Second-Order Bounds for [0,1]-Valued Regression via Betting Loss

Yinan Li YINANLI@ARIZONA.EDU
University of Arizona

Sungjoon Yoon SUNGJOON@ARIZONA.EDU
University of Arizona

Ethan Huang EH3648@NYU.EDU
New York University

Kwang-Sung Jun KWANGSUNGJUN@POSTECH.AC.KR
POSTECH CSE/GSAI

Editors: Steve Hanneke and Tor Lattimore

Abstract

We consider the $[0, 1]$-valued regression problem in the stochastic setting. In a related problem called cost-sensitive classification, Foster and Krishnamurthy (2021) have shown that the log loss minimizer achieves an improved generalization bound compared to that of the squared loss minimizer in the sense that the bound scales with the cost of the best classifier, which can be arbitrarily small depending on the problem instance. Such a result is often called a first-order bound. For $[0, 1]$-valued regression, we first show that the log loss minimizer leads to a similar first-order bound. We then ask if there exists a loss function that achieves a variance-dependent bound (also known as a second-order bound), which is a strict improvement upon first-order bounds. We answer this question in the affirmative by proposing a novel loss function called betting loss. Our result is *variance-adaptive* in the sense that the bound is attained by an algorithm *without any knowledge about the variance*, which is in contrast to the existing works such as weighted least squares with known variances or those that model label variance or its distribution such as distributional reinforcement learning.

Keywords: Learning theory, Regression, Loss functions, Second-order bounds

1. Introduction

We consider the $[0, 1]$-valued regression problem: We are given a dataset $D_n = \{(x_t, y_t)\}_{t=1}^n$where $x_t \in \mathcal{X}$ is the feature of the t-th data point and $y_t \in [0, 1]$ is its label. We assume the data $(x_t, y_t) \sim \mathcal{D}_{X,Y}$ is i.i.d., $\forall t \in [n]$. The goal is to, given a function class $\mathcal{F} \subset \{\mathcal{X} \to [0, 1]\}$, find a function f such that the prediction $f(x)$ is as close as possible to y on average where $(x, y) \sim \mathcal{D}_{X,Y}$.

While being one of the simplest machine learning tasks, this regression task applies to numerous practical applications. First, classification is a special case of this problem where the label space is $\mathcal{Y} = \{0, 1\}$. Second, in Reinforcement Learning (RL), the rewards are typically bounded, and when the episode length is upperbounded, the cumulative reward per episode is also bounded. Thus, in the function approximation setting, one can easily scale the cumulative rewards from each state-action to $[0, 1]$ and perform regression. With this regression, one can construct a policy that chooses the action with the highest predicted value. In goal-oriented RL, regardless of the length of the episode, the rewards are given only at the end of the episode, so, as long as the reward is bounded in a fixed interval, $[0, 1]$-valued regression applies. Finally, human preferences can mostly be expressed as a value in $[0, 1]$. For example, 5-star ratings ($\in \{1, 2, 3, 4, 5\}$) for products can be affine-transformed to $[0, 1]$. Furthermore, datasets commonly used for aligning Large Language Models (LLMs) such

as HelpSteer2 originally contain scores $\{0, 1, 2, 3, 4\}$ (Wang et al., 2024b). Therefore, despite being simple and rather elementary, $[0, 1]$-valued regression is still important, and theoretical and algorithmic advancements can potentially have a huge impact in practice.

However, what do we know about the fundamental performance limits of $[0, 1]$-valued regression? In particular, we ask this question under the realizable setting where the regression function belongs to the given function class $\mathcal{F}$. The standard baseline algorithm for this problem is empirical risk minimization with the squared loss. In the agnostic setting (i.e., no realizability), classic results such as localized Rademacher complexity show that squared loss yields fast rates on the excess loss (Bartlett et al., 2005; Koltchinskii, 2006; Liang et al., 2015). Note that, in the agnostic setting, the loss function is typically a part of the problem definition, and thus existing algorithms naturally use the matching loss function. In the realizable setting, however, the loss function is not part of the problem definition; it is purely an algorithmic choice.

Therefore, in the realizable setting, it is not clear at all if the squared loss is optimal for $[0, 1]$-valued regression. In a related problem called cost-sensitive classification, Foster and Krishnamurthy (2021) have shown that the squared loss is not optimal for $[0, 1]$-valued costs in the realizable setting. Instead, they have shown that the log loss achieves a strictly improved performance bound, a rate that is provably not attained by the squared loss (Foster and Krishnamurthy, 2021, Theorem 2). Specifically, their bound is of the *first-order* type, which means that the performance bound scales with the *magnitude* of the cost/reward being accumulated by the optimal policy. Such a bound is never worse than the standard worst-case bound, yet can be much smaller depending on the problem instance. This has also been called small-loss bound and can be viewed as a problem-dependent accelerated rate.

Such a first-order bound appeared in various machine learning problems (Freund and Schapire, 1997; Foster and Krishnamurthy, 2021; Wagenmaker et al., 2022). In these problems, there is another concept called *second-order* bound (Cesa-Bianchi et al., 2007). While the precise definition can vary across problems, when making stochastic assumptions about how y is related to x, it means that the bound scales with the label's second moment or the variance, which can be much smaller than the magnitude of the label.[1] We elaborate more on second-order bounds in Section 5.

Motivated by the fact that Foster and Krishnamurthy (2021) simply switched a loss function to obtain a first-order bound in cost-sensitive classification, we first report that the same is true in $[0, 1]$-valued regression (see Theorem 1 in Section 2). Given this positive answer, we take a step further and ask the following research question:

Does there exist a loss function whose minimizer leads to a second-order bound?

In this paper, we provide an affirmative answer by proposing a novel loss function inspired by the betting-based confidence bound (Waudby-Smith and Ramdas, 2023; Orabona and Jun, 2024). We emphasize that our algorithm does not require conditional variances as input and allows them to be arbitrarily different depending on x. This is in stark contrast to some existing work that either requires the variance as input (Zhao et al., 2023b) or models variance as part of function approximation (Wang et al., 2024a; Weltz et al., 2023). In some sense, our result shows that obtaining second-order bounds (i.e., adapting to unknown variances) is a free lunch, statistically speaking, in the sense that we do not have to model variance to adapt to it. While there are works that achieve second-order bounds without knowledge of the conditional variances (Zhao et al., 2024; Jun and Kim, 2024; Jia et al.,

1. In the literature outside online learning and reinforcement learning, the so-called 'optimistic rate' can also be seen as a second-order bound. We discuss this in Section 5.

2024; Pacchiano, 2025), the tools therein are specialized for their own contextual bandit problem (in particular, confidence bounds) and do not naturally imply an estimator for $[0,1]$-valued regression.

Our result is primarily statistical. It shows that the true conditional variance can enter the leading finite-sample term through the choice of loss alone, without variance knowledge or explicit variance modeling. This does not by itself resolve the computational problem of efficiently optimizing the proposed min-max objective for large classes. In Section 4, we empirically demonstrate that the proposed loss consistently achieves lower mean absolute error (MAE) than both the log loss and the squared loss.

2. Preliminaries

Notations. We denote $f_x := f(x)$ for any function f and any $x \in \mathcal{X}$. We adopt the nonasymptotic version of $\lesssim$; i.e., $f(x) \lesssim g(x)$ means that there exists a numerical constant $c > 0$, s.t. $f(x) \leq c \cdot g(x),\ \forall x$.

Regression with $[0,1]$-valued labels. We consider the standard supervised learning setting with bounded regression targets. Let $\mathcal{X}$ denote the input space. We observe a dataset $D_n = \{(x_t, y_t)\}_{t=1}^n$ where each pair (x_t, y_t) is drawn i.i.d. from an unknown distribution $\mathcal{D}_{X,Y}$ over $\mathcal{X} \times [0,1]$. We denote by $\mathcal{D}_{Y|X}$ the distribution of the label conditioning on the input, and $\mathcal{D}_X$ the marginal distribution of the input.

Let $\mathcal{F} \subset \{\mathcal{X} \to [0,1]\}$ be a class of prediction functions mapping inputs to the unit interval. We assume *realizability*, i.e., there exists a function $f^* \in \mathcal{F}$ such that:

$$\mathbb{E}_{y \sim \mathcal{D}_{Y|X}}[y \mid x] = f^*(x), \text{ for all } x \in \mathcal{X}.$$

Note that, since we do not have further restrictions on $\mathcal{D}_{X,Y}$, the conditional variance $\sigma_x^2 := \mathbb{E}_{y \sim \mathcal{D}_{Y|X}}[(y - f^*(x))^2 \mid x]$ can vary across $x \in \mathcal{X}$.

Given the observed data D_n, our goal is to find a hypothesis $\hat{f} \in \mathcal{F}$ that achieves low expected absolute error with respect to the ground-truth regression function f^*, which we call the L^1 error:

$$\mathbb{E}_{x \sim \mathcal{D}_X}\left[|f^*(x) - \hat{f}(x)|\right].$$

This differs from minimizing $\mathbb{E}_{x,y \sim \mathcal{D}_{X,Y}}\left[|\hat{f}(x) - y|\right]$, which contains irreducible conditional noise even at $\hat{f} = f^*$. Note that if one has an L^2 bound of $\mathbb{E}_{x \sim \mathcal{D}_X}\left[(f^*(x) - \hat{f}(x))^2\right] \leq B$ then, using Jensen's inequality, one obtains $\mathbb{E}_{x \sim \mathcal{D}_X}\left[|f^*(x) - \hat{f}(x)|\right] \leq \sqrt{B}$. One may argue that analyzing an L^2 bound is better in this sense. However, such an L^1 bound derived from the L^2 bound can be loose. In general, whether one aims to obtain tight bounds on L^1 vs L^2 is a matter of choice. Nevertheless, when one performs regression for costs or rewards in reinforcement learning or bandits, the performance measure therein is bounded in terms of the reward or cost differences without being squared, so having a tight L^1 bound is important for these applications, which is our motivation for choosing the L^1 error. The emphasis on L^1 error is consistent with recent work arguing that mean absolute error is a better prediction objective than mean squared error for controlling the suboptimality gap of greedy policies (Ayoub et al., 2025).

