The Exponential KL Boundary in Robust PAC Learning

Pahan Dewasurendra
Johns Hopkins University

Abstract

Distributionally robust PAC learning under Cressie–Read balls has polynomial sample complexity for every fixed order above one. The Kullback–Leibler endpoint was left open because it amplifies rare events on a different scale. We determine that endpoint.

Let $\rho > 0$ be fixed. For binary classification with zero-one loss, both realizable and agnostic KL-robust learning have sample complexity

$$\Theta_\rho\left(\frac{e^{\rho/\epsilon}}{\epsilon}\big(d + \log(1/\delta)\big)\right)$$

samples, where d is the VC dimension. The realizable statement is sharp without logarithmic factors. It follows from an exact reduction to ordinary PAC learning at the nominal accuracy $A_\rho(\epsilon) \sim e^{-1}\epsilon e^{-\rho/\epsilon}$. The agnostic result requires a different argument. Relative-VC control of ERM loses a factor of order $1/\epsilon$ at this singular endpoint. We combine a recent small-error improper learner with ERM. A new inverse-geometry dichotomy shows that multiplicative comparator slack is harmless exactly where the ERM logarithm is costly, and conversely. Validation between the two candidates gives the same sharp rate as in the realizable case.

We also resolve the singular transition from Cressie–Read to KL. If the order is $k = 1 + \lambda\epsilon/\rho$, then the inverse nominal scale has exponent $(\rho/\epsilon)\log(1+\lambda)/\lambda$. This boundary layer interpolates continuously between KL and every near-KL polynomial regime, and shows that the limits $k \downarrow 1$ and $\epsilon \downarrow 0$ do not commute.

1 Introduction

Distributionally robust learning protects a predictor against changes from a nominal distribution P to any Q in a divergence ball (Ben-Tal et al., 2013; Duchi and Namkoong, 2021). For zero-one loss and an f-divergence ball, the robust risk depends on a classifier only through its ordinary error probability (Hu et al., 2018; Jiang and Guan, 2016). Robust ERM is therefore ordinary ERM. This qualitative equivalence does not imply equal sample complexity. Robustness can strongly amplify a small error in estimating an ordinary error probability.

This amplification has recently been characterized for the Cressie–Read family of orders $k > 1$. The χ^2 case was initiated by Zhou and Liu (2023). General robust-risk estimation bounds were developed by Duchi and Namkoong (2021) and Zhou and Liu (2026). Very recent work gives nearly tight VC sample complexities for every fixed $k > 1$ (Aigner-Horev et al., 2026). At fixed radius, the realizable dependence is $\epsilon^{-k/(k-1)}$. The agnostic exponent is the larger of 2 and $k/(k-1)$.

The KL divergence is the endpoint $k = 1$, but it is not obtained by substituting $k = 1$ into these rates. If an event has nominal probability p, a KL ball of fixed positive radius can inflate it to order $1/\log(1/p)$. Thus polynomially small nominal errors retain constant-order robust risk. The KL endpoint and its possibly non-polynomial PAC complexity were posed explicitly as an open problem by Aigner-Horev et al. (2026, Section 4.1).

Contributions. We solve the KL endpoint and identify the boundary layer that separates it from fixed Cressie–Read orders.

First, we define $A_\rho(q)$ as the smaller Bernoulli probability p satisfying $\mathrm{kl}(q\|p) = \rho$. The realizable KL-robust problem at accuracy ϵ is *exactly* the ordinary realizable PAC problem at accuracy $A_\rho(\epsilon)$. The inverse scale obeys

$$e^{-1}qe^{-\rho/q} \le A_\rho(q) \le qe^{-\rho/q}, \qquad A_\rho(q) \sim e^{-1}qe^{-\rho/q}. \tag{1}$$

Combining this identity with the optimal realizable PAC theorem of Hanneke (2016) gives the sharp exponential rate.

Second, we determine the agnostic complexity with-

Table 1: Fixed-radius dependence on accuracy, suppressing dimension, confidence, and logarithmic factors. Here $k^* = k/(k-1)$.

Uncertainty	Realizable	Agnostic
None	ϵ^{-1}	ϵ^{-2}
Cressie–Read, $k > 1$	ϵ^{-k^*}	$\epsilon^{-(k^* \vee 2)}$
KL, this work	$\epsilon^{-1}e^{\rho/\epsilon}$	$\epsilon^{-1}e^{\rho/\epsilon}$

out a polynomial-prefactor gap. Constant factors in robust accuracy matter exponentially. We therefore work with the full inverse increment of size ϵ. Its normalized square is $\Theta_\rho(A_\rho(\epsilon))$, but ERM alone introduces $\log(1/A_\rho(\epsilon)) = \Theta_\rho(1/\epsilon)$. We prove a stronger dichotomy for the inverse increments and combine two ordinary classifiers. One is the small-error improper learner of Asilis et al. (2025), and the other is ERM. The former wins when the comparator error is negligible relative to the needed increment. The latter wins everywhere else because the increment then absorbs its logarithm. A relative validation step preserves the rate. The realizable subclass supplies the matching lower bound.

Third, we study an order k that approaches one with the target accuracy. For $k = 1 + \lambda\epsilon/\rho$ and fixed $\lambda > 0$, the inverse scale is

$$e^{-1}\epsilon \exp\left\{-\frac{\rho}{\epsilon}\frac{\log(1+\lambda)}{\lambda}\right\}(1 + o(1)). \tag{2}$$

The continuous endpoint $\lambda = 0$ is (1). This gives a phase diagram between polynomial and exponential robust learning.

