(P_j, Q_j) : j \ge 0\}$ has countably many ID laws, singleton fibers of value zero, and is universally learnable. Yet for every learner universally consistent on $\mathcal{D}$ and all $C, c > 0$, the inequality $\mathbb{E}\mathcal{E}_{(P_0,Q_0)}(A(S_n)) \le Ce^{-cn}$ fails for infinitely many n.*

Proof. The p_j are distinct, so the fibers are singletons, with exact detectors h_2 on D_0 and h_3 on D_j, $j \ge 1$; Theorem 4.2 gives universal learnability. All ID risks vanish, $\mathcal{E}_{D_0}(h_3) = 1 = \mathcal{E}_{D_j}(h_2)$, and $\mathcal{E}_{D_0}(h_2) = 0 = \mathcal{E}_{D_j}(h_3)$. For a consistent learner write $\alpha_n = \mathbb{P}_{P_0^n}\{A(S_n) = h_3\} = \mathbb{E}\mathcal{E}_{D_0}(A(S_n))$ and $p_{j,n} = \mathbb{P}_{P_j^n}\{A(S_n) = h_3\}$, so $1 - p_{j,n} = \mathbb{E}\mathcal{E}_{D_j}(A(S_n)) \to 0$. The output is a Markov kernel of the sample, so data processing for relative entropy (Fact B.3) gives

$$n\,\mathrm{KL}(P_j \| P_0) \ge \mathrm{kl}(p_{j,n}, \alpha_n) \ge p_{j,n} \log \frac{1}{\alpha_n} - \log 2.$$

If $\alpha_n \le Ce^{-cn}$ for all $n \ge n_0$, dividing by n and letting $n \to \infty$ gives $\mathrm{KL}(P_j \| P_0) \ge c$ for every $j \ge 1$, contradicting $\mathrm{KL}(P_j \| P_0) = \mathrm{kl}(p_j, \frac{1}{2}) \to 0$. □

Together, Propositions E.3 and E.4 locate the countable-ID boundary: every $e^{-g(n)}$ with $g(n) = o(n)$ is attainable under fiberwise $V = 0$, and $g(n) = cn$ is not in general. The obstruction is the classical one of testing a null law against alternatives that approach it.

E.3 Uncountably many ID laws: the Baire obstruction

Selection is continuous. Fix a learner with kernels κ_n and write μ_n^P for its output law under P^n. For every measurable $B \subset \mathcal{H}$, $|\mu_n^P(B) - \mu_n^{P'}(B)| = |\int \kappa_n(s, B)\, d(P^n - P'^n)(s)| \le \mathrm{TV}(P^n, P'^n) \le n\,\mathrm{TV}(P, P')$ by Fact B.6(i), with total variation normalized as a supremum over events. For a family of ID laws topologized by total variation, each map $P \mapsto \mu_n^P(B)$ is therefore continuous, and a universally consistent learner converges, fiber by fiber, as a pointwise limit of continuous functions of the ID law. On a countable family, the first-plausible-law rule realizes any consistent choice of fiberwise schedules (Appendix E.2). On an uncountable family, pointwise limits of continuous functions are continuous on a comeager set, and a fiberwise choice that is discontinuous everywhere cannot be realized.

Proof of Proposition 4.3. Let $\mathcal{X} = \{0,1,2,3\}$ and, for $t \in (0,1)$, let P_t be the law of $(X, 1)$ with $X \sim \mathrm{Ber}(t)$ on $\{0,1\}$. Distinct t give distinct laws, so each fiber is the single domain $D_t = (P_t, Q_t)$. Both detectors accept 0 and 1 with label 1, so all ID risks vanish; the optimal OOD risk is zero on every domain, and $\mathcal{E}_{D_t}(h_2) = \mathbf{1}\{t \notin \mathbb{Q}\}$, $\mathcal{E}_{D_t}(h_3) = \mathbf{1}\{t \in \mathbb{Q}\}$. For a learner with kernels κ_n, the probability of outputting h_2 is

$$\phi_n(t) = \sum_{s \in \{0,1\}^n} \kappa_n\big((s, 1^n), \{h_2\}\big)\, t^{|s|}(1-t)^{n-|s|},$$

a polynomial in t, where $|s|$ counts the ones in s. The expected excess is $1 - \phi_n(t)$ at rational t and $\phi_n(t)$ at irrational t. Suppose the learner is consistent at every rational, so $\phi_n(q) \to 1$ for each $q \in \mathbb{Q} \cap (0,1)$. For $N, m \in \mathbb{N}$ the set $G_{N,m} = \bigcup_{n \ge N}\{t : \phi_n(t) > 1 - 1/m\}$ is open by continuity and contains every rational, hence is dense. The set $G = \bigcap_{N,m} G_{N,m}$ is comeager in $(0,1)$, and so

is $G \setminus \mathbb{Q}$, which is therefore dense and in particular nonempty by Fact B.8. For $t \in G$ and every m, $\phi_n(t) > 1 - 1/m$ for infinitely many n, so $\limsup_n \phi_n(t) = 1$; at irrational $t \in G$ this is the limsup of the expected excess. Thus a learner is either inconsistent at some rational or has limsup error one on a comeager set of irrationals; no learner is universally consistent. □

The conclusion concerns category, not measure: a comeager set of irrationals can be Lebesgue-null, and Cover (1973) gives tests that are eventually correct at every rational and at Lebesgue-almost every irrational, almost surely. There is no irrational parameter that defeats every learner, since always returning h_3 is exact at all of them. The quantified form above, consistency at the rationals forcing failure at comeager many irrationals, is the precise statement; Dembo & Peres (1994) give the general topological criterion for consistent testing.

F UNRESTRICTED DETECTORS: PROOFS

With $\mathcal{H} = \mathcal{H}_{\text{all}}$, every fiber of $\mathcal{D}^d_{\text{sep}}$ has value zero, attained by the detector that rejects outside the closed set $\operatorname{supp} P_X$ and applies a measurable Bayes rule on it: its ID excess vanishes because $P_X(\operatorname{supp} P_X) = 1$, and its OOD risk vanishes because $Q(\operatorname{supp} P_X) = 0$. This appendix proves the statistical results of Section 5 and the ranking theorem.

F.1 CONSISTENCY: PROOF OF THEOREM 5.1

Fix (P, Q) with $Q(\operatorname{supp} P_X) = 0$. Write $\mathcal{Q}_n$ for the half-open dyadic cubes of side $s_n = 2^{-j_n}$, $j_n = \lceil (\log_2 n)/(2d) \rceil$. Then $s_n \downarrow 0$, $n s_n^d \geq 2^{-d}\sqrt{n} \to \infty$, and the partitions refine. The detector rejects empty cubes and, on an occupied cube, minimizes the empirical conditional loss over all $K + 1$ outputs, with a fixed tie-breaking rule.

OOD error. Almost surely all training points lie in the closed set $S = \operatorname{supp} P_X$, so every accepted point is within $\sqrt{d}\, s_n$ of S. Since rejecting everything has zero OOD risk,

$$\mathbb{E}[E^{\text{out}}_Q(h_n)] = \mathbb{E}[R^{\text{out}}_Q(h_n)] \leq MQ\{x : \operatorname{dist}(x, S) \leq \sqrt{d}\, s_n\} \longrightarrow 0,$$

by continuity from above (Fact B.6(ii)): the neighbourhoods decrease to S, which has Q-mass zero.