The goal is to bound such a generalization error of the learned hypothesis in terms of the sample size n, the function class complexity $\ln|\mathcal{F}|$, and the confidence level δ. The de facto standard

algorithm for regression is the squared loss minimizer:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{(x,y) \in D_n} \frac{1}{2}(f(x) - y)^2. \tag{1}$$

A classical result on the squared loss minimizer yields the following (for completeness, see Proposition 9 in Appendix A):

$$\mathbb{E}_{x \sim \mathcal{D}_X} \left[|f^*(x) - \hat{f}(x)| \right] \lesssim \sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{n}}.$$

While this bound is simple and general, it does not incorporate any notion of conditional variance. It treats all inputs as equally noisy, making it inherently variance-insensitive.

A recent result on the log loss minimizer (Foster and Krishnamurthy, 2021, Theorem 3) immediately implies the following first-order generalization bound, which scales with the magnitude of the target regression function $f^*(x)$ and its complement $1 - f^*(x)$ in expectation:

$$\mathbb{E}_{x \sim \mathcal{D}_X} \left[|f^*(x) - \hat{f}(x)| \right] \lesssim \sqrt{\frac{\left(\mathbb{E}_x[f^*(x)] \wedge \mathbb{E}_x[1 - f^*(x)]\right) \cdot \ln(|\mathcal{F}|/\delta)}{n}} + \frac{\ln(|\mathcal{F}|/\delta)}{n}. \tag{2}$$

In the following theorem, we further improve the bound above to scale with $\mathbb{E}_x[f^*(x)(1 - f^*(x))]$.

Theorem 1 *With probability at least* $1 - \delta$*,*

$$\mathbb{E}_{x \sim \mathcal{D}_X} \left[|f^*(x) - \hat{f}(x)| \right] \lesssim \sqrt{\frac{\mathbb{E}_x[f^*(x)(1 - f^*(x))] \cdot \ln(|\mathcal{F}|/\delta)}{n}} + \frac{\ln(|\mathcal{F}|/\delta)}{n}.$$

The proof is deferred to Appendix B. We note that this bound strictly improves upon the immediate implication of Foster and Krishnamurthy (2021, Eqn. (2)), as $\mathbb{E}_x[f^*(x) \wedge (1 - f^*(x))] \leq \mathbb{E}_x[f^*(x)] \wedge \mathbb{E}_x[1 - f^*(x)]$, the gap between these two quantities can be arbitrarily large as we show in Appendix J.

The bound in Theorem 1 depends on the variance proxy $f^*(x)(1 - f^*(x))$, which upper bounds the conditional variance $\sigma_x^2 := \mathbb{E}[(y - f^*(x))^2 \mid x]$, as indicated by the following Lemma.

Lemma 2 *Let* $Y \in [0, 1]$ *be a random variable. Then* $\mathrm{Var}(Y) \leq \mathbb{E}[Y](1 - \mathbb{E}[Y])$*, and the equality is attained iff* Y *is Bernoulli-distributed.*

For Bernoulli labels, $\sigma_x^2 = f_x^*(1 - f_x^*)$, so Theorem 1 already has the same variance factor as the desired second-order bound. The improvement targeted in this paper is therefore specific to non-Bernoulli [0,1]-valued labels, where the conditional variance can be much smaller than the Bernoulli proxy. For example, labels concentrated near their conditional mean can have $\sigma_x^2 \ll f_x^*(1 - f_x^*)$. In many applications, such as those mentioned in Section 1, the label distributions are often heteroscedastic and non-Bernoulli: some inputs yield lower uncertainty on Y conditioning on X (i.e., low variance), while others are more uncertain (i.e., high variance), and the outputs can take values other than the boundary points 0 and 1. In such settings, first-order bounds may fail to capture the true learnability of the problem.

To address this, our objective in this work is to derive *second-order generalization bounds* that adapt to the true conditional variance. Specifically, we aim to obtain bounds of the form:

$$\mathbb{E}_{x \sim \mathcal{D}_X} \left[|f^*(x) - \hat{f}(x)| \right] \lesssim \sqrt{\frac{\mathbb{E}_x[\sigma_x^2] \cdot \ln(|\mathcal{F}|/\delta)}{n}} + \frac{\ln(|\mathcal{F}|/\delta)}{n}, \tag{3}$$

which provides tighter guarantees in settings where $\mathcal{D}_X$ places a nontrivial probability on x such that the conditional variance σ_x^2 is much smaller than the worst-case upper bound $f^*(x)(1 - f^*(x))$, without requiring the variances as input to the algorithm.

3. Second-Order Bound via Betting Loss

Although being the de facto standard regression algorithm, empirical risk minimization under squared loss admits deficiencies in achieving tight generalization bounds. Theorem 2 of Foster and Krishnamurthy (2021) proves lower bounds showing that in the problem of cost-sensitive classification in statistical learning, the squared loss minimizer fundamentally fails to achieve the first-order regret bound. Inspired by that, we provide Proposition 11 in Appendix C showing that in $[0,1]$ bounded regression, the squared loss minimizer also fails to achieve the first-order L^1 generalization error.

While the log loss achieves the first-order guarantee in Theorem 1, this guarantee cannot in general be upgraded to a second-order bound depending on the true conditional variance $\mathbb{E}_x[\sigma_x^2]$. The following result makes this obstruction formal. The proof is deferred to Appendix D.

Theorem 3 *For every integer $n \geq 2$, there exists a realizable $[0,1]$-valued regression problem with a two-function class $\mathcal{F}_n = \{f^\star, g\}$ such that $\mathbb{E}[\sigma_x^2] = \frac{1}{4n}$, but the log-loss ERM*

$$\hat{f}_{\log} \in \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} \left\{ -Y_i \log f(X_i) - (1 - Y_i) \log(1 - f(X_i)) \right\}$$

satisfies

$$\mathbb{P}\left(\mathbb{E}_x \left| \hat{f}_{\log}(x) - f^\star(x) \right| \geq \frac{1}{8\sqrt{n}} \right) \geq \frac{1}{2e}.$$

Consequently, log-loss ERM cannot satisfy, uniformly over all realizable $[0,1]$-valued regression problems, a high-probability bound of the form in (3). Indeed, for the instance in Theorem 3, the RHS in (3) is $\widetilde{O}(n^{-1})$, whereas the log-loss ERM has $\Omega(n^{-1/2})$ error with constant probability.

This motivates the design of a new objective. In this section, we propose a novel loss function called betting loss that is inspired by the coin-betting framework. We show that, minimizing the betting loss leads to an estimator that adapts to the true conditional variance.

Review of betting-based confidence bounds. Let $Y_1, Y_2, \ldots \sim \nu$ be i.i.d. $[0,1]$-valued random variables from some distribution ν with mean $\mu := \mathbb{E}_{Y \sim \nu}[Y]$. The betting-based confidence bound (Orabona and Jun, 2024) states that, with probability at least $1 - \delta$,

$$\forall n \geq 1, \max_{b \in [0,1]} \sum_{i=1}^{n} \ln(1 + \frac{(b - \mu)(Y_i - \mu)}{\mu(1 - \mu)}) \leq \ln(O(\sqrt{n})/\delta) \,. \tag{4}$$

In practice, one does not know μ, so one can collect all μ's that satisfy the inequality above. This forms a confidence set that contains the mean μ with high probability, and Orabona and Jun (2024) have proven that such a confidence set is essentially a numerically-tight version of the empirical Bernstein bound, which can be seen as a second-order bound. We realized that the LHS above acts as some kind of a measure of goodness of fit and would be a good loss function. Indeed, finding the value of μ that minimizes the LHS equals the sample mean $\hat{\mu}_n$ of $Y_{1:n}$, which is a reasonable estimator. At this point, it seems natural to extend the loss function of μ to the regression setting.

The betting loss for regression. We now investigate into developing a loss function for regression that is inspired by the LHS of (4) and seeing if the resulting loss function would provide a second-order bound. Note that μ that we minimized above should now be $f \in \mathcal{F}$. Our initial attempt was to

minimize the following loss function over f:

$$\max_{h \in \mathcal{F}} \sum_{(x,y) \in D_n} \ln(1 + \frac{(h_x - f_x)(y - f_x)}{f_x(1 - f_x)}). \tag{5}$$

However, our attempts at proving a second-order bound were not successful. We conjecture that the loss above does not lead to our targeted second-order bound. The reason is that the naive objective in Eq. (5) is too rigid to give us a second-order bound because it does not provide sufficient discriminative power. We increase that power by allowing further scaling $h_x - f_x$ by ϕ (thus sensitivity to $h_x - f_x$ increases). This scaling was not needed in a noninductive setting, because f_x was always constant (i.e., μ). Now that it differs across x, some data points require further scaling to have sufficient discriminative power. The role of the clipping is that, without it, the analysis has the coupled term of $\mathbb{E}\,\sigma_x^2(h_x - f_x)^2$. Clipping the betting term $h_x - f_x$ within $[-c, c]$ allows the coupled term $\mathbb{E}\,\sigma_x^2(h_x - f_x)^2$ to be controlled by $c^2\,\mathbb{E}\,\sigma_x^2$, which helps us factor out $\mathbb{E}[\sigma_x^2]$, making a true conditional-variance factor possible. Concretely, we define:

- A fixed parameter $\overline{\phi} := \frac{n}{4}$ that controls the magnitude of perturbation.
- A log wealth function:

$$H_{\phi,c}(h, f) := \sum_{(x,y) \in D_n} \ln\left(1 + (y - f_x)\overline{(\phi(h_x - f_x))}_{[-c,c]}\right),$$

  where $\overline{(x)}_{[a,b]} := \max\{\min\{x, b\}, a\}$ and $c \in [0, \frac{1}{4}]$ is a clipping threshold.

We then define the betting loss as (6) and describe the full algorithm in Algorithm 1. Our betting loss, compared to the initial version (5), has extra maximization over a betting scalar ϕ and also a clipping level c.