2 Problem Setup

Let $\mathcal{H}$ be a class of measurable classifiers $h : \mathcal{X} \to \{-1, 1\}$. For a distribution P on $\mathcal{X} \times \{-1, 1\}$, write

$$p_h = \mathrm{err}_P(h) = P(h(X) \neq Y).$$

The KL uncertainty set and robust zero-one risk are

$$\begin{aligned} \mathcal{U}_\rho(P) &= \{Q \ll P : D_{\mathrm{KL}}(Q\|P) \leq \rho\}, \\ R_\rho(h, P) &= \sup_{Q \in \mathcal{U}_\rho(P)} Q(h(X) \neq Y). \end{aligned} \tag{3}$$

The direction $D_{\mathrm{KL}}(Q\|P)$ allows adversarial reweighting without placing mass outside the support of P.

We use the distribution-free realizable and agnostic sample complexities from Aigner-Horev et al. (2026). In the realizable case, some $h^* \in \mathcal{H}$ has $p_{h^*} = 0$, and the learner must return g with $R_\rho(g, P) \leq \epsilon$ with probability at least $1 - \delta$. In the agnostic case, the requirement is

$$R_\rho(g, P) - \inf_{h \in \mathcal{H}} R_\rho(h, P) \leq \epsilon.$$

The output may be improper unless stated otherwise. We denote the two minimax sample complexities by $M_{\mathrm{KL}}^{\mathrm{real}}$ and $M_{\mathrm{KL}}^{\mathrm{agn}}$.

For $p, q \in [0, 1]$, let

$$\mathrm{kl}(q\|p) = q \log \frac{q}{p} + (1 - q) \log \frac{1 - q}{1 - p}$$

with the usual endpoint conventions.

3 Event Inflation and Its Inverse

The next proposition records the binary reduction and fixes the notation used throughout the paper.

Proposition 1 (KL event inflation)**.** *For $p \in [0, 1]$, define*

$$Q_\rho(p) = \max\{q \in [p, 1] : \mathrm{kl}(q\|p) \leq \rho\}. \tag{4}$$

Then $R_\rho(h, P) = Q_\rho(p_h)$. The map Q_ρ is continuous and increasing, is strictly increasing on $[0, e^{-\rho})$, and equals one on $[e^{-\rho}, 1]$.

The data-processing inequality gives the upper bound in the proposition. Equality is attained by a likelihood ratio that is constant on the error event and its complement. This reduction is known for general f-divergences (Jiang and Guan, 2016; Hu et al., 2018). We include a proof in the supplement.

Let $A_\rho : [0, 1] \to [0, e^{-\rho}]$ be the inverse lower branch, defined by

$$A_\rho(0) = 0, \qquad \mathrm{kl}(q\|A_\rho(q)) = \rho \quad (0 < q \leq 1). \tag{5}$$

Thus $Q_\rho(A_\rho(q)) = q$. This inverse is the nominal error that can be inflated to robust error q.

Lemma 2 (Rare-event scale)**.** *For every $q \in (0, 1)$,*

$$e^{-1}qe^{-\rho/q} \leq A_\rho(q) \leq qe^{-\rho/q}. \tag{6}$$

For fixed $\rho > 0$, as $q \downarrow 0$,

$$A_\rho(q) = e^{-1}qe^{-\rho/q}(1 + o(1)). \tag{7}$$

To see the sandwich, set $p = A_\rho(q)$. The second term in $\mathrm{kl}(q\|p)$ lies between $-q$ and zero. Hence $\rho \leq q \log(q/p) \leq \rho + q$. The asymptotic follows by expanding that second term as $-q + o(q)$.

4 Exact Realizable Complexity

Monotonicity turns the inverse event scale into an exact experiment-level reduction, not only a rate comparison.

Theorem 3 (Exact realizable reduction)**.** *For every hypothesis class $\mathcal{H}$ and every $\epsilon, \delta \in (0,1)$,*

$$M_{\mathrm{KL}}^{\mathrm{real}}(\epsilon, \delta, \mathcal{H}) = M_{\mathrm{PAC}}^{\mathrm{real}}(A_\rho(\epsilon), \delta, \mathcal{H}). \tag{8}$$

If $\mathrm{VC}(\mathcal{H}) = d < \infty$ and $|\mathcal{H}| \geq 3$, then for sufficiently small ϵ and δ,

$$M_{\mathrm{KL}}^{\mathrm{real}}(\epsilon, \delta, \mathcal{H}) = \Theta\left(\frac{d + \log(1/\delta)}{A_\rho(\epsilon)}\right). \tag{9}$$

Consequently, for fixed $\rho > 0$,

$$M_{\mathrm{KL}}^{\mathrm{real}} = \Theta_\rho\left(\frac{e^{\rho/\epsilon}}{\epsilon}[d + \log(1/\delta)]\right). \tag{10}$$

Indeed, $Q_\rho(p_g) \leq \epsilon$ if and only if $p_g \leq A_\rho(\epsilon)$. Thus every learner satisfies one guarantee exactly when it satisfies the other. Equation (9) follows from the optimal classical PAC sample complexity (Hanneke, 2016), including its lower bound. Lemma 2 gives (10).

Theorem 3 shows that finite VC dimension still yields a finite sample bound for every fixed accuracy. It also shows that any definition of robust PAC learnability requiring polynomial dependence on $1/\epsilon$ fails at every fixed positive KL radius for nontrivial classes. This differs from all fixed Cressie–Read orders above one.

5 Sharp Agnostic Complexity

In agnostic learning, the comparator can have any ordinary error. Put

$$b_\epsilon = A_\rho(\epsilon), \qquad \Delta_\epsilon(q) = A_\rho(q + \epsilon) - A_\rho(q) \tag{11}$$

for $0 \leq q \leq 1 - \epsilon$. If the optimal robust risk is q, then $\Delta_\epsilon(q)$ is exactly the additional ordinary error that exhausts the robust excess budget.