Approximating the ID Bayes rule. Let $\eta_y(x) = P(Y = y \mid X = x)$ and $c(a, x) = \sum_{y=1}^{K} \eta_y(x)\ell(a, y)$ for $a \in \{1, \ldots, K + 1\}$. The least ID risk over measurable rules is $\mathbb{E}_{P_X} \min_a c(a, X)$, attained by the smallest minimizer. For a cube B of mass $p_B = P_X(B) > 0$ put $\bar{c}_n(a, B) = \mathbb{E}[c(a, X) \mid X \in B]$, and let $\widetilde{h}_n$ minimize these cube-conditional costs. The sigma fields $\mathcal{F}_n = \sigma(\mathcal{Q}_n)$ increase to the Borel sigma field, so upward martingale convergence (Fact B.4), applied to the finitely many bounded functions $c(a, \cdot)$, gives $\sum_a \mathbb{E}|\mathbb{E}[c(a, X) \mid \mathcal{F}_n] - c(a, X)| \to 0$. The minimum of finitely many coordinates is 1-Lipschitz in the supremum norm, so $R^{\text{in}}_P(\widetilde{h}_n) - \inf_h R^{\text{in}}_P(h) \to 0$.

Empty cubes and estimation. The expected missing mass is $U_n = \sum_{B \in \mathcal{Q}_n} p_B(1 - p_B)^n$. Fix $\varepsilon > 0$ and a box $[-a, a]^d$ with outside mass at most ε; at most $C_a s_n^{-d}$ cubes meet it, and $p(1-p)^n \leq 1/(en)$ gives $U_n \leq C_a/(e n s_n^d) + \varepsilon$, so $U_n \to 0$. On an occupied cube with count N_B and empirical cost $\widehat{c}_n(a, B)$, the labels are conditionally independent with probabilities $q_y = P(Y = y \mid X \in B)$, so $\mathbb{E}[\max_a |\widehat{c}_n(a, B) - \bar{c}_n(a, B)| \mid N_B = N] \leq M \sum_y \sqrt{q_y/N} \leq M\sqrt{K/N}$. For $N_B \sim \operatorname{Bin}(n, p_B)$, $\mathbb{E}[1/(N_B + 1)] \leq 1/((n + 1)p_B)$; with $1/N \leq 2/(N + 1)$ and Jensen's inequality, $\mathbb{E}[p_B \max_a |\widehat{c}_n(a, B) - \bar{c}_n(a, B)| \mathbf{1}\{N_B \geq 1\}] \leq M\sqrt{2Kp_B/(n + 1)}$. The plug-in comparison of empirical and population minimizers costs at most twice this deviation; summing over cubes meeting the box and applying Cauchy–Schwarz bounds the estimation contribution by $2M\sqrt{2KC_a/((n + 1)s_n^d)} \to 0$. Empty cubes contribute at most MU_n and the remaining cubes at most $M\varepsilon$; the comparison may be restricted to occupied cubes because the cube-integrated excess of any fixed action over the cube oracle is nonnegative. Adding the approximation term and letting $\varepsilon \downarrow 0$ gives $\mathbb{E}E^{\text{in}}_P(h_n) \to 0$, and $\mathcal{E}_{(P,Q)} \leq E^{\text{in}}_P + E^{\text{out}}_Q$ proves the theorem.

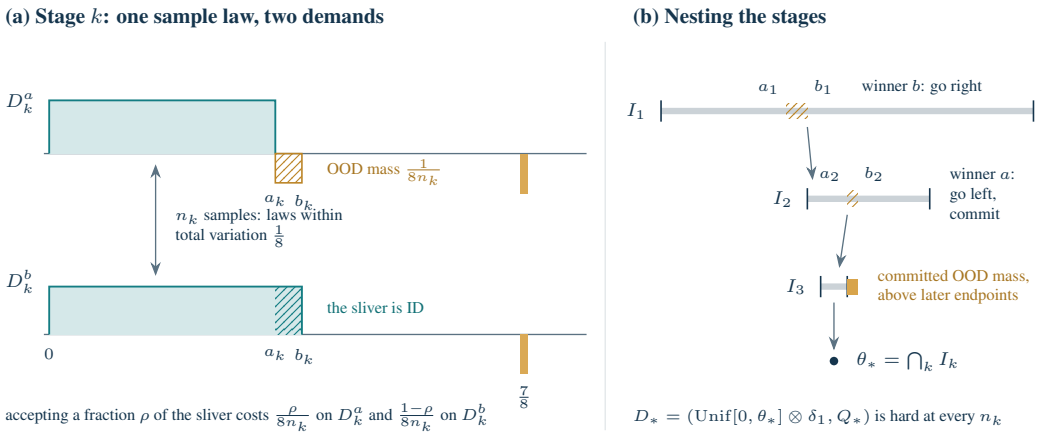

Figure 6: The construction behind Theorem 5.2, not to scale. (a) At stage k, the domains D_k^a and D_k^b have ID endpoints $a_k < b_k$ with $b_k - a_k = \gamma_k$, chosen so that their size-n_k samples have laws within total variation $\frac{1}{8}$. Yet the sliver $(a_k, b_k]$ is OOD for D_k^a, carrying mass $\frac{1}{8n_k}$, and ID for D_k^b, so every rule loses order $1/n_k$ on one of them. (b) The harder of the two decides the next interval, to the left of a_k or to the right of b_k; when it is D_k^a, its sliver is committed as OOD mass, which then lies above every later endpoint. The intervals shrink to θ_*, and the domain with ID endpoint θ_* and all committed mass is within $O(1/n_k)$ in total variation of each stage's harder domain, so it inherits all the stage lower bounds.

F.2 An exact linear rate: proof of Theorem 5.2

Here $K = 1$, zero–one loss, $\mathcal{H} = \mathcal{H}_{\mathrm{all}}$, $P_\theta = \mathrm{Unif}[0, \theta] \otimes \delta_1$ with $\theta \in [\frac{1}{4}, \frac{3}{4}]$, and $Q((\theta, 1]) = 1$; both separate optimal risks are zero. The detector accepting $[0, M_n]$, $M_n = \max_i X_i$, has zero OOD risk and ID risk $(\theta - M_n)/\theta$, so its expected strong excess is exactly $1/(n+1)$, uniformly over the family.

For the fixed-domain lower bound, fix any learner. We construct nested closed intervals $I_k = [l_k, u_k]$, times n_k, and a list of committed OOD subprobability measures. Initially $I_1 = [\frac{1}{4}, \frac{3}{4}]$, $n_0 = 8$, and the list is empty. At the start of stage k every committed measure has full mass in $(u_k, 1]$ and their total mass is $w_k < \frac{1}{8}$. Set $a_k = l_k + |I_k|/3$ and choose an integer $n_k > 10^4 n_{k-1}^2$ large enough that

$$\gamma_k = \frac{a_k}{8n_k - 1} \le |I_k|/3, \qquad b_k = a_k + \gamma_k.$$

Under $\mathrm{Unif}[0, b_k]$ the sliver $(a_k, b_k]$ has mass $1/(8n_k)$; write $\mu_k = (8n_k)^{-1}\,\mathrm{Unif}(a_k, b_k]$. The two candidate OOD laws consist of the existing commitments together with $\mu_k + (1 - w_k - 1/(8n_k))\delta_{7/8}$ or $(1 - w_k)\delta_{7/8}$, with ID endpoints a_k and b_k respectively; call the domains D_k^a and D_k^b. For any rule h with accepted fraction $\rho(h)$ of the sliver, $\mathcal{E}_{D_k^a}(h) \ge \rho(h)/(8n_k)$ and $\mathcal{E}_{D_k^b}(h) \ge (1 - \rho(h))/(8n_k)$. The two size-$n_k$ sample laws are within total variation $\frac{1}{8}$ by Fact B.6(i) and $\gamma_k/b_k = 1/(8n_k)$, so the expected accepted fractions differ by at most $\frac{1}{8}$, and

$$\max\{\mathbb{E}_{D_k^a}\mathcal{E}_{D_k^a}(A(S_{n_k})), \mathbb{E}_{D_k^b}\mathcal{E}_{D_k^b}(A(S_{n_k}))\} \ge \frac{1}{8n_k} \min_{r \in [0,1]} \max\{r - \tfrac{1}{8}, 1 - r\} = \frac{7}{128n_k} \ge \frac{1}{19n_k}.$$