Algorithm 1 Variance-Adaptive Regression via Betting Loss

Require: Dataset $D_n = \{(x_t, y_t)\}_{t=1}^n$, hypothesis class $\mathcal{F}$

1: Define the betting loss

$$L_n(f) := \max_{h \in \mathcal{F}} \max_{\phi \in [0, \overline{\phi}]} \max_{c \in [0, \frac{1}{4}]} \frac{1}{n} H_{\phi,c}(h, f). \tag{6}$$

2: Compute

$$\hat{f} := \arg\min_{f \in \mathcal{F}} L_n(f)\ .$$

3: **Return** $\hat{f}$.

Think of our algorithm (Algorithm 1) as finding a hypothesis f that is "unbeatable" by any other hypothesis h from the same class $\mathcal{F}$. For any data point x, the bettor h makes a "bet" $(h_x - f_x)$, essentially wagering that the true label is in the direction of h_x. The bet is resolved against the "outcome" $(y - f_x)$. The $\ln(1 + \cdots)$ term represents the logarithmic growth of wealth for the bettor. The inner maximization ($\max_h$) represents an adversary attempting to find the most profitable betting strategy against f. The learner's goal ($\min_f$) is to find a function f that leaves no room for any h to consistently accumulate wealth. The key to variance adaptivity lies in the fact that the power of the adversary to increase the loss depends on the variance. In low-variance regions (i.e., x for which

σ_x^2 is small), the residuals $(y - f_x)$ are highly predictable. This makes it significantly easier for an adversary h to find a betting direction that yields massive profits.

Statistically, this creates a steeper curvature in the optimization landscape where the variance is low. This means that the loss function has a higher discriminative power, thus making the learning easier. In contrast, the squared loss treats all residuals equally, failing to adapt to variance. As another comparison, the log loss is designed to capture the Bernoulli-like behavior of $[0, 1]$ labels, so it treats the expected value as the primary signal. In regions where the variance is significantly smaller than $f^*(x)(1 - f^*(x))$, such as a distribution that is highly concentrated near 0.5, the log loss lacks the necessary "pressure" to refine its estimate further.

The following theorem provides a high-probability bound on the expected absolute error of any $f \in \mathcal{F}$ in terms of the suboptimality w.r.t. betting loss.

Theorem 4 (Finite class) *There exist numerical constants c_1, c_2 and c_3, such that with probability at least $1 - \delta$,*

$$\forall f \in \mathcal{F},\ \ \mathbb{E}_x\,|f_x - f_x^*| \le c_1 \cdot \sqrt{\mathbb{E}\,\sigma_x^2 \cdot \left(\frac{1}{n}\ln\left(\frac{|\mathcal{F}|n}{\delta}\right) + \max\{L_n(f) - L_n(f^*), 0\}\right)} + c_2 \cdot \frac{1}{n}\ln\left(\frac{|\mathcal{F}|n}{\delta}\right) + c_3 \cdot (L_n(f) - L_n(f^*)).$$

We provide the full proof in Appendix E.

This result provides a high-probability bound on the prediction error $\mathbb{E}_x\,|f_x - f_x^*|$ of any $f \in \mathcal{F}$, in terms of the difference in empirical betting loss $L_n(f) - L_n(f^*)$. Crucially, the bound adapts to the conditional variance σ_x^2 in the leading term. The excess betting loss $L_n(f) - L_n(f^*)$ directly controls the mean absolute error. In particular, applying the theorem to the output $\hat{f} = \arg\min_{f \in \mathcal{F}} L_n(f)$ of Algorithm 1, we obtain:

$$\mathbb{E}_x\,|\hat{f}_x - f_x^*| \le c_1 \cdot \sqrt{\mathbb{E}\,\sigma_x^2 \cdot \left(\frac{1}{n}\ln\left(\frac{|\mathcal{F}|n}{\delta}\right)\right)} + c_2 \cdot \frac{1}{n}\ln\left(\frac{|\mathcal{F}|n}{\delta}\right).$$

This bound reflects a variance-adaptive fast rate, which improves over convergence bounds that scale with the Bernoulli proxy $f_x^*(1 - f_x^*)$. In particular, if $y = f_x^*$ with probability 1, this is the noiseless case. The conditional variance $\sigma_x^2 = 0$ with probability 1, so the expected variance $\mathbb{E}_x[\sigma_x^2] = 0$. As a result, the term with $\mathbb{E}_x[\sigma_x^2]$ in our bound vanishes, and we are left with the $O(1/n)$ rate. As the noise increases, $\mathbb{E}_x[\sigma_x^2]$ grows from 0, and the first term $\tilde{O}\left(\sqrt{\mathbb{E}_x[\sigma_x^2]/n}\right)$ "grows smoothly" to become the dominant part of the bound. This allows our bound to gracefully and adaptively interpolate between the fast $\tilde{O}(1/n)$ rate for noiseless problems and the variance-dependent $\tilde{O}(1/\sqrt{n})$ rate for noisy problems, all without needing to know the variance $\mathbb{E}_x[\sigma_x^2]$ in advance. This establishes Algorithm 1 as a variance-adaptive learning procedure for [0,1]-valued regression.

Theorem 4 provides a general excess risk bound that holds uniformly for all $f \in \mathcal{F}$ over any finite hypothesis class $\mathcal{F}$. Moreover, by combining this result with complexity control via covering numbers, we can derive concrete generalization bounds for a broader family of hypothesis classes characterized by polynomial covering numbers. Following common terminology (see, e.g., Rakhlin et al. (2017)), we will refer to such VC-type classes as parametric classes.

Definition 5 (Parametric class) *A class of functions $\mathcal{F}$ is a parametric class if there exist positive constants A and v, such that for every $0 < \varepsilon \leq 1$, the covering number $N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)$ satisfies the inequality:*

$$N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \leq (\frac{A}{\varepsilon})^v \,. \tag{7}$$

Theorem 6 (Parametric class) *Assume the covering number of $\mathcal{F}$ satisfies Eqn. (7). Then, there exist constants c_1 and c_2, such that with probability at least $1-\delta$, Algorithm 1 satisfies:*

$$\mathbb{E}_x\,|\hat{f}_x - f_x^*| \leq c_1 \cdot \sqrt{\mathbb{E}_x\,\sigma_x^2 \frac{v}{n}\ln(\frac{n}{\delta})} + c_2 \cdot \frac{v}{n}\ln(\frac{n}{\delta})\,.$$

This result follows from Theorem 4 by applying polynomial covering numbers of parametric classes. We present the full proof in the appendix F.

As our work establishes generalization results for classes with polynomial covering numbers, it is useful to instantiate the abstract condition in standard finite-dimensional models. We give two examples. The first is a bounded affine linear class, where the constraints $\|x\|_2 \leq 1$ and $\|\theta\|_2 \leq 1/2$ ensure that every predictor takes values in $[0,1]$. The second is a bounded logistic class, which is another natural model for $[0,1]$-valued regression because the sigmoid link maps arbitrary linear scores into $(0,1)$. In both cases, the class has polynomial $\|\cdot\|_\infty$-covering number with dimension parameter $v = d$, so Theorem 6 yields the following concrete consequences.

Corollary 7 (Linear class) *Let $\mathcal{F}$ be a linear function class in d-dimensional space: $\mathcal{F} = \{x \mapsto x^\top\theta + \frac{1}{2} : \|\theta\|_2 \leq \frac{1}{2}\}$ and $\mathcal{X}$ be the instance space: $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. Then, there exist constants c_1 and c_2, such that with probability at least $1-\delta$, the output $\hat{f} = \arg\min_{f\in\mathcal{F}} L_n(f)$ satisfies:*

$$\mathbb{E}_x\,|\hat{f}_x - f_x^*| \leq c_1 \cdot \sqrt{\mathbb{E}_x[\sigma_x^2]\frac{d}{n}\ln(\frac{n}{\delta})} + c_2 \cdot \frac{d}{n}\ln(\frac{n}{\delta}).$$

For the linear class, since $Y \in [0,1]$, we always have $\sigma_x^2 \leq 1/4$. Hence the leading term in Corollary 7 is at most $\widetilde{O}(\sqrt{d/n})$, matching the standard worst-case scaling for d-dimensional linear regression up to logarithmic factors. This is consistent with classical minimax lower bounds for linear regression, which show that one cannot improve the $O(\sqrt{d/n})$ worst-case L^1 error in general (Tsybakov, 2004; Wainwright, 2019). Corollary 7 shows that, on easier instances with smaller average conditional variance $\mathbb{E}_x[\sigma_x^2]$, the leading term improves to $\widetilde{O}\left(\sqrt{\mathbb{E}_x[\sigma_x^2]\frac{d}{n}}\right)$, thereby adapting to the true average conditional variance.

Corollary 8 (Logistic class) *Let $\mathcal{F}$ be a logistic function class in d-dimensional space:*

$$\mathcal{F} := \left\{x \mapsto \sigma(\theta^\top x) : \|\theta\|_2 \leq B\right\}, \qquad \mathcal{X} := \left\{x \in \mathbb{R}^d : \|x\|_2 \leq R\right\},$$

where $\sigma(z) := 1/(1+\exp(-z))$. Then there exist universal constants $C_0, C_1, C_2 > 0$ such that, with probability at least $1-\delta$, the output $\hat{f} = \arg\min_{f\in\mathcal{F}} L_n(f)$ of Algorithm 1 satisfies

$$\mathbb{E}_x\,|\hat{f}_x - f_x^*| \leq C_1\sqrt{\mathbb{E}[\sigma_x^2]\,\frac{d}{n}\log\left(\frac{C_0(2+BR)n^5}{\delta}\right)} + C_2\frac{d}{n}\log\left(\frac{C_0(2+BR)n^5}{\delta}\right).$$

Together, these two corollaries illustrate that the variance-adaptive guarantee is not limited to a finite hypothesis class. It applies to standard finite-dimensional model classes through their polynomial covering numbers. In the worst case, the bounds recover the usual $\widetilde{O}(\sqrt{d/n})$ scaling for L^1 error, while in low-variance instances the leading term adapts to the smaller quantity $\mathbb{E}_x\, \sigma_x^2$.