Relative VC bounds control ordinary ERM through

$$p_{\widehat{h}} - p_* \lesssim \sqrt{p_* \gamma_n} + \gamma_n, \quad \gamma_n = \frac{d \log(2en/d) + \log(8/\delta)}{n}. \tag{12}$$

This is the standard relative VC inequality combined with empirical optimality (Boucheron et al., 2005; Aigner-Horev et al., 2026). The inverse-increment modulus relevant to this inequality is

$$\beta_\rho(\epsilon) = \inf_{0 \leq q \leq 1-\epsilon} \frac{\Delta_\epsilon(q)^2}{A_\rho(q + \epsilon)}. \tag{13}$$

Theorem 4 (KL inverse geometry)**.** *Fix $\rho > 0$. For all sufficiently small ϵ,*

$$\beta_\rho(\epsilon) = \Theta_\rho(b_\epsilon). \tag{14}$$

Moreover, for every fixed $\kappa > 0$,

$$\inf_{\substack{0 \leq q \leq 1-\epsilon: \\ A_\rho(q) \geq \kappa \Delta_\epsilon(q)}} \frac{\Delta_\epsilon(q)^2}{A_\rho(q) b_\epsilon \log(e/b_\epsilon)} \longrightarrow \infty. \tag{15}$$

The first statement says that the hardest local variance scale is the realizable endpoint $q = 0$. It would give an ERM bound with an extra factor $\log(e/b_\epsilon) = \Theta_\rho(1/\epsilon)$, so it does not by itself prove the sharp prefactor. The second statement removes this obstruction. It says that wherever the comparator error is not negligible relative to the required increment, the increment has enough slack to absorb the ERM logarithm.

Here is the proof geometry. When $q = O(\epsilon)$, the exponential sandwich (6) makes $A_\rho(q)/A_\rho(q + \epsilon)$ uniformly negligible. For the remaining small-q range, implicit differentiation gives

$$A'_\rho(q) = \frac{p(1-p)}{q-p} \log \frac{q(1-p)}{p(1-q)} \geq \frac{\rho p}{2q^2}, \tag{16}$$

where $p = A_\rho(q)$. Integration and (6) prove both (14) and the divergent ratio in (15). On a compact range away from zero, A'_ρ has a positive minimum while b_ϵ is smaller than every power of ϵ. Complete uniform details are in the supplement.

The other ingredient is recent small-error agnostic learning. A consequence of Asilis et al. (2025, Theorem 1.4) is the following. If $D_\delta = d + \log(1/\delta)$, there is an improper learner producing h_{sm} such that

$$p_{h_{\mathrm{sm}}} - p_* \leq a_0 p_* + C_0 \left(\sqrt{\frac{p_* D_\delta}{n}} + \frac{D_\delta}{n}\right) \tag{17}$$

with probability at least $1 - \delta$, for universal constants a_0, C_0. The multiplicative term is unavoidable in their small-error regime and is not generally an excess-risk guarantee. Earlier work gives exact-comparator bounds away from the smallest error scales (Hanneke et al., 2024). The KL dichotomy lets us use the two types of control where each is strongest.

Theorem 5 (Agnostic KL complexity)**.** *Fix $\rho > 0$ and let $\mathrm{VC}(\mathcal{H}) = d < \infty$. There are $C_\rho, \epsilon_\rho > 0$ such that*

$$M_{\mathrm{KL}}^{\mathrm{agn}}(\epsilon, \delta, \mathcal{H}) \leq C_\rho \frac{d + \log(1/\delta)}{A_\rho(\epsilon)} \tag{18}$$

for $0 < \epsilon \leq \epsilon_\rho$. If $|\mathcal{H}| \geq 3$, a matching lower bound holds. Consequently,

$$M_{\mathrm{KL}}^{\mathrm{agn}}(\epsilon, \delta, \mathcal{H}) = \Theta_\rho\left(\frac{e^{\rho/\epsilon}}{\epsilon}[d + \log(1/\delta)]\right). \tag{19}$$

For the upper bound, split the sample into three constant fractions. Train h_{sm} on the first, ordinary ERM h_{er} on the second, and use the third to choose the lower empirical-error candidate. Let $q = Q_\rho(p_*) \leq 1 - \epsilon$ and $\Delta = \Delta_\epsilon(q)$. At sample size $n \asymp_\rho D_\delta / b_\epsilon$, Theorem 4 controls every square-root term at scale Δ. If $p_* \leq \kappa\Delta$, then (17) controls h_{sm}. Otherwise (15)

makes the ERM quantity in (12) $o(\Delta^2/p_*)$, so h_{er} is controlled. Increasing the sample constant makes the better candidate's excess at most $\Delta/3$. Relative concentration over two fixed candidates makes validation cost another $\Delta/3$. Thus the selected error is below $A_\rho(q+\epsilon)$, which is the required robust guarantee. The case $q > 1-\epsilon$ is automatic.

The lower bound restricts to realizable distributions and invokes Theorem 3. Hence the exact agnostic rate is already forced by the zero-comparator endpoint, even though attaining it requires an improper small-error learner. This differs from fixed orders $k \geq 2$, where interior error scales can determine the agnostic rate (Aigner-Horev et al., 2026).

6 The Cressie–Read Boundary Layer

For $k > 1$, let

$$f_k(t) = \frac{t^k - kt + k - 1}{k(k-1)}$$

and let $A_{k,\rho}(q)$ be the lower nominal event probability whose robust inflation under an f_k ball is q. Binary data processing gives the defining equation

$$q^k p^{1-k} + (1-q)^k (1-p)^{1-k} - 1 = k(k-1)\rho, \quad (20)$$

with $p = A_{k,\rho}(q)$.