Select a candidate attaining this maximum and put $\eta_{k+1} = \min\{|I_k|/3, 10^{-4}n_k^{-2}\}$. If the winner is a, commit μ_k and set $I_{k+1} = [a_k - \eta_{k+1}, a_k]$; if it is b, commit nothing and set $I_{k+1} = [b_k, b_k + \eta_{k+1}]$. These intervals lie in I_k, and every commitment keeps full mass strictly above the new upper endpoint. Let θ_* be the common point of the intervals and Q_* the sum of all commitments with the remaining mass at $\frac{7}{8}$; the committed mass is at most $\sum_k 1/(8n_k) < \frac{1}{8}$, and every commitment lies in $(\theta_*, 1]$, so $D_* = (P_{\theta_*}, Q_*)$ belongs to the family. At stage k, with D_W the winner and θ_W its endpoint, $|\theta_* - \theta_W| \le 10^{-4}n_k^{-2}$, hence $\mathrm{TV}(P_{\theta_*}^{n_k}, P_{\theta_W}^{n_k}) \le 4 \cdot 10^{-4}/n_k$; the ID risks of any rule differ by at most $4 \cdot 10^{-4}n_k^{-2}$, and the OOD laws differ only through later commitments, with total variation at most $2\sum_{j>k} 1/(8n_j) \le 10^{-3}/n_k$. As the separate optimal risks vanish, transferring the sample law

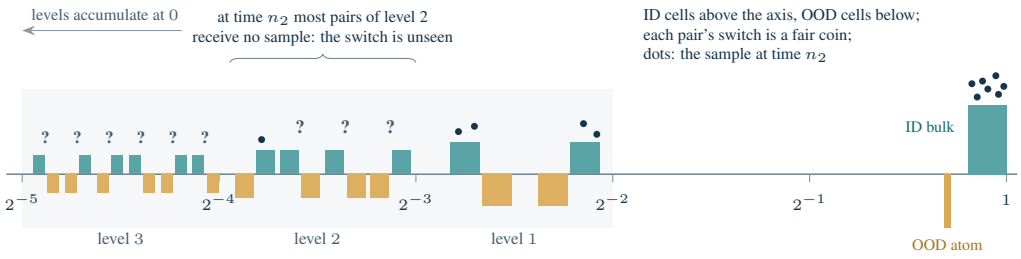

Figure 7: The construction behind Theorem 5.3, on a logarithmic scale and not to scale. Level k occupies the band $(2^{-k-2}, 2^{-k-1})$ and contains m_k pairs of small cells; a hidden fair switch makes one cell of each pair ID (above the axis) and the other OOD (below). At the time n_k chosen for level k, the level carries so little ID mass that each pair is unsampled with probability at least $\frac{3}{4}$ (question marks): for such a pair the sample is independent of the switch, and every rule errs on half of the pair's mass. The levels accumulate only at 0, so every domain of the family is measure-separated, and a limsup argument over the random switches selects one domain that is hard at infinitely many of the times n_k.

and the risks gives

$$\mathbb{E}_{D_*}\mathcal{E}_{D_*}(A(S_{n_k})) \geq \frac{1}{19n_k} - \frac{4\cdot 10^{-4}}{n_k} - \frac{4\cdot 10^{-4}+10^{-3}}{n_k} \geq \frac{1}{25n_k}$$

at every constructed time.

F.3 No rate: proof of Theorem 5.3

For each $k \geq 1$ let $\Lambda_k = (2^{-k-2}, 2^{-k-1})$. A domain of the family is specified by masses $\varepsilon_k > 0$ with $\sum_k \varepsilon_k \leq \frac{1}{8}$, integers $m_k \geq 1$, and m_k pairs of disjoint nondegenerate closed cells in Λ_k, all disjoint within a level. A switch $\sigma_{k,i} \in \{1,2\}$ designates one cell of each pair as ID and the other as OOD. The ID marginal places mass ε_k/m_k uniformly on each ID cell and its remaining mass uniformly on $[\frac{7}{8}, 1]$, with all labels 1; the OOD law places the same mass uniformly on each OOD cell and its remaining mass at $\frac{13}{16}$. Cells of different levels accumulate only at 0, so every OOD law gives zero mass to the ID support; the family is universally learnable by Theorem 5.1, and the detector accepting the ID support has both risks zero, so excesses equal risks.

Fix a learner and a candidate rate, and replace the rate by a positive nonincreasing envelope R with $R(t) \to 0$, which only strengthens the target. Choose increasing times n_k with

$$\varepsilon_k = 16\cdot 2^k R(2^{-k}n_k) \leq 8^{-k-1}, \qquad m_k = \lceil 4n_k\varepsilon_k \rceil;$$

the masses then sum to at most $\frac{1}{56}$. Fix admissible cells and draw all switches independently and fairly; this randomness only selects a hard domain, and the learner receives the usual ID sample. For a rule h let $L_k(h)$ be the sum of its ID and OOD errors on the cells of level k, and $Z_k(\sigma) = \mathbb{E}_{S_{n_k}} L_k(A(S_{n_k}))/\varepsilon_k \in [0,2]$, the expectation including the learner's randomization. A given pair receives no sample with probability $(1-\varepsilon_k/m_k)^{n_k} \geq 1 - n_k\varepsilon_k/m_k \geq \frac{3}{4}$, and on that event the sample is independent of the pair's switch. For a fixed rule with acceptance fractions r_1, r_2 on the two cells, the switch-averaged contribution of the pair is $\frac{\varepsilon_k}{2m_k}((1-r_1)+r_2+(1-r_2)+r_1) = \varepsilon_k/m_k$. Summing over pairs gives $\mathbb{E}[Z_k \mid \sigma_{-k}] \geq \frac{3}{4}$, and since $Z_k \leq 2$, $\mathbb{P}(Z_k \geq \frac{1}{4} \mid \sigma_{-k}) \geq \frac{1}{4}$. For $G_k = \{Z_k \geq \frac{1}{4}\}$, Fact B.6(iii) gives $\mathbb{P}(\limsup_k G_k) \geq \frac{1}{4}$ without any independence. Choose σ_* in this limsup and let D_* be its domain. For infinitely many k,

$$\mathbb{E}\mathcal{E}_{D_*}(A(S_{n_k})) \geq \tfrac{1}{2}\mathbb{E}L_k(A(S_{n_k})) \geq \varepsilon_k/8 = 2\cdot 2^k R(2^{-k}n_k).$$

Given $C, c > 0$, all large k have $2^k \geq C$ and $2^{-k} \leq c$, so by monotonicity the right side exceeds $CR(cn_k)$ at infinitely many of these times. Embedding the construction on a coordinate axis gives every dimension $d \geq 1$.

F.4 Nearest-neighbour ranking

For independent $X \sim P_X$ and $Z \sim Q$ and a score r, let $\mathrm{AUC}(r; D) = \mathbb{P}(r(X) > r(Z)) + \frac{1}{2}\mathbb{P}(r(X) = r(Z))$.

Theorem F.1 (Nearest-neighbour ranking). *On a separable metric space with $Q(\operatorname{supp} P_X) = 0$, let $r_n(x) = -\operatorname{dist}(x, \{X_1, \ldots, X_n\})$. The optimal measurable AUC is 1, $\mathbb{E}[1 - \mathrm{AUC}(r_n; D)] \le \mathbb{E}_{X,Z}\big(1 - P_X(B(X, \operatorname{dist}(Z, \operatorname{supp} P_X)))\big)^n \to 0$, and $\mathrm{AUC}(r_n; D) \to 1$ almost surely along one training sequence.*

Proof. Optimal AUC and the expected bound. The score $-\operatorname{dist}(\cdot, \operatorname{supp} P_X)$ attains AUC 1, since $\operatorname{dist}(Z, \operatorname{supp} P_X) > 0$ almost surely under Q and $\operatorname{dist}(X, \operatorname{supp} P_X) = 0$ almost surely under P_X. Put $g(Z) = \operatorname{dist}(Z, \operatorname{supp} P_X)$. Every sample point lies in $\operatorname{supp} P_X$ almost surely, so $\operatorname{dist}(Z, \{X_1, \ldots, X_n\}) \ge g(Z)$, and a ranking failure or tie requires $\operatorname{dist}(X, \{X_1, \ldots, X_n\}) \ge g(Z)$, an event of conditional probability $(1 - P_X(B(X, g(Z))))^n$ given (X, Z). Every positive-radius ball around a support point has positive mass, so dominated convergence gives $\mathbb{E}[1 - \mathrm{AUC}(r_n; D)] \to 0$.