Does it work for nonparametric classes? It is natural to ask whether our guarantees extend to nonparametric function classes. However, all our attempts based on standard techniques did not lead to a valid bound. The key difficulty seems to be that the loss function L_n can be $O(n)$-Lipschitz in the input f, unlike squared loss, which is $O(1)$-Lipschitz in the input f. This makes standard localization and covering-number arguments unable to be used to derive a non-vacuous generalization bound for nonparametric classes. Relatedly, Srebro et al. (2010) show that for scale-sensitive classes, even smooth and strongly convex losses can exhibit unavoidable $n^{-1/2}$ slow rates for excess risk. Although their setting is not identical to ours, this suggests that obtaining fast variance-adaptive rates for rich nonparametric classes may require additional structure. Appendix I gives a Lipschitz-class construction where $L_n(f^*) - L_n(f_0) > c_1$ for another function f_0 and some constant c_1 with constant probability. We view fast second-order convergence rates for general nonparametric classes as an open problem.

4. Experiments

In this section, we provide the empirical results that confirm our theoretical findings. These experiments are controlled finite-class sanity checks. They isolate the statistical effect predicted by the theory by exactly enumerating a small candidate class. They are not intended to demonstrate scalable optimization of the betting loss for large classes.

Function class. We consider a d-dimensional regression setting with target function f^*. First, we sample $\theta^* \in \mathbb{R}^d$ from an isotropic Gaussian distribution and normalize it to have Euclidean norm $S = 0.5$:

$$\theta^* \sim \mathcal{N}(0, I_d), \qquad \theta^* \leftarrow \tfrac{S}{\|\theta^*\|} \text{ followed by } \theta^*.$$

The target function is then defined as

$$f^* := (x \mapsto \sigma(x^T\theta^*)),$$

where $\sigma(z) := 1/(1+\exp(-z))$ is the sigmoid function.

To avoid the potential complications from the convergence issues from optimization, we consider a finite function class. Specifically, we sample 20 θ's independently by drawing η from the unit sphere followed by setting

$$\theta = \theta^* + \varepsilon \cdot \eta,$$

where $\varepsilon = 0.2$. Let Θ be the set of these θ's and θ^*, which means $|\Theta| = 21$. Finally, we construct our function class as

$$\mathcal{F} := \{x \mapsto \sigma(x^T\theta) : \theta \in \Theta\}\,.$$

Data distribution. Feature vectors are sampled as

$$x \sim \tfrac{1}{\sqrt{d}}\mathcal{N}(0, I_d).$$

For each x, we generate its label from a Beta distribution with mean f_x^* and parameter $\rho \in (0,1)$:

$$y \sim \mathrm{Beta}\Big(f_x^* \cdot \tfrac{1-\rho}{\rho},\ (1-f_x^*) \cdot \tfrac{1-\rho}{\rho}\Big),$$

which satisfies that $y \in [0,1]$ with probability 1. Let

$$\alpha = f_x^* \cdot \frac{1-\rho}{\rho}, \qquad \beta = (1-f_x^*) \cdot \frac{1-\rho}{\rho}.$$

For a $\mathrm{Beta}(\alpha,\beta)$ distribution, the variance is

$$\mathrm{Var}(y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Plugging in $\alpha = f_x^* \frac{1-\rho}{\rho}$ and $\beta = (1-f_x^*)\frac{1-\rho}{\rho}$ gives

$$\alpha\beta = f_x^*(1-f_x^*)\left(\tfrac{1-\rho}{\rho}\right)^2.$$

Also,

$$(\alpha+\beta)^2(\alpha+\beta+1) = \left(\tfrac{1-\rho}{\rho}\right)^2\left(\tfrac{1-\rho}{\rho}+1\right).$$

Therefore,

$$\mathrm{Var}(y) = \frac{f_x^*(1-f_x^*)\left(\tfrac{1-\rho}{\rho}\right)^2}{\left(\tfrac{1-\rho}{\rho}\right)^2\left(\tfrac{1-\rho}{\rho}+1\right)}.$$

By simplifying the expression, we obtain the following result:

$$\mathrm{Var}(y) = f_x^*(1-f_x^*)\rho.$$

Therefore, the variance is proportional to ρ.

Training. Given the candidate set $\mathcal{F}$, we select a function by minimizing either log loss, squared loss, or betting loss on the training data $D_n \sim \mathcal{D}_{X,Y}^n$. 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

$$\hat{f}^{\log} = \arg\min_{f\in\mathcal{F}} L_n^{\log}(f).$$

Similarly, the squared loss of f is

$$L_n^{\text{squared}}(f) = \frac{1}{n}\sum_{(x,y)\in D_n}(y-f_x)^2,$$

and the selected function is

$$\hat{f}^{\text{squared}} = \arg\min_{f\in\mathcal{F}} L_n^{\text{squared}}(f).$$

For betting loss, we define

$$L_n^{\text{betting}}(f) = \max_{h\in\mathcal{F}}\ \max_{\phi\in[0,\overline{\phi}]}\ \max_{c\in[0,1]}\frac{1}{n}\sum_{(x,y)\in D_n}\ln\Big(1+(y-f_x)\,\overline{(\phi(h_x-f_x))}_{[-c,c]}\Big),$$

where $\overline{(z)}_{[a,b]} := \max\{\min\{z, b\}, a\}$, and $\overline{\phi} = n/4$. In our implementation, we discretize $\phi \in \{1, 2, 4, \ldots, \frac{n}{4}\}$ and $c \in \{\frac{1}{n}, \frac{2}{n}, \frac{4}{n}, \ldots, 1\}$ to solve the inner maximization, and select

$$\hat{f}^{\text{betting}} = \arg\min_{f \in \mathcal{F}} L_n^{\text{betting}}(f).$$

Evaluation. We evaluate the performance of the trained function $\hat{f}$ using the mean absolute error (MAE):

$$\text{MAE}(\hat{f}) := \mathbb{E}_{x \sim \mathcal{D}_X}[|\hat{f}_x - f_x^*|].$$

We perform Monte Carlo estimate of the MAE using a test set of size $m = 10{,}000$.

We fix the feature dimension to $d = 2$, vary the training sample-to-dimension ratio $n/d \in \{2, 4, 8\}$, and try $\rho \in \{0.01, 0.02, 0.04\}$. For each configuration $(n/d, \rho)$, we repeat each experiment 1000 times, and we report the average MAE and its standard error.

n/d	ρ	Log loss MAE	Squared loss MAE	Betting loss MAE
2	0.01	0.02289 ± 0.00030	0.02291 ± 0.00030	**0.02045** ± 0.00036
	0.02	0.02509 ± 0.00021	0.02514 ± 0.00021	**0.02340** ± 0.00029
	0.04	0.02601 ± 0.00015	0.02606 ± 0.00015	**0.02539** ± 0.00020
4	0.01	0.01824 ± 0.00040	0.01829 ± 0.00040	**0.01391** ± 0.00043
	0.02	0.02192 ± 0.00033	0.02182 ± 0.00033	**0.01899** ± 0.00039
	0.04	0.02398 ± 0.00027	0.02413 ± 0.00026	**0.02212** ± 0.00033
8	0.01	0.01203 ± 0.00042	0.01203 ± 0.00042	**0.00654** ± 0.00037
	0.02	0.01739 ± 0.00041	0.01741 ± 0.00041	**0.01349** ± 0.00043
	0.04	0.02162 ± 0.00034	0.02177 ± 0.00033	**0.01838** ± 0.00040

Table 1: Comparison of average mean absolute error (MAE) obtained under log loss, squared loss and betting loss across varying n/d and ρ. For each configuration, the reported value corresponds to the average MAE over repeated trials, and the value after the $\pm$ denotes the standard error. Best results are marked in **bold**. Across all comparisons, the betting loss consistently performs better than the other losses in paired t-tests ($p < 0.01$).

Results. Table 1 and Figure 1 summarize the average MAE with standard error from function selection with log loss, squared loss and betting loss across different values of n/d and ρ. The results show that all three losses decrease as n/d increases, with betting loss consistently achieving the lowest MAE. The real-world experiments are described in the Appendix L.

5. Related Work

Regression with heteroscedastic noise. Regression with heteroscedastic noise can be dated back to (Aitkin, 1935), and has been further developed into Gaussian processes (Kersting et al., 2007; Goldberg et al., 1997). However, these works typically model the input-dependent noise variance or impose distributional structure on the noise process. Our setting is different: the learner is not given a variance model, and the guarantee depends on the true average conditional variance for arbitrary $[0, 1]$-valued conditional label distributions.

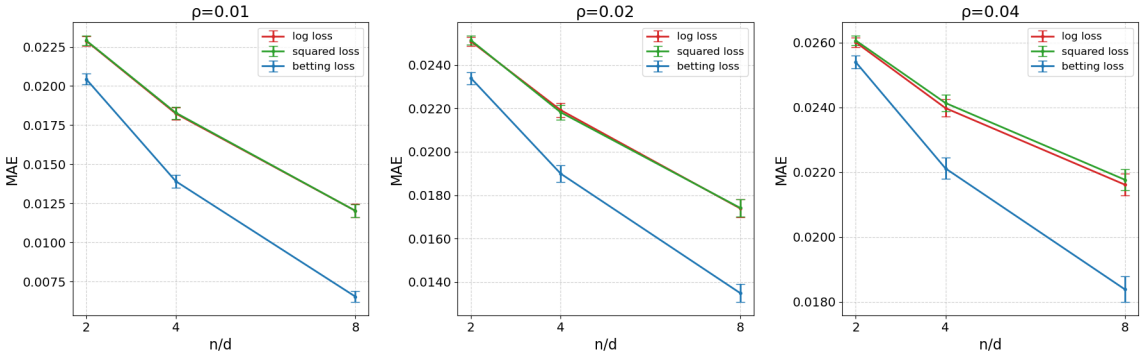

Figure 1: Comparison of average mean absolute error (MAE) obtained under log loss, squared loss and betting loss across varying n/d and ρ. Error bars denote the standard error.