Theorem 6 (Near-KL boundary)**.** *Fix $\rho > 0$ and $\lambda > 0$. Let $k_q = 1 + \lambda q/\rho$. As $q \downarrow 0$,*

$$A_{k_q,\rho}(q) = e^{-1} q \exp\left\{-\frac{\rho}{q} h(\lambda)\right\}(1 + o(1)), \qquad h(\lambda) = \frac{\log(1+\lambda)}{\lambda}. \quad (21)$$

At the continuous endpoint $\lambda = 0$, define $k_q = 1$ and $h(0) = 1$. Then (21) remains valid by Lemma 2.

For the proof, put $k = 1+\eta$ and $p = A_{k,\rho}(q)$. Expanding the second term of (20), with $\eta = \lambda q/\rho$, yields

$$q^k p^{1-k} = q(1+\lambda)[1 + \eta + o(\eta)].$$

Since $q^k p^{1-k} = q \exp\{\eta \log(q/p)\}$, taking logarithms proves (21). The second-order term $1+\eta$ produces the prefactor e^{-1}.

For comparison, if $k > 1$ is fixed, (20) gives

$$A_{k,\rho}(q) \sim q^{k/(k-1)} [k(k-1)\rho]^{-1/(k-1)}. \quad (22)$$

This is the inverse form of the known polynomial realizable rate. The boundary parameter $\lambda = (k-1)\rho/q$ distinguishes the two limits. In particular,

$$\lim_{q \downarrow 0} \frac{q}{\rho} \log \frac{q}{A_{k,\rho}(q)} = \begin{cases} 1, & k = 1, \\ 0, & k > 1 \text{ fixed}, \end{cases} \quad (23)$$

while $k = 1 + \lambda q/\rho$ gives the full interpolation $h(\lambda)$. Thus the order-one endpoint is a singular boundary, rather than a fixed-order rate with a large exponent.

7 Discussion and Limitations

KL robustness makes rare classification errors extraordinarily expensive. At fixed radius, driving robust risk below ϵ requires a nominal error near $\epsilon e^{-\rho/\epsilon}$. The exponential sample law is therefore not caused by a loose robust-risk estimator. It follows from an exact reduction to ordinary PAC learning. The agnostic theorem shows that an interior comparator does not create a harder exponential scale.

The result also clarifies the role of the ambiguity model. A KL ball permits arbitrary reweighting of the joint input-label distribution on its support. This is intentionally conservative and may not represent structured covariate or label shifts. The exponential law is a property of that model, not a claim that every practical distribution shift requires exponentially many samples.

Our results are information-theoretic. ERM can be computationally hard for a general VC class. The paper treats zero-one loss and a fixed positive radius. When ρ decreases with ϵ, the inverse map itself gives the relevant interpolation, but a complete uniform phase diagram in (ρ, ϵ, k) is beyond our fixed-radius theorems. The agnostic upper bound is improper and uses the small-error learner of Asilis et al. (2025) as a black box. Determining whether a simpler learner attains the same sharp rate is open.

A Binary Event Inflation

We first prove the reduction from robust zero-one risk to a binary event.

Proposition 7 (Event inflation)**.** *Let E be a measurable event with $P(E) = p$. Then*

$$\sup_{Q \ll P : D_{\mathrm{KL}}(Q\|P) \le \rho} Q(E) = \max\{q \in [p, 1] : \mathrm{kl}(q\|p) \le \rho\}. \tag{24}$$

The right side is continuous and increasing in p, strictly increasing for $p < e^{-\rho}$, and equal to one for $p \ge e^{-\rho}$.

Proof. Apply the data-processing inequality for KL divergence to the measurable map $z \mapsto \mathbf{1}\{z \in E\}$. Every feasible Q with $Q(E) = q$ satisfies $\mathrm{kl}(q\|p) \le D_{\mathrm{KL}}(Q\|P) \le \rho$. This proves the upper bound.

Conversely, for $p, q \in (0, 1)$ define

$$\frac{dQ_q}{dP} = \frac{q}{p}\mathbf{1}_E + \frac{1-q}{1-p}\mathbf{1}_{E^c}.$$

Then $Q_q(E) = q$ and $D_{\mathrm{KL}}(Q_q\|P) = \mathrm{kl}(q\|p)$. Endpoint cases follow by limits. This proves equality.

For $p \in (0, 1)$, the function $q \mapsto \mathrm{kl}(q\|p)$ is continuous and strictly increasing on $[p, 1]$. Its value at one is $\log(1/p)$. Thus the upper root is one exactly when $p \ge e^{-\rho}$. Below this threshold, implicit differentiation of $\mathrm{kl}(q\|p) = \rho$ gives

$$\frac{dq}{dp} = \frac{q - p}{p(1-p)\log\{q(1-p)/[p(1-q)]\}} > 0. \tag{25}$$

Continuity at the threshold and endpoints follows from the defining binary constraint. □

The same proof applies to the error event of any classifier and establishes Proposition 1 in the main paper.

B The KL Inverse Scale

Recall that $A_\rho(q)$ is the unique $p \in (0, q)$ satisfying $\mathrm{kl}(q\|p) = \rho$, with endpoint values $A_\rho(0) = 0$ and $A_\rho(1) = e^{-\rho}$.

Lemma 8 (Exponential sandwich and asymptotic)**.** *For $q \in (0, 1)$,*

$$e^{-1} q e^{-\rho/q} \le A_\rho(q) \le q e^{-\rho/q}. \tag{26}$$

For fixed $\rho > 0$,

$$A_\rho(q) = e^{-1} q e^{-\rho/q}(1 + o(1)) \quad \text{as } q \downarrow 0. \tag{27}$$

Proof. Put $p = A_\rho(q)$. Since $0 < p < q < 1$,

$$(1-q)\log(1-q) \le (1-q)\log\frac{1-q}{1-p} \le 0.$$

The elementary inequality $\log(1-q) \ge -q/(1-q)$ bounds the left side by $-q$. Using $\mathrm{kl}(q\|p) = \rho$ gives

$$\rho \le q \log\frac{q}{p} \le \rho + q.$$

Solving these inequalities for p proves (26).

