SIGN BETTING FOR MISSPECIFIED REGRESSION: VARIANCE ADAPTIVITY AND OPTIMAL APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study bounded regression under model misspecification, evaluating a function f by its $L_1(P_X)$ distance to the conditional mean $\eta(x) = \mathbb{E}[Y \mid X = x]$, rather than the standard $L_2(P_X)$. This L_1 metric is natural for downstream decision problems such as bandits and reinforcement learning. We analyze the min–max estimator of Baraud & Maillard (2025), which we call *sign-betting regression*. Our main results show that sign-betting regression achieves high-probability L_1 oracle inequalities under misspecification. Under realizability, our bound adapts to the average conditional variance and recovers the variance-adaptive rate of Li et al. (2026). We further characterize the optimal universal approximation factor for proper learning by showing that sign betting achieves factor 3, while no proper learner can satisfy a distribution-free, high-probability L_1 oracle inequality with any universal approximation factor strictly below 3. We extend the upper-bound analysis beyond finite classes and complement the theory with experiments in synthetic and real-world data model-selection settings.

1 INTRODUCTION

Regression is a fundamental problem in statistics and machine learning and often an intermediate step in a downstream decision problem, with a quantity of interest being the underlying conditional mean. We study bounded regression, where (X, Y) denotes a covariate–response pair with $Y \in [0, 1]$, and evaluate a predictor f by its expected absolute discrepancy from the conditional mean, $\varepsilon_1(f) := \mathbb{E}\big[|f(X) - \eta(X)|\big]$, where $\eta(x) := \mathbb{E}[Y \mid X = x]$.

The L_1 accuracy of a conditional-mean predictor has several downstream interpretations. For binary responses, it controls the excess risk of the plug-in classifier $g_f(x) := \mathbf{1}\{f(x) \geq 1/2\}$ relative to the Bayes classifier $g^\star(x) := \mathbf{1}\{\eta(x) \geq 1/2\}$. Writing $R_{\mathrm{cls}}(g) := \mathbb{P}(g(X) \neq Y)$ for the classification risk, the standard comparison inequality gives $R_{\mathrm{cls}}(g_f) - R_{\mathrm{cls}}(g^\star) \leq 2\varepsilon_1(f)$ (Audibert & Tsybakov, 2007). The L_1 metric is also natural for downstream decision problems such as bandits and reinforcement learning, where policy suboptimality is typically controlled through differences in expected rewards or costs rather than their squares (Ayoub et al., 2025). Finally, when f is used as a reward model to select the highest-scoring response among N candidates, each marginally distributed according to P_X, the expected reward loss relative to the best candidate in the same set is at most $N\varepsilon_1(f)$. This connects L_1 reward-model accuracy to recent analyses of Best-of-N and related inference-time selection procedures (Huang et al., 2025). See derivations in Appendix A.

Regression under misspecification. Let $\mathcal{F}$ be a class of $[0, 1]$-valued predictors and define $\varepsilon_{1,*} := \inf_{f \in \mathcal{F}} \varepsilon_1(f)$. Under misspecification, we do not assume that $\eta \in \mathcal{F}$. We ask whether a learner can compete with the best achievable L_1 error within the class. Ideally, an oracle inequality of the form

$$\varepsilon_1(\widehat{f}) \leq C_{\mathrm{aprx}}\, \varepsilon_{1,*} + r_n, \qquad r_n \to 0 \quad \text{as } n \to \infty, \tag{1}$$

where C_{aprx} is a universal constant. We call C_{aprx} the *approximation factor* and r_n the *estimation term*. Beyond requiring r_n to vanish, we seek guarantees that adapt to the expected conditional variance $\mathbb{E}[\mathrm{Var}(Y \mid X)]$ and become sharper when this quantity is small. This leads to two central questions:

> *What approximation and noise-adaptive guarantees are achievable for L_1 conditional-mean estimation under model misspecification? When the learner is required to return a member of $\mathcal{F}$, what is the smallest universal approximation factor in a distribution-free high-probability oracle inequality?*

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.

Why standard regression objectives do not resolve the problem. A loss that identifies the conditional mean under unrestricted optimization can select an approximation to η that differs from the best L_1 approximation when the model class is misspecified. Squared loss is the canonical example: minimizing its population risk over $\mathcal{F}$ selects the best L_2 approximation to η, rather than the best L_1 approximation targeted by (1). More generally, standard losses such as squared and log loss induce their own notions of projection onto $\mathcal{F}$ under misspecification (Mey & Loog, 2021). We show that this mismatch is fundamental: e.g., squared-loss ERM admits an L_1 guarantee whose approximation term scales as $\sqrt{\varepsilon_{1,*}}$, and this dependence is unavoidable even in a noiseless problem with a function class of size 2. Consequently, squared-loss ERM cannot satisfy (1) with C_{aprx} being a universal constant. A second issue is statistical adaptation. Although an $n^{-1/2}$ estimation term is unavoidable in the worst case over distributions of (X, Y) with $Y \in [0, 1]$ (Tsybakov & Zaiats, 2009; Wainwright, 2019), it can be overly pessimistic when the conditional variance of the labels is small (Foster & Krishnamurthy, 2021; Li et al., 2026). Our goal is to achieve linear dependence on the best L_1 approximation error $\varepsilon_{1,*}$ together with adaptation to low noise. As we show in Section 3, the standard regression objectives considered therein do not, in general, provide either of these desiderata.

Sign-Betting Regression. We study a min-max estimator introduced by Baraud & Maillard (2025) for fixed-design regression with possibly heteroscedastic and heavy-tailed errors. We refer to this procedure as *Sign-Betting Regression (SBR)*, reflecting the directional betting interpretation developed in Section 4.2.

Our results. Our first main result is a high-probability generalization inequality for SBR. In the general misspecified setting, it implies $\varepsilon_1(\widehat{f}) \le 3\varepsilon_{1,*} + r_n$, where the estimation term r_n adapts to the residual second moment of an L_1-optimal comparator. Under realizability, the approximation terms disappear and the leading statistical term adapts directly to the average conditional variance $\mathbb{E}[\mathrm{Var}(Y \mid X)]$, recovering the variance-adaptive rate of Li et al. (2026). Our second main result identifies the optimal universal approximation factor for proper learning. We prove that, for proper learning, (1) cannot hold with any universal constant $C_{\mathrm{aprx}} < 3$ and any distribution-independent r_n. Since SBR achieves factor 3, this characterizes the optimal universal proper approximation factor under our learning criterion. Finally, we extend the upper-bound analysis beyond finite classes and complement our theoretical results with experiments on synthetic data, real-world datasets, and LLM model selection.

Our setting and questions differ from those of Baraud & Maillard (2025) in several respects. They study fixed-design regression and expected L_1 error evaluated on the observed design points. We instead consider i.i.d. random design and study high-probability oracle inequalities for the population error $\|f - \eta\|_{L_1(P_X)}$. A direct black-box conversion—conditioning on the covariates, applying their fixed-design result, and then using global uniform convergence—yields a population guarantee in expectation but incurs a variance-independent $n^{-1/2}$ generalization term, which can be substantially larger than our bound on low-noise instances. Our direct random-design analysis achieves variance adaptation. See Appendix I for details.

Paper organization. Section 2 formalizes the learning problem, the L_1 oracle criterion, and the noise quantities used throughout the paper. Section 3 shows how standard regression objectives can fail both the desired approximation and noise-adaptation guarantees. Section 4 develops the population geometry of sign betting and establishes our main variance-adaptive oracle inequality. Section 5 characterizes the optimal universal approximation factor for proper learning, and Section 6 extends the upper-bound framework beyond finite classes. Appendix B discusses related work and connections to prior results, and Appendix M presents the experiments. The remaining appendices contain proofs and supporting technical material.

2 Problem Setting and Learning Objective

Setting and notation. Let $\mathcal{X}$ be a measurable input space, and let P be an unknown distribution over $\mathcal{X} \times [0, 1]$. We observe n i.i.d. samples $(X_i, Y_i)_{i=1}^n \sim P^n$. We write P_X for the marginal distribution of X and $\eta(x) := \mathbb{E}[Y \mid X = x]$ for the regression function. We consider a nonempty class $\mathcal{F} \subseteq \{f : \mathcal{X} \to [0, 1]\}$ of measurable predictors, without assuming that $\eta \in \mathcal{F}$. For a measurable $g : \mathcal{X} \to \mathbb{R}$ and $p \in [1, \infty)$, we write $\|g\|_p := (\mathbb{E}_{X \sim P_X} |g(X)|^p)^{1/p}$. For a measurable function $Z = Z(X, Y)$, we use the shorthand $PZ := \mathbb{E}[Z(X, Y)]$ and $P_n Z := n^{-1} \sum_{i=1}^n Z(X_i, Y_i)$.

Approximation error. For $p \in [1, \infty)$ and $f \in \mathcal{F}$, define $\varepsilon_p(f) := \|f - \eta\|_p$ and $\varepsilon_{p,*} := \inf_{f \in \mathcal{F}} \varepsilon_p(f)$. Our main performance criterion is the L_1 conditional-mean approximation error

$\varepsilon_1(f) = \|f - \eta\|_1 = \mathbb{E}\,|f(X) - \eta(X)|$, and $\varepsilon_{1,*}$ is the smallest such error attainable within $\mathcal{F}$. When the infimum exists, we let $f_\star \in \arg\min_{f\in\mathcal{F}} \varepsilon_1(f)$ denote a best L_1 approximation to η in $\mathcal{F}$. In particular, such a minimizer always exists for the finite classes considered in our main results. We say that $(P, \mathcal{F})$ is *realizable* (or *well-specified*) if $\eta \in \mathcal{F}$, and *misspecified* otherwise. For a finite class, realizability is equivalent to $\varepsilon_{1,*} = 0$. Our results do not assume realizability. Under misspecification, our guarantees compare the learned predictor with arbitrary comparators $f_0 \in \mathcal{F}$, allowing the bound to trade off their approximation error (how well f_0 approximates η) and comparator-dependent statistical terms. We distinguish this criterion in Appendix E from the absolute-loss risk $\mathbb{E}\,|Y - f(X)|$, which is generally a different population objective.

Noise and approximation decomposition. Let $\xi := Y - \eta(X)$, $\sigma^2(X) := \mathrm{Var}(Y \mid X) = \mathbb{E}[\xi^2 \mid X]$, and $\bar{\sigma}^2 := \mathbb{E}[\sigma^2(X)]$. Then $\mathbb{E}[\xi \mid X] = 0$, and for every predictor f, $\mathbb{E}[(Y - f(X))^2 \mid X] = \sigma^2(X) + (\eta(X) - f(X))^2$. The first term is the conditional label variance, while the second is the squared approximation error. This decomposition will enter the variance-dependent bounds under misspecification.

Proper learning. A learner is *proper* with respect to $\mathcal{F}$ if it always returns a predictor $\hat{f} \in \mathcal{F}$; otherwise it is *improper*. Our main results focus on proper learning: our main algorithm is proper, and our lower bound characterizes the best universal approximation factor achievable by proper learners.

3 Limitations of standard regression objectives

We first ask whether familiar regression objectives already provide the L_1 oracle guarantee in (1). We focus here on squared loss and log loss, while deferring absolute loss to Appendix E and betting loss of Li et al. (2026) to Appendix F. These objectives fail for two conceptually different reasons. Absolute loss is misaligned with the target itself: its population minimizer is generally a conditional median rather than the conditional mean η. By contrast, squared loss, log loss, and betting loss are all aligned with conditional-mean regression under realizability, with betting loss enjoying variance-adaptive guarantees (Li et al., 2026). Nevertheless, under model misspecification, squared loss, log loss, and betting loss may not produce a predictor whose L_1 error is within any universal constant factor of $\varepsilon_{1,*}$. These limitations motivate an objective that remains aligned with the conditional mean under misspecification while enjoying variance-adaptive statistical guarantees.

3.1 Squared-loss ERM

We begin with the familiar squared-loss ERM, $\widehat{f}_{\mathrm{sq}} \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \big(Y_i - f(X_i)\big)^2$. Squared loss is calibrated to the conditional mean when optimization is unrestricted. Under misspecification, however, minimizing squared loss over $\mathcal{F}$ does not yield the best L_1 approximation to η. Indeed, for every $f \in \mathcal{F}$,

$$\mathbb{E}\big[(Y - f(X))^2\big] = \mathbb{E}\big[\mathrm{Var}(Y \mid X)\big] + \|f - \eta\|_2^2 . \tag{2}$$

Thus, at the population level, squared loss selects an L_2-optimal approximation to η, whereas our benchmark $\varepsilon_{1,*}$ is defined by the best L_1 approximation. The following proposition quantifies the resulting finite-sample guarantee. The proof is deferred to Appendix C.1.

Proposition 1 (L_1 guarantee for squared-loss ERM)**.** *Suppose $Y \in [0,1]$ and $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ is finite. There exists a universal constant $C > 0$ such that, with probability at least $1 - \delta$,*

$$\varepsilon_1(\widehat{f}_{\mathrm{sq}}) \le \varepsilon_{2,*} + C\sqrt{\frac{\log(2|\mathcal{F}|^2/\delta)}{n}} \le \sqrt{\varepsilon_{1,*}} + C\sqrt{\frac{\log(2|\mathcal{F}|^2/\delta)}{n}}. \tag{3}$$

The first inequality in (3) reflects the natural geometry of squared loss: apart from estimation error, squared-loss ERM competes with $\varepsilon_{2,*}$. The second inequality follows from $\varepsilon_{2,*} \le \sqrt{\varepsilon_{1,*}}$ for $[0,1]$-valued hypothesis class $\mathcal{F}$. Next, we show in Proposition 2 that this square-root dependence on $\varepsilon_{1,*}$ is not merely an artifact of converting an L_2 bound to L_1.

Proposition 2 (Squared loss has no constant L_1 approximation factor)**.** *For every $p \in (0, 1/4)$, there exists a noiseless regression problem and a two-element class $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ such that $\varepsilon_{1,*} = p$, while the unique population squared-loss minimizer, $f^*_{\mathrm{sq}} \in \arg\min_{f\in\mathcal{F}} \mathbb{E}\big[(Y - f(X))^2\big]$, satisfies*

$$\varepsilon_1(f^*_{\mathrm{sq}}) = \frac{1}{2}\sqrt{\varepsilon_{1,*}}. \tag{4}$$

Consequently, squared-loss ERM cannot satisfy the basic goal (1) *for any finite universal approximation factor.*

Proposition 2 shows that the $\sqrt{\varepsilon_{1,*}}$ dependence in (3) is unavoidable up to a universal constant. This is a population-level obstruction: even with noiseless labels and arbitrarily many samples, squared loss can prefer an L_2-better predictor whose L_1 error is of order $\sqrt{\varepsilon_{1,*}}$ rather than $\varepsilon_{1,*}$. Appendix C.2 gives a sharper result (with proof) that, for any $\gamma \in (0,1)$, one can construct an instance satisfying $\varepsilon_1(f^*_{\mathrm{sq}}) = (1-\gamma)\sqrt{\varepsilon_{1,*}}$. Proposition 2 is a specialization with $\gamma = 1/2$.

The $n^{-1/2}$ statistical term in (3) cannot be improved in general. Standard minimax lower bounds show an $\Omega(n^{-1/2})$ error for arbitrary estimators in simple bounded mean-estimation problems; see, e.g., Tsybakov & Zaiats (2009); Wainwright (2019). The limitation of the estimation error from squared-loss ERM is the failure to adapt to easier instances. In realizable full-information cost-sensitive classification, Foster & Krishnamurthy (2021, Theorem 2) show that least-squares plug-in prediction can incur $\Omega(n^{-1/2})$ excess risk even when the first-order rate is $O(1/n)$. Analogously, in realizable $[0,1]$-valued regression, Li et al. (2026, Proposition 8) show that squared-loss ERM can incur $\Omega(n^{-1/2})$ L_1 error even when the first-order rate is $O(1/n)$.

Taken together, these results identify two distinct limitations of squared-loss ERM for our objective. Under misspecification, its L_2 approximation can be arbitrarily worse than the best L_1 approximation in multiplicative terms; even under realizability, squared-loss ERM does not adapt to favorable conditional variance.

3.2 LOG-LOSS ERM

Another natural choice for conditional-mean estimation is log loss, which is widely used for learning probabilistic predictions. Recall the log-loss ERM

$$\widehat{f}_{\log} \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell_{\log}\big(Y_i, f(X_i)\big), \qquad \ell_{\log}(y,q) := -y\log q - (1-y)\log(1-q). \tag{5}$$

For any $Y \in [0,1]$, conditioning on $X = x$ gives

$$\mathbb{E}[\ell_{\log}(Y,q) \mid X = x] = -\eta(x)\log q - (1-\eta(x))\log(1-q),$$

which is uniquely minimized at $q = \eta(x)$. Thus, like squared loss, log loss is calibrated to the conditional mean when optimization is unrestricted.

Log loss is particularly relevant to our comparison because, under realizability, it can exhibit favorable first-order behavior that squared loss fails to achieve. In the cost-sensitive classification setting, Foster & Krishnamurthy (2021) show that log-loss regression yields a first-order guarantee. More directly for our regression setting, Li et al. (2026) establish a first-order guarantee for log-loss ERM measured in the same L_1 prediction error considered in this paper. This motivates asking whether it also avoids the other deficiency of squared loss under misspecification: the mismatch between its population objective and the best L_1 approximation in $\mathcal{F}$. We show that it does not. Even when all predictions are uniformly bounded away from 0 and 1, log-loss ERM can incur L_1 error of order $\sqrt{\varepsilon_{1,*}}$, and therefore cannot achieve a universal constant-factor L_1 oracle guarantee.

For $p \in [0,1]$ and $q \in (0,1)$, let $\mathrm{kl}(p,q) := p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$, with the convention $0\log 0 = 0$. At the population level (see Appendix D.1 for the derivation),

$$\mathbb{E}\big[\ell_{\log}(Y, f(X)) - \ell_{\log}(Y, \eta(X))\big] = \mathbb{E}\Big[\mathrm{kl}\big(\eta(X), f(X)\big)\Big]. \tag{6}$$

Thus, under misspecification, population log-loss minimization selects a KL-optimal approximation to η, which may not coincide with the best L_1 approximation.

Because log loss becomes increasingly sensitive as predictions approach 0 or 1, an L_1 guarantee in terms of $\varepsilon_{1,*}$ requires explicit control of the prediction range. The following proposition gives the resulting guarantee under a restricted boundary condition.

Proposition 3 (L_1 guarantee for log-loss ERM)**.** *Suppose $Y \in [0,1]$, $\mathcal{F}$ is finite, and, for some $\gamma \in (0, 1/2)$, $\mathcal{F} \subseteq [\gamma, 1-\gamma]^{\mathcal{X}}$. Then, with probability at least $1-\delta$,*

$$\varepsilon_1(\widehat{f}_{\log}) \le \sqrt{\frac{\log(1/\gamma)}{1-\gamma}\,\varepsilon_{1,*}} + \sqrt{\left(2 + \frac{2}{3}\log\frac{1}{\gamma}\right)\frac{\log(2|\mathcal{F}|^2/\delta)}{n}}. \tag{7}$$

In particular, for every fixed γ, $\varepsilon_1(\widehat{f}_{\log}) \lesssim_\gamma \sqrt{\varepsilon_{1,*}} + \sqrt{\log(|\mathcal{F}|/\delta)/n}$.

The proof is deferred to Appendix D.2. Proposition 3 shows that, even under a restricted boundary condition, the natural L_1 guarantee for log-loss ERM scales as $\sqrt{\varepsilon_{1,*}}$ rather than linearly with $\varepsilon_{1,*}$. As with squared loss, one might wonder whether this $\sqrt{\varepsilon_{1,*}}$ dependence is merely an artifact of the analysis. The next proposition shows that it is unavoidable: even for a two-element class uniformly bounded away from 0 and 1, the population log-loss minimizer can have L_1 error of order $\sqrt{\varepsilon_{1,*}}$.

Proposition 4 (Log loss has no constant L_1 approximation factor)**.** *For every $p \in (0,1)$, there exists a noiseless regression problem and a two-element class $\mathcal{F} \subseteq [1/4, 3/4]^{\mathcal{X}}$ such that $\varepsilon_{1,*} = p/4$, while the unique population log-loss minimizer $f^*_{\log} \in \arg\min_{f\in\mathcal{F}} \mathbb{E}[\ell_{\log}(Y, f(X))]$ satisfies*

$$\varepsilon_1(f^*_{\log}) = \frac{1}{2}\sqrt{\varepsilon_{1,*}}. \tag{8}$$

Consequently, log-loss ERM cannot satisfy the basic goal (1) *for any finite universal approximation factor, even when all predictions are uniformly bounded away from* 0 *and* 1.

Proposition 4 shows that the $\sqrt{\varepsilon_{1,*}}$ dependence in Proposition 3 is unavoidable up to a universal constant. Even when the hypothesis class is uniformly bounded away from 0 and 1, the geometry of KL divergence can make the population log-loss minimizer have L_1 error of order $\sqrt{\varepsilon_{1,*}}$, even though $\mathcal{F}$ contains another hypothesis with L_1 error $\varepsilon_{1,*}$. Thus, as with squared loss, log-loss ERM cannot in general achieve linear dependence on $\varepsilon_{1,*}$ or a finite universal L_1 approximation factor.

The restriction $\mathcal{F} \subseteq [\gamma, 1-\gamma]^{\mathcal{X}}$ is essential for Proposition 3. Without requiring predictions to stay uniformly away from 0 and 1, log-loss ERM can fail much more severely: its population minimizer may have L_1 error bounded away from zero even as $\varepsilon_{1,*} \to 0$. We give this stronger boundary-sensitive obstruction in Appendix D.4. Thus, even in the more favorable setting where all hypotheses are bounded away from 0 and 1, the worst-case approximation dependence of log-loss ERM remains of order $\sqrt{\varepsilon_{1,*}}$, matching the qualitative limitation of squared-loss ERM.

4 Sign betting loss and variance-adaptive bounds

4.1 Sign betting for misspecified regression

Motivated by the limitations of standard regression objectives established in Section 3, we turn to the min–max pairwise estimator introduced by Baraud & Maillard (2025). We show that this procedure, which we call *Sign-Betting Regression (SBR)*, aligns more with the best L_1 approximation to the conditional mean within $\mathcal{F}$ and yields strong oracle guarantees under misspecification. Given n i.i.d. samples $(X_i, Y_i)_{i=1}^n$, for $f, h \in \mathcal{F}$, define

$$\widehat{T}(f,h) := \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - f(X_i)\big)\,\mathrm{sign}\big(h(X_i) - f(X_i)\big), \tag{9}$$

where $\mathrm{sign}(0) = 0$. Our *Sign-Betting Regression (SBR)* algorithm returns

$$\hat{f} \in \arg\min_{f\in\mathcal{F}} \sup_{h\in\mathcal{F}} \widehat{T}(f,h). \tag{10}$$

For the finite classes considered in our main results, the supremum above is a maximum.