First- and second-order bounds. From the adversarial setting, to our knowledge, the first appearance of the first-order bound is from the prediction with expert advice setting Freund and Schapire (1997), which is an adversarial (i.e., nonstochastic) setting. In the same setting, Cesa-Bianchi et al. (2007) developed a second-order bound with the prod algorithm. Note that the notion of second-order can be defined in various ways; e.g., Hazan and Kale (2010). In K-armed bandits, Stoltz (2005) and Allenberg et al. (2006) have shown first-order bound. In linear bandits, obtaining a first-order regret was an open problem (Agarwal et al., 2017), which was later resolved by Allen-Zhu et al. (2018). Second-order bounds were developed by Hazan and Kale (2011) and improved by Ito et al. (2020). We refer to (Neu) for a review of the first-/second-order bounds in adversarial settings.

Optimistic rates and localized generalization bounds. Our work is also closely related to the literature on "optimistic rates" and "localized generalization bounds," which primarily focuses on bounding the excess risk (i.e., $L(\hat{f}) - L^\star$) and seeks to improve upon worst-case slow rates (typically $\tilde{O}(n^{-1/2})$) by exploiting low variance of the excess loss. Foundational work by Bartlett et al. (2005) and Koltchinskii (2006) developed the machinery of local Rademacher complexities, showing that under suitable curvature/self-bounding conditions, specifically, if the L^2 distance between any function f and the optimal function $f^\star$ is upper bounded by the excess risk (a condition satisfied by squared loss), faster rates can be obtained. Srebro et al. (2010) proved that for smooth non-negative loss functions, the excess risk $L(\hat{f}) - L^\star$ scales with the optimal risk $L^\star$. This can be interpreted as a variance-adaptive bound: in our realizable setting ($f^* \in \mathcal{F}$), the best achievable squared loss is precisely the expected conditional variance: $L^* := \min_{f\in\mathcal{F}} \mathbb{E}[(f(x) - y)^2] = \mathbb{E}[(f^*(x) - y)^2] = \mathbb{E}[\sigma_x^2]$. In conjunction with the fact that the excess risk under squared loss is exactly the L^2 generalization error: $L(\hat{f}) - L(f^*) = \mathbb{E}[(\hat{f} - f^*)^2]$, Srebro et al. (2010)'s main theorem, which bounds the excess risk for smooth losses, implies a variance-adaptive $\tilde{O}(\sqrt{L^*/n})$ bound on the L^2 generalization error: $\mathbb{E}[(\hat{f} - f^*)^2] \le \widetilde{O}\left(\sqrt{\frac{\mathbb{E}[\sigma_x^2]}{n}}\right)$. After applying Jensen's inequality, this yields only $\widetilde{O}\left(\left(\frac{\mathbb{E}[\sigma_x^2]}{n}\right)^{1/4}\right)$ for the L^1 error, which is much slower than the $\widetilde{O}(\sqrt{\mathbb{E}[\sigma_x^2]/n})$ rate targeted here.

Srebro et al. (2010) (Section 3 therein) further summarized the landscape of achievable rates, showing that the $O(1/\sqrt{n})$ dependence is generally unavoidable for non-parametric or non-strongly

convex settings. They demonstrate that a fast rate of $O(1/n)$ that is independent of $L^\star$ is possible only for smooth and strongly convex losses (like squared loss) in parametric settings.

In this vein, Liang et al. (2015) introduced "Offset Rademacher Complexity" to study the excess risk under squared loss in agnostic settings. They confirmed (e.g., in their Lemma 10) that for parametric regression, the excess risk scales as $O(1/n)$, recovering the results of Rakhlin et al. (2017) without assuming boundedness of the noise or functions. Results based on offset Rademacher complexity and related localized analyses imply fast excess squared-loss rates for parametric regression. In the realizable setting, these guarantees can be read as $\mathbb{E}_x[(\hat{f}_n^{\mathrm{sq}}(x) - f^*(x))^2] = O(1/n)$ where the big-O here is w.r.t. n only and $f^*(x) = \mathbb{E}[y|x]$ is the regression function. Applying Jensen's inequality, we obtain an L^1 bound: $\mathbb{E}_x[|\hat{f}_n^{\mathrm{sq}}(x) - f^*(x)|] = O(1/\sqrt{n})$. Although such results achieve a fast L^1 rate, they do not adapt to the ease of low label noise.

Stochastic bandits with function approximation. We now discuss first-/second-order bounds in the stochastic bandit problem with function approximation (also known as structured bandits). Hereafter, unless noted otherwise, the noise model is such that the reward (label) is bounded with a known range, which can be easily translated to $[0, 1]$-valued reward. The first-order bound was first obtained by Foster and Krishnamurthy (2021) for generic function classes. We classify second-order bounds as follows:

- **With known variance**: Based on weighted linear regression, Zhou et al. (2021); Zhou and Gu (2022); Zhao et al. (2023b) have obtained second-order bounds in linear models.
- **Unknown variances but with models of variance or distribution**: In the pure exploration setting, Weltz et al. (2023) have considered modeling the variance explicitly with a specific function class in order to obtain improved sample complexity. Wang et al. (2024a) have shown that modeling not just mean or variance but the noise distribution itself leads to a second-order bound. However, note that modeling variance or distribution has a price to pay due to the extra modeling.
- **Unknown variances**: The last set of works do not make any effort in modeling the variance or distribution, and thus there is no extra price to pay, at least in the statistical sense. For the linear model, Zhang et al. (2022) proposed a second-order regret bound, which was further improved by Kim et al. (2022). The optimal rate in this setting was first obtained by Zhao et al. (2023a), and Jun and Kim (2024) obtained the same bound but with improved numerical performance along with removal of an unnatural technical assumption on the noise. For generic function class, Jia et al. (2024) and Pacchiano (2025) both independently developed a second-order bound where the dependence of the function class appears as the eluder dimension (Russo and Van Roy, 2013).

While the work with unknown variances are the closest to our work, we emphasize that the tools developed therein do not directly imply any meaningful result for the regression setting, to our knowledge. Delineating the challenges is left as future work. That said, we believe the estimator might be useful in obtaining an improved second-order regret bound in bandits with general function classes, just in the same way that the log loss has played a role in obtaining a first-order bound (Foster and Krishnamurthy, 2021).

There is another set of work that considers sub-Gaussian noise, which is more general than the bounded reward. Kirschner and Krause (2018) consider the heteroscedastic noise in linear bandits for the first time, to our knowledge. Their work assumes that the noise is $\sigma^2(x)$-sub-Gaussian when pulling arm x and that the value of $\sigma^2(x)$ is known to the algorithm. Jun and Kim (2024) considered a

further generalized setting where the noise is σ_t^2-sub-Gaussian at time step t, and σ_t^2 can be dependent on anything that happened up to choosing the arm x_t at time t. Furthermore, they assume that the algorithm does not have access to σ_t^2 but rather an upper bound σ_0^2 and have shown that there exists a computationally efficient algorithm whose performance provably adapts to $\max_t \sigma_t^2$ for the leading term (though there is a lower order term with a σ_0^2 dependence).

6. Conclusion

We have introduced a new approach to regression that achieves second-order generalization guarantees by minimizing a novel *betting loss* function inspired by the betting-based confidence bounds. Our analysis establishes that minimizing this loss yields estimators whose guarantee adapts to the conditional variance of the data – without requiring any prior knowledge. Our bound is first-of-its-kind, to our knowledge.

We further demonstrate that our generalization error bounds scale favorably with the local noise level and that the guarantee extends from finite to parametric classes, with concrete instantiations for bounded linear and logistic classes. These results show that, for finite classes and parametric classes, variance adaptivity is statistically attainable without explicit variance estimation, through a carefully chosen loss function alone and without adding extra assumptions. The current work leaves open important algorithmic and statistical questions. The betting loss has a nested min-max structure, and scalable optimization for large classes remains unresolved. Moreover, our positive infinite-class result does not cover general nonparametric classes. We view the contribution as a statistical attainability theorem and as evidence that carefully designed losses can encode variance adaptivity beyond what is obtained by standard squared or log losses.

This insight suggests several promising directions for extending the betting loss framework to other domains where adapting to noise is critical, such as active learning and exploration in reinforcement learning.

Acknowledgments

Yinan Li, Sungjoon Yoon, and Kwang-Sung Jun were supported in part by the National Science Foundation under grant CCF-2327013 and Meta Platforms, Inc. This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH)).

References

Alekh Agarwal, Akshay Krishnamurthy, John Langford, Haipeng Luo, et al. Open problem: First-order regret bounds for contextual bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 4–7, 2017.

AC Aitkin. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1935.

Zeyuan Allen-Zhu, Sébastien Bubeck, and Yuanzhi Li. Make the minority great again: First-order regret bound for contextual bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 186–194, 2018.

Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan Consistency in On-Line Learning in Case of Unbounded Losses Under Partial Monitoring. In *Algorithmic Learning Theory (ALT)*, pages 229–243. 2006.

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

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, 2005.

Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352, 2007.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Wine quality. UCI Machine Learning Repository, 2009. URL https://doi.org/10.24432/C56S3T.

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

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Paul Goldberg, Christopher Williams, and Christopher Bishop. Regression with input-dependent noise: A gaussian process treatment. *Advances in Neural Information Processing Systems (NeurIPS)*, 1997.

Elad Hazan and Satyen Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80:165–188, 2010.

Elad Hazan and Satyen Kale. Better algorithms for benign bandits. *Journal of Machine Learning Research*, 12(4), 2011.

Shinji Ito, Shuichi Hirahara, Tasuku Soma, and Yuichi Yoshida. Tight first- and second-order regret bounds for adversarial linear bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 2028–2038, 2020.