Almost sure convergence. On one infinite i.i.d. sequence, let $S = \operatorname{supp} P_X$; separability gives $P_X(S) = 1$. For a countable base (U_j), every U_j of positive mass is hit almost surely, since avoiding it for n draws has probability $(1 - P_X(U_j))^n \to 0$. On the intersection Ω_P of these countably many events with the event that every sample point lies in S, the sample is dense in S: for $x \in S$ and $\varepsilon > 0$ a basic neighbourhood $x \in U_j \subset B(x, \varepsilon)$ has positive mass and contains a sample point. Hence $\operatorname{dist}(x, \{X_1, \ldots, X_n\}) \to 0$ for every $x \in S$, while $\operatorname{dist}(z, \{X_1, \ldots, X_n\}) \ge \operatorname{dist}(z, S) > 0$ for every $z \notin S$. For each $(x, z) \in S \times S^c$, eventually $r_n(x) > r_n(z)$; dominated convergence for each training sequence in Ω_P, with $P_X \times Q$ giving full mass to $S \times S^c$, yields $\mathrm{AUC}(r_n; D) \to 1$. □

The event Ω_P depends only on the ID law, so one training sequence serves every hidden law with $Q(S) = 0$, though not uniformly; the theorem supplies neither a rate nor a certificate, and it concerns a fixed metric rather than a learned representation.

G Where the framework ends

Three examples mark the boundary of the theory: an uncountable fiber whose countable parts are all consistent while the whole is not; a diagonal fiber that separates mixed from deterministic compatibility and corrects a published necessity claim; and overlap, where a single domain is already an obstruction.

G.1 An uncountable fiber without a consistent schedule

For a fiber $\mathcal{C}$, zero lies in the product closure of its profiles iff every window has value zero, while a consistent schedule is a sequence converging to zero coordinatewise (Proposition C.1). For countable $\mathcal{C}$ the two coincide (Theorem 2.3); for uncountable fibers they need not.

Give $\mathcal{X} = \{z\} \sqcup [0, 1]$ the disjoint-union separable metric topology, use zero–one loss, and let $P = \delta_z \otimes \delta_1$. For each finite $F \subset [0, 1]$ let h_F accept z and $[0, 1] \setminus F$ and reject F, and parameterize these detectors by the disjoint union of the empty tuple and the Borel sets $\{(x_1, \ldots, x_k) \in [0, 1]^k : x_1 < \cdots < x_k\}$, $k \ge 1$. The incidence relation $x \in F$ is Borel, so evaluation is jointly measurable for all Borel laws on the parameter space. Let the fiber contain every OOD law Q on $[0, 1]$. If $a(Q)$ is the total mass of the atoms of Q, then

$$\inf_{F \text{ finite}} R^{\mathrm{out}}_Q(h_F) = 1 - a(Q), \qquad \mathcal{E}_{(P,Q)}(h_F) = \sum_{x \text{ an atom of } Q,\ x \notin F} Q\{x\},$$

and the ID excess is zero. An atomless Q has zero excess for every h_F despite absolute OOD error one; excess is relative to the class.

Every window is compatible: for finitely many domains and $\varepsilon > 0$, reject finitely many atoms carrying all but ε of each domain's atomic mass. Every countable subfamily is consistent: the union

of its atoms is countable, and rejecting its first n elements (repeating a finite set if the union is finite) is a consistent schedule. The whole fiber is not. For any schedule, write F_n for its random finite rejection set; by Tonelli, $\int_0^1 \mathbb{P}(x \in F_n)\,dx = \mathbb{E}\,\mathrm{Leb}(F_n) = 0$, so for each n, $\mathbb{P}(x \in F_n) = 0$ for Lebesgue-almost every x. Intersecting these countably many full-measure sets gives a point x_* with $\mathbb{P}(x_* \in F_n) = 0$ for all n, and the domain (P, δ_{x_*}) has expected excess one at every time. All domains have $Q(\mathrm{supp}\, P_X) = 0$, and unrestricted detectors would solve them by accepting only z: the obstruction is sequential reach within this class, not uncountability alone.

G.2 Deterministic compatibility and a published necessity claim

Take $\mathcal{X} = \{-2, -1\} \cup \{1/j : j \geq 1\} \subset \mathbb{R}$, $a_0 = -2$, $a_1 = -1$, $x_j = 1/j$, the ID law $P = \frac{1}{2}\delta_{(a_0,1)} + \frac{1}{2}\delta_{(a_1,1)}$, the domains $D_j = (P, \delta_{x_j})$, zero–one loss, and $\mathcal{H} = \{h_k : k \geq 1\}$, where h_k accepts a_0, a_1, x_k and rejects every other x_j. Every domain has disjoint ID and OOD supports, both separate optimal risks are zero, and

$$\mathcal{E}_{D_j}(h_k) = \mathbf{1}\{j = k\}.$$

No deterministic detector is within any $\varepsilon < 1$ of both optima on the whole fiber. Yet the fiber has value zero, since the uniform law on $h_1, \ldots, h_m$ has worst-case excess $1/m$, and a deterministic learner attains it: reading the fair bits $B_i = \mathbf{1}\{X_i = a_1\}$ from the sample and outputting h_{K_n} with $K_n = 1 + \sum_{i=1}^{n} B_i 2^{i-1}$, uniform on $\{1, \ldots, 2^n\}$, gives $\sup_j \mathbb{E}\mathcal{E}_{D_j}(h_{K_n}) = 2^{-n}$.

Fang et al. (2024, Condition 4 and Theorem 13, p. 14) formulate a deterministic compatibility condition for finite-ID spaces and state that it is necessary for learnability. The example has one ID law, a bounded instance space, a countable measurable class, and zero–one loss. For every test prior $\alpha \in [0, 1)$ the prior mixture $(1 - \alpha)P + \alpha(\delta_{x_j} \otimes \delta_2)$ has zero optimal risk and expected risk at most $\alpha 2^{-n}$, so the example also meets their prior-unknown formulation. Condition 4 fails for every tolerance below one, because its OOD component fails on D_k for the chosen h_k. The necessity proof in their Appendix K.1 first obtains expected bounds separately for every domain sharing an ID law, of the form $\forall D\ \mathbb{E}_S L_D(S) \leq \varepsilon$, and then infers one sample realization good for all of them, $\exists S\ \forall D\ L_D(S) \leq \varepsilon$; the example shows that this exchange fails, since every sample selects an index whose domain has loss one. The repair is local: with randomized learners, uniform learnability over finitely many ID laws is equivalent to $V = 0$ on every fiber (Proposition E.1), and the counterexample itself uses a deterministic learner, so it does not rely on the change of convention. Randomization cannot always be extracted from the sample: with a single known ID point and the same diagonal profile, the value is still zero, but a deterministic learner returns one index at each n and has worst-case excess one at every time.

G.3 Overlap and genuine conflicts

The overlap floor. Let $K = 1$, zero–one loss, and $\mathcal{H} = \mathcal{H}_{\mathrm{all}}$; the two constant detectors show that both separate optimal risks are zero. For a measurable acceptance set A and densities p, q with respect to a common dominating measure λ,

$$R_P^{\mathrm{in}}(h) + R_Q^{\mathrm{out}}(h) = P_X(A^c) + Q(A) = \int [p\mathbf{1}_{A^c} + q\mathbf{1}_A]\,d\lambda \geq \int \min(p, q)\,d\lambda = 1 - \mathrm{TV}(P_X, Q).$$

Half this sum bounds the maximum, and averaging over detectors gives $\widehat{\mathcal{E}}_D(\mu) \geq \frac{1}{2}(1 - \mathrm{TV}(P_X, Q))$. If $P_X \perp Q$, a Borel set A with $P_X(A) = 1$ and $Q(A) = 0$ gives an exact detector; otherwise positive overlap prevents convergence even for a single domain. The floor is a lower bound, not the value of every domain.