The terminology "sign betting" comes from viewing h as a challenger that places a directional bet against the residual of f. At an input x, the challenger bets $+1$ when $h(x) > f(x)$ and -1 when $h(x) < f(x)$. Its payoff is $(Y - f(X))\,\mathrm{sign}(h(X) - f(X))$. Thus, if h lies above f, the challenger benefits when the observed response is also above f; if h lies below f, it benefits when the observed response is below f. A large value of $\widehat{T}(f,h)$ therefore indicates that the direction from f toward h is systematically supported by the residuals.

The inner supremum in (10) searches for the challenger $h \in \mathcal{F}$ that achieves the largest empirical sign-betting value $\widehat{T}(f,h)$ against a given candidate f. The outer minimization then selects a predictor whose residuals are difficult to exploit by any $h \in \mathcal{F}$. Thus, unlike ordinary empirical risk minimization with a pointwise loss $\ell(Y, f(X))$, the objective associated with f depends on how f compares with the other predictors in $\mathcal{F}$.

We next study the population geometry of sign betting. The key identity shows that, for any $f, h \in \mathcal{F}$, the two betting quantities associated with (f,h) and (h,f) sum exactly to $\|f - h\|_1$. This identity will allow us to translate control of the sign-betting objective into control of L_1 error.

4.2 POPULATION GEOMETRY OF SIGN BETTING

We first study the population behavior of sign betting. We will show that, for any pair $f, h \in \mathcal{F}$, the two population quantities $T_P(f, h)$ and $T_P(h, f)$ sum exactly to the $L_1(P_X)$ distance between f and h. This identity leads to a population performance guarantee for sign betting and provides the basic mechanism for the high-probability guarantees developed in the next subsection. For $f, h \in \mathcal{F}$, define

$$T_P(f, h) := \mathbb{E}\big[(Y - f(X))\operatorname{sign}(h(X) - f(X))\big] = \mathbb{E}\big[(\eta(X) - f(X))\operatorname{sign}(h(X) - f(X))\big]. \tag{11}$$

The second equality follows by conditioning on X and using $\mathbb{E}[Y \mid X] = \eta(X)$. We also define the population sign-betting objective

$$B_P(f) := \sup_{h \in \mathcal{F}} T_P(f, h). \tag{12}$$

The basic geometry of $T_P(\cdot, \cdot)$ follows from a pairwise identity:

Lemma 5 (Pairwise identity for sign betting)**.** *For any $f, h \in \mathcal{F}$,*

$$T_P(f, h) + T_P(h, f) = \|f - h\|_1. \tag{13}$$

Proof. Using $\operatorname{sign}(f - h) = -\operatorname{sign}(h - f)$,

$$\begin{aligned} T_P(f, h) + T_P(h, f) &= \mathbb{E}\big[(\eta - f)\operatorname{sign}(h - f) + (\eta - h)\operatorname{sign}(f - h)\big] \\ &= \mathbb{E}\big[(h - f)\operatorname{sign}(h - f)\big] \\ &= \mathbb{E}\,|h - f| = \|f - h\|_1. \end{aligned}$$

□

The population objective $B_P(f)$ is also directly controlled by the L_1 approximation error:

Lemma 6 (Population sign-betting objective)**.** *For every $f \in \mathcal{F}$,*

$$0 \le B_P(f) \le \varepsilon_1(f). \tag{14}$$

Proof. Since $h = f$ is an admissible choice in the definition of $B_P(f)$ in (12) and $T_P(f, f) = 0$, we have $B_P(f) \ge 0$. Moreover, for every $h \in \mathcal{F}$,

$$T_P(f, h) = \mathbb{E}\big[(\eta(X) - f(X))\operatorname{sign}(h(X) - f(X))\big] \le \mathbb{E}\,|\eta(X) - f(X)| = \varepsilon_1(f).$$

Taking the supremum over $h \in \mathcal{F}$ yields $B_P(f) \le \varepsilon_1(f)$. □

Lemma 5 and Lemma 6 yield a population performance guarantee for sign betting:

Proposition 7 (Population guarantee for sign betting)**.** *Suppose the minimum of B_P over $\mathcal{F}$ is attained, and let $f_B \in \arg\min_{f \in \mathcal{F}} B_P(f)$. Then, for every $f_0 \in \mathcal{F}$,*

$$\varepsilon_1(f_B) \le \varepsilon_1(f_0) + 2B_P(f_0). \tag{15}$$

Consequently, $\varepsilon_1(f_B) \le 3\varepsilon_1(f_0)$.

Proof. By Lemma 5,

$$\|f_B - f_0\|_1 = T_P(f_B, f_0) + T_P(f_0, f_B) \le B_P(f_B) + B_P(f_0) \le 2B_P(f_0),$$

where the last inequality follows from the optimality of f_B. Therefore,

$$\varepsilon_1(f_B) \le \varepsilon_1(f_0) + \|f_B - f_0\|_1 \le \varepsilon_1(f_0) + 2B_P(f_0).$$

The second claim follows from $B_P(f_0) \le \varepsilon_1(f_0)$ by Lemma 6. □

Equation (15) is more informative than the worst-case factor-three bound: the additional approximation penalty is governed by the population sign-betting value $B_P(f_0)$, which can be much smaller than $\varepsilon_1(f_0)$. We next show that the empirical Sign-Betting Regression algorithm inherits this population guarantee, up to f_0-dependent estimation terms that decay with n and adapt to the label variance.

4.3 MAIN VARIANCE-ADAPTIVE GUARANTEE

We now turn to the finite-sample performance of SBR. As we will see, the leading estimation term is controlled by an f_0-dependent residual second moment. For $f_0 \in \mathcal{F}$, define

$$\begin{aligned} v(f_0) &:= \sup_{h \in \mathcal{F}} \mathbb{E}\Big[(Y - f_0(X))^2 \mathbf{1}\{h(X) \ne f_0(X)\}\Big] \\ &= \sup_{h \in \mathcal{F}} \mathbb{E}\left[\Big(\sigma^2(X) + (\eta(X) - f_0(X))^2\Big)\mathbf{1}\{h(X) \ne f_0(X)\}\right]. \end{aligned} \tag{16}$$

The second equality follows from the noise–approximation decomposition in Section 2. In particular, $v(f_0) \le \bar{\sigma}^2 + \varepsilon_2(f_0)^2$. Thus, under misspecification, $v(f_0)$ captures both the conditional label variance and the squared approximation residual, restricted to regions where f_0 disagrees with other predictors in $\mathcal{F}$. This localization can make $v(f_0)$ smaller than the corresponding global residual second moment $\bar{\sigma}^2 + \varepsilon_2(f_0)^2$. Our main finite-class guarantee is the following.

Theorem 8 (Variance-adaptive guarantee for SBR)**.** *Suppose $\mathcal{F}$ is finite, and let $\hat{f}$ be the output of SBR in* (10)*. For $\rho > 0$, let $C_\rho := \frac{13}{3} + \frac{1}{\rho} + \frac{23}{12}\rho$. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds simultaneously for every $\rho > 0$:*

$$\varepsilon_1(\hat{f}) \le \inf_{f_0 \in \mathcal{F}} \left\{ \varepsilon_1(f_0) + (2+\rho) B_P(f_0) + (2+\rho)\sqrt{\frac{2v(f_0)\log(2|\mathcal{F}|^2/\delta)}{n}} \right\} + C_\rho \frac{\log(2|\mathcal{F}|^2/\delta)}{n}. \tag{17}$$

The parameter ρ is used only in the analysis and is not an input to SBR. Since the event in Theorem 8 holds simultaneously for all $\rho > 0$, ρ can be chosen as a function of f_0 and the other quantities when interpreting the bound. In particular, the theorem implies that, for every $f_0 \in \mathcal{F}$,

$$\varepsilon_1(\hat{f}) \le \varepsilon_1(f_0) + 2B_P(f_0) + r_n(f_0, \delta), \tag{18}$$

where $r_n(f_0, \delta) \to 0$ as $n \to \infty$ with P, f_0, $\mathcal{F}$, and δ fixed. Thus, the finite-sample guarantee recovers the population bound in Proposition 7 up to a vanishing estimation term. An explicit form of $r_n(f_0, \delta)$ is given in Corollary 18 in the appendix.

The full proof of Theorem 8 is given in Appendix G. At a high level, its proof follows the ingredients developed in the previous subsection. To elaborate, the pairwise identity in Lemma 5 continues to hold exactly on the sample:

$$\widehat{T}(f,h) + \widehat{T}(h,f) = P_n|f-h|.$$

Hence, mirroring the proof of Proposition 7, for any $f_0 \in \mathcal{F}$, the empirical optimality of $\hat{f}$ implies

$$P_n|\hat{f} - f_0| \le 2 \sup_{h \in \mathcal{F}} \widehat{T}(f_0, h). \tag{19}$$

Bernstein concentration then upper bounds the right-hand side in terms of $B_P(f_0)$ and $v(f_0)$, and a further concentration step converts the empirical distance on the left-hand side in (19) into $\|\hat{f} - f_0\|_1$.

4.4 Consequences of the main guarantee

We now specialize Theorem 8 to three regimes: the general agnostic setting, realizability, and favorable misspecification. Each specialization represents a different aspect of the guarantee.

General agnostic guarantee. We first consider the general agnostic setting, where we impose no additional relationship between η and $\mathcal{F}$ or any additional condition on the L_1-optimal predictor $f_\star \in \arg\min_{f \in \mathcal{F}} \varepsilon_1(f)$, beyond the basic assumptions of Section 2. By Lemma 6, $B_P(f_\star) \le \varepsilon_1(f_\star) = \varepsilon_{1,*}$. Combining this bound with (18) gives the following guarantee.

Corollary 9 (General agnostic guarantee)**.** *Under the assumptions of Theorem 8, with probability at least $1-\delta$,*

$$\varepsilon_1(\hat{f}) \le 3\varepsilon_{1,*} + r_n(f_\star, \delta), \tag{20}$$

where $r_n(f_\star, \delta)$ is given explicitly in Corollary 18 and satisfies $r_n(f_\star, \delta) \to 0$ as $n \to \infty$ with P, $\mathcal{F}$, and δ fixed.

Thus, without assuming realizability or any additional property of the L_1-optimal predictor, SBR achieves the basic goal (1) with approximation factor 3. The estimation term is $f_\star$-dependent and retains variance adaptation through the residual second moment $v(f_\star)$. The factor 3 in (20) is obtained by optimizing the analysis parameter ρ in Theorem 8, as detailed in Corollary 18. We return to the optimality of the factor 3 in Section 5.

Realizable case. We next consider the realizable setting, where $\eta \in \mathcal{F}$. Taking $f_0 = \eta$ in Theorem 8 gives $\varepsilon_1(\eta) = 0$ and $B_P(\eta) = 0$. Moreover, by (16),

$$v(\eta) = \sup_{h \in \mathcal{F}} \mathbb{E}\Big[\sigma^2(X)\mathbf{1}\{h(X) \ne \eta(X)\}\Big] \le \bar{\sigma}^2,$$

where $\bar{\sigma}^2 = \mathbb{E}[\sigma^2(X)]$ is the average conditional label variance. Thus, under realizability, the approximation terms vanish and the estimation error adapts directly to the conditional label variance.

Corollary 10 (Realizable guarantee)**.** *Under the assumptions of Theorem 8, suppose that $(P, \mathcal{F})$ is realizable. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for every $\rho > 0$,*

$$\begin{aligned}\varepsilon_1(\hat{f}) &\le (2+\rho)\sqrt{\frac{2v(\eta)\log(2|\mathcal{F}|^2/\delta)}{n}} + C_\rho \frac{\log(2|\mathcal{F}|^2/\delta)}{n} \\ &\le (2+\rho)\sqrt{\frac{2\bar{\sigma}^2\log(2|\mathcal{F}|^2/\delta)}{n}} + C_\rho \frac{\log(2|\mathcal{F}|^2/\delta)}{n}.\end{aligned} \tag{21}$$

The second bound in (21) recovers, up to constants, the realizable variance-adaptive rate established by Li et al. (2026), where the estimation error for their betting loss scales as $\sqrt{\bar{\sigma}^2\log(|\mathcal{F}|/\delta)/n} + \log(|\mathcal{F}|/\delta)/n$. Moreover, the localized quantity $v(\eta)$ can be smaller than the global average conditional variance $\bar{\sigma}^2$, since it includes label variance only on regions where some predictor in $\mathcal{F}$ disagrees with η. Thus, our Sign-Betting Regression retains the favorable noise adaptation available under realizability while extending the guarantee to misspecified classes.

Favorable misspecification. Finally, because the approximation term in Theorem 8 depends on $B_P(f_0)$ rather than only on $\varepsilon_1(f_0)$, the guarantee can be substantially sharper on favorable misspecified instances. Suppose that an L_1-optimal predictor $f_\star \in \arg\min_{f\in\mathcal{F}} \varepsilon_1(f)$ satisfies $B_P(f_\star) = 0$. Taking $f_0 = f_\star$ gives the following result.

Corollary 11 (Favorable misspecification)**.** *Under the assumptions of Theorem 8, suppose that $f_\star \in \arg\min_{f\in\mathcal{F}} \varepsilon_1(f)$ satisfies $B_P(f_\star) = 0$. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for every $\rho > 0$,*

$$\varepsilon_1(\hat{f}) \le \varepsilon_{1,*} + (2+\rho)\sqrt{\frac{2v(f_\star)\log(2|\mathcal{F}|^2/\delta)}{n}} + C_\rho \frac{\log(2|\mathcal{F}|^2/\delta)}{n}. \tag{22}$$

Thus, even when $\varepsilon_{1,*} > 0$, SBR achieves approximation factor 1 on such misspecified instances while retaining a variance-adaptive estimation term. We give a sufficient condition under which an L_1-optimal predictor satisfies $B_P(f_\star) = 0$ under genuine misspecification in Appendix J.

5 Optimality of the universal proper approximation factor

Corollary 9 shows that SBR achieves approximation factor 3 while retaining the variance-adaptive, comparator-dependent estimation term inherited from Theorem 8. To characterize whether this factor can be improved by a different proper learner, it is useful to first record a coarser consequence in which the estimation term is uniform over the underlying distribution. This places the upper bound in the same distribution-free framework as the lower bound developed below: for any target accuracy $\varepsilon > 0$, the sample size required to guarantee $\varepsilon_1(\hat{f}) \le 3\varepsilon_{1,*} + \varepsilon$ depends only on $|\mathcal{F}|$, ε, and the confidence level δ, and not on the underlying distribution P. We then show that no proper learner can replace the factor 3 by any universal constant strictly smaller than 3.

Corollary 12 (Uniform achievability of approximation factor three)**.** *Under the assumptions of Theorem 8, there exists a universal constant $C > 0$ such that, for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\varepsilon_1(\hat{f}) \le 3\varepsilon_{1,*} + C\sqrt{\frac{\log(2|\mathcal{F}|^2/\delta)}{n}}. \tag{23}$$

We defer the proof to Appendix K.1. Note that, unlike Corollary 9, the estimation term in (23) depends only on n, $|\mathcal{F}|$, and δ, and is therefore distribution-independent. For fixed $\mathcal{F}$ and δ, it vanishes as $n \to \infty$ uniformly over all distributions P. Corollary 12 shows that approximation factor 3 is achievable by a proper learner with a distribution-independent estimation term. We next show that this factor is sharp: no proper learning algorithm can achieve a universal approximation factor strictly smaller than 3, even if the estimation term is allowed to vanish arbitrarily slowly with the sample size.

Theorem 13 (Sharp lower bound for proper learning)**.** *Fix any $C \in [1,3)$. There exist a measurable input space $\mathcal{X}$, a fixed two-element class $\mathcal{F}_C \subseteq \{f : \mathcal{X} \to [0,1]\}$, independent of the sample size, and a constant $\varepsilon_C > 0$ such that the following holds. For every $n \ge 1$ and every possibly randomized*

proper learner $\widehat{f}_n \in \mathcal{F}_C$ based on n i.i.d. observations, there exists a distribution P on $\mathcal{X} \times \{0,1\}$, with finitely supported marginal P_X, such that

$$\mathbb{P}\left(\varepsilon_1(\widehat{f}_n) > C\varepsilon_{1,*} + \varepsilon_C\right) \geq \frac{3}{8}. \tag{24}$$

The probability in (24) *includes any internal randomization of the learner.*

Theorem 13 is a lower bound on proper learning itself, rather than on SBR or on any particular proper learning algorithm. It rules out any uniform high-probability guarantee of the form

$$\varepsilon_1(\widehat{f}_n) \leq C\varepsilon_{1,*} + r_n(\mathcal{F}, \delta), \qquad C \in [0, 3), \tag{25}$$

where the estimation term $r_n(\mathcal{F}, \delta)$ is independent of P and satisfies $r_n(\mathcal{F}, \delta) \to 0$ as $n \to \infty$ for every fixed $\mathcal{F}$ and δ. Notably, the impossibility holds regardless of how slowly the estimation term $r_n(\mathcal{F}, \delta)$ vanishes with n. Together with Corollary 12, this shows that 3 is the smallest universal approximation factor achievable by proper learners under the distribution-free, high-probability criterion considered here. We defer the proof of Theorem 13, together with the formal derivation of the impossibility of (25), to Appendix K.2 and Appendix K.3, respectively.

6 Beyond finite classes

The variance-adaptive guarantee of SBR extends beyond finite hypothesis classes. The extension is governed by the pseudo-dimension of the hypothesis class $\mathcal{F}$ and the VC dimension of its induced pairwise comparison class. Define $\mathcal{C}_{\mathcal{F}} := \{x \mapsto \mathbb{1}\{h(x) > f(x)\} : f, h \in \mathcal{F}\}$, and let $d_{\mathrm{pred}} := \mathrm{Pdim}(\mathcal{F})$ be the pseudo-dimension of $\mathcal{F}$, $d_{\mathrm{cmp}} := \mathrm{VCdim}(\mathcal{C}_{\mathcal{F}})$ be the VC dimension of $\mathcal{C}_{\mathcal{F}}$.

Theorem 14 (Variance-adaptive guarantee beyond finite classes)**.** *Let $\mathcal{F}$ be a class of measurable functions from $\mathcal{X}$ to $[0,1]$, suppose $d_{\mathrm{pred}}, d_{\mathrm{cmp}} < \infty$, and suppose the minimum in* (10) *is attained. Write $d := 1 + d_{\mathrm{pred}} + d_{\mathrm{cmp}}$. For $\rho > 0$, let C_ρ be as in Theorem 8. There exists a universal constant $c > 0$ such that, for every $\delta \in (0,1)$, letting $\alpha_\delta := c\,\frac{d\log(en)+\log(2/\delta)}{n}$, with probability at least $1-\delta$, the following holds simultaneously for every $\rho > 0$:*

$$\varepsilon_1(\hat{f}) \leq \inf_{f_0 \in \mathcal{F}} \left\{\varepsilon_1(f_0) + (2+\rho)B_P(f_0) + (2+\rho)\sqrt{2v(f_0)\alpha_\delta}\right\} + C_\rho\alpha_\delta. \tag{26}$$

Thus, the same approximation terms and comparator-dependent variance adaptation of Theorem 8 continue to hold for infinite classes satisfying these complexity conditions, with the finite-class log-cardinality dependence replaced by a complexity term d, which captures both prediction and comparison complexity of the hypothesis class $\mathcal{F}$. Appendix L provides the proof and verifies the required complexity bounds for bounded linear, logistic, and sparse prediction classes as examples.

7 Conclusion

We studied conditional-mean regression under model misspecification with population L_1 error as the target metric. We showed that several standard regression objectives do not, in general, provide a constant-factor L_1 oracle guarantee under misspecification. In contrast, the min–max pairwise estimator of Baraud & Maillard (2025), which we call *Sign-Betting Regression*, admits high-probability L_1 oracle inequalities with comparator-dependent estimation terms. Under realizability, these bounds recover the variance-adaptive rates established by Li et al. (2026). We further showed that the resulting approximation factor 3 is optimal for proper learning: no proper learner can achieve any universal approximation factor strictly below 3 in a distribution-free high-probability L_1 oracle inequality with a vanishing estimation term. We also extended the upper bound beyond finite classes using prediction and pairwise-comparison complexity measures. Experiments in Appendix M illustrate the behavior of Sign-Betting Regression in synthetic and real-data model-selection settings. Together, these results provide a sharp statistical characterization of proper L_1 conditional-mean learning under misspecification.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code generation and debugging for parts of the experimental code, checking the correctness of proofs, checking minor technical details such as constants and algebraic calculations, and polishing the writing for clarity and presentation. All LLM-generated or AI-assisted code was manually reviewed, tested, and verified by the authors before use. All AI-assisted mathematical suggestions and proof checks were independently reviewed and verified by the authors, and all AI-assisted writing was reviewed and revised as needed. The authors take full responsibility for the correctness, originality, and final content of this work, including any text, claims, code, or other artifacts produced or revised with the aid of generative AI.

REFERENCES

Mehmet Eren Ahsen and Mathukumalli Vidyasagar. An approach to one-bit compressed sensing based on probably approximately correct learning theory. *Journal of Machine Learning Research*, 20(11):1–23, 2019.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608 – 633, 2007. doi: 10.1214/009053606000001217. URL https://doi.org/10.1214/009053606000001217.

Alex Ayoub, David Szepesvári, Alireza Bakhtiari, Csaba Szepesvári, and Dale Schuurmans. Rectifying regression in reinforcement learning. *arXiv preprint arXiv:2510.00885*, 2025.

Yannick Baraud and Guillaume Maillard. Estimating a regression function under possible heteroscedastic and heavy-tailed errors. application to shape-restricted regression. *arXiv preprint arXiv:2506.00852*, 2025.

Yannick Baraud, Hélène Halconruy, and Guillaume Maillard. Robust density estimation with the l 1-loss. applications to the estimation of a density on the line satisfying a shape constraint. In *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, volume 61, pp. 2098–2148. Institut Henri Poincaré, 2025.

Olivier Bousquet, Daniel Kane, and Shay Moran. The optimal approximation factor in density estimation. In *Conference on Learning Theory*, pp. 318–341. PMLR, 2019.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. ULTRAFEEDBACK: Boosting language models with scaled AI feedback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9722–9744. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cui24f.html.

Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 18907–18919, 2021.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence

Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh

Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International conference on machine learning*, pp. 1939–1948. PMLR, 2018.

Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. doi: 10.1109/TIP.2020.2967829.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J Foster. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. *arXiv preprint arXiv:2503.21878*, 2025.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

Yinan Li, Sungjoon Yoon, Ethan Huang, and Kwang-Sung Jun. Second-order bounds for $[0, 1]$-valued regression via betting loss. In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 4674–4721. PMLR, 29 Jun–03 Jul 2026. URL https://proceedings.mlr.press/v336/li26a.html.

Alexander Mey and Marco Loog. Consistency and finite sample behavior of binary class probability estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8967–8974, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Alexandre B Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

Vladimir Vapnik. On the uniform convergence of relative frequencies of events to their probabilities. In *Doklady Akademii Nauk USSR*, volume 181, pp. 781–787, 1968.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Jon Wellner et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

Yannis G Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, 13(2):768–774, 1985.

Appendix

Table of Contents

A Downstream implications of L_1 conditional-mean accuracy **15**

B Related work **17**

C Proofs and Additional Results for Squared-Loss ERM **19**
C.1 Proof of Proposition 1 19
C.2 A population lower bound for squared loss 19

D Proofs and Additional Results for Log-Loss ERM **20**
D.1 Proof of Equation (6) 20
D.2 Proof of Proposition 3 21
D.3 Proof of Proposition 4 23
D.4 Failure without uniform boundary control 25

E Why not absolute loss? **26**

F Failure of the original betting loss under misspecification **28**

G Proof of the finite-class variance-adaptive guarantee **31**

H The finite-sample guarantee recovers the population guarantee **32**

I Relation to the fixed-design guarantee of Baraud and Maillard **33**

J A sufficient condition for favorable misspecification **35**

K Proofs for the sharp factor-three characterization **35**
K.1 Proof of the uniform factor-three guarantee 35
K.2 Proof of Theorem 13 36
K.3 Impossibility of sub-three uniform guarantees 38

L Beyond finite classes: proof and examples **38**
L.1 Prediction and comparison complexity 39
L.2 A uniform Bernstein inequality 40
L.3 Proof of Theorem 14 43
L.4 Examples of infinite hypothesis classes 44

M Experiments **46**
M.1 Synthetic Experiment 46
M.2 Real-World Experiment 48
M.3 LLM Experiment 50

A Downstream implications of L_1 conditional-mean accuracy

For completeness, we record several elementary consequences of controlling the L_1 error

$$\varepsilon_1(f) := \mathbb{E}\big[|f(X) - \eta(X)|\big], \qquad \eta(x) := \mathbb{E}[Y \mid X = x].$$

These observations motivate our choice of performance criterion.

Binary classification. Suppose $Y \in \{0,1\}$, so that $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$. Let

$$g_f(x) := \mathbf{1}\{f(x) \geq 1/2\}, \qquad g^\star(x) := \mathbf{1}\{\eta(x) \geq 1/2\}.$$

Then

$$R_{\mathrm{cls}}(g_f) - R_{\mathrm{cls}}(g^\star) \leq 2\varepsilon_1(f), \tag{27}$$

where $R_{\mathrm{cls}}(g) := \mathbb{P}(g(X) \neq Y)$. Indeed, the standard excess-risk identity gives

$$R_{\mathrm{cls}}(g_f) - R_{\mathrm{cls}}(g^\star) = \mathbb{E}\big[|2\eta(X) - 1|\mathbf{1}\{g_f(X) \neq g^\star(X)\}\big].$$

Whenever $g_f(X) \neq g^\star(X)$, the values $f(X)$ and $\eta(X)$ lie on opposite sides of $1/2$, and hence

$$|\eta(X) - 1/2| \leq |\eta(X) - f(X)|.$$

Therefore,

$$|2\eta(X) - 1|\mathbf{1}\{g_f(X) \neq g^\star(X)\} \leq 2|\eta(X) - f(X)|,$$

which proves (27); see also Audibert & Tsybakov (2007).

Calibration and subpopulation prediction errors. Consider the population L_1 calibration error

$$\mathrm{CE}_1(f) := \mathbb{E}\Big[\big|\mathbb{E}[Y \mid f(X)] - f(X)\big|\Big].$$

Since

$$\mathbb{E}[Y \mid f(X)] = \mathbb{E}[\eta(X) \mid f(X)],$$

conditional Jensen's inequality yields

$$\begin{aligned}\mathrm{CE}_1(f) &= \mathbb{E}\Big[\big|\mathbb{E}[\eta(X) - f(X) \mid f(X)]\big|\Big] \\ &\leq \mathbb{E}\,|\eta(X) - f(X)| = \varepsilon_1(f).\end{aligned} \tag{28}$$

More generally, for every measurable subpopulation $G \subseteq \mathcal{X}$,

$$\begin{aligned}\Big|\mathbb{E}\big[(Y - f(X))\mathbf{1}\{X \in G\}\big]\Big| &= \Big|\mathbb{E}\big[(\eta(X) - f(X))\mathbf{1}\{X \in G\}\big]\Big| \\ &\leq \mathbb{E}\big[|\eta(X) - f(X)|\mathbf{1}\{X \in G\}\big] \\ &\leq \varepsilon_1(f).\end{aligned} \tag{29}$$

Equivalently, whenever $\mathbb{P}(X \in G) > 0$,

$$\big|\mathbb{E}[Y - f(X) \mid X \in G]\big| \leq \frac{\varepsilon_1(f)}{\mathbb{P}(X \in G)}.$$

Thus, small L_1 conditional-mean error controls both global calibration error and aggregate residual errors over arbitrary subpopulations. These properties are closely related to the objectives studied in multiaccuracy and multicalibration (Hébert-Johnson et al., 2018; Kim et al., 2019).

Best-of-N selection. We next consider a setting in which f serves as a reward model for selecting among a collection of candidate responses. Let $X_1, \ldots, X_N$ be candidate responses, and assume that each X_i has marginal distribution P_X. Independence among the candidates is not required. Define

$$i^\star \in \arg\max_{i \in [N]} \eta(X_i), \qquad \widehat{i} \in \arg\max_{i \in [N]} f(X_i),$$

with arbitrary tie-breaking. Then

$$\mathbb{E}\big[\eta(X_{i^\star}) - \eta(X_{\widehat{i}})\big] \leq N\varepsilon_1(f). \tag{30}$$

To see this, by the definition of $\widehat{i}$,

$$f(X_{i^\star}) - f(X_{\widehat{i}}) \leq 0.$$

Hence, samplewise,

$$\begin{aligned}\eta(X_{i^\star}) - \eta(X_{\widehat{i}}) &= \eta(X_{i^\star}) - f(X_{i^\star}) + f(X_{i^\star}) - f(X_{\widehat{i}}) + f(X_{\widehat{i}}) - \eta(X_{\widehat{i}}) \\ &\leq |\eta(X_{i^\star}) - f(X_{i^\star})| + |\eta(X_{\widehat{i}}) - f(X_{\widehat{i}})| \\ &\leq \sum_{i=1}^{N} |\eta(X_i) - f(X_i)|.\end{aligned}$$

Taking expectations and using the common marginal distribution of the X_i's gives

$$\mathbb{E}\big[\eta(X_{i^\star}) - \eta(X_{\widehat{i}})\big] \le \sum_{i=1}^{N} \mathbb{E}\,|\eta(X_i) - f(X_i)| = N\varepsilon_1(f).$$

Thus, L_1 accuracy of a reward model yields a direct guarantee for Best-of-N selection. The bound requires only control of the marginal prediction error of each candidate and does not require independence among the candidates. This provides one connection between the regression criterion studied here and recent analyses of inference-time selection with imperfect reward models (Huang et al., 2025).

B Related work

Closest prior work: fixed-design sign-based regression. The closest work to ours is Baraud & Maillard (2025), who introduce the same minimax signed-residual estimator, up to normalization and sign conventions, in a fixed-design regression setting. Their design points are deterministic, while the errors are independent but may be non-identically distributed, heteroscedastic, and heavy-tailed; only a finite moment of order $p \in [1, 2]$ is required. They establish nonasymptotic oracle inequalities for the expected L_1 error evaluated on the observed design points. Their bounds already accommodate model misspecification, incur an approximation factor three, and adapt to averaged noise moments and to comparator-dependent complexity of the associated comparison sets.

Our statistical target and questions are different. We consider i.i.d. random design and study the population error $\|f - \eta\|_{L_1(P_X)}$ rather than error restricted to the observed design points. We establish distribution-free, high-probability population oracle inequalities whose second-order terms retain both directional approximation geometry and disagreement-localized residual moments. In addition, we study the information-theoretic approximation factor achievable by arbitrary proper learners and prove that no proper learner can universally improve on factor three under our high-probability learning criterion. Appendix I gives a more detailed comparison. In particular, conditioning on the random covariates and applying the fixed-design result as a black box, followed by ordinary uniform convergence, yields a population bound with an additional variance-independent $n^{-1/2}$ generalization term. Our direct random-design analysis avoids this crude global penalty through a localized treatment of the empirical-to-population transition.

Conditional-mean estimation and regression under misspecification. Our target,

$$\varepsilon_1(f) = \mathbb{E}\,|f(X) - \eta(X)|, \qquad \eta(x) = \mathbb{E}[Y \mid X = x],$$

should be distinguished from minimizing an observable loss such as $\mathbb{E}\,|Y - f(X)|$. The distinction between a statistical functional and the loss used to elicit it is classical: squared loss elicits the conditional mean, whereas absolute loss elicits a conditional median (Gneiting, 2011). More generally, losses that correctly identify the conditional mean under unrestricted optimization can induce a different approximation geometry once optimization is restricted to a misspecified class. For example, squared-loss minimization projects η onto the model class in squared L_2 distance rather than in L_1 distance.

Finite-sample L_1 estimation of binary class probabilities through empirical risk minimization with probability-estimation losses was studied by Mey & Loog (2021), including a discussion of model misspecification. Their perspective starts from a surrogate loss and studies the probability estimate induced by minimizing that loss. Under misspecification, the resulting population optimizer is therefore determined by the geometry of the chosen surrogate. Our benchmark instead fixes the target metric first and asks to compete directly with

$$\varepsilon_{1,*} = \inf_{f \in \mathcal{F}} \|\eta - f\|_{L_1(P_X)}.$$

This distinction motivates our negative results for standard regression objectives: a loss may be well suited for estimating the conditional mean under realizability while failing to provide any universal constant-factor approximation to the best L_1 predictor in a misspecified class.

First- and second-order regression guarantees. A related line of work studies statistical guarantees that improve when the underlying prediction or decision problem is favorable. In cost-sensitive classification and contextual bandits, Foster & Krishnamurthy (2021) show that log-loss regression

and triangular discrimination can yield first-order guarantees whose leading term adapts to the performance of the optimal decision rule. For bounded regression, Li et al. (2026) introduce a betting loss and establish variance-adaptive, or second-order, guarantees, improving on first-order dependence by adapting directly to label variance without requiring that variance to be known in advance. The present work asks how such favorable second-order behavior interacts with *model misspecification*. In particular, the relevant second-order quantity under misspecification is no longer determined by label variance alone: our bound also reflects the approximation residual and the regions on which competing predictors disagree with a comparator. At the same time, we require linear control of the best L_1 approximation error. As shown in Section 3, these two goals need not be inherited by regression objectives that have favorable guarantees in realizable settings.

Although we use the term *sign betting*, SBR should also be distinguished from the betting loss of Li et al. (2026). The two objectives are different. SBR compares a candidate f against competitors h through signed residual statistics of the form

$$(Y - f(X))\,\mathrm{sign}(h(X) - f(X)),$$

and minimizes the worst such comparison. The resulting population identity directly exposes the L_1 distance between competing predictors. This pairwise comparison structure is what permits the misspecified oracle analysis developed here.

Minimum-distance estimation, pairwise testing, and the factor three. The pairwise-comparison structure of SBR has close antecedents in minimum-distance and testing-based estimation. In density estimation, Yatracos (1985) developed minimum-distance procedures based on sets that distinguish pairs of candidate densities, leading to constant-factor L_1 or total-variation oracle guarantees. Related testing-based ideas underlie more recent robust L_1 density estimators, including Baraud et al. (2025), and form part of the methodological lineage of the regression estimator of Baraud & Maillard (2025). These methods share the principle that, rather than minimizing the target loss directly, one constructs pairwise tests or signed statistics whose expectations reveal which candidate is closer to the target.

The appearance of approximation factor three also has a particularly close parallel in density hypothesis selection. Bousquet et al. (2019) show that, when an estimator is required to output one of the candidate densities, factor three is the optimal universal approximation factor in total variation, even for a class containing only two candidates; if improper outputs are allowed, the optimal factor improves to two. There is a useful metric correspondence between their problem and ours. For a fixed covariate marginal P_X, associate a predictor f with the joint distribution

$$Q_f(dx, dy) := P_X(dx)\,\mathrm{Bernoulli}(f(x))(dy).$$

Then

$$\mathrm{TV}(Q_f, Q_\eta) = \mathbb{E}_{P_X}\,|f(X) - \eta(X)| = \varepsilon_1(f).$$

Thus the same factor-three geometry appears naturally in the two problems.

The statistical models are nevertheless different. Density hypothesis selection assumes access to fully specified candidate distributions, whereas our candidates specify conditional-mean functions and the covariate marginal P_X is itself unknown and sampled as part of the regression problem. Moreover, our lower bound concerns a uniform, high-probability regression oracle inequality with a distribution-independent vanishing remainder. Accordingly, our proper-learning lower bound is proved directly rather than by invoking the density-estimation result. The connection to Bousquet et al. (2019) should therefore be viewed as identifying a common proper-versus-improper approximation phenomenon rather than as an equivalence of the two learning problems.

Calibration, subgroup accuracy, and downstream decision making. Finally, the L_1 conditional-mean criterion has connections to several other prediction objectives. Small $\varepsilon_1(f)$ controls aggregate residuals over every measurable subpopulation, relating it to multiaccuracy (Kim et al., 2019) and multicalibration (Hébert-Johnson et al., 2018). For binary responses it also controls the excess error of the plug-in classifier, while when f serves as a reward model it controls the regret of selecting the largest predicted reward among a finite collection of candidates. The latter connects to recent work on the statistical behavior of Best-of-N and other inference-time selection procedures under imperfect reward models (Huang et al., 2025). These connections motivate the choice of $L_1(P_X)$ error as our primary statistical target; their elementary derivations are collected in Appendix A.

C Proofs and Additional Results for Squared-Loss ERM

C.1 Proof of Proposition 1

We note that a generic uniform-convergence analysis of squared loss followed by the conversion $\varepsilon_1(f) \le \varepsilon_2(f)$ yields only an $O(n^{-1/4})$ statistical term. To show $O(n^{-1/2})$ rate in Proposition 1, our proof proceeds by comparing empirical squared risks pairwise and exploiting the fact that the variance of a squared-loss difference is controlled by the L_2 distance between the corresponding predictors.

Proof. For $f, g \in \mathcal{F}$, define the pairwise loss difference $\Delta_{f,g}(X,Y) := (Y-f(X))^2-(Y-g(X))^2$. Since $Y, f(X), g(X) \in [0,1]$, we have $|\Delta_{f,g}| \le 1$, and

$$\Delta_{f,g}^2 = (f-g)^2(f+g-2Y)^2 \le 4(f-g)^2.$$

Hence $\mathrm{Var}(\Delta_{f,g}) \le 4\|f-g\|_2^2$. A one-sided Bernstein inequality and a union bound over all ordered pairs $(f,g) \in \mathcal{F}^2$ imply that, with probability at least $1-\delta$, simultaneously for every $f, g \in \mathcal{F}$,

$$(P-P_n)\Delta_{f,g} \le 2\sqrt{2}\,\|f-g\|_2 \sqrt{\frac{\log(2|\mathcal{F}|^2/\delta)}{n}} + \frac{4}{3}\frac{\log(2|\mathcal{F}|^2/\delta)}{n}. \tag{31}$$

Let $f_{\mathrm{sq}}^* \in \arg\min_{f\in\mathcal{F}}\|f-\eta\|_2$, so that $\left\|f_{\mathrm{sq}}^*-\eta\right\|_2 = \varepsilon_{2,*}$. By the conditional-mean decomposition in (2), for every $f, g \in \mathcal{F}$,

$$P\Delta_{f,g} = \|f-\eta\|_2^2 - \|g-\eta\|_2^2.$$

Moreover, empirical optimality of $\widehat{f}_{\mathrm{sq}}$ gives $P_n\Delta_{\widehat{f}_{\mathrm{sq}},f_{\mathrm{sq}}^*} \le 0$. Therefore, on the event (31), writing $r := \left\|\widehat{f}_{\mathrm{sq}}-\eta\right\|_2$ and $e := \varepsilon_{2,*}$, we obtain

$$r^2 - e^2 \le 2\sqrt{2}(r+e)\sqrt{\frac{\log(2|\mathcal{F}|^2/\delta)}{n}} + \frac{4}{3}\frac{\log(2|\mathcal{F}|^2/\delta)}{n},$$

where we used $\left\|\widehat{f}_{\mathrm{sq}}-f_{\mathrm{sq}}^*\right\|_2 \le r+e$.

If $r \le e$, the desired conclusion is immediate. Otherwise, let $q := r-e > 0$. Since $r^2-e^2 = q(r+e)$ and $r+e \ge q$, the preceding inequality yields

$$q^2 \le 2\sqrt{2}\,q\sqrt{\frac{\log(2|\mathcal{F}|^2/\delta)}{n}} + \frac{4}{3}\frac{\log(2|\mathcal{F}|^2/\delta)}{n}.$$

Solving this quadratic inequality gives

$$q \le \left(\sqrt{2}+\sqrt{\frac{10}{3}}\right)\sqrt{\frac{\log(2|\mathcal{F}|^2/\delta)}{n}}.$$

Consequently,

$$\left\|\widehat{f}_{\mathrm{sq}}-\eta\right\|_2 \le \varepsilon_{2,*} + \left(\sqrt{2}+\sqrt{\frac{10}{3}}\right)\sqrt{\frac{\log(2|\mathcal{F}|^2/\delta)}{n}}.$$

Since $\varepsilon_1(\widehat{f}_{\mathrm{sq}}) \le \left\|\widehat{f}_{\mathrm{sq}}-\eta\right\|_2$, this proves the first inequality in (3).

For the second inequality, let $f_1^* \in \arg\min_{f\in\mathcal{F}}\varepsilon_1(f)$. Because f_1^* and η are $[0,1]$-valued, $|f_1^*-\eta| \le 1$, and hence $\|f_1^*-\eta\|_2^2 \le \|f_1^*-\eta\|_1 = \varepsilon_{1,*}$. Therefore $\varepsilon_{2,*} \le \sqrt{\varepsilon_{1,*}}$, which completes the proof.

□

C.2 A population lower bound for squared loss

The following generalizes Proposition 2 and implies that the $\sqrt{\varepsilon_{1,*}}$ dependence in Eqn. (3) is unavoidable.

Proposition 15 (Sharpness of the $\sqrt{\varepsilon_{1,*}}$ dependence)**.** *For every $\gamma \in (0,1)$ and every $p \in (0,(1-\gamma)^2)$, there exists a noiseless regression problem and a two-element class $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ such that $\varepsilon_{1,*} = p$, while the unique population squared-loss minimizer f^*_{sq} satisfies*

$$\varepsilon_1(f^*_{\mathrm{sq}}) = (1-\gamma)\sqrt{\varepsilon_{1,*}}. \tag{32}$$

Moreover, squared-loss ERM selects f^*_{sq} eventually almost surely as $n \to \infty$.

Proof. Let $\mathcal{X} = \{a, b\}$ with $\mathbb{P}(X = b) = p$ and $\mathbb{P}(X = a) = 1 - p$, and let $Y = 0$ almost surely. Thus $\eta \equiv 0$. Consider the two predictors $f_{\mathrm{spike}}, f_{\mathrm{flat}} \in [0,1]^{\mathcal{X}}$ defined by $f_{\mathrm{spike}}(a) = 0$, $f_{\mathrm{spike}}(b) = 1$, and $f_{\mathrm{flat}}(x) = (1-\gamma)\sqrt{p}$ for both $x \in \mathcal{X}$. The latter is $[0,1]$-valued because $p < 1$ and $\gamma \in (0,1)$.

Their L_1 approximation errors are

$$\varepsilon_1(f_{\mathrm{spike}}) = p, \qquad \varepsilon_1(f_{\mathrm{flat}}) = (1-\gamma)\sqrt{p}.$$

Since $p < (1-\gamma)^2$, we have $p < (1-\gamma)\sqrt{p}$. Hence f_{spike} is the unique L_1-optimal element of $\mathcal{F} := \{f_{\mathrm{spike}}, f_{\mathrm{flat}}\}$ and $\varepsilon_{1,*} = p$.

On the other hand, because the regression problem is noiseless with $\eta \equiv 0$, population squared risk equals squared L_2 error. For the two predictors,

$$\mathbb{E}\big[(Y - f_{\mathrm{spike}}(X))^2\big] = p, \qquad \mathbb{E}\big[(Y - f_{\mathrm{flat}}(X))^2\big] = (1-\gamma)^2 p < p.$$

Therefore f_{flat} is the unique population squared-loss minimizer, so $f^*_{\mathrm{sq}} = f_{\mathrm{flat}}$. Consequently,

$$\varepsilon_1(f^*_{\mathrm{sq}}) = (1-\gamma)\sqrt{p} = (1-\gamma)\sqrt{\varepsilon_{1,*}},$$

which proves (32).

It remains to verify the statement for empirical squared-loss ERM. The empirical squared loss of f_{flat} is deterministic and equal to $(1-\gamma)^2 p$. For f_{spike}, it is

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i = b\},$$

which converges almost surely to p by the strong law of large numbers. Since $(1-\gamma)^2 p < p$, squared-loss ERM therefore uniquely selects $f_{\mathrm{flat}} = f^*_{\mathrm{sq}}$ for all sufficiently large n, almost surely. □

In particular, taking $\gamma = 1/2$ recovers Proposition 2. More generally, because γ can be arbitrarily small, Proposition 15 shows that the coefficient in front of $\sqrt{\varepsilon_{1,*}}$ can be made arbitrarily close to one. It also implies the absence of any finite universal approximation factor: for any fixed $C < \infty$, choosing, for example, $\gamma = 1/2$ and then $p > 0$ sufficiently small gives $\varepsilon_1(f^*_{\mathrm{sq}}) > C\varepsilon_{1,*}$, and empirical squared-loss ERM converges almost surely to this population minimizer.