Zeyu Jia, Jian Qian, Alexander Rakhlin, and Chen-Yu Wei. How does variance shape the regret in contextual bandits? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Kwang-Sung Jun and Jungtaek Kim. Noise-adaptive confidence sets for linear bandits and application to bayesian optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Kristian Kersting, Christian Plagemann, Patrick Pfaff, and Wolfram Burgard. Most likely heteroscedastic gaussian process regression. In *ACM International Conference Proceeding Series*, volume 227, 2007.

Yeoneung Kim, Insoon Yang, and Kwang-Sung Jun. Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture mdps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 358–384. PMLR, 2018.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. 2006.

Tor Lattimore and Csaba Szepesvári. Bandit Algorithms. 2018. URL http://downloads.tor-lattimore.com/book.pdf.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285. PMLR, 2015.

Gergely Neu. URL https://cs.bme.hu/~gergo/files/tutorial.pdf.

Francesco Orabona and Kwang-Sung Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *IEEE Transactions on Information Theory*, 70(1):436–455, 2024. doi: 10.1109/TIT.2023.3330187.

Aldo Pacchiano. Second order bounds for contextual bandits with function approximation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. 2017.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2256–2264, 2013.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss. *arXiv preprint arXiv:1009.3896*, 2010.

Gilles Stoltz. *Incomplete information and internal regret in prediction of individual sequences*. PhD thesis, Université Paris Sud-Paris XI, 2005.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2004.

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 22384–22429, 2022.

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

Kaiwen Wang, Owen Oertell, Alekh Agarwal, Nathan Kallus, and Wen Sun. More benefits of being distributional: Second-order bounds for reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024a.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*, 2024b.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2023.

Justin Weltz, Tanner Fiez, Alexander Volfovsky, Eric Laber, Blake Mason, Lalit Jain, et al. Experimental designs for heteroskedastic variance. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhiyu Zhang, Ashok Cutkosky, and Yannis Paschalidis. Optimal comparator adaptive online learning with switching cost. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 23936–23950, 2022.

Heyang Zhao, Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. In *Proceedings of the Conference on Learning Theory (COLT)*, volume 195 of *Proceedings of Machine Learning Research*, pages 4977–5020. PMLR, 12–15 Jul 2023a.

Heyang Zhao, Dongruo Zhou, Jiafan He, and Quanquan Gu. Optimal online generalized linear regression with stochastic noise and its application to heteroscedastic bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 42259–42279, 2023b.

Yao Zhao, Kwang-Sung Jun, Tanner Fiez, and Lalit Jain. Adaptive experimentation when you can't experiment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Dongruo Zhou and Quanquan Gu. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 36337–36349, 2022.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 4532–4576. PMLR, 2021.

Appendix

Table of Contents

Appendix A. Squared Loss ERM Bound

Proposition 9 *Let $\mathcal{F} \subseteq \{f : \mathcal{X} \to [0,1]\}$ be finite. Suppose the regression problem is realizable, i.e., there exists $f^\star \in \mathcal{F}$ such that*

$$f^\star(x) = \mathbb{E}[Y \mid X = x] \qquad \text{for all } x \in \mathcal{X}.$$

Let

$$\hat{f}_{\mathrm{sq}} \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \big(f(X_i) - Y_i\big)^2.$$

Then with probability at least $1-\delta$*,*

$$\mathbb{E}_x\Big[\big(\hat{f}_{\mathrm{sq}}(x)-f^\star(x)\big)^2\Big] \le \frac{28\log(|\mathcal{F}|/\delta)}{3n}.$$

Consequently,

$$\mathbb{E}_x\Big[\big|\hat{f}_{\mathrm{sq}}(x)-f^\star(x)\big|\Big] \le \sqrt{\frac{28\log(|\mathcal{F}|/\delta)}{3n}}.$$

Proof Let $L(f)=\frac{1}{n}\sum_{t=1}^n (f(x_t)-y_t)^2$.

Let $\hat{f}$ be the ERM estimator under squared loss, and for any $f\in\mathcal{F}$, let $\mathcal{E}(f):=\mathbb{E}[|f-f^*|^2]$.

Our goal is to find an ε such that $\mathbb{P}(\mathcal{E}(\hat{f})>\varepsilon)\le\delta$.

If $\mathcal{E}(\hat{f})>\varepsilon$, it must be that we picked a function f with $\mathcal{E}(f)>\varepsilon$ that looks better than f^* on the training set: $L(f)\le L(f^*)$. The idea is to bound $\mathbb{P}(L(f)-L(f^*)\le 0)$, for any "bad" function f with $\mathcal{E}(f)>\varepsilon$.

Fix one such "bad" f. For each datapoint (x_i,y_i), define:

$$Z_i=(f^*(x_i)-y_i)^2-(f(x_i)-y_i)^2$$

The event $L(f)-L(f^*)\le 0$ is equivalent to $\frac{1}{n}\sum_{i=1}^n Z_i\ge 0$. Note that $\mathbb{E}\,Z_i=L(f^*)-L(f)=-\mathcal{E}(f)$. So we are to bound

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (Z_i-\mathbb{E}\,Z_i)\ge\mathcal{E}(f)\right)$$

To obtain a fast rate, we aim to apply Bernstein inequality. For independent, zero-mean random variables $X_1,\dots,X_n$ such that $|X_i|\le M$ and with the total variance $\sum\mathbb{E}\,X_i^2\le V$, we have:

$$\mathbb{P}\left(\sum_{i=1}^n X_i\ge t\right)\le\exp\left(-\frac{t^2/2}{V+Mt/3}\right)$$

To apply Bernstein inequality, we need to find M and V for the variables $Z_i-\mathbb{E}\,Z_i$.

- Since $f(x)$ and y are all in $[0,1]$, the squared loss is also in $[0,1]$. Hence $Z_i\in[-1,1]$, and we can set $M=2$.
- Since $\mathrm{Var}(Z_i)\le\mathbb{E}\,Z_i^2$, it suffices to upper bound $\mathbb{E}\,Z_i^2$. Note that

$$\begin{aligned}Z_i^2&=[(f^*(x_i)-y_i)^2-(f(x_i)-y_i)^2]^2\\&=(f^*(x_i)-f(x_i))^2(f^*(x_i)+f(x_i)-2y_i)^2\\&\le 4(f^*(x_i)-f(x_i))^2\end{aligned}$$

  We take the expectation on both sides,

$$\begin{aligned}\mathbb{E}\,Z_i^2&\le 4\,\mathbb{E}(f^*(x_i)-f(x_i))^2\\&=4\mathcal{E}(f)\end{aligned}$$

  So we can set $V=4n\mathcal{E}(f)$.

Applying Bernstein inequality with $t = n\mathcal{E}(f)$, $M = 2$ and $V = 4n\mathcal{E}(f)$, we have

$$\begin{aligned}
\mathbb{P}(L(f) - L(f^*) \leq 0) &= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}\, Z_i) \geq \mathcal{E}(f)\right) \\
&\leq \exp\left(-\frac{(n\mathcal{E}(f))^2/2}{4n\mathcal{E}(f) + 2(n\mathcal{E}(f))/3}\right) \\
&= \exp\left(-\frac{(n\mathcal{E}(f))}{28/3}\right) \\
&\leq \exp\left(-\frac{3n\varepsilon}{28}\right) && (\mathcal{E}(f) > \varepsilon)
\end{aligned}$$

Taking a union bound,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}(\hat{f}) > \varepsilon) &\leq \sum_{f \in \mathcal{F}: \mathcal{E}(f) > \varepsilon} \mathbb{P}(L(f) - L(f^*) \leq 0) \\
&\leq |\mathcal{F}| \exp\left(-\frac{3n\varepsilon}{28}\right)
\end{aligned}$$

Solving $|\mathcal{F}| \exp\left(-\frac{3n\varepsilon}{28}\right) = \delta$ for ε, we get

$$\varepsilon = \frac{28 \ln(|\mathcal{F}|/\delta)}{3n}$$

That is, with probability at least $1 - \delta$,

$$\mathbb{E}[|\hat{f} - f^*|^2] \leq \frac{28 \ln(|\mathcal{F}|/\delta)}{3n}$$

The L^1 bound follows from Jensen's inequality:

$$\mathbb{E}[|\hat{f} - f^*|] \leq \sqrt{\mathbb{E}[|\hat{f} - f^*|^2]} \leq \sqrt{\frac{28 \log(|\mathcal{F}|/\delta)}{3n}}.$$

■

Appendix B. Proof of Theorem 1

Note that if f_x^* is 0 or 1, realizability and $Y \in [0, 1]$ imply $Y = f_x^\star$ almost surely ($Y = 0$ a.s. or $Y = 1$ a.s.). All likelihood-ratio expressions below are then interpreted by continuity, equivalently with the standard conventions $0 \log(0/q) = 0$ and $a \log(a/0) = +\infty$ for $a > 0$. The displayed quantities extend to the endpoints by this limit argument.