The sandwich implies $p/q \to 0$. Therefore

$$\begin{aligned}(1-q)\log\frac{1-q}{1-p} &= (1-q)\{-q + p + O(q^2 + p^2)\} \\ &= -q + o(q).\end{aligned}$$

Substitution into the binary constraint yields

$$\log\frac{q}{p} = \frac{\rho}{q} + 1 + o(1),$$

which proves (27). □

Implicit differentiation will also be useful.

Lemma 9 (Inverse derivative)**.** *The inverse A_ρ is continuously differentiable on $(0, 1)$ and, with $p = A_\rho(q)$,*

$$A'_\rho(q) = \frac{p(1-p)}{q-p}\log\frac{q(1-p)}{p(1-q)}. \tag{28}$$

If $q \le 1/2$, then

$$A'_\rho(q) \ge \frac{\rho p}{2q^2} \ge \frac{\rho}{2eq}e^{-\rho/q}. \tag{29}$$

Proof. For $F(q, p) = \mathrm{kl}(q\|p) - \rho$, direct differentiation gives

$$\begin{aligned}\partial_q F &= \log\frac{q(1-p)}{p(1-q)}, \\ \partial_p F &= \frac{p-q}{p(1-p)}.\end{aligned}$$

Both are nonzero on $0 < p < q < 1$, so the implicit function theorem proves (28). The logarithm in that equation is at least $\log(q/p)$, which is at least ρ/q by (26). Also $q - p \le q$ and $1 - p \ge 1/2$ when $q \le 1/2$. These facts prove the first lower bound. The second follows from the lower side of (26). □

C Realizable Reduction

Proof of Theorem 3 in the main paper. For every classifier g, Proposition 7 gives robust risk $Q_\rho(\mathrm{err}_P(g))$. Since Q_ρ is increasing and $Q_\rho(A_\rho(\epsilon)) = \epsilon$,

$$Q_\rho(\mathrm{err}_P(g)) \le \epsilon \iff \mathrm{err}_P(g) \le A_\rho(\epsilon).$$

This equivalence holds pointwise for every data set, every internal random seed of the learner, and every realizable distribution. Taking the infimum over sample sizes and learners proves the exact sample-complexity identity.

For a class of VC dimension d with at least three members, the optimal ordinary realizable sample complexity is

$$\Theta\left(\frac{d+\log(1/\delta)}{a}\right)$$

at ordinary accuracy a (Hanneke, 2016). Set $a = A_\rho(\epsilon)$ and apply Lemma 8. □

D A Monotone-Transform ERM Bound

We give a self-contained reduction from a standard relative VC inequality to the inverse-increment modulus. Let p_h and $\widehat{p}_h$ denote population and empirical error.

Lemma 10 (Localized ERM error)**.** *There is a universal C_0 such that the following holds. Let $\widehat{h}$ be an empirical risk minimizer over a class of VC dimension d, and let $p_* = \inf_{h\in\mathcal{H}} p_h$. With probability at least $1-\delta$,*

$$\begin{aligned} p_{\widehat{h}} - p_* &\le C_0\{\sqrt{p_*\gamma_n} + \gamma_n\}, \\ \gamma_n &= \frac{d\log(2en/d) + \log(8/\delta)}{n}. \end{aligned} \tag{30}$$

The statement holds for an approximate population minimizer by taking a limit if the infimum is not attained.

Proof. A standard relative VC bound states that, after increasing a universal constant, simultaneously for all $h \in \mathcal{H}$,

$$|p_h - \widehat{p}_h| \le 2\sqrt{p_h\gamma_n} + 2\gamma_n. \tag{31}$$

See, for example, Boucheron et al. (2005). Choose h_* with p_{h_*} arbitrarily close to p_*. Empirical optimality and (31) imply, with $x = p_{\widehat{h}}$,

$$x \le p_* + 2\sqrt{p_*\gamma_n} + 2\sqrt{x\gamma_n} + 4\gamma_n,$$

up to an arbitrarily small approximation term. Solving this quadratic inequality in $\sqrt{x}$ gives $x - p_* \le C_0(\sqrt{p_*\gamma_n} + \gamma_n)$ for a universal C_0. □

Theorem 11 (Inverse-increment bound)**.** *Let $Q : [0,1] \to [0,1]$ be continuous, satisfy $Q(0) = 0$, and be strictly increasing until a possible terminal plateau at one. Let A be its inverse lower branch. Assume a measurable empirical risk minimizer exists. Define*

$$\beta_Q(\epsilon) = \inf_{0\le q\le 1-\epsilon} \frac{[A(q+\epsilon) - A(q)]^2}{A(q+\epsilon)}.$$

There is a universal C such that ERM has transformed excess risk at most ϵ with probability at least $1-\delta$ if

$$n \ge \frac{C}{\beta_Q(\epsilon)}\left[d\log\frac{e}{\beta_Q(\epsilon)} + \log\frac{1}{\delta}\right]. \tag{32}$$

Proof. If $Q(p_*) > 1-\epsilon$, every classifier has transformed excess at most ϵ. Otherwise put $q = Q(p_*)$ and

$$\Delta = A(q+\epsilon) - A(q).$$

Then $p_* = A(q)$ and $p_* + \Delta = A(q+\epsilon)$. By definition,

$$\beta_Q(\epsilon) \le \frac{\Delta^2}{p_* + \Delta} \le \Delta. \tag{33}$$

Suppose $\gamma_n \le c\beta_Q(\epsilon)$, where $c > 0$ is a sufficiently small universal constant. Equations (30) and (33) give

$$\begin{aligned} p_{\widehat{h}} - p_* &\le C_0\left(\sqrt{cp_*\frac{\Delta^2}{p_*+\Delta}} + c\Delta\right) \\ &\le C_0(\sqrt{c} + c)\Delta \le \Delta. \end{aligned}$$

Monotonicity now gives $Q(p_{\widehat{h}}) - Q(p_*) \le \epsilon$.