D Proofs and Additional Results for Log-Loss ERM

D.1 Proof of Equation (6)

By conditioning on X and using $\mathbb{E}[Y \mid X] = \eta(X)$,

$$\mathbb{E}\big[\ell_{\log}(Y, f(X)) \,\big|\, X\big] = -\eta(X)\log f(X) - (1-\eta(X))\log(1-f(X)),$$
$$\mathbb{E}\big[\ell_{\log}(Y, \eta(X)) \,\big|\, X\big] = -\eta(X)\log \eta(X) - (1-\eta(X))\log(1-\eta(X)).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\big[\ell_{\log}(Y, f(X)) - \ell_{\log}(Y, \eta(X)) \,\big|\, X\big] \\
&\quad = \eta(X)\log\frac{\eta(X)}{f(X)} + (1-\eta(X))\log\frac{1-\eta(X)}{1-f(X)} \\
&\quad = \mathrm{kl}\big(\eta(X), f(X)\big).
\end{aligned}$$

Taking expectation over X proves (6).

The same calculation remains valid when $\eta(X) \in \{0,1\}$, with the standard convention $0\log 0 = 0$.

D.2 Proof of Proposition 3

Proof. Let $M := |\mathcal{F}|$ and define

$$\alpha := \frac{\log(2M^2/\delta)}{n}.$$

For $f \in \mathcal{F}$, write

$$K(f) := \mathbb{E}\Big[\mathrm{kl}\big(\eta(X), f(X)\big)\Big],$$

and define the squared Hellinger-type distance

$$H^2(f,\eta) := \mathbb{E}\Big[\big(\sqrt{f(X)} - \sqrt{\eta(X)}\big)^2 + \big(\sqrt{1-f(X)} - \sqrt{1-\eta(X)}\big)^2\Big].$$

The elementary relation between total variation and Hellinger distance for Bernoulli distributions implies

$$\varepsilon_1(f) = \mathbb{E}\,|f(X) - \eta(X)| \le H(f,\eta). \tag{33}$$

We first relate the Hellinger distance of the empirical minimizer to the empirical excess log loss. For $f \in \mathcal{F}$, let

$$\Delta_f(X,Y) := \ell_{\log}(Y, f(X)) - \ell_{\log}(Y, \eta(X)).$$

Fix $X = x$, and abbreviate $p = \eta(x)$ and $q = f(x)$. When $p \in (0,1)$,

$$\exp\left(-\frac{1}{2}\Delta_f(x,y)\right) = \left(\frac{q}{p}\right)^{y/2}\left(\frac{1-q}{1-p}\right)^{(1-y)/2}.$$

The right-hand side is a convex function of $y \in [0,1]$. Since $\mathbb{E}[Y \mid X = x] = p$, convexity and the chord joining its values at 0 and 1 give

$$\begin{aligned}\mathbb{E}\Big[e^{-\Delta_f/2}\,\Big|\,X = x\Big] &\le p\sqrt{\frac{q}{p}} + (1-p)\sqrt{\frac{1-q}{1-p}} \\ &= \sqrt{pq} + \sqrt{(1-p)(1-q)}.\end{aligned}$$

The same inequality holds directly when $p \in \{0,1\}$, since then $Y = p$ almost surely conditional on $X = x$. Therefore,

$$\mathbb{E}\, e^{-\Delta_f/2} \le 1 - \frac{1}{2}H^2(f,\eta) \le \exp\left(-\frac{1}{2}H^2(f,\eta)\right).$$

By independence,

$$\mathbb{E}\exp\left\{\frac{n}{2}H^2(f,\eta) - \frac{1}{2}\sum_{i=1}^{n}\Delta_f(X_i,Y_i)\right\} \le 1.$$

Hence, by Markov's inequality,

$$\mathbb{P}\left(H^2(f,\eta) > P_n\Delta_f + 2\alpha\right) \le e^{-n\alpha}.$$

A union bound over $f \in \mathcal{F}$ shows that, with probability at least $1 - \delta/2$,

$$H^2(f,\eta) \le P_n\Delta_f + 2\alpha \qquad \text{for every } f \in \mathcal{F}. \tag{34}$$

Indeed, $Me^{-n\alpha} = \delta/(2M) \le \delta/2$.

On this event, empirical optimality of $\widehat{f}_{\log}$ implies that, for every $f_0 \in \mathcal{F}$,

$$H^2(\widehat{f}_{\log}, \eta) \le P_n\Delta_{\widehat{f}_{\log}} + 2\alpha \le P_n\Delta_{f_0} + 2\alpha. \tag{35}$$

It remains to control the empirical excess loss of the comparator. Fix $f_0 \in \mathcal{F}$ and write

$$K_0 := K(f_0), \qquad B := \log\frac{1}{\gamma}.$$

Since $f_0(X) \in [\gamma, 1-\gamma]$, we have

$$\Delta_{f_0} \le B \qquad \text{and} \qquad P\Delta_{f_0} = K_0. \tag{36}$$

We next show that

$$P\Delta_{f_0}^2 \le (B+2)K_0. \tag{37}$$

To prove (37), again condition on $X = x$ and write $p = \eta(x)$ and $q = f_0(x)$. Since $\Delta_{f_0}(x, y)$ is affine in y, its square is convex, and therefore

$$\mathbb{E}\Big[\Delta_{f_0}^2 \,\Big|\, X = x\Big] \le p \log^2 \frac{p}{q} + (1-p)\log^2 \frac{1-p}{1-q}.$$

We use the elementary inequality

$$t \log^2 t \le (B+2)\big(t \log t - t + 1\big), \qquad 0 \le t \le e^B. \tag{38}$$

For completeness, if $t \le 1$, writing $u = -\log t$ and using $e^u \ge 1 + u + u^2/2$ gives

$$t \log^2 t \le 2(t \log t - t + 1).$$

If $t \ge 1$, write $u = \log t \in [0, B]$. Since

$$u - 2 + (u+2)e^{-u} \ge 0,$$

we have

$$u^2 \le (u+2)(u - 1 + e^{-u}) \le (B+2)(u - 1 + e^{-u}),$$

which is equivalent to (38).

Applying (38) to $t = p/q$ and $t = (1-p)/(1-q)$, which are both at most $1/\gamma = e^B$, yields

$$p \log^2 \frac{p}{q} \le (B+2)\left[p \log \frac{p}{q} - p + q\right],$$

$$(1-p)\log^2 \frac{1-p}{1-q} \le (B+2)\left[(1-p)\log \frac{1-p}{1-q} - (1-p) + (1-q)\right].$$

Summing the two inequalities gives

$$\mathbb{E}\Big[\Delta_{f_0}^2 \,\Big|\, X = x\Big] \le (B+2)\mathrm{kl}(p, q).$$

Taking expectation over X proves (37).

A one-sided Bernstein inequality, followed by a union bound over $f_0 \in \mathcal{F}$, now implies that, with probability at least $1 - \delta/2$, simultaneously for every $f_0 \in \mathcal{F}$,

$$P_n \Delta_{f_0} \le K_0 + \sqrt{2(B+2)K_0\alpha} + \frac{2B}{3}\alpha. \tag{39}$$

Here we used $\mathrm{Var}(\Delta_{f_0}) \le P\Delta_{f_0}^2$ together with (37).

Combining (35) and (39), we obtain, simultaneously for every $f_0 \in \mathcal{F}$,

$$H^2(\widehat{f}_{\log}, \eta) \le K_0 + \sqrt{2(B+2)K_0\alpha} + \left(2 + \frac{2B}{3}\right)\alpha. \tag{40}$$

Since

$$2 + \frac{2B}{3} \ge \frac{B+2}{2},$$

the right-hand side of (40) is at most

$$\left(\sqrt{K_0} + \sqrt{\left(2 + \frac{2B}{3}\right)\alpha}\right)^2.$$

Together with (33), this yields

$$\varepsilon_1(\widehat{f}_{\log}) \le \sqrt{K(f_0)} + \sqrt{\left(2 + \frac{2B}{3}\right)\alpha} \qquad \text{for every } f_0 \in \mathcal{F}. \tag{41}$$

Finally, we relate the KL approximation error to the L_1 approximation error. For any $p \in [0,1]$ and $q \in [\gamma, 1-\gamma]$, convexity of $p \mapsto \mathrm{kl}(p,q)$, together with $\mathrm{kl}(q,q) = 0$, gives

$$\mathrm{kl}(p,q) \le |p-q| \max\left\{ \frac{-\log q}{1-q}, \frac{-\log(1-q)}{q} \right\}.$$

Because $q, 1-q \ge \gamma$,

$$\max\left\{ \frac{-\log q}{1-q}, \frac{-\log(1-q)}{q} \right\} \le \frac{\log(1/\gamma)}{1-\gamma}.$$

Hence

$$K(f) \le \frac{\log(1/\gamma)}{1-\gamma} \varepsilon_1(f) \qquad \text{for every } f \in \mathcal{F}. \tag{42}$$

Let $f_* \in \arg\min_{f \in \mathcal{F}} \varepsilon_1(f)$, which exists because $\mathcal{F}$ is finite. Substituting $f_0 = f_*$ into (41) and applying (42) gives

$$\varepsilon_1(\widehat{f}_{\log}) \le \sqrt{\frac{\log(1/\gamma)}{1-\gamma}\, \varepsilon_{1,*}} + \sqrt{\left(2 + \frac{2}{3}\log\frac{1}{\gamma}\right) \frac{\log(2|\mathcal{F}|^2/\delta)}{n}}.$$

This proves (7). For fixed γ, the display immediately implies

$$\varepsilon_1(\widehat{f}_{\log}) \lesssim_\gamma \sqrt{\varepsilon_{1,*}} + \sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}}.$$

□

D.3 Proof of Proposition 4

Proof. Fix $p \in (0,1)$. Let $\mathcal{X} = \{a, b\}$ with

$$\mathbb{P}(X = b) = p, \qquad \mathbb{P}(X = a) = 1 - p,$$

and let

$$Y = \frac{1}{2} \qquad \text{almost surely.}$$

Thus the problem is noiseless and

$$\eta(a) = \eta(b) = \frac{1}{2}.$$

Consider the two hypotheses f_{spike} and f_{flat} defined by

$$f_{\mathrm{spike}}(a) = \frac{1}{2}, \qquad f_{\mathrm{spike}}(b) = \frac{3}{4},$$

and

$$f_{\mathrm{flat}}(x) = \frac{1}{2} + \frac{\sqrt{p}}{4}, \qquad x \in \{a, b\}.$$

Since $p \in (0,1)$, both hypotheses take values in $[1/4, 3/4]$. Let $\mathcal{F} = \{f_{\mathrm{spike}}, f_{\mathrm{flat}}\}$.

Their L_1 errors are

$$\varepsilon_1(f_{\mathrm{spike}}) = p\left|\frac{3}{4} - \frac{1}{2}\right| = \frac{p}{4},$$

and

$$\varepsilon_1(f_{\mathrm{flat}}) = \frac{\sqrt{p}}{4}.$$

Because $p < \sqrt{p}$ for $p \in (0,1)$, f_{spike} is the unique L_1-optimal hypothesis, and hence

$$\varepsilon_{1,*} = \frac{p}{4}. \tag{43}$$

We next compare the population log risks. Since $\eta \equiv 1/2$, the excess population log risk of any f is $\mathbb{E}[\mathrm{kl}(1/2, f(X))]$. For f_{spike},

$$\mathbb{E}\left[\mathrm{kl}\left(\frac{1}{2}, f_{\mathrm{spike}}(X)\right)\right] = p\,\mathrm{kl}\left(\frac{1}{2}, \frac{3}{4}\right) = -\frac{p}{2}\log\frac{3}{4}. \tag{44}$$

For f_{flat}, using

$$4f_{\mathrm{flat}}(1 - f_{\mathrm{flat}}) = 1 - \frac{p}{4},$$

we obtain

$$\mathbb{E}\left[\mathrm{kl}\left(\frac{1}{2}, f_{\mathrm{flat}}(X)\right)\right] = \mathrm{kl}\left(\frac{1}{2}, \frac{1}{2} + \frac{\sqrt{p}}{4}\right) = -\frac{1}{2}\log\left(1 - \frac{p}{4}\right). \tag{45}$$

To compare the two quantities, consider $g(u) := -\log(1 - u/4)$ on $[0, 1]$. The function g is strictly convex and $g(0) = 0$. Therefore, for every $p \in (0, 1)$,

$$-\log\left(1 - \frac{p}{4}\right) = g(p) < pg(1) = -p\log\frac{3}{4}.$$

Combining this inequality with (44)–(45) shows that

$$\mathbb{E}\big[\ell_{\log}(Y, f_{\mathrm{flat}}(X))\big] < \mathbb{E}\big[\ell_{\log}(Y, f_{\mathrm{spike}}(X))\big].$$

Hence f_{flat} is the unique population log-loss minimizer:

$$f^*_{\log} = f_{\mathrm{flat}}.$$

Its L_1 error is therefore

$$\varepsilon_1(f^*_{\log}) = \frac{\sqrt{p}}{4}.$$

Using (43),

$$\frac{\sqrt{p}}{4} = \frac{1}{2}\sqrt{\frac{p}{4}} = \frac{1}{2}\sqrt{\varepsilon_{1,*}},$$

which proves (8).

For completeness, this population obstruction also applies to empirical log-loss ERM. Since $\mathcal{F}$ contains only two hypotheses and both are uniformly bounded away from 0 and 1, their log losses are bounded. The strong law of large numbers therefore implies that their empirical risks converge almost surely to their respective population risks. Because the population inequality above is strict,

$$\widehat{f}_{\log} = f_{\mathrm{flat}}$$

eventually almost surely. Consequently,

$$\varepsilon_1(\widehat{f}_{\log}) \longrightarrow \frac{\sqrt{p}}{4} = \frac{1}{2}\sqrt{\varepsilon_{1,*}} \qquad \text{almost surely.}$$

Finally,

$$\frac{\varepsilon_1(f^*_{\log})}{\varepsilon_{1,*}} = \frac{1}{\sqrt{p}},$$

which diverges as $p \downarrow 0$. Thus, for any finite constant C, one can choose p sufficiently small so that $\varepsilon_1(f^*_{\log}) > C\varepsilon_{1,*}$. Since empirical log-loss ERM eventually selects $f^*_{\log}$ almost surely, it cannot satisfy a universal constant-factor L_1 oracle guarantee of the form (1). This failure occurs even though $\mathcal{F} \subseteq [1/4, 3/4]^{\mathcal{X}}$ and the labels are noiseless. □

D.4 FAILURE WITHOUT UNIFORM BOUNDARY CONTROL

We next show that the restriction that all hypotheses stay uniformly away from 0 and 1 is essential for Proposition 3. Without such a restriction, the failure of log-loss ERM can be substantially stronger than the $\sqrt{\varepsilon_{1,*}}$ phenomenon in Proposition 4: the best L_1 approximation can have arbitrarily small error, while the population log-loss minimizer has L_1 error bounded away from zero.

Proposition 16 (Failure of log-loss ERM without uniform boundary control)**.** *For every $p \in (0, 1/4]$, there exists a noiseless regression problem and a two-element class $\mathcal{F} \subseteq (0,1)^{\mathcal{X}}$ such that*

$$\varepsilon_{1,*} = p\left(\frac{1}{2} - e^{-1/p}\right) \le \frac{p}{2},$$

while the unique population log-loss minimizer $f^*_{\log}$ satisfies

$$\varepsilon_1(f^*_{\log}) = \frac{1}{4}.$$

Consequently, there is no function $\Psi : [0,1] \to [0,\infty)$ satisfying $\Psi(\varepsilon) \to 0$ as $\varepsilon \to 0$ for which population log-loss minimization obeys

$$\varepsilon_1(f^*_{\log}) \le \Psi(\varepsilon_{1,*})$$

uniformly over finite classes $\mathcal{F} \subseteq (0,1)^{\mathcal{X}}$. Moreover, empirical log-loss ERM eventually selects $f^*_{\log}$ almost surely.

Proof. Fix $p \in (0, 1/4]$. Let $\mathcal{X} = \{a, b\}$ with

$$\mathbb{P}(X = b) = p, \qquad \mathbb{P}(X = a) = 1 - p,$$

and let

$$Y = \frac{1}{2} \qquad \text{almost surely.}$$

Thus

$$\eta(a) = \eta(b) = \frac{1}{2},$$

and the regression problem is noiseless.

Consider the two hypotheses f_0 and g given by

$$f_0(a) = \frac{1}{2}, \qquad f_0(b) = e^{-1/p},$$

and

$$g(x) = \frac{3}{4}, \qquad x \in \{a, b\}.$$

Both hypotheses take values strictly between 0 and 1, so all population log risks are finite.

Their L_1 errors are

$$\varepsilon_1(f_0) = p\left(\frac{1}{2} - e^{-1/p}\right) \le \frac{p}{2}$$

and

$$\varepsilon_1(g) = \frac{1}{4}.$$

Since $p \le 1/4$ and $e^{-1/p} > 0$,

$$\varepsilon_1(f_0) < \frac{1}{4} = \varepsilon_1(g),$$

so f_0 is the unique L_1-optimal hypothesis and

$$\varepsilon_{1,*} = p\left(\frac{1}{2} - e^{-1/p}\right). \tag{46}$$

In particular, $\varepsilon_{1,*} \to 0$ as $p \downarrow 0$.

We now compare the population log risks. Since $\eta \equiv 1/2$, their excess log risks are given by

$$K(f) := \mathbb{E}\left[\mathrm{kl}\left(\frac{1}{2}, f(X)\right)\right].$$

For f_0,

$$\begin{aligned} K(f_0) &= p\,\mathrm{kl}\left(\frac{1}{2}, e^{-1/p}\right) \\ &= p\left[\frac{1}{2}\log\frac{1/2}{e^{-1/p}} + \frac{1}{2}\log\frac{1/2}{1-e^{-1/p}}\right] \\ &= \frac{1}{2} - p\log 2 - \frac{p}{2}\log\left(1-e^{-1/p}\right) \\ &\geq \frac{1}{2} - p\log 2. \end{aligned} \tag{47}$$

For g,

$$K(g) = \mathrm{kl}\left(\frac{1}{2}, \frac{3}{4}\right) = \frac{1}{2}\log\frac{4}{3}. \tag{48}$$

For $p \leq 1/4$,

$$\frac{1}{2} - p\log 2 \geq \frac{1}{2} - \frac{1}{4}\log 2 > \frac{1}{2}\log\frac{4}{3}.$$

Combining this with (47) and (48) yields

$$K(f_0) > K(g).$$

Hence g is the unique population log-loss minimizer:

$$f^*_{\log} = g.$$

Therefore,

$$\varepsilon_1(f^*_{\log}) = \frac{1}{4}, \tag{49}$$

whereas $\varepsilon_{1,*} \to 0$ by (46).

This already rules out any uniform population guarantee of the form $\varepsilon_1(f^*_{\log}) \leq \Psi(\varepsilon_{1,*})$ with $\Psi(\varepsilon) \to 0$ as $\varepsilon \to 0$.

Finally, both hypotheses have bounded log loss on this fixed instance, and the population log-risk gap is strictly positive. By the strong law of large numbers, their empirical log risks converge almost surely to their respective population risks. Thus empirical log-loss ERM selects g eventually almost surely, and consequently

$$\varepsilon_1(\widehat{f}_{\log}) \longrightarrow \frac{1}{4} \qquad \text{almost surely.}$$

□

Proposition 16 shows that some uniform control away from 0 and 1 is necessary for an L_1 oracle guarantee of the form in Proposition 3. The failure is not caused by allowing predictions to equal exactly 0 or 1: for every fixed instance in the construction, all predictions lie strictly inside $(0, 1)$ and all log risks are finite. Rather, the problem is that the L_1-optimal hypothesis is allowed to approach the boundary as $\varepsilon_{1,*} \to 0$. In that regime, an arbitrarily small L_1 approximation error can still incur a nonvanishing log-loss penalty, causing log-loss minimization to prefer a hypothesis whose L_1 error stays bounded away from zero.

E WHY NOT ABSOLUTE LOSS?

Since our target is the L_1 error $\|f - \eta\|_1 = \mathbb{E}\,|f(X) - \eta(X)|$, it is natural to ask whether one can simply minimize the empirical absolute loss $\frac{1}{n}\sum_{i=1}^n |Y_i - f(X_i)|$. Despite the superficial similarity in L_1 measure, this approach is fundamentally misaligned with our objective: its population counterpart

$\mathbb{E}\,|Y - f(X)|$ targets the conditional median of $Y \mid X$, whereas our target is the conditional mean $\eta(x) = \mathbb{E}[Y \mid X = x]$.

Indeed, for each x, the conditional absolute risk $t \mapsto \mathbb{E}[|Y - t| \mid X = x]$ is minimized by a conditional median of $Y \mid X = x$. Thus, even with infinite data, absolute-loss minimization may not recover the best approximation to η in L_1. In fact, the failure can occur even under realizability.

Consider the simplest example where $X \equiv x_0$ and $Y \sim \mathrm{Bernoulli}(p)$ for some $0 < p < 1/2$. Let $\mathcal{F} = \{0, p\}$ consist of two constant predictors. Since $\eta(x_0) = p \in \mathcal{F}$, we have $\varepsilon_{1,*} := \inf_{f \in \mathcal{F}} \|f - \eta\|_1 = 0$. On the other hand, for any $t \in [0, 1]$, the population absolute risk is

$$\mathbb{E}\,|Y - t| = p(1 - t) + (1 - p)t = p + (1 - 2p)t.$$

Since $p < 1/2$, this risk is strictly increasing in t, and hence its unique minimizer over $\mathcal{F}$ is 0, rather than the true conditional mean p.

This example highlights the basic distinction between absolute loss and our sign-betting objective. Absolute loss compares a predictor directly with the observed response Y and therefore inherits the median-seeking geometry of absolute deviation. In contrast, sign betting uses pairwise signed residual comparisons designed to recover the geometry of distance to the conditional mean. We next make this distinction more explicit at the population level.

The sign in absolute loss depends on the observed response. Recall that $|Y - f(X)| = (Y - f(X))\,\mathrm{sign}(Y - f(X))$. Thus absolute loss can equivalently be written as $\mathbb{E}[\sup_{s \in [-1,1]} s(Y - f(X))]$. Importantly, the maximizing direction is chosen separately for each realization of (X, Y) and therefore may depend on the observed response Y. Conditioning on $X = x$, the resulting objective is $t \mapsto \mathbb{E}[|Y - t| \mid X = x]$, whose minimizers are exactly the conditional medians of $Y \mid X = x$. Hence the full conditional distribution of $Y \mid X$ matters, rather than only its conditional mean.