Two individually solvable domains in conflict. Take $\mathcal{X} = \{-1, 0, 1\}$, $P = \delta_0 \otimes \delta_1$, $D_- = (P, \delta_{-1})$, $D_+ = (P, \delta_1)$, and two detectors: h_- accepts $\{0, 1\}$ and h_+ accepts $\{-1, 0\}$. Both have zero ID risk and each domain has a perfect detector, but the profiles are $(0, 1)$ and $(1, 0)$, so every law has maximal excess at least $\frac{1}{2}$ and the fair mixture attains it. The pair is a window of value $\frac{1}{2}$ with no obstructed singleton, and the optimal OOD risk of the fair mixture of δ_{-1} and δ_1 is $\frac{1}{2}$ although each component has optimum zero. Accepting only 0 would solve both, so the conflict is a property of the class.

Ranking conflicts. Let $P_X = \frac{1}{2}(\delta_0 + \delta_1)$ and compare $D_0 = (P, \delta_0)$ with $D_1 = (P, \delta_1)$. For a ranking r put $s(r) = \mathbf{1}\{r(1) > r(0)\} + \frac{1}{2}\mathbf{1}\{r(1) = r(0)\}$; then $\mathrm{AUC}(r; D_0) = \frac{1}{4} + s(r)/2$ and

$\mathrm{AUC}(r; D_1) = \frac{1}{4} + (1 - s(r))/2$. Each optimum is $\frac{3}{4}$, so the two AUC excesses sum to $\frac{1}{2}$ for every ranking and every randomized ranking, and by decoupling every learner has one of these fixed domains with limsup expected excess at least $\frac{1}{4}$. Overlap alone is not a singleton AUC obstruction: if $P_X = Q$ every ranking has AUC $\frac{1}{2}$ and excess zero, and for any single known domain a sequence approaching the optimal AUC is a data-free consistent learner. The strong-risk overlap floor therefore does not transfer literally to AUC.

H Related work in detail

Uniform OOD learnability. Fang et al. (2022; 2024) formulated OOD learning from ID-only data with class-relative excess risk, proved uniform impossibility results under overlap and for restricted classes, and identified positive geometric settings, for risk and for AUC. Their equivalence relation between domains with the same ID law is the fiber of this paper, and their finite-mixture linearity conditions are the ancestors of mixture completion. We keep their observation structure and add three things: the reduction of every learner on a fiber to a schedule, which turns rates into a question about sequences of laws on detectors; the completion and affinity results, which turn mixing into a structural operation on a fiber; and fixed-domain lower bounds. The correction in Appendix G.2 concerns the necessity direction of one condition and leaves their positive results untouched.

Adaptive stopping. Garov & Chaudhuri (2025) allow an almost surely finite, data-dependent sample count with a guarantee at a fixed known mixture weight, and show that no such certificate exists on all disjoint-support domains, including a convex-ID version for $d \geq 2$. Strong excess dominates fixed-weight excess, so these impossibilities transfer to our certificate. In their binary, zero–one, unrestricted, separated setting both optima are zero and $\max\{R_P^{\mathrm{in}}(h), R_Q^{\mathrm{out}}(h)\} \leq 2R_D^{1/2}(h)$, so a fixed-weight certificate at tolerance $\varepsilon/2$ yields the strong one at ε; their family with a known positive support margin is therefore certifiable but not uniformly learnable. Theorem 2.4 explains why such a separation needs infinitely many ID laws.

Universal rates and individual lower bounds. Classical consistency theory separates convergence from distribution-free rates (Stone, 1977; Devroye, 1982; Devroye et al., 1996). The universal-rate program of Bousquet et al. (2021) classifies realizable binary learning with distribution-dependent constants and time rescaling; Hanneke et al. (2023) treat multiclass learning and Hanneke & Moran (2026) the agnostic case. We adopt their rate convention without importing their rate categories, since the OOD experiment observes only one of the two populations. A minimax lower bound may use a different hard distribution at each n; an individual lower bound keeps one along infinitely many sample sizes (Antos & Lugosi, 1998). Every lower bound here is individual. Within a fiber the witness is a mixture that condenses blind spots; across fibers it is a limit of nested ID laws (Theorem 5.2) or a generic switch sequence (Theorem 5.3).

The analysis behind the main theorem. Banach & Steinhaus (1927) published the uniform boundedness principle as a principle of condensation of singularities. Deutsch & Hundal (2010) proved its quantitative form: pointwise convergence of bounded linear operators that is not uniform is almost arbitrarily slow. Appendix D.7 explains how decoupling, completion, and affinity place OOD rates in this setting, and Appendix E.3 how Baire category governs selection over uncountably many ID laws, in the spirit of the discernibility results of Cover (1973); Dembo & Peres (1994).

Support estimation, novelty detection, and open-set recognition. Devroye & Wise (1980, Theorem 1) analyze a detector built from shrinking neighbourhoods of the sample; applied with P_X as the training measure and $(P_X + Q)/2$ as the test measure, their result gives binary consistency whenever $Q(\mathrm{supp}\, P_X) = 0$, and Theorem 5.1 extends it to finite label sets and bounded losses. Cuevas & Fraiman (1997) study plug-in support estimators, Schölkopf et al. (2001) estimate a region containing a prescribed fraction of the ID law, and Steinwart et al. (2005) formulate anomaly detection as density-level-set classification; these targets control the ID side of the error, while the strong excess also asks about every admissible hidden law. Scheirer et al. (2013) formalize open-set recognition, where classes absent from training appear at test time, through a constrained recognition objective; controlling an accepted region geometrically and controlling its mass under every admissible hidden law require different assumptions.

Neural scores and likelihood. Maximum-softmax (Hendrycks & Gimpel, 2017), ODIN (Liang et al., 2018), Mahalanobis (Lee et al., 2018), energy (Liu et al., 2020), and deep nearest-neighbour (Sun

et al., 2022) scores turn a trained network into a detector, with representation, score, and threshold as separate design choices. Our results concern existence and impossibility over specified classes and domain families and do not rank these scores; Theorem F.1 concerns a fixed metric, not a learned representation. Nalisnick et al. (2019) observed that deep generative models can assign higher likelihood to OOD data, and Zhang et al. (2021, Proposition 1) construct alternatives with the same score distribution, an obstruction for a chosen statistic even with the true ID law. Observation fibers instead concern indistinguishability of the training experiment; both views require an explicit family of hidden alternatives.

Additional observations. Ben-David et al. (2010) prove impossibility theorems for domain adaptation with unlabeled target data, whose indistinguishability arguments are close in spirit to fibers; that target sample is absent from our experiment. Semi-supervised novelty detection observes a contaminated unlabeled sample (Blanchard et al., 2010), and Outlier Exposure trains against auxiliary outliers (Hendrycks et al., 2019). Observations whose law depends on Q refine the fibers, while an auxiliary law fixed independently of Q leaves them unchanged; in both cases the analysis of Section 2 applies to the resulting fibers.

Conformal methods. Bates et al. (2023) construct conformal outlier p-values with marginal and calibration-conditional false-alarm guarantees, and Tibshirani et al. (2019) extend conformal prediction to covariate shift with a known likelihood ratio. These are finite-sample controls of one error type or of coverage; the strong-excess certificate also bounds missed detections relative to the best detector in a class, so the collapse of Theorem 2.4 leaves conformal false-alarm control intact.

Benchmarks. Yang et al. (2021) organize anomaly, novelty, open-set, and OOD detection in one taxonomy, and OpenOOD (Yang et al., 2022) provides a common benchmark. A finite suite of test distributions compares concrete methods; a universal-rate lower bound quantifies over a whole family and keeps one witness along infinitely many sample sizes. The point-mass fiber and its completion have the same ID observations and the same class but different rate behaviour, which is why benchmark performance, consistency, and rates are complementary forms of evidence.