Theorem 10 (Restatement of Theorem 1) *Under log loss, we define:*

$$L_n^{\log}(f) := \sum_{(x,y) \in D_n} y \ln(\frac{1}{f_x}) + (1 - y) \ln(\frac{1}{1 - f_x}).$$

Let $\hat{f} = \arg\min_{f \in \mathcal{F}} L_n^{\log}(f)$. *Then, with probability at least* $1 - \delta$,

$$\mathbb{E}_x\, |\hat{f}_x - f_x^*| \leq 8\sqrt{\mathbb{E}[f_x^*(1 - f_x^*)]\frac{\ln(|\mathcal{F}|/\delta)}{n}} + 4\frac{\ln(|\mathcal{F}|/\delta)}{n}\,.$$

Proof For any $f \in \mathcal{F}$, define

$$H(f) := \frac{1}{2}(L_n^{\log}(f^*) - L_n^{\log}(f)) = \sum_{(x,y)\in D_n} \frac{1}{2} y \ln(\frac{f_x}{f_x^*}) + \frac{1}{2}(1-y)\ln(\frac{1-f_x}{1-f_x^*}) .$$

Inspired by Foster and Krishnamurthy (2021), for a fixed $f \in \mathcal{F}$, consider the martingale of

$$\frac{\exp(H(f))}{\mathbb{E}[\exp(H(f))]}$$

and apply Markov's inequality to obtain that

$$1 - \frac{\delta}{|\mathcal{F}|} \le \mathbb{P}\left(\frac{1}{n}H(f) \le \frac{1}{n}\ln(\mathbb{E}[\exp(H(f))]) + \frac{\ln(|\mathcal{F}|/\delta)}{n}\right) .$$

Note that

$$\begin{aligned}\frac{1}{n}\ln(\mathbb{E}[\exp(H(f))]) &= \frac{1}{n}\ln\left(\mathbb{E}[\prod_{(x,y)\in D_n}(\frac{f_x}{f_x^*})^{\frac{1}{2}y}(\frac{1-f_x}{1-f_x^*})^{\frac{1}{2}(1-y)}]\right)\\ &= \ln\left(\mathbb{E}[(\frac{f_x}{f_x^*})^{\frac{1}{2}y}(\frac{1-f_x}{1-f_x^*})^{\frac{1}{2}(1-y)}]\right) \qquad \text{(independence)}\end{aligned}$$

Taking a union bound over $f \in \mathcal{F}$,

$$1-\delta \le \mathbb{P}\left(\forall f \in \mathcal{F},\ \frac{1}{n}H(f) \le \ln\left(\mathbb{E}[(\frac{f_x}{f_x^*})^{\frac{1}{2}y}(\frac{1-f_x}{1-f_x^*})^{\frac{1}{2}(1-y)}]\right) + \frac{\ln(|\mathcal{F}|/\delta)}{n}\right) .$$

The rest of the proof conditions on the event that

$$\forall f \in \mathcal{F},\ \frac{1}{n}H(f) \le \ln\left(\mathbb{E}[(\frac{f_x}{f_x^*})^{\frac{1}{2}y}(\frac{1-f_x}{1-f_x^*})^{\frac{1}{2}(1-y)}]\right) + \frac{\ln(|\mathcal{F}|/\delta)}{n}.$$

Now taking $f = \hat{f}$. By the definition of $\hat{f}$, $H(\hat{f}) \ge 0$, which implies that

$$0 \le \ln\left(\mathbb{E}[(\frac{\hat{f}_x}{f_x^*})^{\frac{1}{2}y}(\frac{1-\hat{f}_x}{1-f_x^*})^{\frac{1}{2}(1-y)}]\right) + \frac{\ln(|\mathcal{F}|/\delta)}{n}. \tag{8}$$

Next, we upper bound $\mathbb{E}[(\frac{\hat{f}_x}{f_x^*})^{\frac{1}{2}y}(\frac{1-\hat{f}_x}{1-f_x^*})^{\frac{1}{2}(1-y)}]$.

$$\begin{aligned}\mathbb{E}[(\frac{\hat{f}_x}{f_x^*})^{\frac{1}{2}y}(\frac{1-\hat{f}_x}{1-f_x^*})^{\frac{1}{2}(1-y)}] &= \mathbb{E}\exp\left(\frac{1}{2}y\ln(\frac{\hat{f}_x}{f_x^*}) + \frac{1}{2}(1-y)\ln(\frac{1-\hat{f}_x}{1-f_x^*})\right)\\ &= \mathbb{E}\exp\left(\mathbb{E}'\frac{1}{2}y'\ln(\frac{\hat{f}_x}{f_x^*}) + \frac{1}{2}(1-y')\ln(\frac{1-\hat{f}_x}{1-f_x^*})\right)\\ &\qquad (y' \sim \text{Bernoulli}(y))\\ &\le \mathbb{E}\,\mathbb{E}'\exp\left(\frac{1}{2}y'\ln(\frac{\hat{f}_x}{f_x^*}) + \frac{1}{2}(1-y')\ln(\frac{1-\hat{f}_x}{1-f_x^*})\right)\\ &\qquad \text{(Jensen's inequality)}\\ &= \mathbb{E}_x\, f_x^* \cdot \sqrt{\frac{\hat{f}_x}{f_x^*}} + (1-f_x^*)\cdot\sqrt{\frac{1-\hat{f}_x}{1-f_x^*}}\\ &= \mathbb{E}_x[\sqrt{f_x^*\hat{f}_x} + \sqrt{(1-f_x^*)(1-\hat{f}_x)}]\end{aligned}$$

Combining with Eqn. (8),

$$\begin{aligned}
\frac{\ln(|\mathcal{F}|/\delta)}{n} &\geq -\ln(\mathbb{E}_x[\sqrt{f_x^*\hat{f}_x} + \sqrt{(1-f_x^*)(1-\hat{f}_x)}]) \\
&= -\ln(1 - \mathbb{E}[1 - \sqrt{f_x^*\hat{f}_x} - \sqrt{(1-f_x^*)(1-\hat{f}_x)}]) \\
&\geq \mathbb{E}[1 - \sqrt{f_x^*\hat{f}_x} - \sqrt{(1-f_x^*)(1-\hat{f}_x)}] && (\ln(1+x) \leq x) \\
&= \mathbb{E}[\frac{1}{2}(\sqrt{f_x^*} - \sqrt{\hat{f}_x})^2 + \frac{1}{2}(\sqrt{1-f_x^*} - \sqrt{1-\hat{f}_x})^2] \\
&= \mathbb{E}[D^2(f_x^*, \hat{f}_x)], && (9)
\end{aligned}$$

where $D^2(p,q)$ for scalers $p, q \in [0,1]$ denotes the Hellinger distance between two Bernoulli distributions with parameters p and q: i.e., $D^2(p,q) = \frac{1}{2}(\sqrt{p} - \sqrt{q})^2 + \frac{1}{2}(\sqrt{1-p} - \sqrt{1-q})^2$.

From the proof of Proposition 3 of Foster and Krishnamurthy (2021), we know that

$$\begin{aligned}
D^2(p,q) &= \frac{1}{2}(\sqrt{p} - \sqrt{q})^2 + \frac{1}{2}(\sqrt{1-p} - \sqrt{1-q})^2 \\
&= \frac{(p-q)^2}{2} \cdot \left(\frac{1}{(\sqrt{p} + \sqrt{q})^2} + \frac{1}{(\sqrt{1-p} + \sqrt{1-q})^2} \right) \\
&\geq \frac{(p-q)^2}{2} \cdot \left(\frac{1}{(\sqrt{p} + \sqrt{q})^2 \wedge (\sqrt{1-p} + \sqrt{1-q})^2} \right) \\
&\geq \frac{(p-q)^2}{4} \cdot \left(\frac{1}{(p+q) \wedge (1-p+1-q)} \right). && ((a+b)^2 \leq 2a^2 + 2b^2)
\end{aligned}$$

Let $g(p,q) = (p+q) \wedge (1-p+1-q)$. Then, by Eqn. (9),

$$\mathbb{E}\left[(f_x^* - \hat{f}_x)^2 \cdot \frac{1}{2g(f_x^*, \hat{f}_x)} \right] \leq 2\frac{\ln(|\mathcal{F}|/\delta)}{n}.$$

Using $\frac{A^2}{2B} = \max_{\eta > 0} \eta A - \frac{\eta^2}{2} B$ for $A, B > 0$, we have, for any $\eta > 0$,

$$\begin{aligned}
2\frac{\ln(|\mathcal{F}|/\delta)}{n} &\geq \mathbb{E}[\max_\eta \eta |f_x^* - \hat{f}_x| - \frac{\eta^2}{2} g(f_x^*, \hat{f}_x)] \\
&\geq \max_\eta \eta\, \mathbb{E}[|f_x^* - \hat{f}_x|] - \frac{\eta^2}{2}\, \mathbb{E}[g(f_x^*, \hat{f}_x)] && \text{(Jensen)} \\
\implies \mathbb{E}[|f_x^* - \hat{f}_x|] &\leq \min_\eta \frac{\eta}{2}\, \mathbb{E}[g(f_x^*, \hat{f}_x)] + \frac{1}{\eta} \frac{2\ln(|\mathcal{F}|/\delta)}{n}.
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}\, g(f_x^*, \hat{f}_x) &= \mathbb{E}[(f_x^* + \hat{f}_x) \wedge (1 - f_x^* + 1 - \hat{f}_x)] \\
&\leq \mathbb{E}[(|f_x^* - \hat{f}_x| + 2f_x^*) \wedge (|f_x^* - \hat{f}_x| + 2(1 - f_x^*))] \\
&= \mathbb{E}[|f_x^* - \hat{f}_x| + (2f_x^* \wedge 2(1 - f_x^*))] \\
&= \mathbb{E}[|f_x^* - \hat{f}_x|] + 2\,\mathbb{E}[f_x^* \wedge (1 - f_x^*)] \\
&\leq \mathbb{E}[|f_x^* - \hat{f}_x|] + 4\,\mathbb{E}[f_x^*(1 - f_x^*)].
\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{E}[|f_x^* - \hat{f}_x|] &\leq \frac{\eta}{2}\mathbb{E}[|f_x^* - \hat{f}_x|] + 2\eta\,\mathbb{E}[f_x^*(1-f_x^*)] + \frac{1}{\eta}\frac{2\ln(|\mathcal{F}|/\delta)}{n} \\ &\leq \frac{1}{2}\mathbb{E}[|f_x^* - \hat{f}_x|] + 2\eta\,\mathbb{E}[f_x^*(1-f_x^*)] + \frac{1}{\eta}\frac{2\ln(|\mathcal{F}|/\delta)}{n} \qquad \text{(assume } \eta \leq 1\text{)} \\ \implies \mathbb{E}[|f_x^* - \hat{f}_x|] &\leq 4\eta\,\mathbb{E}[f_x^*(1-f_x^*)] + \frac{4}{\eta}\frac{\ln(|\mathcal{F}|/\delta)}{n}.\end{aligned}$$

We can choose $\eta = 1 \wedge \sqrt{\frac{\ln(|\mathcal{F}|/\delta)/n}{\mathbb{E}[f_x^*(1-f_x^*)]}}$, which satisfies the assumption above, to arrive at

$$\mathbb{E}[|f_x^* - \hat{f}_x|] \leq 8\sqrt{\mathbb{E}[f_x^*(1-f_x^*)]\frac{\ln(|\mathcal{F}|/\delta)}{n}} + 4\frac{\ln(|\mathcal{F}|/\delta)}{n}\,.$$

■

Appendix C. Squared Loss Fails to Achieve First-Order Bound

Proposition 11 *There exists a $[0,1]$-valued regression task where the squared loss minimizer fails to achieve the first-order bound, with at least constant probability.*

Proof Consider a dataset $D_n = \{(x_t, y_t)\}_{t=1}^n$ and a function class $\mathcal{F} = \{f^*, \tilde{f}\}$. With the realizability condition: $f^*(x) = \mathbb{E}[y|x]$.