It remains to make the sample condition explicit. A standard self-consistency calculation shows that (32), after increasing C, implies $n \ge d$ and $\gamma_n \le c\beta_Q(\epsilon)$. For completeness, write $b = \beta_Q(\epsilon)$. The premise gives $n/d \ge (C/b)\log(e/b)$. Hence $\log(2en/d) \le C'\log(e/b)$, after increasing C. The confidence term is controlled directly by the second term in (32). □

E The KL Modulus

Define

$$\beta_\rho(\epsilon) = \inf_{0\le q\le 1-\epsilon} \frac{[A_\rho(q+\epsilon) - A_\rho(q)]^2}{A_\rho(q+\epsilon)}.$$

Theorem 12 (KL inverse-increment modulus)**.** *For every fixed $\rho > 0$, there are $c_\rho > 0$ and $\epsilon_\rho > 0$ such that*

$$c_\rho A_\rho(\epsilon) \le \beta_\rho(\epsilon) \le A_\rho(\epsilon) \tag{34}$$

for $0 < \epsilon \le \epsilon_\rho$.

Proof. The choice $q = 0$ gives the upper bound. We prove the reverse bound uniformly over q. Put $s = q+\epsilon$ and

$$B(q) = \frac{[A_\rho(s) - A_\rho(q)]^2}{A_\rho(s)}.$$

Choose

$$\eta = \min\{\rho/4, 1/4\}.$$

We consider three ranges.

Range 1: $0 \le q \le \sqrt{2}\epsilon$. For $q > 0$, Lemma 8 gives

$$\frac{A_\rho(q)}{A_\rho(s)} \le e\frac{q}{s}\exp\left\{-\rho\left(\frac{1}{q}-\frac{1}{s}\right)\right\} \le e\exp\left\{-\frac{\rho}{\sqrt{2}(1+\sqrt{2})\epsilon}\right\} = o(1). \quad (35)$$

The case $q = 0$ has ratio zero. For all sufficiently small ϵ, the ratio in (35) is at most $1/2$. Therefore

$$B(q) \ge \frac{1}{4}A_\rho(s) \ge \frac{1}{4}A_\rho(\epsilon).$$

Range 2: $q \ge \sqrt{2}\epsilon$ and $s \le \eta$. The function $u \mapsto u^{-1}e^{-\rho/u}$ is increasing on $(0, \eta]$. Lemma 9 therefore gives

$$A_\rho(s) - A_\rho(q) \ge \frac{\rho\epsilon}{2eq}e^{-\rho/q}. \quad (36)$$

Use the upper sandwich bounds for $A_\rho(s)$ and $A_\rho(\epsilon)$. With $c = q/\epsilon \ge \sqrt{2}$, we obtain

$$\frac{B(q)}{A_\rho(\epsilon)} \ge \frac{\rho^2}{4e^2c^2(c+1)\epsilon^2}\exp\left\{\frac{\rho}{\epsilon}\left(1-\frac{2}{c}+\frac{1}{c+1}\right)\right\}. \quad (37)$$

The exponent in parentheses is nonnegative at $c = \sqrt{2}$ and is increasing for larger c. On $[\sqrt{2}, 2]$, the polynomial prefactor diverges. On $[2, \eta/\epsilon]$, the exponent is at least $1/3$, while the reciprocal polynomial factor is at worst order ϵ. The exponential term then dominates uniformly. Thus the right side of (37) is at least one for all sufficiently small ϵ.

Range 3: $s > \eta$. For sufficiently small ϵ, Range 1 is excluded and $q \ge \eta/2$. Equation (28) and the exponential sandwich give, uniformly for $u \in [\eta/2, 1)$,

$$A'_\rho(u) \ge \rho A_\rho(\eta/2)(1-e^{-\rho}) =: m_\rho > 0.$$

Consequently $A_\rho(s) - A_\rho(q) \ge m_\rho\epsilon$. Since $A_\rho(s) \le 1$,

$$B(q) \ge m_\rho^2\epsilon^2.$$

By Lemma 8, $A_\rho(\epsilon) = o(\epsilon^m)$ for every fixed m. In particular, this last display dominates $A_\rho(\epsilon)$ for small enough ϵ.

Combining the three ranges proves (34), with a constant that may depend on ρ. □

The sharp agnostic upper bound needs a stronger consequence of the same range analysis. Write

$$b_\epsilon = A_\rho(\epsilon), \qquad \Delta_\epsilon(q) = A_\rho(q+\epsilon) - A_\rho(q).$$

Lemma 13 (Logarithmic separation)**.** *For every fixed $\rho > 0$ and $\kappa > 0$,*

$$\inf_{\substack{0\le q\le 1-\epsilon:\\ A_\rho(q)\ge\kappa\Delta_\epsilon(q)}} \frac{\Delta_\epsilon(q)^2}{A_\rho(q)b_\epsilon\log(e/b_\epsilon)} \longrightarrow \infty \quad (38)$$

as $\epsilon \downarrow 0$.