Our sign-betting objective instead takes the form

$$\widehat{L}_{\mathrm{SB}}(f) := \sup_{h \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))\,\mathrm{sign}(h(X_i) - f(X_i)).$$

For every fixed pair (f, h), the direction $\mathrm{sign}(h(X) - f(X))$ is a function of X alone and does not depend on the realized response. Consequently,

$$\mathbb{E}[(Y - f(X))\,\mathrm{sign}(h(X) - f(X))] = \mathbb{E}[(\eta(X) - f(X))\,\mathrm{sign}(h(X) - f(X))],$$

where the equality follows from $\mathbb{E}[Y - \eta(X) \mid X] = 0$. Thus the population sign-betting objective

$$B_P(f) := \sup_{h \in \mathcal{F}} \mathbb{E}[(Y - f(X))\,\mathrm{sign}(h(X) - f(X))]$$

depends on the conditional distribution of $Y \mid X$ only through its mean $\eta(X)$.

This difference can also be viewed as a difference in the order of the supremum and expectation. In the scalar case, absolute loss has the form

$$\mathbb{E}\,|Y - f| = \mathbb{E}\left[\sup_{s \in [-1,1]} s(Y - f)\right],$$

where the sign may adapt to each realization of Y. In contrast, if the direction must be chosen before observing the response, then

$$\sup_{s \in [-1,1]} \mathbb{E}[s(Y - f)] = |\,\mathbb{E}[Y] - f|.$$

The former is minimized by a median of Y, whereas the latter is minimized by its mean. Sign betting implements the analogous distinction in regression by restricting the betting direction to be X-measurable.

A direct comparison on the realizable Bernoulli example. Consider again the example at the beginning of this appendix section: $X \equiv x_0$, $Y \sim \mathrm{Bernoulli}(p)$ with $0 < p < 1/2$, and $\mathcal{F} = \{0, p\}$. Here $\eta = p$. As shown at the beginning of this section, absolute loss strictly prefers the predictor 0, since $\mathbb{E}\,|Y - t| = p + (1 - 2p)t$ is increasing in t.

The sign-betting population objective makes the opposite choice. For $f = p$, choosing $h = p$ gives the zero direction, while choosing $h = 0$ gives

$$\mathbb{E}[(Y - p)\,\mathrm{sign}(0 - p)] = -\,\mathbb{E}[Y - p] = 0.$$

Hence $B_P(p) = 0$. For $f = 0$, choosing $h = p$ gives

$$\mathbb{E}[(Y - 0)\,\mathrm{sign}(p - 0)] = \mathbb{E}[Y] = p,$$

and therefore $B_P(0) = p$. Thus sign betting uniquely prefers $p = \eta$, while absolute loss uniquely prefers the conditional median 0. In particular,

$$\arg\min_{f \in \mathcal{F}} \mathbb{E}\,|Y - f(X)| = \{0\}, \qquad \arg\min_{f \in \mathcal{F}} B_P(f) = \{\eta\}.$$

More generally, whenever $\eta \in \mathcal{F}$, we have $B_P(\eta) = 0$. For any $f \in \mathcal{F}$, using $h = \eta$ as the challenger gives

$$B_P(f) \geq \mathbb{E}[(\eta(X) - f(X))\,\mathrm{sign}(\eta(X) - f(X))] = \|f - \eta\|_1 .$$

Hence η is the unique population minimizer whenever $\|f - \eta\|_1 > 0$ for every $f \neq \eta$. This is fundamentally different from absolute loss, which can fail to recover η even under realizability.

Connection to the dual representation of the L_1 error. The distinction above can also be understood through

$$\|f - \eta\|_1 = \sup_{\substack{s:\mathcal{X}\to[-1,1] \\ s \text{ measurable}}} \mathbb{E}[(\eta(X) - f(X))s(X)] = \sup_{\substack{s:\mathcal{X}\to[-1,1] \\ s \text{ measurable}}} \mathbb{E}[(Y - f(X))s(X)].$$

The optimal direction is $s(X) = \mathrm{sign}(\eta(X) - f(X))$. Our sign-betting objective is a restricted version of this dual formulation, using the comparison directions $\{\mathrm{sign}(h - f) : h \in \mathcal{F}\}$ available from the model class. In contrast, absolute loss uses the response-dependent direction $\mathrm{sign}(Y - f(X))$, which lead to a target of conditional median.

F Failure of the original betting loss under misspecification

Betting loss was introduced in Li et al. (2026) to obtain variance-adaptive guarantees in the realizable setting. Under misspecification, however, it suffers from a similar approximation-error obstruction as squared and log loss: its L_1 error can scale as $\sqrt{\varepsilon_{1,*}}$, rather than linearly in $\varepsilon_{1,*}$. This is not merely a limitation of the existing realizable analysis. The following proposition shows that the original betting loss can have asymptotic L_1 error of order $\sqrt{\varepsilon_{1,*}}$, ruling out any universal constant-factor agnostic L_1 oracle guarantee.

Recall the betting objective of Li et al. (2026). For $f, h \in \mathcal{F}$, $\phi \in [0, n/4]$, and $c \in [0, 1/4]$, define

$$b_{h,f}^{\phi,c}(x) := \overline{\big(\phi(h(x) - f(x))\big)}_{[-c,c]},$$

and

$$\widehat{L}_{\mathrm{bet,n}}(f) := \max_{h \in \mathcal{F}} \max_{\phi \in [0,n/4]} \max_{c \in [0,1/4]} \frac{1}{n} \sum_{i=1}^{n} \log\Big(1 + (Y_i - f(X_i)) b_{h,f}^{\phi,c}(X_i)\Big).$$

The original betting estimator is any empirical minimizer $\hat{f}_{\mathrm{bet}} \in \arg\min_{f \in \mathcal{F}} \widehat{L}_{\mathrm{bet,n}}(f)$. For the proof, it is useful to introduce the corresponding population objective

$$L_{\mathrm{bet,n}}(f) := \max_{h \in \mathcal{F}} \max_{\phi \in [0,n/4]} \max_{c \in [0,1/4]} \mathbb{E}\left[\log\Big(1 + (Y - f(X)) b_{h,f}^{\phi,c}(X)\Big)\right].$$

We keep the subscript n because the admissible range of ϕ depends on the sample size.

Proposition 17. *For every $p \in (0, 1/64]$, there exist a distribution on a two-point input space and a two-element class $\mathcal{F}$ such that $\varepsilon_{1,*} = p$, but every empirical minimizer $\hat{f}_{\mathrm{bet}}$ of the original betting objective satisfies*

$$\left\| \hat{f}_{\mathrm{bet}} - \eta \right\|_1 \longrightarrow \frac{1 - p}{4}\sqrt{p} \qquad \textit{almost surely as } n \to \infty.$$

Consequently, there is no universal constant $C < \infty$ such that the original betting estimator satisfies

$$\left\| \hat{f}_{\mathrm{bet}} - \eta \right\|_1 \le C\varepsilon_{1,*} + o_n(1)$$

over arbitrary misspecified finite classes.

Proof. Fix $p \in (0, 1/64]$ and let $\mathcal{X} = \{a, b\}$, with $\mathbb{P}(X = b) = p$ and $\mathbb{P}(X = a) = 1 - p$. Let

$$Y \mid X = a \sim \mathrm{Bernoulli}(1/2), \qquad Y \mid X = b = 0 \quad \text{almost surely.}$$

Thus $\eta(a) = 1/2$ and $\eta(b) = 0$.

Set $t := \sqrt{p}/4$ and consider the two-element class $\mathcal{F} = \{f_0, g\}$ defined by

	a	b
η	$1/2$	0
f_0	$1/2$	1
g	$1/2 + t$	$0.$

Their L_1 errors are

$$\|f_0 - \eta\|_1 = p, \qquad \|g - \eta\|_1 = (1-p)t = \frac{1-p}{4}\sqrt{p}.$$

Since $p \le 1/64$, we have $4\sqrt{p} \le 1/2 < 1 - p$, and therefore $(1-p)t > p$. Hence f_0 is the unique L_1 oracle and

$$\varepsilon_{1,*} = p.$$

We next show that the population betting objective strictly prefers g.

Step 1: the L_1 oracle has population betting loss larger than $p/5$. For $f = f_0$, choose the challenger $h = g$ and take $\phi = c = 1/4$, which is admissible for every $n \ge 1$. At $X = a$, the bet is $t/4$, whereas at $X = b$ the bet is $-1/4$. Therefore,

$$\begin{aligned} L_{\mathrm{bet,n}}(f_0) &\ge p \log \frac{5}{4} + \frac{1-p}{2} \log\left(1 - \frac{t^2}{64}\right) \\ &= p \log \frac{5}{4} + \frac{1-p}{2} \log\left(1 - \frac{p}{1024}\right). \end{aligned}$$

Using $\log(5/4) \ge 7/32$ and $\log(1-u) \ge -2u$ for $u \in [0, 1/2]$,

$$\begin{aligned} L_{\mathrm{bet,n}}(f_0) &\ge \frac{7p}{32} - \frac{(1-p)p}{1024} \\ &\ge \left(\frac{7}{32} - \frac{1}{1024}\right) p > \frac{p}{5}. \end{aligned}$$

This bound holds for every $n \ge 1$.

Step 2: the worse L_1 predictor has population betting loss at most $p/8$. For $f = g$, the challenger $h = g$ gives betting loss zero, so it remains to consider $h = f_0$. At $X = b$, we have $Y - g(b) = 0$, and hence this point contributes zero regardless of the bet.

At $X = a$, write $u := \min\{\phi t, c\} \in [0, 1/4]$. Since $f_0(a) - g(a) = -t$, the bet equals $-u$. Conditional on $X = a$, the expected log payoff is

$$\begin{aligned} &\frac{1}{2} \log\left(1 + u\left(\frac{1}{2} + t\right)\right) + \frac{1}{2} \log\left(1 - u\left(\frac{1}{2} - t\right)\right) \\ &= \frac{1}{2} \log\left(1 + 2tu - \left(\frac{1}{4} - t^2\right) u^2\right). \end{aligned}$$

Completing the square gives

$$1 + 2tu - \left(\frac{1}{4} - t^2\right) u^2 \le \frac{1}{1 - 4t^2}.$$

Consequently,

$$L_{\mathrm{bet,n}}(g) \le -\frac{1-p}{2}\log(1-4t^2) = -\frac{1-p}{2}\log\left(1-\frac{p}{4}\right).$$

Using $-\log(1-u) \le u/(1-u)$,

$$L_{\mathrm{bet,n}}(g) \le \frac{p(1-p)}{8(1-p/4)} < \frac{p}{8}.$$

Thus, uniformly over n,

$$L_{\mathrm{bet,n}}(f_0) - L_{\mathrm{bet,n}}(g) > \frac{p}{5} - \frac{p}{8} = \frac{3p}{40}.$$

In particular, the population betting objective strictly prefers g, despite f_0 being the unique L_1 oracle.

Step 3: the empirical objective eventually makes the same choice. The distribution of (X, Y) is supported on only three atoms,

$$\{(a,0),(a,1),(b,0)\}.$$

For any $f, h \in \mathcal{F}$ and any admissible (ϕ, c), define

$$\ell_{f,h,\phi,c}(x,y) := \log\left(1 + (y - f(x))b_{h,f}^{\phi,c}(x)\right).$$

Since $|y - f(x)| \le 1$ and $|b_{h,f}^{\phi,c}(x)| \le 1/4$, we have

$$\frac{3}{4} \le 1 + (y - f(x))b_{h,f}^{\phi,c}(x) \le \frac{5}{4},$$

and hence $|\ell_{f,h,\phi,c}(x,y)| \le \log(4/3)$ uniformly over all f, h, ϕ, c and all n.

Let $\widehat{q}_n(z)$ and $q(z)$ denote the empirical and population probabilities of an atom z. Then

$$\sup_{f,h,\phi,c} \left|(P_n - P)\ell_{f,h,\phi,c}\right| \le \log(4/3) \sum_{z\in\{(a,0),(a,1),(b,0)\}} |\widehat{q}_n(z) - q(z)| \longrightarrow 0 \qquad \text{almost surely.}$$

Importantly, this bound is uniform over all admissible betting parameters, so the fact that the range $\phi \in [0, n/4]$ grows with n causes no difficulty.

It follows that

$$\max_{f\in\mathcal{F}} \left|\widehat{L}_{\mathrm{bet,n}}(f) - L_{\mathrm{bet,n}}(f)\right| \longrightarrow 0 \qquad \text{almost surely.}$$

Combining this with the uniform population gap above, we obtain

$$\widehat{L}_{\mathrm{bet,n}}(g) < \widehat{L}_{\mathrm{bet,n}}(f_0)$$

for all sufficiently large n, almost surely. Hence every empirical minimizer satisfies $\hat{f}_{\mathrm{bet}} = g$ eventually almost surely. Therefore,

$$\left\|\hat{f}_{\mathrm{bet}} - \eta\right\|_1 \longrightarrow \|g - \eta\|_1 = \frac{1-p}{4}\sqrt{p} \qquad \text{almost surely.}$$

Finally,

$$\frac{\|g - \eta\|_1}{\varepsilon_{1,*}} = \frac{1-p}{4\sqrt{p}} \longrightarrow \infty \qquad \text{as } p \downarrow 0.$$

Thus, for any fixed $C < \infty$, one can choose $p > 0$ sufficiently small such that $\|g - \eta\|_1 > C\varepsilon_{1,*}$. Since every empirical minimizer converges almost surely to g, no universal bound of the form

$$\left\|\hat{f}_{\mathrm{bet}} - \eta\right\|_1 \le C\varepsilon_{1,*} + o_n(1)$$

can hold over arbitrary misspecified finite classes. □

The obstruction is already present at the population level. The proof establishes more than asymptotic failure of the empirical estimator. For every n, the population objective satisfies

$$L_{\mathrm{bet,n}}(g) < L_{\mathrm{bet,n}}(f_0),$$

with a gap bounded below by a positive multiple of p. Thus the failure is not caused by sampling fluctuations or by insufficient concentration of the empirical betting loss. Even with exact knowledge of the population objective, minimizing the original betting loss selects the predictor with the larger L_1 error on this instance.

Proposition 17 shows that the original betting loss ERM can asymptotically select a predictor with error $\Theta(\sqrt{\varepsilon_{1,*}})$, and this deficiency is not merely an artifact of that proof strategy. This motivates the use of a different objective in the misspecified setting.

G Proof of the finite-class variance-adaptive guarantee

Proof of Theorem 8. Let $M := |\mathcal{F}|$ and, for brevity in the proof, write

$$\alpha := \frac{\log(2M^2/\delta)}{n}.$$

For $f, h \in \mathcal{F}$, define

$$Z_{f,h}(X,Y) := (Y - f(X))\,\mathrm{sign}(h(X) - f(X)), \qquad D_{f,h}(X) := |f(X) - h(X)|.$$

Thus $\widehat{T}(f,h) = P_n Z_{f,h}$ and $T_P(f,h) = PZ_{f,h}$.

We first establish a uniform concentration event. Since $Y, f(X) \in [0,1]$, we have $|Z_{f,h}| \le 1$, and hence $|Z_{f,h} - PZ_{f,h}| \le 2$ and $\mathrm{Var}(Z_{f,h}) \le PZ_{f,h}^2$. A one-sided Bernstein inequality therefore gives, for each fixed pair (f,h),

$$(P_n - P)Z_{f,h} \le \sqrt{2PZ_{f,h}^2\,\alpha} + \frac{4}{3}\alpha \tag{50}$$

except on an event of probability at most $e^{-n\alpha}$.

Similarly, since $0 \le D_{f,h} \le 1$,

$$\mathrm{Var}(D_{f,h}) \le PD_{f,h}^2 \le PD_{f,h}.$$

The lower-tail Bernstein inequality gives

$$(P - P_n)D_{f,h} \le \sqrt{2PD_{f,h}\,\alpha} + \frac{2}{3}\alpha \tag{51}$$

except on an event of probability at most $e^{-n\alpha}$.

There are at most M^2 ordered pairs in each of (50) and (51). Therefore, by a union bound, both inequalities hold simultaneously for every $f, h \in \mathcal{F}$ with probability at least

$$1 - 2M^2 e^{-n\alpha} = 1 - \delta.$$

We work on this event throughout the remainder of the proof.

Fix an arbitrary comparator $f_0 \in \mathcal{F}$. By the definition of $B_P(f_0)$, for every $h \in \mathcal{F}$,

$$PZ_{f_0,h} = T_P(f_0,h) \le B_P(f_0).$$

Moreover,

$$PZ_{f_0,h}^2 = \mathbb{E}\Big[(Y - f_0(X))^2 \mathbf{1}\{h(X) \ne f_0(X)\}\Big] \le v(f_0).$$

Hence (50) implies

$$\sup_{h \in \mathcal{F}} \widehat{T}(f_0,h) \le B_P(f_0) + \sqrt{2v(f_0)\alpha} + \frac{4}{3}\alpha. \tag{52}$$

The key deterministic step is the sample analogue of Lemma 5. Pointwise,

$$Z_{f,h} + Z_{h,f} = |f - h|,$$

and therefore

$$\begin{aligned} P_n|\hat{f} - f_0| &= \widehat{T}(\hat{f}, f_0) + \widehat{T}(f_0, \hat{f}) \\ &\le \sup_{h\in\mathcal{F}} \widehat{T}(\hat{f}, h) + \sup_{h\in\mathcal{F}} \widehat{T}(f_0, h) \\ &\le 2 \sup_{h\in\mathcal{F}} \widehat{T}(f_0, h), \end{aligned} \tag{53}$$

where the last inequality follows from the empirical optimality of $\hat{f}$ in (10). Combining (52) and (53) gives

$$P_n|\hat{f} - f_0| \le 2B_P(f_0) + 2\sqrt{2v(f_0)\alpha} + \frac{8}{3}\alpha. \tag{54}$$

It remains to convert the empirical distance to its population counterpart. Let

$$d := \left\|\hat{f} - f_0\right\|_1 = PD_{\hat{f},f_0}.$$

Because (51) holds simultaneously for all pairs in $\mathcal{F}$, it may be applied to the data-dependent pair $(\hat{f}, f_0)$. Thus

$$d \le P_n|\hat{f} - f_0| + \sqrt{2d\alpha} + \frac{2}{3}\alpha.$$

Using (54),

$$d \le 2B_P(f_0) + 2\sqrt{2v(f_0)\alpha} + \sqrt{2d\alpha} + \frac{10}{3}\alpha. \tag{55}$$

Fix any $\rho > 0$ and set

$$\theta := \frac{\rho}{2+\rho} \in (0, 1).$$

Young's inequality yields

$$\sqrt{2d\alpha} \le \theta d + \frac{\alpha}{2\theta}.$$

Substituting this into (55) and using $1/(1-\theta) = (2+\rho)/2$, we obtain

$$\begin{aligned} d \le (2+\rho)\left(B_P(f_0) + \sqrt{2v(f_0)\alpha}\right) \\ + \left(\frac{13}{3} + \frac{1}{\rho} + \frac{23}{12}\rho\right)\alpha. \end{aligned} \tag{56}$$

Finally, by the triangle inequality,

$$\varepsilon_1(\hat{f}) = \left\|\hat{f} - \eta\right\|_1 \le \varepsilon_1(f_0) + \left\|\hat{f} - f_0\right\|_1.$$

Combining this with (56) gives, for every $f_0 \in \mathcal{F}$,

$$\begin{aligned} \varepsilon_1(\hat{f}) \le \varepsilon_1(f_0) + (2+\rho)B_P(f_0) + (2+\rho)\sqrt{2v(f_0)\alpha} \\ + \left(\frac{13}{3} + \frac{1}{\rho} + \frac{23}{12}\rho\right)\alpha. \end{aligned}$$

The concentration event above is uniform over all pairs in $\mathcal{F}$ and does not depend on either f_0 or ρ. Hence the bound holds simultaneously for every comparator $f_0 \in \mathcal{F}$ and every $\rho > 0$. Taking the infimum over $f_0 \in \mathcal{F}$ and substituting $\alpha = \log(2|\mathcal{F}|^2/\delta)/n$ completes the proof. □

H THE FINITE-SAMPLE GUARANTEE RECOVERS THE POPULATION GUARANTEE

Corollary 18 (Optimized finite-sample bound)**.** *Under the assumptions of Theorem 8, with probability at least* $1 - \delta$*, simultaneously for every* $f_0 \in \mathcal{F}$*,*

$$\varepsilon_1(\hat{f}) \le \varepsilon_1(f_0) + 2B_P(f_0) + r_n(f_0, \delta), \tag{57}$$

where, writing

$$\alpha := \frac{\log(2|\mathcal{F}|^2/\delta)}{n}, \qquad s(f_0) := \sqrt{2v(f_0)\alpha},$$

we may take

$$r_n(f_0, \delta) := 2s(f_0) + \frac{13}{3}\alpha + 2\sqrt{\alpha\left(B_P(f_0) + s(f_0) + \frac{23}{12}\alpha\right)}. \tag{58}$$

Consequently, for every fixed $f_0 \in \mathcal{F}$ and $\delta \in (0,1)$, $r_n(f_0, \delta) \to 0$ as $n \to \infty$.

Proof. Fix $f_0 \in \mathcal{F}$. On the event of Theorem 8, for every $\rho > 0$,

$$\varepsilon_1(\hat{f}) \le \varepsilon_1(f_0) + (2+\rho)B_P(f_0) + (2+\rho)s(f_0) + \left(\frac{13}{3} + \frac{1}{\rho} + \frac{23}{12}\rho\right)\alpha.$$

Rearranging the ρ-dependent terms gives

$$\varepsilon_1(\hat{f}) \le \varepsilon_1(f_0) + 2B_P(f_0) + 2s(f_0) + \frac{13}{3}\alpha + \rho\left(B_P(f_0) + s(f_0) + \frac{23}{12}\alpha\right) + \frac{\alpha}{\rho}.$$

Since the theorem holds simultaneously for every $\rho > 0$, we may choose

$$\rho = \sqrt{\frac{\alpha}{B_P(f_0) + s(f_0) + \frac{23}{12}\alpha}}.$$

Using

$$\inf_{\rho>0}\left\{\rho A + \frac{\alpha}{\rho}\right\} = 2\sqrt{\alpha A}, \qquad A > 0,$$

yields (57)–(58). The same high-probability event is uniform over $f_0 \in \mathcal{F}$, so the result holds simultaneously for all comparators f_0.

Finally, for fixed f_0, $\mathcal{F}$, and δ, we have $\alpha \to 0$ and $s(f_0) = \sqrt{2v(f_0)\alpha} \to 0$. Hence every term in (58) converges to zero. □

Corollary 18 makes precise the connection between the finite-sample result and the population guarantee in Proposition 7. For every fixed comparator f_0, SBR satisfies the same leading population bound $\varepsilon_1(f_0) + 2B_P(f_0)$, up to an estimation term that vanishes with the sample size. The optimizing value of ρ is used only in the analysis and does not need to be known or supplied to SBR.