I Open problems and directions from applications

Every OOD guarantee in this paper is paid for in one of three currencies: time within a fiber, which buys consistency and, once one rate is attained, every rate (Proposition 2.2); selection across ID laws, which buys the nontrivial rates of Sections 4 and 5; and, when only ID data are observed, an assumption about the hidden family, the only currency that buys a rate against hidden laws that can be mixed (Theorem 3.3). Applications differ in the hidden family they face, the extra observations they have, and the criterion they score. Each subsection reads one application through fibers, states what the paper already implies, and ends with an open problem; Table 5 collects them.

I.1 Deployments are mixtures

A benchmark of distribution shifts spans hospitals, camera traps and locations (Koh et al., 2021); OOD benchmarks score a detector on several OOD test sets (Yang et al., 2022); and detecting the long tail of rare skin conditions has been posed as OOD detection (Guha Roy et al., 2022). In each case the deployed hidden law is a mixture $\sum_k w_k Q_k$ of sources with unknown weights, and when every mixture of admissible sources is admissible, the fiber is mixture-complete.

What the theory implies. On a mixture-complete fiber a learner attains a rate only if it attains it uniformly (Theorem 3.3), for the strong excess, the excess at a fixed OOD proportion, and AUC (Proposition D.1). A learning curve measured on each fixed OOD set, however fast, transfers to mixtures of the sets only if it holds uniformly over them, which requires $V(\mathcal{C}) = 0$.

Observation (a condensed test law). Let $h_{n_1}, h_{n_2}, \dots$ be detectors trained at sizes $n_1 < n_2 < \cdots$, let Q_k be the available OOD test law on which h_{n_k} misses most, and put $Q_* = \sum_k 2^{-k} Q_k$. The raw OOD risk, the miss rate under zero–one loss, is linear in the hidden law, so for every k

$$R^{\mathrm{out}}_{Q_*}(h_{n_k}) = \textstyle\sum_j 2^{-j} R^{\mathrm{out}}_{Q_j}(h_{n_k}) \;\geq\; 2^{-k}\, R^{\mathrm{out}}_{Q_k}(h_{n_k}).$$

One test law thus exposes the blind spot at every chosen size, discounted by 2^{-k}, and needs no new data, being a weighted average of per-set miss rates. With finitely many sets its curve is dominated

Table 5: Applications read through observation fibers: what the learner never sees, what else it observes, what the paper already implies, and the open problem each raises.

Application	What stays hidden	Extra observation	What the theory implies	Open problem
pooled deployments, OOD benchmarks	mixture weights of the sources	none	a rate on $\sigma\mathcal{C}$ is a uniform rate on $\mathcal{C}$ (Theorem 3.3); condensed test law	Open problem I.1
near OOD in medicine, driving	hidden mass near the ID support	none, or an audit sample	margin: exact; envelope ψ: $\psi(1/n)$; accumulation: no rate (Proposition D.2)	Open problem I.2
pretrained and foundation models	the deployed Q	auxiliary outliers; wild sample	pretraining changes $\mathcal{H}$ and V; auxiliary data keep the fibers, wild data refine them	Open problem I.3
security, fraud, jailbreaks	Q, chosen after the learner	none	excess at least $V(\mathcal{C}) - \varepsilon$ at every n, or $V^{\mathrm{P}}(\mathcal{C}) - \varepsilon$ if the deployed detector is seen	Open problem I.4
model-agnostic searches	the signal law	two mixed samples	with background data alone and $V > 0$, no rate without an assumption on the signal	Open problem I.5
continuous environments	the environment θ and its Q	the ID sample selects θ	gluing can fail (Proposition 4.3)	Open problem I.6

by the worst case over the sets at each size, which is the curve to report; the point of Q_* is that one fixed test law already carries the slow decay. With finitely many sizes, renormalizing the weights only increases them. Along a schedule consistent on every Q_k, the same inequality holds for the class-relative excess (Lemma 3.2) and for AUC (Proposition D.1). Like the witness in the proof of Theorem 3.3, Q_* is chosen after the detectors; it is a stress test, not an estimate of deployed error.

Open problem I.1 (Rates from incidence knowledge)**.** Incidence knowledge, such as epidemiological bounds on rare conditions, constrains the mixture weights. Let $\mathcal{C} = \{(P, Q_k) : k \in \mathbb{N}\}$ be a consistent fiber with $V(\mathcal{C}) > 0$, let W be a known set of probability vectors on $\mathbb{N}$, and let $\mathcal{C}_W = \{(P, \sum_k w_k Q_k) : w \in W\}$; the basic case is a tail envelope $W_\tau = \{w : \sum_{k>r} w_k \le \tau(r) \text{ for all } r\}$ with $\tau \downarrow 0$. Within one known fiber time is free (Proposition 2.2), so price the class instead: fix a filtration $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots$ with $\bigcup_m \mathcal{H}_m = \mathcal{H}$, for instance by model size, and let $V_m(\mathcal{C}_W)$ be the infimum of $\sup_{D \in \mathcal{C}_W} \widehat{\mathcal{E}}_D(\mu)$ over laws μ on $\mathcal{H}_m$. Determine the decay of $V_m(\mathcal{C}_W)$ as $m \to \infty$, up to constants, as a functional of W and of $(E_P^{\mathrm{in}}(h), (E_{Q_k}^{\mathrm{out}}(h))_{k\in\mathbb{N}})_{h\in\mathcal{H}_m}$; the strong excesses alone do not determine it, since the maximum in the strong excess does not commute with mixing. On the prefix fiber with $\mathcal{H}_m = \{h_1, \dots, h_m\}$, (4) gives h_m the worst-case excess $\sup_{w\in W_\tau} \sum_{k\ge m} w_k \le \tau(m-1)$, the tail bound of Section 6; general classes are open.

The prefix case is easy because its blind spots are nested, so by (4) the largest detector of $\mathcal{H}_m$ has the least excess on every hidden law; in general blind spots overlap, randomization changes values (Proposition C.3), and $V_m(\mathcal{C}_W)$ has no evident closed form.

I.2 Near OOD in safety-critical perception

Medical imaging has a public benchmark for OOD detection and localization (Zimmerer et al., 2022), dermatology must flag unseen conditions (Guha Roy et al., 2022), driving has a benchmark for anomaly segmentation in road scenes (Chan et al., 2021), and large pretrained models have been applied to near-OOD tasks (Fort et al., 2021). In fiber terms, near OOD is hidden mass close to a boundary of the ID support that the class cannot draw exactly; Appendix D.6 makes this a fiber with a continuous blind spot.

What the theory implies. Proposition D.2 sorts hidden families by the mass they may put in the blind spot. With a margin the schedule h_n is exact after finite time; an envelope $Q((1, 1+s]) \le \psi(s)$ gives it worst-case excess at most $\psi(1/n)$; without control, mixtures load the blind spot at every scale and no learner attains any rate. The three fibers share their ID law (Appendix D.6), so a near-OOD guarantee is a statement about hidden mass near the ID support, which ID data cannot certify. They

can certify the false-alarm side, as conformal outlier tests do (Bates et al., 2023), but a certificate for the strong excess exists on a fiber only when $V = 0$ (Theorem 2.4).

Open problem I.2 (Auditing the blind spot)**.** Let $\mathcal{X} = [0, 2]$, $K = 1$ and zero–one loss, let the ID laws be $P_\theta = \mathrm{Unif}[0, \theta] \otimes \delta_1$ for θ in a set $\Theta \subset (0, 1]$, and let the detectors accept $[0, t]$ with label 1, $t \in (0, 2]$. The hidden laws of P_θ live on $(\theta, 2]$ and are constrained by a margin, $Q((\theta, \theta + \tau]) = 0$, or by an envelope, $Q((\theta, \theta + s]) \le \psi(s)$ for all s. A learner observes $S_n \sim P_\theta^n$ and an audit sample $T_m \sim Q^m$ of the deployed hidden law, and returns a detector $\widehat{h}$ and a bound $\widehat{B}$; it certifies at level δ if $\mathbb{P}_D\{\mathcal{E}_D(\widehat{h}) > \widehat{B}\} \le \delta$ for every domain D. Determine, up to constants and as functions of (n, m), of τ or ψ, and of Θ, the minimax expected strong excess and the least attainable $\sup_D \mathbb{E}\widehat{B}$.