Proof. Retain $s = q + \epsilon$ and $\eta = \min\{\rho/4, 1/4\}$. In Range 1 of the preceding proof, equation (35) implies

$$\frac{A_\rho(q)}{\Delta_\epsilon(q)} = o(1)$$

uniformly. Hence the constraint in (38) is empty in that range for all sufficiently small ϵ.

Suppose next that $q \ge \sqrt{2}\epsilon$ and $s \le \eta$. Combine (36) with the upper sandwich bounds on $A_\rho(q)$ and b_ϵ to get

$$\frac{\Delta_\epsilon(q)^2}{A_\rho(q)b_\epsilon} \ge \frac{\rho^2\epsilon}{4e^2q^3}\exp\left\{\frac{\rho}{\epsilon}-\frac{\rho}{q}\right\}. \quad (39)$$

The right side is increasing in q on this range because $q \le \rho/4$. At $q = \sqrt{2}\epsilon$, it is at least

$$\frac{c_\rho}{\epsilon^2}\exp\left\{\frac{\rho}{\epsilon}(1-1/\sqrt{2})\right\}.$$

Also, Lemma 8 gives $\log(e/b_\epsilon) = O_\rho(1/\epsilon)$. Thus (39), divided by $\log(e/b_\epsilon)$, diverges uniformly.

Finally, if $s > \eta$, the compact-range argument from Range 3 gives $\Delta_\epsilon(q) \ge m_\rho\epsilon$. Since $A_\rho(q) \le 1$,

$$\frac{\Delta_\epsilon(q)^2}{A_\rho(q)b_\epsilon\log(e/b_\epsilon)} \ge \frac{m_\rho^2\epsilon^2}{b_\epsilon\log(e/b_\epsilon)} \longrightarrow \infty.$$

The three ranges prove the claim. □

F The Sharp Hybrid Agnostic Bound

We first record the ordinary-learning ingredient used in the main paper.

Lemma 14 (Small-error candidate)**.** *Let $D_\delta = d + \log(1/\delta)$. There are universal constants a_0, C_0 and an improper learner such that, for every binary class of VC dimension d, its output h_{sm} satisfies*

$$p_{h_{\mathrm{sm}}} - p_* \le a_0 p_* + C_0\left(\sqrt{\frac{p_*D_\delta}{n}} + \frac{D_\delta}{n}\right) \quad (40)$$

with probability at least $1 - \delta$.

Proof. The first branch of Asilis et al. (2025, Theorem 1.4) states that one learner has ordinary error at most

$$2.1p_* + C_0\left(\sqrt{\frac{p_*D_\delta}{n}} + \frac{D_\delta}{n}\right).$$

Subtracting p_* proves the claim with $a_0 = 1.1$. If the infimum defining p_* is not attained, use an arbitrarily close comparator and take a limit. □

The final sample split selects between two classifiers. The following two-point form of the localized bound will be used conditionally on the training blocks.

Lemma 15 (Relative validation)**.** *Let g_1, g_2 be fixed classifiers, and let $\widehat{g}$ minimize empirical error over them on m fresh examples. With probability at least $1-\delta$,*

$$p_{\widehat{g}} - \min_{j\in\{1,2\}} p_{g_j} \le C_1 \left(\sqrt{\frac{p_{\min}\log(4/\delta)}{m}} + \frac{\log(4/\delta)}{m} \right), \tag{41}$$

where $p_{\min} = \min_j p_{g_j}$.

Proof. Apply the relative Bernoulli deviation inequality to each of the two fixed losses and take a union bound. Empirical optimality yields the same quadratic inequality as in the proof of Lemma 10, now with VC complexity replaced by $\log 2$. Solving it proves (41) after changing a universal constant. □

Theorem 16 (Sharp agnostic upper bound)**.** *Fix $\rho > 0$. There are $C_\rho, \epsilon_\rho > 0$ such that every binary class of finite VC dimension satisfies*

$$M_{\mathrm{KL}}^{\mathrm{agn}}(\epsilon, \delta, \mathcal{H}) \le C_\rho \frac{d + \log(1/\delta)}{A_\rho(\epsilon)} \tag{42}$$

for $0 < \epsilon \le \epsilon_\rho$.

Proof. The case $d = 0$ is immediate, so assume $d \ge 1$. Set

$$b = b_\epsilon, \qquad D = d + \log(24/\delta).$$

Split the sample into three independent blocks of a common size m, ignoring at most two examples. On the first block run the learner from Lemma 14. On the second block compute an ordinary ERM. On the third block choose the candidate with smaller empirical error.

Let $p_* = \inf_{h\in\mathcal{H}} p_h$. If $Q_\rho(p_*) > 1-\epsilon$, every output already has robust excess at most ϵ. Otherwise put

$$q = Q_\rho(p_*), \qquad \Delta = \Delta_\epsilon(q).$$

Then $p_* = A_\rho(q)$ and $p_* + \Delta = A_\rho(q+\epsilon)$. Theorem 12 gives

$$b\{p_* + \Delta\} \le C_\rho \Delta^2. \tag{43}$$

In particular, $p_* b \le C_\rho \Delta^2$ and $b \le C_\rho \Delta$.

Take $m \ge K_\rho D/b$, where K_ρ will be a sufficiently large constant. Fix $\kappa = (12a_0)^{-1}$. If $p_* \le \kappa\Delta$, then Lemma 14 and (43) give

$$p_{h_{\mathrm{sm}}} - p_* \le \frac{\Delta}{12} + C_\rho \left(\frac{\Delta}{\sqrt{K_\rho}} + \frac{\Delta}{K_\rho} \right).$$

Thus this excess is at most $\Delta/3$ for large enough K_ρ.