I Relation to the fixed-design guarantee of Baraud and Maillard

We clarify here the relation between our random-design guarantees and the fixed-design result of Baraud & Maillard (2025). Their estimator coincides with SBR up to normalization when the optimization is exact, but their statistical setting and risk criterion are different. In particular, they consider deterministic design points and control the expected L_1 error evaluated on those design points. Our results instead concern high-probability population error under i.i.d. random design.

We record the consequence obtained by applying their theorem conditionally on the random covariates and then using a standard uniform-convergence argument. This provides a natural black-box baseline against which to compare our direct random-design analysis.

An application of the fixed-design result. Suppose that $\mathcal{F}$ is finite, with $M := |\mathcal{F}|$, and define

$$R(f) := \varepsilon_1(f) = \mathbb{E}\,|f(X) - \eta(X)|, \qquad R_n(f) := \frac{1}{n}\sum_{i=1}^{n}|f(X_i) - \eta(X_i)|.$$

Let

$$\bar{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^{n}\sigma^2(X_i), \qquad \bar{\sigma}^2 := \mathbb{E}[\sigma^2(X)].$$

Condition on $X_{1:n} := (X_1, \dots, X_n)$. The responses then form a fixed-design regression problem,

$$Y_i = \eta(X_i) + \xi_i, \qquad \mathbb{E}[\xi_i \mid X_{1:n}] = 0, \qquad \mathbb{E}[\xi_i^2 \mid X_{1:n}] = \sigma^2(X_i).$$

Hence the $p = 2$ specialization of the oracle inequality of Baraud & Maillard (2025) applies conditionally on $X_{1:n}$.

For a finite class, the comparison-set families appearing in their complexity measure contain at most M sets, so their relevant VC complexity is at most of order $\log M$. Consequently, their theorem implies, for a universal constant $\kappa > 0$,

$$\mathbb{E}\Big[R_n(\hat{f}) \,\Big|\, X_{1:n}\Big] \le 3 \inf_{f\in\mathcal{F}} R_n(f) + \kappa\sqrt{\frac{\bar{\sigma}_n^2 \log M}{n}}. \tag{59}$$

We suppress inessential modifications of the logarithmic complexity term.

Taking expectations over $X_{1:n}$, using

$$\mathbb{E}\left[\inf_{f\in\mathcal{F}} R_n(f)\right] \le \inf_{f\in\mathcal{F}} \mathbb{E}\, R_n(f) = \varepsilon_{1,*},$$

and Jensen's inequality gives

$$\mathbb{E}\, R_n(\hat{f}) \le 3\varepsilon_{1,*} + \kappa\sqrt{\frac{\bar{\sigma}^2 \log M}{n}}. \tag{60}$$

Thus the fixed-design result already yields a variance-sensitive bound for the expected error of SBR evaluated on the observed covariates.

A direct conversion to population risk. To pass from the empirical-design error to population error, define

$$\Delta_n := \sup_{f\in\mathcal{F}} |R_n(f) - R(f)|.$$

Since $\hat{f} \in \mathcal{F}$,

$$R(\hat{f}) \le R_n(\hat{f}) + \Delta_n.$$

Moreover, because $|f(X) - \eta(X)| \in [0, 1]$, standard finite-class uniform convergence gives

$$\mathbb{E}[\Delta_n] \lesssim \sqrt{\frac{\log M}{n}}.$$

Combining this with (60) yields the population bound

$$\mathbb{E}\,\varepsilon_1(\hat{f}) \le 3\varepsilon_{1,*} + \kappa\sqrt{\frac{\bar{\sigma}^2 \log M}{n}} + C\sqrt{\frac{\log M}{n}} \tag{61}$$

for a universal constant C.

The last term in (61) is independent of the conditional label variance. In particular, even in the realizable noiseless case $\varepsilon_{1,*} = \bar{\sigma}^2 = 0$, this direct black-box argument leaves an $O(\sqrt{\log(M)/n})$ population-generalization term.

This calculation highlights the additional issue created by random design: a variance-sensitive fixed-design bound does not by itself provide a variance-sensitive population guarantee. Our analysis does not treat the fixed-design theorem as a black box followed by global uniform convergence. Instead, it analyzes the signed-residual process and the empirical-to-population transition jointly, leading to the localized, high-probability guarantees of Theorem 8. In particular, our bound retains comparator-dependent approximation and second-order quantities and can achieve substantially smaller statistical terms on favorable instances.

J A SUFFICIENT CONDITION FOR FAVORABLE MISSPECIFICATION

Proposition 19 (A sufficient condition for favorable misspecification)**.** *Let*

$$D_{\mathcal{F}} := \{x \in \mathcal{X} : \exists f, h \in \mathcal{F},\ f(x) \neq h(x)\}$$

be the disagreement region of $\mathcal{F}$. Suppose there exists $f_\star \in \mathcal{F}$ such that

$$f_\star(x) = \eta(x) \qquad \textit{for every } x \in D_{\mathcal{F}}.$$

Then $f_\star$ is an L_1-optimal predictor and $B_P(f_\star) = 0$.

Proof. Outside $D_{\mathcal{F}}$, all predictors in $\mathcal{F}$ agree, and hence every $f \in \mathcal{F}$ satisfies $f = f_\star$ on $D_{\mathcal{F}}^c$. On $D_{\mathcal{F}}$, we have $f_\star = \eta$. Therefore, for every $f \in \mathcal{F}$,

$$\varepsilon_1(f) = \varepsilon_1(f_\star) + \|f - f_\star\|_1 \geq \varepsilon_1(f_\star),$$

so $f_\star$ is L_1-optimal.

Moreover, for every $h \in \mathcal{F}$, $\operatorname{sign}(h - f_\star) = 0$ outside $D_{\mathcal{F}}$, whereas $\eta - f_\star = 0$ on $D_{\mathcal{F}}$. Hence

$$T_P(f_\star, h) = \mathbb{E}[(\eta - f_\star)\operatorname{sign}(h - f_\star)] = 0.$$

Taking the supremum over $h \in \mathcal{F}$ gives $B_P(f_\star) = 0$. □

K PROOFS FOR THE SHARP FACTOR-THREE CHARACTERIZATION

K.1 PROOF OF THE UNIFORM FACTOR-THREE GUARANTEE

Proof of Corollary 12. Let

$$\alpha_\delta := \frac{\log(2|\mathcal{F}|^2/\delta)}{n},$$

and fix

$$f_\star \in \arg\min_{f \in \mathcal{F}} \varepsilon_1(f).$$

Since $Y \in [0,1]$ and every $f \in \mathcal{F}$ is $[0,1]$-valued,

$$\varepsilon_{1,*} \leq 1, \qquad B_P(f_\star) \leq \varepsilon_{1,*} \leq 1, \qquad v(f_\star) \leq 1. \tag{62}$$

Suppose first that $\alpha_\delta \leq 1$. Since the event in Theorem 8 holds simultaneously for every $\rho > 0$, we may choose

$$\rho = \sqrt{\alpha_\delta}.$$

Applying Theorem 8 with $f_0 = f_\star$ and using (62) gives

$$\begin{aligned}
\varepsilon_1(\hat{f}) &\leq \varepsilon_{1,*} + (2+\rho)B_P(f_\star) + (2+\rho)\sqrt{2v(f_\star)\alpha_\delta} + C_\rho\alpha_\delta \\
&\leq 3\varepsilon_{1,*} + \rho\varepsilon_{1,*} + (2+\rho)\sqrt{2\alpha_\delta} + \left(\frac{13}{3} + \frac{1}{\rho} + \frac{23}{12}\rho\right)\alpha_\delta \\
&\leq 3\varepsilon_{1,*} + \sqrt{\alpha_\delta} + 3\sqrt{2\alpha_\delta} + \frac{13}{3}\alpha_\delta + \sqrt{\alpha_\delta} + \frac{23}{12}\alpha_\delta^{3/2}.
\end{aligned}$$

Because $\alpha_\delta \leq 1$,

$$\alpha_\delta \leq \sqrt{\alpha_\delta}, \qquad \alpha_\delta^{3/2} \leq \sqrt{\alpha_\delta}.$$

Hence

$$\varepsilon_1(\hat{f}) \leq 3\varepsilon_{1,*} + \left(\frac{33}{4} + 3\sqrt{2}\right)\sqrt{\alpha_\delta}. \tag{63}$$

If instead $\alpha_\delta > 1$, boundedness gives

$$\varepsilon_1(\hat{f}) \leq 1.$$

Since

$$\frac{33}{4} + 3\sqrt{2} > 1 \qquad \text{and} \qquad \sqrt{\alpha_\delta} > 1,$$

the bound in (63) also holds trivially in this case.

Therefore, for all $n \geq 1$ and $\delta \in (0,1)$,

$$\varepsilon_1(\hat{f}) \leq 3\varepsilon_{1,*} + \left(\frac{33}{4} + 3\sqrt{2}\right)\sqrt{\frac{\log(2|\mathcal{F}|^2/\delta)}{n}}. \tag{64}$$

The estimation term in (64) depends only on n, $|\mathcal{F}|$, and δ, and hence is uniform over the underlying distribution P. □

K.2 Proof of Theorem 13

Proof. Fix $C \in [1,3)$. Take $\mathcal{X} = \mathbb{N}$, set $a := (C+1)/(C+5)$, and define

$$f_-(x) \equiv a, \qquad f_+(x) \equiv 1-a, \qquad \mathcal{F}_C := \{f_-, f_+\}.$$

Also set

$$q := \frac{1}{2(1-a)} = \frac{C+5}{8}, \qquad e := \frac{1}{2} - a, \qquad r := \frac{1+a}{1-a} = \frac{C+3}{2}.$$

Then $a \in [1/3, 1/2)$, $q \in (0,1)$, $e > 0$, and $r > C$. Define $g := (r-C)e > 0$ and

$$\varepsilon_C := \frac{g}{2} = \frac{(3-C)^2}{8(C+5)}.$$

Both $\mathcal{F}_C$ and ε_C depend only on C.

Now fix $n \geq 1$ and a possibly randomized proper learner with output $\widehat{f}_n \in \mathcal{F}_C$. All probabilities involving its output include the learner's internal randomization.

For an integer N to be chosen below, construct two priors Π_- and Π_+ over vectors $\theta = (\theta_1, \ldots, \theta_N)$ of conditional means. Under each prior, the coordinates are independent, with

$$\Pi_- : \quad \theta_j = \begin{cases} a, & \text{with probability } q, \\ 1, & \text{with probability } 1-q, \end{cases}$$

$$\Pi_+ : \quad \theta_j = \begin{cases} 1-a, & \text{with probability } q, \\ 0, & \text{with probability } 1-q. \end{cases}$$

For each fixed θ, let P_θ be the distribution defined by

$$X \sim \mathrm{Unif}([N]), \qquad Y \mid X = j \sim \mathrm{Bernoulli}(\theta_j).$$

Its regression function satisfies $\eta_\theta(j) = \theta_j$ for $j \in [N]$. For $f \in \mathcal{F}_C$, write

$$d_\theta(f) := \mathbb{E}_{P_\theta}\,|f(X) - \eta_\theta(X)| = \frac{1}{N}\sum_{j=1}^{N} |f(j) - \theta_j|.$$

Under Π_-, direct calculation gives

$$\mathbb{E}_{\Pi_-}\, d_\theta(f_-) = (1-q)(1-a) = e,$$
$$\mathbb{E}_{\Pi_-}\, d_\theta(f_+) = q(1-2a) + (1-q)a = re.$$

By symmetry, $\mathbb{E}_{\Pi_+}\, d_\theta(f_+) = e$ and $\mathbb{E}_{\Pi_+}\, d_\theta(f_-) = re$.

Let $t := g/[4(1+C)]$. For $s \in \{-,+\}$, let $\bar{s}$ denote the opposite sign and define

$$G_s := \left\{|d_\theta(f_s) - e| \leq t, \quad |d_\theta(f_{\bar{s}}) - re| \leq t\right\}.$$

Each risk is an average of N independent $[0,1]$-valued random variables under the corresponding prior. Hoeffding's inequality and a union bound therefore yield

$$\Pi_s(G_s^c) \leq 4\exp(-2Nt^2).$$

On G_s,

$$\begin{aligned} d_\theta(f_{\bar{s}}) - C d_\theta(f_s) &\geq re - t - C(e+t) \\ &= g - (1+C)t = \frac{3g}{4} > \varepsilon_C. \end{aligned}$$

Since $C \geq 1$, this also implies $d_\theta(f_{\bar{s}}) > d_\theta(f_s)$. Thus f_s is the unique L_1-optimal member of $\mathcal{F}_C$ on G_s, and

$$d_\theta(f_{\bar{s}}) > C \min_{f \in \mathcal{F}_C} d_\theta(f) + \varepsilon_C. \tag{65}$$

We next show that the two priors induce nearly indistinguishable sample distributions. Let

$$Q_s := \int P_\theta^{\otimes n} \, \Pi_s(d\theta), \qquad s \in \{-, +\},$$

be the mixture law of the entire training sample. In particular, Q_s is a mixture of product distributions, not a product of single-observation mixtures. Under both priors,

$$\mathbb{E}_{\Pi_-} \theta_j = qa + (1-q) = \frac{1}{2}, \qquad \mathbb{E}_{\Pi_+} \theta_j = q(1-a) = \frac{1}{2}.$$

Let $D := \{X_1, \ldots, X_n \text{ are all distinct}\}$. Conditional on any distinct input sequence, the sampled coordinates $\theta_{X_1}, \ldots, \theta_{X_n}$ are independent under either prior. After integrating out these coordinates, the conditional label law is $\mathrm{Bernoulli}(1/2)^{\otimes n}$ under both Q_- and Q_+. The input law is also identical under the two mixtures. Hence their restrictions to D coincide.

Writing $\mathrm{TV}(Q_-, Q_+) := \sup_A |Q_-(A) - Q_+(A)|$, we obtain

$$\mathrm{TV}(Q_-, Q_+) \leq Q_-(D^c) \leq \sum_{1 \leq i < j \leq n} \mathbb{P}(X_i = X_j) = \frac{n(n-1)}{2N},$$

where the collision probabilities are computed under the common input law. Choose an integer

$$N \geq \max\left\{1, \, 6n(n-1), \, \frac{\log 48}{2t^2}\right\}.$$

The preceding bounds then give

$$\mathrm{TV}(Q_-, Q_+) \leq \frac{1}{12}, \qquad \Pi_s(G_s^c) \leq \frac{1}{12} \quad (s \in \{-, +\}). \tag{66}$$

Let $\mathbb{P}_s$ denote the joint experiment that first draws $\theta \sim \Pi_s$, then draws the sample $S \sim P_\theta^{\otimes n}$, and includes the learner's internal randomization. For each sample S, let $\phi(S) \in [0, 1]$ be the probability, over this internal randomization, that the learner returns f_+. Properness implies that it returns f_- with probability $1 - \phi(S)$. Consequently,

$$\begin{aligned}
\mathbb{P}_-(\widehat{f}_n = f_+) + \mathbb{P}_+(\widehat{f}_n = f_-) &= \int \phi \, dQ_- + \int (1 - \phi) \, dQ_+ \\
&= 1 + \int \phi \, d(Q_- - Q_+) \\
&\geq 1 - \mathrm{TV}(Q_-, Q_+) \\
&\geq \frac{11}{12}.
\end{aligned}$$

The first inequality follows from $0 \leq \phi \leq 1$ and the definition of total variation; for example, write $\phi(S) = \int_0^1 \mathbf{1}\{\phi(S) > u\} \, du$. Removing the bad environments and using (66) yields

$$\begin{aligned}
\frac{1}{2} \sum_{s \in \{-,+\}} \mathbb{P}_s\Big(\widehat{f}_n = f_{\bar{s}}, \, G_s\Big) &\geq \frac{11}{24} - \frac{1}{2} \sum_{s \in \{-,+\}} \Pi_s(G_s^c) \\
&\geq \frac{11}{24} - \frac{1}{12} = \frac{3}{8}.
\end{aligned} \tag{67}$$

The left-hand side of (67) equals

$$\frac{1}{2} \sum_{s \in \{-,+\}} \int \mathbf{1}_{G_s}(\theta) \, \mathbb{P}_{S \sim P_\theta^{\otimes n}}\Big(\widehat{f}_n = f_{\bar{s}}\Big) \, \Pi_s(d\theta).$$

Since both priors have finite support, this is a finite weighted average. It follows that there exist $s \in \{-,+\}$ and a fixed $\theta \in G_s$ such that

$$\mathbb{P}_{S\sim P_\theta^{\otimes n}}\left(\widehat{f}_n = f_{\bar{s}}\right) \geq \frac{3}{8}.$$

For this θ, (65) implies

$$\mathbb{P}_{S\sim P_\theta^{\otimes n}}\left(d_\theta(\widehat{f}_n) > C \min_{f\in\mathcal{F}_C} d_\theta(f) + \varepsilon_C\right) \geq \frac{3}{8}.$$

Taking $P = P_\theta$ proves (24). The final counterexample is the fixed distribution P_θ, not either sample mixture Q_s; its observations are i.i.d. and its input marginal has finite support. □

The construction above uses the same input marginal $P_X = \mathrm{Unif}([N])$ in the two hard experiments. Hence the proof in fact continues to hold even if the learner is given P_X.

K.3 Impossibility of sub-three uniform guarantees

We now derive the impossibility of (25). Fix $C \in [1,3)$, and let $\mathcal{F}_C$ and $\varepsilon_C > 0$ be supplied by Theorem 13. Suppose, toward a contradiction, that a possibly randomized proper learning rule satisfies

$$\mathbb{P}_{S\sim P^{\otimes n}}\left(\varepsilon_1(\widehat{f}_n) \leq C\varepsilon_{1,*} + r_n(\mathcal{F}_C,\delta)\right) \geq 1-\delta \tag{68}$$

for every $n \geq 1$, every $\delta \in (0,1)$, and every distribution P on $\mathcal{X}\times[0,1]$. Here the errors are evaluated under P relative to $\mathcal{F}_C$, and the learner may depend on δ. Assume that the estimation term is deterministic, independent of P, and satisfies $r_n(\mathcal{F}_C,\delta)\to 0$ for each fixed δ.

Set $\delta = 1/4$. Since $\mathcal{F}_C$ and ε_C are fixed independently of n, there exists n sufficiently large that $r_n(\mathcal{F}_C,1/4) \leq \varepsilon_C$. For this n, (68) implies, for every P,

$$\begin{aligned}&\mathbb{P}_{S\sim P^{\otimes n}}\left(\varepsilon_1(\widehat{f}_n) > C\varepsilon_{1,*} + \varepsilon_C\right)\\ &\quad\leq \mathbb{P}_{S\sim P^{\otimes n}}\left(\varepsilon_1(\widehat{f}_n) > C\varepsilon_{1,*} + r_n(\mathcal{F}_C,1/4)\right) \leq \frac{1}{4}.\end{aligned}$$

This contradicts Theorem 13, which supplies a distribution P for which the first probability is at least $3/8$. The same contradiction applies if the proposed guarantee is required only for all sufficiently large n, provided the sample-size threshold is independent of P.

For $C < 1$, the argument is trivial. Take the singleton class $\mathcal{F} = \{f \equiv 0\}$ and a distribution with $X \equiv 1$ and $Y \equiv 1$. Every proper learner then satisfies $\varepsilon_1(\widehat{f}_n) = \varepsilon_{1,*} = 1$ deterministically. Thus (25) fails with probability one whenever $r_n(\mathcal{F},1/4) < 1 - C$, which holds for all sufficiently large n if $r_n(\mathcal{F},1/4)\to 0$.

Therefore, for every $C \in [0,3)$, no proper learner can satisfy a distribution-free high-probability oracle inequality of the form

$$\varepsilon_1(\widehat{f}_n) \leq C\varepsilon_{1,*} + r_n(\mathcal{F},\delta)$$

with a distribution-independent estimation term satisfying $r_n(\mathcal{F},\delta)\to 0$ for each fixed $\mathcal{F}$ and δ. In particular, the conclusion is unaffected by how slowly the estimation term vanishes with n.

Together with Corollary 12, which establishes such a uniform guarantee with approximation factor 3, this proves that 3 is the smallest universal approximation factor achievable by proper learners under the distribution-free, high-probability criterion considered in Section 5.

L Beyond finite classes: proof and examples

This appendix proves Theorem 14 and verifies its complexity conditions for several standard infinite hypothesis classes. The proof has two ingredients. First, prediction and comparison complexity control the metric entropy of the empirical processes induced by SBR. Second, a uniform Bernstein inequality converts these entropy bounds into variance-sensitive concentration.

Throughout this appendix, covering numbers are taken with centers in the class. We suppress standard measurability issues throughout.

L.1 Prediction and comparison complexity

Recall the pairwise comparison class

$$\mathcal{C}_{\mathcal{F}} := \left\{ x \mapsto \mathbb{1}\{h(x) > f(x)\} : f, h \in \mathcal{F} \right\},$$

and let $d_{\mathrm{pred}} := \mathrm{Pdim}(\mathcal{F})$, $d_{\mathrm{cmp}} := \mathrm{VCdim}(\mathcal{C}_{\mathcal{F}})$, and $d := 1 + d_{\mathrm{pred}} + d_{\mathrm{cmp}}$. For $f, h \in \mathcal{F}$, define

$$Z_{f,h}(x, y) := (y - f(x))\,\mathrm{sign}(h(x) - f(x)), \qquad D_{f,h}(x) := |f(x) - h(x)|,$$

and introduce the corresponding function classes

$$\mathcal{Z} := \{Z_{f,h} : f, h \in \mathcal{F}\}, \qquad \mathcal{D} := \{D_{f,h} : f, h \in \mathcal{F}\}.$$

The next proposition shows that the two complexity measures in Theorem 14 imply uniform polynomial covering bounds for the function classes needed in the proof.

Proposition 20 (Prediction and comparison complexity imply uniform entropy)**.** *Suppose $\mathcal{F} \subseteq \{f : \mathcal{X} \to [0,1]\}$ and $d_{\mathrm{pred}}, d_{\mathrm{cmp}} < \infty$. There exist universal constants $A_0 \geq e$ and $c_0 > 0$ such that, for every probability measure Q on $\mathcal{X} \times [0,1]$ and every $u \in (0,1]$,*

$$\log N\big(u, \mathcal{Z}, L_2(Q)\big) \leq c_0 d \log(A_0/u),$$

and

$$\log N\big(u, \mathcal{D}, L_2(Q_X)\big) \leq c_0 d \log(A_0/u),$$

where Q_X denotes the $\mathcal{X}$-marginal of Q.