Audit samples depend on Q and refine the fibers. For a single known θ, Proposition D.2 settles the constrained fibers, whose value is zero; the unknown boundary is what couples the two samples, the ID sample locating θ and the audit seeing the hidden mass above it. Without the audit, the unconstrained family has worst-case expected excess one at every n (Lemma 2.1 and Proposition D.2).

I.3 Foundation models and observations that see the hidden law

Outlier exposure trains against auxiliary outliers (Hendrycks et al., 2019); other methods train on unlabeled wild data that mixes ID and OOD inputs (Katz-Samuels et al., 2022), with a theoretical analysis in Du et al. (2024); zero-shot detectors are built from vision-language representations (Ming et al., 2022); and OOD detection with large language models has been studied empirically (Liu et al., 2024).

Representations change the class. A pretrained representation fixed before deployment, with a score such as maximum softmax (Hendrycks & Gimpel, 2017) or energy (Liu et al., 2020), defines the detector class. It leaves the fibers unchanged and moves the benchmark and the value. It can remove a blind spot by supplying the missing limit: adjoining h_∞ to the prefix class gives $V(\sigma\mathcal{C}_{\mathrm{atom}}) = 0$ (Section 3), and adjoining the boundary detector does the same for near OOD (Proposition D.2). On a mixture-complete fiber its effect on rates is all or none (Theorem 3.3).

Auxiliary and wild data. Auxiliary outliers drawn from a law fixed independently of Q leave the fibers unchanged (Appendix H), so Theorem 3.3 applies as stated: whatever the auxiliary law, a mixture-complete fiber with $V > 0$ has no rate. A wild sample from $(1 - \pi)P_X + \pi Q$ depends on Q and refines the fibers, as in semi-supervised novelty detection (Blanchard et al., 2010) and unsupervised domain adaptation (Ben-David et al., 2010). Decoupling and Theorem 2.3 hold on the refined fibers, but mixtures of hidden laws leave them (Section 6), so condensation, as proved, no longer applies.

Observation (wild data on the prefix fiber). On the prefix fiber of Proposition 3.5, where ID data alone admit no rate on the completion, a wild sample removes the blind spot. Let J be the largest index j with $1/j$ among the wild draws, and use h_{J+1}, which rejects $1/1, \dots, 1/J$; its excess is $Q\{1/j : j > J\}$ (Equation (4)), and a fresh draw from Q has a larger index than N earlier draws with probability at most $1/(N + 1)$. With $N \sim \mathrm{Bin}(n, \pi)$ hidden draws, $\mathbb{E}[1/(N + 1)] \le 1/((n + 1)\pi)$, so every hidden law on $\{1/j : j \in \mathbb{N}\}$, mixtures included, has expected excess at most $1/((n + 1)\pi)$.

Open problem I.3 (Universal rates with wild data)**.** Fix one ID law P and let the learner observe $S_n \sim P^n$ and n wild draws from $(1 - \pi)P_X + \pi Q$, with $\pi \in (0, 1]$ unknown. Which fibers become learnable at which universal rates, with constants depending on (Q, π)? In particular, is every fiber that is consistent in this experiment learnable at some universal rate, or can the slowness move into estimating the hidden law from the wild sample and become statistical, as in Section 5? An answer classifies the attainable rates through the class and the hidden family, in the spirit of Bousquet et al. (2021).

A lower bound must condense blind spots into one hidden law that the learner now observes through the wild sample, so the witness of Theorem 3.3 changes the data it is meant to defeat.

I.4 Adaptive adversaries

In security monitoring, fraud detection and jailbreak detection, OOD inputs come from an attacker who can study the deployed system; OOD detection has been certified robust to bounded perturbations (Bitterwolf et al., 2020). The hidden law is then not fixed in advance.

Observation (the value is a floor). For every learner and every n, the output law μ_n is one law on $\mathcal{H}$ for the whole fiber (Lemma 2.1), and $V(\mathcal{C})$ is the infimum over such laws of the worst case: the upper value of the game in which the learner commits first. Hence $\sup_{D\in\mathcal{C}} \widehat{\mathcal{E}}_D(\mu_n) \geq V(\mathcal{C})$, and a hidden law chosen after μ_n is fixed forces expected excess above $V(\mathcal{C}) - \varepsilon$ at every time, for every $\varepsilon > 0$. An attacker who sees the deployed detector itself, not only its law, faces the pure value $V^{\mathrm{p}}(\mathcal{C}) = \inf_{h\in\mathcal{H}} \sup_{D\in\mathcal{C}} \mathcal{E}_D(h) \geq V(\mathcal{C})$ of Appendix C.3: randomization no longer helps, and on the diagonal fiber of Appendix G.2 the pure value is one although $V = 0$.

What the theory implies. Against adaptive attackers consistency buys nothing: a moving blind spot is exactly what they exploit. The target is $V = 0$ against an attacker who knows the learner, and $V^{\mathrm{p}} = 0$ against one who sees the deployed detector. A perturbation budget is an assumption about the hidden family of the kind the takeaway calls for: it bounds what the attacker can reach.

Open problem I.4 (Drifting hidden laws)**.** Let $\mathcal{C}$ be a fiber with hidden family $\mathcal{Q}$, d a metric on laws on $\mathcal{X}$, and $B : \mathbb{N} \to [0, \infty)$ nondecreasing. A drift with budget B is a sequence (Q_n) in $\mathcal{Q}$, fixed in advance, with $\sum_{k<n} d(Q_k, Q_{k+1}) \leq B(n)$ for all n; a schedule (μ_n) is consistent against it if $\widehat{\mathcal{E}}_{(P,Q_n)}(\mu_n) \to 0$. The budget $B \equiv 0$ gives the consistency of Section 2, decided window by window on countable fibers (Theorem 2.3); without a budget, a drift chosen after the learner forces excess above $V(\mathcal{C}) - \varepsilon$ at every time. For d the total variation the question reduces to known notions. The excess is $2M$-Lipschitz in total variation, uniformly over detectors, so a drift with bounded budget converges to a law, possibly outside the hidden family, at which every schedule consistent on $\mathcal{C}$ is consistent by the same bound, while an unbounded budget allows infinitely many jumps into the blind spot, which forces $V(\mathcal{C}) = 0$. For countable fibers and d a transport distance, determine the budgets for which some schedule is consistent against every drift with that budget, and the rates it attains: a dynamic analogue of Theorem 2.3, stated through the hidden laws reachable within the budget.

Transport distances are the interesting case because they price a move by the distance travelled: a hidden law can then drift into the blind spot only as fast as the budget allows.

I.5 Scientific discovery

Model-agnostic searches for new physics look for signals that depart from the background without a model of the signal; a community challenge benchmarks anomaly detectors for them on simulated collider events (Kasieczka et al., 2021), and classification without labels trains a classifier to distinguish mixed samples (Metodiev et al., 2017). The background is the ID law and the signal the hidden law.

What the theory implies. A model-agnostic search is the fiber of all signals, and a mixture of signals is a signal, so the fiber is mixture-complete. If it has a moving blind spot for the criterion scored, Theorem 3.3 and Proposition D.1 leave no rate for the risk, the excess at a fixed signal fraction, or AUC; with unrestricted detectors, a background law learned from data, and signals separated from it, Theorem 5.3 leaves none either. By the takeaway, a sensitivity rate must come from an assumption on the signal, such as localization in a known region, or from observations that see it. Two mixtures of background and signal in different proportions are such an observation.