Suppose instead that $p_* > \kappa\Delta$. For the ERM block, define

$$\gamma_m = \frac{d\log(2em/d) + \log(48/\delta)}{m}.$$

Since $m \ge K_\rho D/b$ and $(d/D)\log(D/d)$ is bounded uniformly for $D/d \ge 1$,

$$\gamma_m \le \frac{C}{K_\rho} b \log(e/b). \tag{44}$$

Lemma 13 makes the right side $o_\rho(\Delta^2/p_*)$, uniformly in the present regime. Lemma 10 therefore yields $p_{h_{\mathrm{er}}} - p_* \le \Delta/3$ for sufficiently small ϵ, after increasing K_ρ if needed.

We have proved that in either regime one candidate has error $p_{\min} \le p_* + \Delta/3 \le p_* + \Delta$. Conditional on the first two blocks, Lemma 15 and (43) give

$$p_{\widehat{g}} - p_{\min} \le C_\rho \left(\frac{\Delta}{\sqrt{K_\rho}} + \frac{\Delta}{K_\rho} \right) \le \frac{\Delta}{3}.$$

The three events hold simultaneously with probability at least $1-\delta$. Consequently $p_{\widehat{g}} \le p_* + 2\Delta/3 < p_* + \Delta$, and monotonicity gives

$$Q_\rho(p_{\widehat{g}}) - Q_\rho(p_*) \le \epsilon.$$

Since the total sample size is at most $3m + 2$, this proves (42). □

Every realizable distribution is admissible in the agnostic model. The exact realizable lower bound therefore also lower-bounds agnostic sample complexity. Combining it with Theorem 16 and Lemma 8 proves the matching rate in the main paper.

G Cressie–Read Boundary Layer

For $k > 1$, binary data processing and the two-level likelihood ratio from Proposition 7 show that the inverse lower branch $A_{k,\rho}(q)$ satisfies

$$\begin{aligned} q^k p^{1-k} + (1-q)^k (1-p)^{1-k} - 1 &= k(k-1)\rho, \\ p &= A_{k,\rho}(q). \end{aligned} \tag{45}$$

Theorem 17 (Boundary asymptotic)**.** *Fix $\rho > 0$ and $\lambda > 0$. Set $k_q = 1 + \lambda q/\rho$. Then*

$$A_{k_q,\rho}(q) = e^{-1} q \exp\left\{ -\frac{\rho}{q} \frac{\log(1+\lambda)}{\lambda} \right\} (1 + o(1)). \tag{46}$$

For fixed $k > 1$,

$$A_{k,\rho}(q) \sim q^{k/(k-1)} [k(k-1)\rho]^{-1/(k-1)}. \tag{47}$$

Proof. Write $k = 1 + \zeta$ and first take $\zeta = \lambda q/\rho$. Let

$$T = q^{1+\zeta} p^{-\zeta}, \qquad C = (1-q)^{1+\zeta}(1-p)^{-\zeta}.$$

Equation (45) gives

$$T = 1 + (1+\zeta)\zeta\rho - C. \tag{48}$$

Since $0 < p < q$ and $\zeta = O(q)$,

$$\begin{aligned}\log C &= (1+\zeta)\log(1-q) - \zeta\log(1-p)\\ &= -q - \zeta q - \frac{q^2}{2} + \zeta p + O(q^3).\end{aligned}$$

Exponentiating cancels the displayed $q^2/2$ term and yields

$$C = 1 - q - \zeta q + \zeta p + O(q^3). \tag{49}$$

The first-order versions of (48) and (49) show that $T/q \to 1 + \lambda > 1$. Since $T/q = \exp\{\zeta \log(q/p)\}$, this implies $p/q \to 0$ faster than any power. Hence $\zeta p = o(q^2)$ in (49). Using $\zeta = \lambda q/\rho$ in (48) now gives

$$\begin{aligned}T &= q(1+\lambda) + \frac{q^2}{\rho}(\lambda + \lambda^2) + o(q^2)\\ &= q(1+\lambda)\{1 + \zeta + o(\zeta)\}.\end{aligned}$$

Therefore

$$\zeta \log \frac{q}{p} = \log(1+\lambda) + \zeta + o(\zeta).$$

Solving for p proves (46).

For fixed $k > 1$, the common-event term in (45) tends to one. Thus $q^k p^{1-k} \to k(k-1)\rho$. Solving for p gives (47). □

At $k = 1$, Lemma 8 gives the same boundary formula with $\log(1+\lambda)/\lambda$ continuously extended to one at $\lambda = 0$. The logarithmic phase statement in the main paper follows immediately.

References

Aigner-Horev, E., Rosenberg, D., and Weiss, R. (2026). The sample complexity of distributionally robust PAC learning under Cressie–Read divergences. *arXiv preprint arXiv:2608.04686*.

Asilis, J., Høgsgaard, M. M., and Velegkas, G. (2025). On agnostic PAC learning in the small error regime. In *Advances in Neural Information Processing Systems*, volume 38.

Ben-Tal, A., den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357.

Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375.

Duchi, J. C. and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406.

Hanneke, S. (2016). The optimal sample complexity of PAC learning. *Journal of Machine Learning Research*, 17(38):1–15.

Hanneke, S., Larsen, K. G., and Zhivotovskiy, N. (2024). Revisiting agnostic PAC learning. In *Proceedings of the 65th IEEE Symposium on Foundations of Computer Science*, pages 1968–1982. IEEE.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. (2018). Does distributionally robust supervised learning give robust classifiers? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2029–2037.

Jiang, R. and Guan, Y. (2016). Data-driven chance constrained stochastic program. *Mathematical Programming*, 158(1–2):291–327.

Zhou, Z. and Liu, W. (2023). Sample complexity for distributionally robust learning under chi-square divergence. *Journal of Machine Learning Research*, 24(230):1–27.

Zhou, Z. and Liu, W. (2026). Rademacher complexity for distributionally robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29098–29106.