Proof. We first record the standard entropy bounds that will be used throughout. Since $\mathrm{Pdim}(\mathcal{F}) = d_{\mathrm{pred}}$, the class $\mathcal{F}$ is a VC-subgraph class with VC index at most $d_{\mathrm{pred}} + 1$ (the shift by one is only a matter of convention). Because all functions in $\mathcal{F}$ take values in $[0,1]$, Theorem 2.6.7 of Wellner et al. (2013) therefore implies that there exist universal constants $a_0 \geq e$ and $c_1 > 0$ such that, for every probability measure R on $\mathcal{X}$ and every $u \in (0,1]$,

$$\log N\big(u, \mathcal{F}, L_2(R)\big) \leq c_1(1 + d_{\mathrm{pred}}) \log(a_0/u). \tag{69}$$

Likewise, since $\mathcal{C}_{\mathcal{F}}$ is a $\{0,1\}$-valued class of VC dimension d_{cmp}, the uniform entropy bound for VC classes of sets (Wellner et al., 2013, Theorem 2.6.4) gives

$$\log N\big(u, \mathcal{C}_{\mathcal{F}}, L_2(R)\big) \leq c_1(1 + d_{\mathrm{cmp}}) \log(a_0/u). \tag{70}$$

The precise universal constants in (69)–(70) will not be important.

We first bound the entropy of $\mathcal{Z}$. For $f, h \in \mathcal{F}$, let

$$c_{h,f}(x) := \mathbb{1}\{h(x) > f(x)\}.$$

Under the convention $\mathrm{sign}(0) = 0$,

$$\mathrm{sign}(h - f) = c_{h,f} - c_{f,h},$$

and hence every $Z_{f,h} \in \mathcal{Z}$ is contained in the larger product class

$$\mathcal{H} := \left\{ (x, y) \mapsto (y - f(x))(c_1(x) - c_2(x)) : f \in \mathcal{F},\ c_1, c_2 \in \mathcal{C}_{\mathcal{F}} \right\}.$$

The map defining $\mathcal{H}$ is Lipschitz in its three functional arguments. Indeed, for $f, f' \in \mathcal{F}$ and $c_1, c_2, c_1', c_2' \in \mathcal{C}_{\mathcal{F}}$,

$$\begin{aligned} &\big|(y - f)(c_1 - c_2) - (y - f')(c_1' - c_2')\big| \\ &\quad \leq |f - f'|\,|c_1 - c_2| + |y - f'|\big(|c_1 - c_1'| + |c_2 - c_2'|\big) \\ &\quad \leq |f - f'| + |c_1 - c_1'| + |c_2 - c_2'|, \end{aligned} \tag{71}$$

where the last inequality uses $f', y \in [0,1]$ and $c_1, c_2 \in \{0,1\}$.

Fix a probability measure Q on $\mathcal{X} \times [0,1]$, and let Q_X denote its $\mathcal{X}$-marginal. Set $\delta := u/6$. Choose δ-covers of $\mathcal{F}$ and $\mathcal{C}_{\mathcal{F}}$ in $L_2(Q_X)$. By (71) and Minkowski's inequality, approximating each of the

three components within δ produces an $L_2(Q)$ error of at most

$$3\delta = \frac{u}{2}.$$

Thus the corresponding product functions form an external $u/2$-cover of $\mathcal{Z}$. If covering centers are required to belong to $\mathcal{Z}$, then for every covering ball that intersects $\mathcal{Z}$ choose one element of that intersection as its new center. By the triangle inequality this produces an internal u-cover with no increase in cardinality. Consequently,

$$N\big(u, \mathcal{Z}, L_2(Q)\big) \le N\big(u/6, \mathcal{F}, L_2(Q_X)\big) N\big(u/6, \mathcal{C}_{\mathcal{F}}, L_2(Q_X)\big)^2. \tag{72}$$

Applying (69)–(70) to (72) gives

$$\begin{aligned}\log N\big(u, \mathcal{Z}, L_2(Q)\big) &\le c_1 \Big[(1 + d_{\mathrm{pred}}) + 2(1 + d_{\mathrm{cmp}})\Big] \log(6a_0/u) \\ &\le 3c_1 d \log(6a_0/u),\end{aligned}$$

since $d = 1 + d_{\mathrm{pred}} + d_{\mathrm{cmp}}$.

It remains to control $\mathcal{D}$. The absolute-value map satisfies

$$\big||f - h| - |f' - h'|\big| \le |f - f'| + |h - h'|.$$

Therefore, taking $u/2$-covers of the two copies of $\mathcal{F}$ and applying Minkowski's inequality yields

$$N\big(u, \mathcal{D}, L_2(Q_X)\big) \le N\big(u/2, \mathcal{F}, L_2(Q_X)\big)^2. \tag{73}$$

Here the resulting centers are themselves of the form $|f' - h'|$ and hence belong to $\mathcal{D}$ when the covers of $\mathcal{F}$ are chosen internally.

Using (69) once more,

$$\log N\big(u, \mathcal{D}, L_2(Q_X)\big) \le 2c_1(1 + d_{\mathrm{pred}}) \log(2a_0/u) \le 2c_1 d \log(2a_0/u).$$

Choosing universal constants, for example $A_0 \ge \max\{e, 6a_0\}$ and $c_0 \ge 3c_1$, proves both asserted bounds. □

L.2 A UNIFORM BERNSTEIN INEQUALITY

We next give the empirical-process inequality used to replace the finite union bound in the proof of Theorem 8. The important feature is that the leading deviation term depends on the individual second moment Pg^2.

Lemma 21 (Uniform Bernstein inequality)**.** *Let $W_1, \dots, W_n$ be i.i.d. from a probability measure P, and let $\mathcal{G}$ be a class of measurable functions $g : \mathcal{W} \to [-1, 1]$. Suppose there exist $A \ge e$ and $\nu \ge 1$ such that, for every probability measure Q on $\mathcal{W}$ and every $u \in (0, 1]$,*

$$\log N\big(u, \mathcal{G}, L_2(Q)\big) \le \nu \log(A/u). \tag{74}$$

For $\zeta \in (0, 1)$, define

$$\beta := \frac{\nu \log(An) + \log(2/\zeta)}{n}.$$

Then, with probability at least $1 - \zeta$,

$$|(P_n - P)g| \le 33 \left(\sqrt{Pg^2\,\beta} + \beta\right) \qquad \textit{simultaneously for all } g \in \mathcal{G}. \tag{75}$$

Proof. The proof has two steps. We first derive a uniform deviation bound for a fixed tuning parameter λ, and then optimize λ on a finite grid.

Step 1: A uniform bound for fixed λ. We begin with the following elementary inequality: for every $a, b \in [-1, 1]$ and $0 < \lambda \le 1/8$,

$$\frac{1}{2} e^{\lambda(a-b) - 8\lambda^2 b^2} + \frac{1}{2} e^{\lambda(b-a) - 8\lambda^2 a^2} \le 1. \tag{76}$$

Indeed, the left-hand side equals

$$e^{-4\lambda^2(a^2+b^2)} \cosh\Big(\lambda(a - b) + 4\lambda^2(a^2 - b^2)\Big).$$

Since $\cosh(t) \le e^{t^2/2}$ and

$$\Big|\lambda(a-b) + 4\lambda^2(a^2-b^2)\Big| = \lambda|a-b|\,|1+4\lambda(a+b)| \le 2\lambda|a-b|,$$

the preceding display is at most

$$\exp\Big(-4\lambda^2(a^2+b^2) + 2\lambda^2(a-b)^2\Big) = e^{-2\lambda^2(a+b)^2} \le 1.$$

Now fix $0 < \lambda \le 1/8$ and $\delta \in (0,1)$, and write

$$H := \nu \log(An).$$

We claim that, with probability at least $1-\delta$,

$$|(P_n - P)g| \le 8\lambda P g^2 + \frac{H + 1 + \log(2/\delta)}{\lambda n} \qquad \text{for every } g \in \mathcal{G}. \tag{77}$$

Let $W_1', \dots, W_n'$ be an independent copy of the sample. This is the usual ghost-sample symmetrization device; see, e.g., Wellner et al. (2013, Section 2.3.1). Define

$$D_\lambda(g) := \lambda \sum_{i=1}^n \big(g(W_i) - g(W_i')\big) - 8\lambda^2 \sum_{i=1}^n g(W_i')^2$$

and

$$Z_\lambda := \sup_{g \in \mathcal{G}} \Big\{ \lambda n (P_n - P)g - 8\lambda^2 n P g^2 \Big\}.$$

Conditional on the original sample,

$$\mathbb{E}_{W'}\, D_\lambda(g) = \lambda n (P_n - P)g - 8\lambda^2 n P g^2.$$

Hence Jensen's inequality gives

$$\begin{aligned} e^{Z_\lambda} &= \sup_{g \in \mathcal{G}} \exp\big\{\mathbb{E}_{W'}\, D_\lambda(g)\big\} \\ &\le \sup_{g \in \mathcal{G}} \mathbb{E}_{W'}\, e^{D_\lambda(g)} \\ &\le \mathbb{E}_{W'} \sup_{g \in \mathcal{G}} e^{D_\lambda(g)}. \end{aligned} \tag{78}$$

Taking expectation over the original sample,

$$\mathbb{E}\, e^{Z_\lambda} \le \mathbb{E}_{W,W'} \left[\sup_{g \in \mathcal{G}} e^{D_\lambda(g)} \right]. \tag{79}$$

Fix the two samples and let

$$Q := \frac{1}{2n} \sum_{i=1}^n \big(\delta_{W_i} + \delta_{W_i'}\big)$$

be their pooled empirical measure. By (74), there exists an $L_2(Q)$-cover $\{h_1, \dots, h_M\} \subseteq \mathcal{G}$ of radius $1/n$ satisfying

$$M \le (An)^\nu = e^H.$$

Thus, for every $g \in \mathcal{G}$, some h_j satisfies

$$\|g - h_j\|_{L_2(Q)} \le \frac{1}{n}, \qquad Q|g - h_j| \le \frac{1}{n}, \tag{80}$$

where the second inequality follows from Cauchy–Schwarz.

Next, independently for each i, swap W_i and W_i' with probability $1/2$. Denote the resulting ordered pair by $(\widetilde{W}_i, \widetilde{W}_i')$, and define

$$D_\lambda^\varepsilon(g) := \lambda \sum_{i=1}^n \big(g(\widetilde{W}_i) - g(\widetilde{W}_i')\big) - 8\lambda^2 \sum_{i=1}^n g(\widetilde{W}_i')^2.$$

The pooled empirical measure Q is unchanged by the swaps, so the same cover applies to every swap configuration.

Let h_j satisfy (80). Since $|g^2 - h_j^2| \le 2|g - h_j|$, we have

$$\begin{aligned}\left|D_\lambda^\varepsilon(g) - D_\lambda^\varepsilon(h_j)\right| &\le \lambda \sum_{i=1}^{n} \Big(|g - h_j|(\widetilde{W}_i) + |g - h_j|(\widetilde{W}_i') \Big) \\ &\quad + 16\lambda^2 \sum_{i=1}^{n} |g - h_j|(\widetilde{W}_i') \\ &\le (2\lambda + 32\lambda^2) n Q|g - h_j| \\ &\le 2\lambda + 32\lambda^2 \le 1. \end{aligned} \tag{81}$$

Consequently,

$$\sup_{g \in \mathcal{G}} e^{D_\lambda^\varepsilon(g)} \le e \sum_{j=1}^{M} e^{D_\lambda^\varepsilon(h_j)}. \tag{82}$$

Conditional on the two samples, fix one cover element h_j. For the ith pair, averaging over the two possible swap configurations gives

$$\frac{1}{2} e^{\lambda(a-b) - 8\lambda^2 b^2} + \frac{1}{2} e^{\lambda(b-a) - 8\lambda^2 a^2}, \qquad a = h_j(W_i), \quad b = h_j(W_i'),$$

which is at most 1 by (76). Since the swaps are independent across i,

$$\mathbb{E}_\varepsilon \, e^{D_\lambda^\varepsilon(h_j)} \le 1. \tag{83}$$

Taking expectation in (82) therefore yields

$$\mathbb{E}_\varepsilon \left[\sup_{g \in \mathcal{G}} e^{D_\lambda^\varepsilon(g)} \right] \le eM \le e^{H+1}. \tag{84}$$

Finally, since each pair (W_i, W_i') consists of two i.i.d. observations, independently swapping its coordinates does not change the joint distribution. Hence

$$\mathbb{E}_{W,W'} \left[\sup_{g \in \mathcal{G}} e^{D_\lambda(g)} \right] = \mathbb{E}_{W,W',\varepsilon} \left[\sup_{g \in \mathcal{G}} e^{D_\lambda^\varepsilon(g)} \right] \le e^{H+1}.$$

Together with (79), this gives

$$\mathbb{E}\, e^{Z_\lambda} \le e^{H+1}. \tag{85}$$

By Markov's inequality,

$$\Pr\big(Z_\lambda > H + 1 + \log(2/\delta)\big) \le \frac{\delta}{2}.$$

Thus, except on an event of probability at most $\delta/2$,

$$(P_n - P)g \le 8\lambda P g^2 + \frac{H + 1 + \log(2/\delta)}{\lambda n} \qquad \forall g \in \mathcal{G}.$$

Applying the same argument to $-\mathcal{G}$ and taking a union bound proves (77).

Step 2: Simultaneously optimize λ. Consider the finite grid

$$\Lambda := \left\{ \frac{j}{8n} : j = 1, \dots, n \right\}.$$

Apply (77) to each $\lambda \in \Lambda$ with $\delta = \zeta/n$, and take a union bound. With probability at least $1 - \zeta$,

$$|(P_n - P)g| \le 8\lambda P g^2 + \frac{H + 1 + \log(2n/\zeta)}{\lambda n} \qquad \forall g \in \mathcal{G}, \quad \forall \lambda \in \Lambda. \tag{86}$$

By definition,

$$n\beta = H + \log(2/\zeta).$$

Moreover, since $A \geq e$ and $\nu \geq 1$,

$$H = \nu \log(An) \geq 1 + \log n.$$

Therefore

$$H + 1 + \log(2n/\zeta) = n\beta + 1 + \log n \leq 2n\beta.$$

Hence, on the same event,

$$|(P_n - P)g| \leq 8\lambda P g^2 + \frac{2\beta}{\lambda} \qquad \forall g \in \mathcal{G}, \quad \forall \lambda \in \Lambda. \tag{87}$$

Fix $g \in \mathcal{G}$ and write

$$v := Pg^2 \in [0, 1].$$

If $v \leq \beta$, choose $\lambda = 1/8$ in (87). Then

$$|(P_n - P)g| \leq v + 16\beta \leq 17\beta.$$

Now suppose $v > \beta$, and define the continuous optimizer

$$\lambda_\star := \frac{1}{8}\sqrt{\frac{\beta}{v}}.$$

Then $\lambda_\star < 1/8$. Also, $\beta \geq 1/n$ and $v \leq 1$ imply

$$\lambda_\star \geq \frac{1}{8\sqrt{n}} \geq \frac{1}{8n}.$$

Let λ be the largest element of Λ satisfying $\lambda \leq \lambda_\star$. Since the grid spacing is $1/(8n)$,

$$\frac{\lambda_\star}{2} \leq \lambda \leq \lambda_\star.$$

Using (87),

$$\begin{aligned} |(P_n - P)g| &\leq 8\lambda_\star v + \frac{4\beta}{\lambda_\star} \\ &= \sqrt{v\beta} + 32\sqrt{v\beta} \\ &= 33\sqrt{v\beta}. \end{aligned}$$

Combining the two cases yields

$$|(P_n - P)g| \leq 33\left(\sqrt{Pg^2\,\beta} + \beta\right) \qquad \forall g \in \mathcal{G}.$$

Since the event in (87) is common to all $g \in \mathcal{G}$, this proves the result. □

L.3 Proof of Theorem 14

We now combine Proposition 20 and Lemma 21 with the same pairwise argument used for the finite-class result.

Proof of Theorem 14. Let

$$\widehat{L}(f) := \sup_{h \in \mathcal{F}} \widehat{T}(f, h).$$

By Proposition 20, the classes $\mathcal{Z}$ and $\mathcal{D}$ satisfy the entropy condition in Lemma 21 with parameters of order d. Applying Lemma 21 separately to $\mathcal{Z}$ and $\mathcal{D}$, each with failure probability $\delta/2$, and increasing the universal constant c in the definition of α_δ if necessary, we obtain an event of probability at least $1 - \delta$ on which, simultaneously for every $f, h \in \mathcal{F}$,

$$(P_n - P)Z_{f,h} \leq \sqrt{2PZ_{f,h}^2\,\alpha_\delta} + \frac{4}{3}\alpha_\delta, \tag{88}$$

$$(P - P_n)D_{f,h} \leq \sqrt{2PD_{f,h}\,\alpha_\delta} + \frac{2}{3}\alpha_\delta. \tag{89}$$

For (89), we used $PD_{f,h}^2 \le PD_{f,h}$, since $0 \le D_{f,h} \le 1$. We work on this event for the remainder of the proof.

Fix any comparator $f_0 \in \mathcal{F}$. By (88), for every $h \in \mathcal{F}$,

$$P_n Z_{f_0,h} \le PZ_{f_0,h} + \sqrt{2PZ_{f_0,h}^2\,\alpha_\delta} + \frac{4}{3}\alpha_\delta.$$

Since $PZ_{f_0,h} \le B_P(f_0)$ and $PZ_{f_0,h}^2 \le v(f_0)$, taking the supremum over h yields

$$\widehat{L}(f_0) \le B_P(f_0) + \sqrt{2v(f_0)\alpha_\delta} + \frac{4}{3}\alpha_\delta. \tag{90}$$

The empirical analogue of Lemma 5 holds exactly:

$$\widehat{T}(f,h) + \widehat{T}(h,f) = P_n|f-h|.$$

Hence, by the empirical optimality of $\hat{f}$,

$$\begin{aligned} P_n|\hat{f} - f_0| &= \widehat{T}(\hat{f}, f_0) + \widehat{T}(f_0, \hat{f}) \\ &\le \widehat{L}(\hat{f}) + \widehat{L}(f_0) \le 2\widehat{L}(f_0). \end{aligned} \tag{91}$$

Let $r := \left\|\hat{f} - f_0\right\|_1$. Since (89) holds uniformly over all pairs, it applies to the data-dependent pair $(\hat{f}, f_0)$:

$$r \le P_n|\hat{f} - f_0| + \sqrt{2r\alpha_\delta} + \frac{2}{3}\alpha_\delta.$$

Combining this inequality with (90) and (91) gives

$$r \le 2B_P(f_0) + 2\sqrt{2v(f_0)\alpha_\delta} + \sqrt{2r\alpha_\delta} + \frac{10}{3}\alpha_\delta. \tag{92}$$

Fix $\rho > 0$ and let $\theta := \rho/(2+\rho)$. By Young's inequality,

$$\sqrt{2r\alpha_\delta} \le \theta r + \frac{\alpha_\delta}{2\theta}.$$

Substituting this into (92) and using $1/(1-\theta) = (2+\rho)/2$ gives

$$r \le (2+\rho)\left(B_P(f_0) + \sqrt{2v(f_0)\alpha_\delta}\right) + C_\rho \alpha_\delta,$$

where C_ρ is the same constant as in Theorem 8.

Finally, by the triangle inequality,

$$\varepsilon_1(\hat{f}) \le \varepsilon_1(f_0) + \left\|\hat{f} - f_0\right\|_1.$$

The concentration event is independent of f_0 and ρ, so the resulting inequality holds simultaneously for every $f_0 \in \mathcal{F}$ and every $\rho > 0$. Taking the infimum over f_0 proves (26). □

L.4 Examples of infinite hypothesis classes

We conclude by verifying the complexity conditions in Theorem 14 for several standard prediction classes. We use standard results on the VC dimension of linear threshold classes and the pseudo-dimension of real-valued linear classes; see, e.g., Vapnik (1968); Anthony & Bartlett (2009); Shalev-Shwartz & Ben-David (2014). Recall that the relevant complexity is $d = 1 + d_{\mathrm{pred}} + d_{\mathrm{cmp}}$.

Bounded linear prediction classes. Let $\phi : \mathcal{X} \to \mathbb{R}^m$ be a fixed measurable feature map and $a : \mathcal{X} \to \mathbb{R}$ a fixed offset. Consider

$$\mathcal{F}_{\mathrm{lin}} := \left\{ f_\theta(x) = a(x) + \langle \theta, \phi(x) \rangle : \theta \in \Theta \right\} \subseteq \{ f : \mathcal{X} \to [0,1] \},$$

where $\Theta \subseteq \mathbb{R}^m$ is any parameter set for which the predictions lie in $[0,1]$.

For a subgraph threshold t,

$$f_\theta(x) > t \quad \Longleftrightarrow \quad \big\langle (\theta, 1), (\phi(x), a(x) - t) \big\rangle > 0.$$

Thus the subgraph class is a subclass of the class of homogeneous halfspaces in $\mathbb{R}^{m+1}$, and hence

$$d_{\text{pred}} \le m + 1.$$

Moreover,

$$f_\gamma(x) > f_\theta(x) \quad \Longleftrightarrow \quad \langle \gamma - \theta, \phi(x) \rangle > 0.$$

Therefore $\mathcal{C}_{\mathcal{F}_{\text{lin}}}$ is a subclass of the class of homogeneous halfspaces in $\mathbb{R}^m$, and hence

$$d_{\text{cmp}} \le m.$$

Consequently, $d = O(m)$, and Theorem 14 applies with

$$\alpha_\delta \lesssim \frac{m \log(en) + \log(2/\delta)}{n}.$$

Logistic and other monotone-link prediction classes. Let $\ell : \mathbb{R} \to (0, 1)$ be a fixed strictly increasing bijection, and consider

$$\mathcal{F}_\ell := \left\{ f_\theta(x) = \ell\big(a(x) + \langle \theta, \phi(x) \rangle\big) : \theta \in \Theta \right\}.$$

This includes logistic prediction with $\ell(z) = (1 + e^{-z})^{-1}$ and probit prediction with $\ell(z) = \Phi(z)$. Since ℓ is strictly increasing, for every $t \in (0, 1)$,

$$f_\theta(x) > t \quad \Longleftrightarrow \quad \Big\langle (\theta, 1), (\phi(x), a(x) - \ell^{-1}(t)) \Big\rangle > 0.$$

Thus the subgraph class is again a subclass of the class of homogeneous halfspaces in $\mathbb{R}^{m+1}$, yielding

$$d_{\text{pred}} \le m + 1.$$

Similarly,

$$f_\gamma(x) > f_\theta(x) \quad \Longleftrightarrow \quad \langle \gamma - \theta, \phi(x) \rangle > 0,$$

so $\mathcal{C}_{\mathcal{F}_\ell}$ is a subclass of the class of homogeneous halfspaces in $\mathbb{R}^m$, and therefore

$$d_{\text{cmp}} \le m.$$

Consequently, $d = O(m)$ and

$$\alpha_\delta \lesssim \frac{m \log(en) + \log(2/\delta)}{n}.$$

Sparse linear and sparse monotone-link prediction classes. Let $\phi : \mathcal{X} \to \mathbb{R}^p$, let $1 \le s \le p$, and restrict parameter vectors to satisfy $\|\theta\|_0 \le s$. Consider either the bounded linear class

$$f_\theta(x) = a(x) + \langle \theta, \phi(x) \rangle$$

or the monotone-link class

$$f_\theta(x) = \ell\big(a(x) + \langle \theta, \phi(x) \rangle\big),$$

with the linear class restricted as needed so that its predictions lie in $[0, 1]$.