Open problem I.5 (Rates from two mixtures)**.** Let the experiment observe n draws from each of $M_i = (1 - \pi_i)P_X + \pi_i Q$, $i = 1, 2$, with unknown $0 \leq \pi_1 < \pi_2 \leq 1$, where P is the background law and Q ranges over a signal family $\mathcal{Q}$ closed under countable mixtures. For which classes of detectors or rankings is $\{(P, Q) : Q \in \mathcal{Q}\}$ learnable at some universal rate in this experiment, for the strong excess or the AUC excess, and which rates occur? An answer characterizes the attainable rates through the class, $\mathcal{Q}$ and the gap $\pi_2 - \pi_1$, and reduces to the wild-data experiment of Open problem I.3 when $K = 1$ and $\pi_1 = 0$ is known.

Mixing signals changes both observed laws, so condensation must defeat a learner that reacts to the mixture, and unknown fractions add an identification problem.

I.6 Continuous families of environments

A driving system meets a continuum of weather and lighting conditions, and a clinical model a continuum of sites and scanners. Each environment θ has its own ID law P_θ and its own fiber, and the learner must select the fiber from the sample.

What the theory implies. Selection costs an exponentially small error over finitely many ID laws (Theorem 4.1) and a sub-exponential one over countably many (Theorem 4.2); over uncountably many it can defeat consistency even when every fiber is a single domain with an exact detector (Proposition 4.3), because a universal learner makes its fiberwise choice as a pointwise limit of continuous functions of the ID law (Section 4). Per-environment detectors with a selector therefore need not glue over a continuum.

Open problem I.6 (Regular selection)**.** Let Θ be a Polish space, $\theta \mapsto P_\theta$ injective and continuous for total variation, and every fiber $\mathcal{C}_\theta$ universally learnable. Find conditions on (P_θ) and on fiberwise consistent schedules (ν_n^θ), such as continuity of $\theta \mapsto \nu_n^\theta$ at each n, under which $\bigcup_\theta \mathcal{C}_\theta$ is universally learnable; σ-compactness of Θ alone does not suffice, since $\Theta = (0, 1)$ in Proposition 4.3. Every universal learner yields fiberwise consistent schedules that are n-Lipschitz in P_θ for total variation (Section 4). When $\mathcal{X}$ is finite this necessary condition is also sufficient: a plug-in learner that selects $\widehat{\theta}$ with $P_{\widehat{\theta}}$ close to the empirical law in total variation and runs $\nu^{\widehat{\theta}}$ at a slowly growing time $k(N)$ adds at most $M\, k(N)$ times the estimation error. Is it sufficient in general, where the empirical law does not converge in total variation? Borel dependence alone cannot suffice: the fiberwise choice in Proposition 4.3, h_2 at rational and h_3 at irrational t, is Borel. An answer is a criterion, in the spirit of that of Dembo & Peres (1994) for consistent testing, that reduces to Theorem 4.2 for countable Θ and excludes the family of Proposition 4.3. This expands the first open problem of Section 6.

The obstruction is category, not measure: the failure set in Proposition 4.3 may be Lebesgue-null. A criterion must therefore be topological, and unlike a test, a learner outputs a law on detectors rather than one bit.

I.7 From theorems to practice

- Name the hidden family that makes $V = 0$: a margin or an envelope near the boundary (Proposition D.2), a tail or incidence bound (Appendix D.4, Open problem I.1), or an audit (Open problem I.2). On a hidden family closed under mixing, any rate is a uniform rate, which requires $V = 0$ (Theorem 3.3).
- Separate the three costs: time within a fiber, which buys every rate or none (Proposition 2.2); selection across ID laws, exponential, sub-exponential, or failing (Theorems 4.1 and 4.2 and Proposition 4.3); and statistics, from an exact $1/n$ to none (Theorems 5.2 and 5.3).
- Report the worst case over OOD test sets at each training size, not per-set curves or their average: on hidden families closed under mixing, only a uniform curve is a rate (Theorem 3.3 and Appendix I.1).
- Read AUC and the excess at a fixed OOD proportion the same way: the mixing theorem holds for them verbatim (Proposition D.1).

J On the epigraph

Sextus Empiricus states the problem of induction as a dilemma (Sextus Empiricus, 1933, II.204). To establish a universal from particulars, one reviews either some of them or all of them. Reviewing some is insecure, because an omitted particular may contravene the universal; reviewing all is impossible, because the particulars are infinite and indefinite. In OOD detection the particulars are the hidden laws of a fiber and the universal is a guarantee that holds for all of them. The paper can be read as an exact account of the two horns and of what lies between them (Table 6).

Table 6: A dictionary between the dilemma of Sextus Empiricus (1933, II.204) and the results of the paper.

Sextus Empiricus	This paper
particulars	the hidden laws Q of a fiber $\mathcal{C}_P$
the universal	a guarantee that holds for every hidden law of the fiber
reviewing some	the output of a schedule at one time n (Lemma 2.1)
an omitted particular contravenes	the moving blind spot, $V(\mathcal{C}) > 0$ (Proposition 3.5)
reviewing all	uniform learnability, $V(\mathcal{C}) = 0$; for a finite window, compatibility $V_F = 0$
not considered	reviewing in order: consistency (Theorem 2.3), and its pace (Theorem 3.3)
infinite	countable fibers: every rate but not uniformity; mixing loses the rate (Corollary 3.4)
indefinite	uncountable families: consistency can fail (Appendix G.1 and Proposition 4.3)

Some. By decoupling (Lemma 2.1), a learner reviews the hidden laws of a fiber only through a schedule fixed in advance, and at each time it has reviewed some of them. The first horn is the value: if $V(\mathcal{C}) > 0$, then whatever the learner, at every time and for every $\varepsilon > 0$ some hidden law has excess above $V(\mathcal{C}) - \varepsilon$. On the prefix fiber (Proposition 3.5) the schedule h_{n+1} has reviewed the points $1, \ldots, n$ and the omitted points $j > n$ contravene, which is every row of Figure 1(b). The OOD learner is worse placed than the inductivist of Sextus: its sample comes from the ID law alone, so it never observes the particulars it reviews, and its review is blind as well as partial.

All. Reviewing all particulars at one time is uniform learnability, $V(\mathcal{C}) = 0$. A finite window F can be served at one time to any accuracy exactly when it is compatible, $V_F = 0$, and some finite windows are not (Appendix G.3); when every window is compatible, the impossibility of the second horn belongs to infinite families, as Sextus says. The prefix fiber has compatible windows and value one.

Between the horns. The dilemma omits a third option, which learning theory supplies: review the particulars one after another, so that each is reviewed eventually. This is consistency. On a countable fiber it is decided window by window (Theorem 2.3), and the prefix schedule has it although $V = 1$; the gap between the two notions is the gap between universal and uniform guarantees in learning theory (Bousquet et al., 2021). The third option raises a question of its own, how fast the review proceeds. On a countable fiber with compatible windows it can proceed at any pace (Corollary 3.4); once hidden laws can be mixed, it has no pace at all unless it is uniform: a learner attains a rate only if it attains that rate uniformly (Theorem 3.3). A countable mixture is a single particular that carries omitted particulars into infinitely many stages of the review, so the insecurity of the first horn returns inside one hidden law (Figure 1(c)).

Infinite and indefinite. The two words mark two different obstacles. A countably infinite family with compatible windows can still be reviewed in order, at any pace (Corollary 3.4); what infinity costs is uniformity, and what mixing costs is the rate. An uncountable family cannot be enumerated, and then consistency itself can fail: every countable subfamily of the fiber of all OOD laws on $[0, 1]$ in Appendix G.1 has a consistent schedule, but the fiber has none, and over uncountably many ID laws selection can defeat consistency (Proposition 4.3). In the terms of this paper, the particulars being infinite costs uniformity, their being mixed costs the rate, and their being indefinite can cost consistency itself.

What remains. The takeaway of the main text answers the dilemma in the only way the setting allows. Time secures each particular, the sample secures selection among ID laws, and a rate against hidden laws that can be mixed must be bought with an assumption about them, which no ID sample can check.