There are at most

$$\sum_{j=0}^{s} \binom{p}{j} \le \left(\frac{ep}{s}\right)^s$$

possible supports of size at most s, and conditional on a fixed support the prediction class has pseudo-dimension $O(s)$. Combining the corresponding growth-function bound with a union bound over supports gives

$$d_{\text{pred}} = O\left(s \log \frac{ep}{s}\right).$$

See, e.g., Anthony & Bartlett (2009); Shalev-Shwartz & Ben-David (2014) for the standard growth-function and pseudo-dimension machinery.

For pairwise comparisons, if θ and γ are both s-sparse, then $\gamma - \theta$ is at most $2s$-sparse. Hence $\mathcal{C}_{\mathcal{F}}$ is a subclass of the class of $2s$-sparse homogeneous halfspaces in $\mathbb{R}^p$. Applying the same

support-counting and growth-function argument gives

$$d_{\mathrm{cmp}} = O\left(s \log \frac{ep}{s}\right).$$

Related explicit lower and upper bounds on the VC dimension of sparse halfspaces are given by Ahsen & Vidyasagar (2019). Therefore

$$d = O\left(s \log \frac{ep}{s}\right),$$

and Theorem 14 applies with

$$\alpha_\delta \lesssim \frac{s \log(ep/s) \log(en) + \log(2/\delta)}{n}.$$

Remark 22 (Realizable specialization)**.** *For any of the classes above, if $\eta \in \mathcal{F}$, then $\varepsilon_1(\eta) = B_P(\eta) = 0$ and $v(\eta) \le \bar{\sigma}^2$. Hence Theorem 14 implies, up to constants depending only on ρ,*

$$\varepsilon_1(\hat{f}) \lesssim_\rho \sqrt{\bar{\sigma}^2 \frac{d \log(en) + \log(1/\delta)}{n}} + \frac{d \log(en) + \log(1/\delta)}{n}.$$

Thus the same variance-adaptive interpolation as in the finite-class result continues to hold for these infinite classes.

M Experiments

M.1 Synthetic Experiment

In this section, we provide empirical results that validate our theoretical findings. Our synthetic experiment is a controlled finite-class sanity check. It isolates the approximation behavior predicted by the theory by exactly enumerating a small candidate class. It is not intended to demonstrate scalable optimization of the sign-betting objective over large hypothesis classes.

Function class. We consider a d-dimensional regression setting with target regression function η. We use the logistic prediction family

$$f_\theta(x) := \sigma(x^\top \theta),$$

where $\sigma(z) := 1/(1 + \exp(-z))$ is the sigmoid function. We define the target regression function as

$$\eta(x) = \frac{1}{2}, \qquad \forall x \in \mathcal{X}.$$

To study model misspecification in a controlled setting, we consider a finite function class that does not contain the target regression function. Specifically, we choose two inputs

$$x_c = e_1, \qquad x_r = e_2,$$

where $e_1, e_2 \in \mathbb{R}^d$ are the first two standard basis vectors, and construct two groups of candidate functions.

The first group consists of *local-error* candidates that agree with the target function on x_c but deviate from it on x_r:

$$f_j^L(x_c) = \frac{1}{2}, \qquad f_j^L(x_r) = \frac{1}{2} + A_j.$$

Their corresponding parameters are

$$\theta_j^L = \left(0,\, \mathrm{logit}\left(\frac{1}{2} + A_j\right), 0, \ldots, 0\right).$$

The second group consists of *distributed-error* candidates that deviate from the target function on both inputs:

$$f_j^D(x_c) = f_j^D(x_r) = \frac{1}{2} + B_j.$$

Their corresponding parameters are

$$\theta_j^D = \left(\mathrm{logit}\left(\frac{1}{2} + B_j\right), \mathrm{logit}\left(\frac{1}{2} + B_j\right), 0, \ldots, 0 \right).$$

Let $\mathcal{F}_L$ and $\mathcal{F}_D$ denote the two groups of candidate functions. We finally construct the finite function class as

$$\mathcal{F} := \mathcal{F}_L \cup \mathcal{F}_D.$$

The target regression function η is not contained in $\mathcal{F}$.

Data distribution. We consider a two-point input distribution with

$$x_c = e_1 = (1, 0)^\top, \qquad x_r = e_2 = (0, 1)^\top,$$

and sample

$$X = \begin{cases} x_c, & \text{with probability } 1 - p, \\ x_r, & \text{with probability } p. \end{cases}$$

The target regression function is fixed as

$$\eta(x) = \mathbb{E}[Y \mid X = x] = \frac{1}{2}.$$

We use a parameter $\rho \in [0, 1)$ to control the conditional label noise. This experimental noise parameter is unrelated to the parameter ρ appearing in our theoretical bounds. When $\rho = 0$, the label is deterministic:

$$Y \mid X = x = \frac{1}{2}.$$

For $\rho > 0$, we generate the label as

$$Y \mid X = x \sim \mathrm{Beta}\left(\eta(x)\frac{1-\rho}{\rho}, (1 - \eta(x))\frac{1-\rho}{\rho}\right).$$

Since $\eta(x) = 1/2$, this reduces to

$$Y \mid X = x \sim \mathrm{Beta}\left(\frac{1-\rho}{2\rho}, \frac{1-\rho}{2\rho}\right).$$

In both cases,

$$\mathbb{E}[Y \mid X = x] = \frac{1}{2}, \qquad \mathrm{Var}(Y \mid X = x) = \frac{\rho}{4},$$

where $\rho = 0$ corresponds to the deterministic label setting.

Training. Given the candidate set $\mathcal{F}$, we select a function by minimizing either log loss, squared loss, or sign betting loss on the validation data D_n^{val}. The log loss of f is

$$L_n^{\log}(f) = -\frac{1}{n} \sum_{(x,y) \in D_n} \big(y \log f(x) + (1 - y) \log(1 - f(x)) \big),$$

and the selected function is

$$\widehat{f}^{\log} = \arg\min_{f \in \mathcal{F}} L_n^{\log}(f).$$

Similarly, the squared loss of f is

$$L_n^{\mathrm{sq}}(f) = \frac{1}{n} \sum_{(x,y) \in D_n} \big(y - f(x) \big)^2,$$

and the selected function is

$$\widehat{f}^{\mathrm{sq}} = \arg\min_{f \in \mathcal{F}} L_n^{\mathrm{sq}}(f).$$

Table 1: Comparison of average mean absolute error (MAE) under log loss, squared loss, and sign betting loss across varying n/d. Reported values are the mean MAE over repeated trials, with $\pm$ denoting the standard error. Best results are shown in **bold**. Sign betting loss achieves the lowest MAE for all tested values of n/d, tying with the other methods at $n/d = 2$.

n/d	Log loss	Squared loss	Sign betting loss
2	$\mathbf{0.05320 \pm 0.00082}$	$\mathbf{0.05320 \pm 0.00082}$	$\mathbf{0.05320 \pm 0.00082}$
4	0.06629 ± 0.00086	0.06629 ± 0.00086	$\mathbf{0.04490 \pm 0.00067}$
8	0.07960 ± 0.00068	0.06068 ± 0.00087	$\mathbf{0.04605 \pm 0.00070}$
16	0.08142 ± 0.00063	0.06937 ± 0.00084	$\mathbf{0.03945 \pm 0.00047}$
32	0.08411 ± 0.00054	0.07850 ± 0.00071	$\mathbf{0.03714 \pm 0.00034}$
64	0.08879 ± 0.00026	0.08428 ± 0.00053	$\mathbf{0.03538 \pm 0.00015}$

For sign betting loss, for each $f, h \in \mathcal{F}$, we define

$$\widehat{T}(f,h) = \frac{1}{n} \sum_{(x,y)\in D_n} \big(y - f(x)\big)\,\mathrm{sign}\big(h(x) - f(x)\big)\,,$$

where $\mathrm{sign}(0) = 0$, and select

$$\widehat{f}^{\mathrm{SB}} = \arg\min_{f\in\mathcal{F}} \max_{h\in\mathcal{F}} \widehat{T}(f,h).$$

Evaluation. For each selected function $\widehat{f}$, we evaluate its population mean absolute error (MAE),

$$\varepsilon_1(\widehat{f}) = \mathbb{E}_X\left[\left|\widehat{f}(X) - \eta(X)\right|\right].$$

Since the input distribution is supported on only two points, this expectation can be computed exactly as

$$\varepsilon_1(\widehat{f}) = (1-p)\left|\widehat{f}(x_c) - \frac{1}{2}\right| + p\left|\widehat{f}(x_r) - \frac{1}{2}\right|.$$

We fix the feature dimension to $d = 2$ and the rare-input probability to $p = 0.1$, vary the training sample-to-dimension ratio $n/d \in \{2, 4, 8, 16, 32, 64\}$, and consider different noise levels $\rho \in \{0, 0.02, 0.04\}$. For each configuration $(n/d, \rho)$, we repeat the experiment $B = 1000$ times. If $\widehat{f}_b$ denotes the selected function in repetition b, we report the average MAE

$$\bar{\varepsilon}_1 = \frac{1}{B}\sum_{b=1}^{B} \varepsilon_1(\widehat{f}_b),$$

with standard error

$$\mathrm{SE} = \frac{\mathrm{SD}\left(\varepsilon_1(\widehat{f}_1), \ldots, \varepsilon_1(\widehat{f}_B)\right)}{\sqrt{B}}.$$

Results. Table 1 reports the average MAE with standard error for the noiseless setting ($\rho = 0$), while Figure 1 summarizes the corresponding results across different noise levels $\rho \in \{0, 0.02, 0.04\}$ and varying values of sample-to-dimension ratio n/d. In the noiseless setting, all three methods achieve the same average MAE at the smallest sample size. For all larger values of n/d, sign betting consistently achieves the lowest MAE, and its MAE decreases toward the best population MAE attainable within the function class as the sample size increases. Figure 1 shows that this advantage persists in the presence of label noise. Across both noisy settings, sign betting consistently achieves the lowest MAE over all tested values of n/d, and its MAE decreases as the sample size increases.

M.2 Real-World Experiment

Dataset. We use the KonIQ-10k (Hosu et al., 2020) image-quality dataset, in which each image is associated with multiple human quality ratings on a five-point scale. We rescale each rating

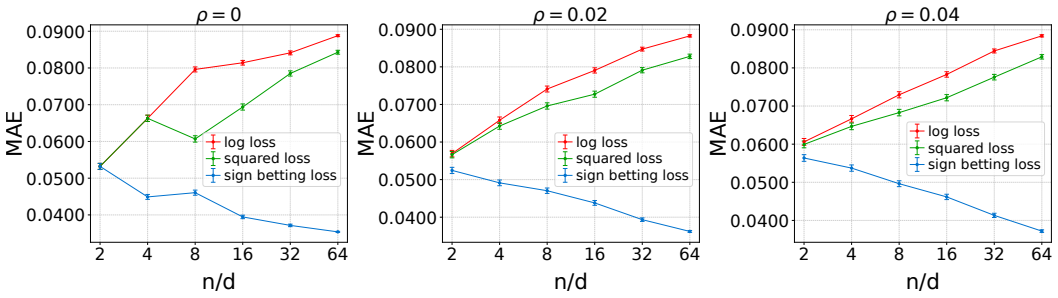


Figure 1: Comparison of average mean absolute error (MAE) obtained under log loss, squared loss, and sign betting loss in the synthetic experiment across varying n/d and noise levels $\rho \in \{0, 0.02, 0.04\}$. Error bars denote the standard error.

$r \in \{1,2,3,4,5\}$ to $Y = \frac{r-1}{4} \in \left\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\right\}$, so that the responses lie in $[0,1]$, consistent with our regression setting.

For an image x, let $c_k(x)$ denote the number of observed ratings at level $k \in \{1,\dots,5\}$. We define the empirical rating distribution by

$$\widehat{q}_k(x) = \frac{c_k(x)}{\sum_{\ell=1}^{5} c_\ell(x)},$$

and its empirical mean by

$$\bar{\eta}(x) = \sum_{k=1}^{5} \widehat{q}_k(x) \frac{k-1}{4}.$$

Here $\eta(x) = \mathbb{E}[Y \mid X = x]$ denotes the population conditional mean from our theoretical formulation, while $\bar{\eta}(x)$ is its empirical counterpart computed from all available human ratings for image x.

Function class. We extract image features using a ResNet-18 (He et al., 2016) backbone pretrained on ImageNet (Russakovsky et al., 2015), with all backbone parameters frozen. Based on these features, we train five families of regression models on the training set using different model configurations and random seeds.Together, these models form a finite hypothesis class containing 38 predictors,

$$\mathcal{F} = \{f_1, \dots, f_{38}\}.$$

Once the candidate models are trained, all model parameters are frozen during model selection.

Training. Given the hypothesis class $\mathcal{F}$, we select a function using log loss, squared loss, or sign betting loss on the validation set. For each repetition, we use the same n distinct validation images for all three methods and sample one human rating for each selected image, forming

$$D_n^{\mathrm{val}} = \{(X_i, Y_i)\}_{i=1}^{n}.$$

The log loss of f is

$$L_n^{\log}(f) = -\frac{1}{n} \sum_{(x,y)\in D_n^{\mathrm{val}}} \left(y \log f(x) + (1-y)\log(1-f(x))\right),$$

and the selected function is

$$\widehat{f}^{\log} = \arg\min_{f\in\mathcal{F}} L_n^{\log}(f).$$

Similarly, the squared loss of f is

$$L_n^{\mathrm{sq}}(f) = \frac{1}{n} \sum_{(x,y)\in D_n^{\mathrm{val}}} \left(y - f(x)\right)^2,$$

and the selected function is

$$\widehat{f}^{\mathrm{sq}} = \arg\min_{f\in\mathcal{F}} L_n^{\mathrm{sq}}(f).$$

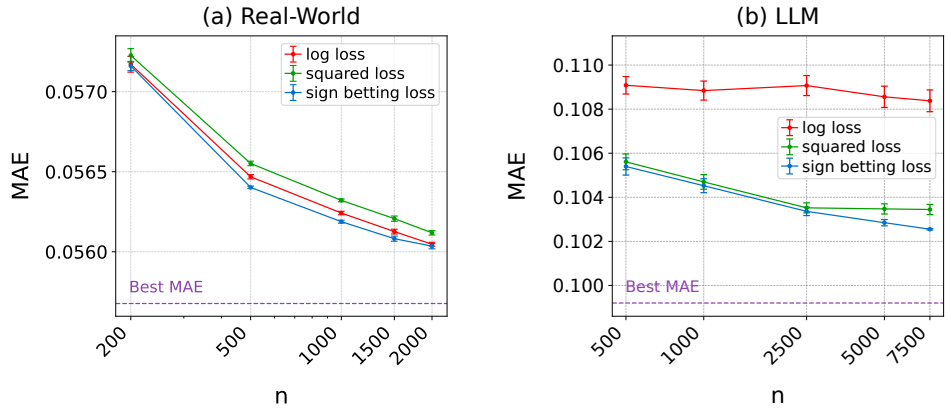


Figure 2: Comparison of average test MAE obtained under log loss, squared loss, and sign betting loss in the real-world and LLM experiments across varying validation sizes n. Error bars denote the standard error, and dashed horizontal lines indicate the MAE of the best model in each function class.

For sign betting, we use

$$\widehat{T}(f,h) = \frac{1}{n} \sum_{(x,y)\in D_n^{\mathrm{val}}} \big(y - f(x)\big)\,\mathrm{sign}\big(h(x) - f(x)\big)\,,$$

and select

$$\widehat{f}^{\mathrm{SB}} = \arg\min_{f\in\mathcal{F}} \max_{h\in\mathcal{F}} \widehat{T}(f,h),$$

where $\mathrm{sign}(0) = 0$.

Evaluation. We evaluate each selected function $\widehat{f}$ on the test set using mean absolute error (MAE) with respect to the empirical mean human score $\bar{\eta}(x)$:

$$\widehat{R}_{\mathrm{test}}(\widehat{f}) = \frac{1}{m} \sum_{i=1}^{m} \left| \widehat{f}(X_i^{\mathrm{test}}) - \bar{\eta}(X_i^{\mathrm{test}}) \right|,$$

where $m = 2015$ is the number of test images.

We also report the model-selection regret relative to the best model in $\mathcal{F}$. Specifically, we define

$$R^{\mathrm{best}} = \min_{f\in\mathcal{F}} \widehat{R}_{\mathrm{test}}(f),$$

and the regret of a selected function $\widehat{f}$ as

$$\mathrm{Regret}(\widehat{f}) = \widehat{R}_{\mathrm{test}}(\widehat{f}) - R^{\mathrm{best}}.$$

Here, R^{best} denotes the test MAE of the best model among the 38 candidates in $\mathcal{F}$.

Results. Figure 2(a) summarizes the average test MAE with standard error from function selection with log loss, squared loss, and sign betting loss across different validation sizes n, together with the test MAE of the best model in each hypothesis class. The results show that the test MAE decreases as the validation size increases for all three selection methods, with sign betting consistently achieving the lowest MAE and approaching the best model as validation size increases. Similarly, the model selection regret decreases with increasing n, with sign betting consistently attaining the lowest regret among the three methods.

M.3 LLM Experiment

Dataset. We use the UltraFeedback (Cui et al., 2024) dataset, which contains prompts paired with multiple responses. In our experiments, each prompt is associated with 4 responses, and we treat each prompt-response pair as one regression sample. Specifically, we represent an input as

$$x = (q, r),$$

where q denotes the prompt and r denotes one of its responses.

To obtain reference rewards, we use Llama-3.3-70B-Instruct (Grattafiori et al., 2024) as the teacher model. Given a prompt-response pair x, the teacher assigns a reward

$$s_{\text{teacher}}(x) \in \{0, 1, 2, 3, 4\}.$$

We normalize the score to $[0, 1]$ and define

$$Y = g(x) = \frac{s_{\text{teacher}}(x)}{4}.$$

Thus, each regression example consists of a prompt-response pair x and its normalized teacher reward Y. We use $g(x)$ as the teacher reference reward for both model selection and evaluation.

Function class. We construct a finite class of candidate reward models using Llama-3.2-3B-Instruct (Grattafiori et al., 2024) as the backbone. We attach a five-class scoring head that predicts

$$p_\theta(k \mid x), \qquad k \in \{0, 1, 2, 3, 4\}.$$

The models are trained on the training set using the teacher reward as labels. We use seven different random seeds and retain five checkpoints along each training trajectory, resulting in $7 \times 5 = 35$ reward models. Although the models are trained as five-class classifiers, each model outputs a prediction reward in $[0, 1]$ using the normalized expected score

$$f_j(x) = \sum_{k=0}^{4} p_j(k \mid x) \frac{k}{4}.$$

We therefore define the finite hypothesis class as

$$\mathcal{F} = \{f_1, \ldots, f_{35}\}.$$

All reward models are frozen before the model selection stage, and the teacher model is not included in $\mathcal{F}$.

Training. Given the candidate set $\mathcal{F}$, we select a function by minimizing either log loss, squared loss, or sign betting loss on the validation data. For each repetition, we sample n prompt-response pairs from the validation set, forming

$$D_n^{\text{val}} = \{(X_i, Y_i)\}_{i=1}^{n}.$$

The log loss of f is

$$L_n^{\log}(f) = -\frac{1}{n} \sum_{(x,y) \in D_n^{\text{val}}} \left[y \log \widetilde{f}(x) + (1 - y) \log(1 - \widetilde{f}(x)) \right],$$

where

$$\widetilde{f}(x) = \text{clip}\Big(f(x), 10^{-6}, 1 - 10^{-6} \Big),$$

and the selected function is

$$\widehat{f}^{\log} = \arg\min_{f \in \mathcal{F}} L_n^{\log}(f).$$

Similarly, the squared loss of f is

$$L_n^{\text{sq}}(f) = \frac{1}{n} \sum_{(x,y) \in D_n^{\text{val}}} \big(y - f(x) \big)^2,$$

and the selected function is

$$\widehat{f}^{\text{sq}} = \arg\min_{f \in \mathcal{F}} L_n^{\text{sq}}(f).$$

For sign betting, we use

$$\widehat{T}(f, h) = \frac{1}{n} \sum_{(x,y) \in D_n^{\text{val}}} \big(y - f(x) \big) \, \text{sign}\big(h(x) - f(x) \big),$$

and select

$$\widehat{f}^{\mathrm{SB}} = \arg\min_{f\in\mathcal{F}} \max_{h\in\mathcal{F}} \widehat{T}(f,h),$$

where $\mathrm{sign}(0) = 0$.

Evaluation. We evaluate each selected reward model $\widehat{f}$ on the test set using mean absolute error (MAE) with respect to the teacher reference reward:

$$R_{\mathrm{test}}(\widehat{f}) = \frac{1}{m}\sum_{i=1}^{m}\left|\widehat{f}(x_i^{\mathrm{test}}) - y_i^{\mathrm{test}}\right|,$$

where the test set contains $m = 4000$ prompt-response pairs and y_i^{test} is the normalized reward assigned by the teacher model. Lower MAE indicates that the selected reward model more closely matches the teacher scoring function on the test set.

We also report the model selection regret relative to the best model in the function class. Specifically, we define

$$R^{\mathrm{best}} = \min_{f\in\mathcal{F}} R_{\mathrm{test}}(f),$$

and

$$\mathrm{Regret}(\widehat{f}) = R_{\mathrm{test}}(\widehat{f}) - R^{\mathrm{best}}.$$

Here, R^{best} is the test MAE of the best model among the 35 candidates in $\mathcal{F}$.

Results. Figure 2(b) summarizes the average test MAE and model selection regret, with standard error, for function selection with log loss, squared loss, and sign betting loss across different validation sizes n. The test MAE of the best model in the function class is also shown as a reference. Across all validation sizes, sign betting consistently achieves lower test MAE than both log loss and squared loss. As the validation size increases, the test MAE of sign betting decreases and moves closer to the best model. Similarly, the model selection regret of sign betting decreases with increasing n and remains the lowest among the three methods across all validation sizes.