Let $\mathcal{X} = \{x^{(1)}, x^{(2)}\}$. We define the distribution $\mathcal{D}_X$ as:

- $\mathbb{P}(X = x^{(2)}) = p_n = \frac{1}{n}$
- $\mathbb{P}(X = x^{(1)}) = 1 - \frac{1}{n}$

The Conditional Distribution of Y:

- At $x^{(1)}$: $Y \sim$ Bernoulli(μ_n), where $\mu_n = \frac{128}{n}$.
- At $x^{(2)}$: $Y \sim$ Bernoulli$(1/2)$.

The Function Class $\mathcal{F}$:

- $f^*(x^{(1)}) = \mu_n$, $f^*(x^{(2)}) = 1/2$.
- $\tilde{f}(x^{(1)}) = \sqrt{\frac{1}{16n}}$, $\tilde{f}(x^{(2)}) = 0$.

The expected value of the target function) is:

$$L^* = \mathbb{E}_x[f^*(x)] = (1-p_n)\mu_n + p_n(1/2) = \left(1 - \frac{1}{n}\right)\frac{128}{n} + \frac{1}{2n}.$$

For large n, $L^* = \mathcal{O}(1/n)$. A first-order bound requires the absolute error to satisfy:

$$\mathbb{E}_x[|f^*(x) - \hat{f}(x)|] \leq \mathcal{O}\left(\sqrt{\frac{L^*\ln(|\mathcal{F}|/\delta)}{n}} + \frac{\ln(|\mathcal{F}|/\delta)}{n}\right).$$

With $L^* = \mathcal{O}(1/n)$, the required first-order rate for this instance is $\mathcal{O}(1/n)$.

We define the empirical squared loss as $\hat{L}_{LS}(f) = \frac{1}{n}\sum_{t=1}^n (f(x_t) - y_t)^2$. Let $n_1 = |\{t : x_t = x^{(1)}\}|$ and $n_2 = |\{t : x_t = x^{(2)}\}|$ be the number of realizations in the samples for $x^{(1)}$ and $x^{(2)}$. Following the logic in Foster and Krishnamurthy (2021), consider the event $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ where:

1. $\mathcal{E}_1$: Exactly one sample is drawn at $x^{(2)}$ ($n_2 = 1$).

2. $\mathcal{E}_2$: The label for that sample is $y = 0$.
3. $\mathcal{E}_3$ The empirical mean at $x^{(1)}$, $\hat{\mu}_1 = \frac{1}{n_1}\sum_{t:x_t=x^{(1)}} y_t$, is such that $\hat{\mu}_1 \leq \frac{3}{2}\mu_n$

For any function f, we can decompose the sum of squares as

$$\sum_{t:x_t=x^{(1)}} (y_t - f(x^{(1)}))^2 = n_1(\hat{\mu}_1 - f(x^{(1)}))^2 + \sum_{t:x_t=x^{(1)}} (y_t - \hat{\mu}_1)^2.$$

The second term is a constant C for a fixed dataset. We have

- Loss of f^*:

$$\hat{L}_{LS}(f^*) = \frac{1}{n}\left[n_1(\hat{\mu}_1 - \mu_n)^2 + C + (1/2 - 0)^2\right] \geq \frac{1}{4n} + \frac{C}{n}$$

- Loss of $\tilde{f}$:

$$\hat{L}_{LS}(\tilde{f}) = \frac{1}{n}\left[n_1(\hat{\mu}_1 - \sqrt{1/16n})^2 + C + (0 - 0)^2\right] = \frac{1}{n}\left[n_1(\hat{\mu}_1 - \sqrt{1/16n})^2\right] + \frac{C}{n}$$

by Lemma 12, $\mathbb{P}(\mathcal{E}) > \frac{1}{10}$. Conditioning on $\mathcal{E}$, $\hat{\mu}_1 \leq \frac{3}{2}\mu_n$, we observe that for sufficiently large n, $\hat{\mu}_1$ is negligible compared to $\sqrt{1/16n}$. Thus:

$$\hat{L}_{LS}(\tilde{f}) = \frac{n_1}{n}\left(\hat{\mu}_1 - \sqrt{\frac{1}{16n}}\right)^2 + \frac{C}{n} < \frac{1}{4n} + \frac{C}{n}.$$

Because $\hat{L}_{LS}(\tilde{f}) < \hat{L}_{LS}(f^*)$, the squared loss minimizer picks $\hat{f}_{LS} = \tilde{f}$. The resulting expected absolute error for picking $\tilde{f}$ is:

$$\mathbb{E}_x[|f^*(x) - \tilde{f}(x)|] = (1-p_n)|\mu_n - \sqrt{1/16n}| + p_n|1/2 - 0| > \left(1 - \frac{1}{n}\right)\left|\frac{128}{n} - \frac{1}{4\sqrt{n}}\right| = \Omega\left(\frac{1}{\sqrt{n}}\right).$$

While a first-order bound demands a rate of $\mathcal{O}(1/n)$, the squared loss achieves only $\Omega(1/\sqrt{n})$. This gap proves that squared loss cannot achieve first-order bounds in this setting. ■

Lemma 12 *Recall the event defined in Proposition 11. We have, or any $n > 256$,*

$$\mathbb{P}(\mathcal{E}) > \frac{1}{10}.$$

Proof The number of samples n_2 follows a Binomial distribution $B(n, 1/n)$.

$$\mathbb{P}(n_2 = 1) = \binom{n}{1}\left(\frac{1}{n}\right)^1\left(1 - \frac{1}{n}\right)^{n-1} = \left(1 - \frac{1}{n}\right)^{n-1}.$$

For $n \geq 2$, the sequence $(1 - 1/n)^{n-1}$ is monotonically decreasing toward $1/e$. Thus:

$$\mathbb{P}(n_2 = 1) > \frac{1}{e}.$$

Since $Y|x^{(2)} \sim$ Bernoulli$(1/2)$, the label $y = 0$ occurs with probability $1/2$ independently of the context selection. Therefore:

$$\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) > \frac{1}{2e} \approx 0.1839.$$

We use the multiplicative Chernoff bound for the sum of $n_1 = n - 1$ independent Bernoulli random variables. For $\delta > 0$,

$$\mathbb{P}(\hat{\mu}_1 \geq (1 + \delta)\mu_n) \leq e^{-\frac{\delta^2\mu_n n_1}{2+\delta}}.$$

Setting $\delta = 1/2$,

$$\mathbb{P}(\hat{\mu}_1 \geq \frac{3}{2}\mu_n) \leq e^{-\frac{(1/4)\mu_n(n-1)}{2.5}} = e^{-\frac{\mu_n(n-1)}{10}}.$$

Substituting $\mu_n = 128/n$:

$$e^{-\frac{128(n-1)}{10n}} = e^{-12.8(1-1/n)}.$$

For $n \geq 256$, $(1-1/n) \geq 255/256$. The exponent is roughly -12.75, yielding $e^{-12.75} \approx 0.0000029$. Thus:

$$\mathbb{P}(\mathcal{E}_3) = 1 - \mathbb{P}(\hat{\mu}_1 \geq \frac{3}{2}\mu_n) \geq 0.9999.$$

Taking a union bound,

$$\mathbb{P}(\mathcal{E}) = \mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) > \frac{1}{10}.$$

■

Appendix D. Log Loss Fails to Achieve Second-Order Bound

Proof [Proof of Theorem 3]

Fix $n \geq 2$. Let the instance space be

$$\mathcal{X} = \{a, b\},$$

and define the marginal distribution of X by

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

The conditional distribution of $Y \in [0, 1]$ is defined as follows:

$$Y \mid X = a = \frac{1}{2} \quad \text{almost surely},$$

and

$$Y \mid X = b \sim \text{Bernoulli}\left(\frac{1}{2}\right).$$

Thus the regression function is

$$f^\star(a) = f^\star(b) = \frac{1}{2}.$$

The conditional variances are

$$\sigma_a^2 = 0, \qquad \sigma_b^2 = \frac{1}{4},$$

and hence

$$\mathbb{E}[\sigma_x^2] = \left(1 - \frac{1}{n}\right) \cdot 0 + \frac{1}{n} \cdot \frac{1}{4} = \frac{1}{4n}.$$

Now define a competing function g by

$$g(a) = \frac{1}{2} + \eta_n, \qquad g(b) = \frac{3}{4},$$

where

$$\eta_n := \frac{1}{4\sqrt{n}}.$$

Let

$$\mathcal{F}_n = \{f^\star, g\}.$$

Both functions take values in $(0, 1)$, so the log loss is finite.

We will show that g has strictly smaller empirical log loss than $f^\star$ with constant probability. Consider the event

$$E := \{\text{exactly one sample has } X = b, \text{ and its label is } Y = 1\}.$$

Since $Y \mid X = b \sim \text{Bernoulli}(1/2)$,

$$\mathbb{P}(E) = \binom{n}{1}\frac{1}{n}\left(1 - \frac{1}{n}\right)^{n-1} \cdot \frac{1}{2} = \frac{1}{2}\left(1 - \frac{1}{n}\right)^{n-1}.$$

Using the standard inequality

$$\left(1 - \frac{1}{n}\right)^{n-1} \geq e^{-1}, \qquad n \geq 2,$$

we get

$$\mathbb{P}(E) \geq \frac{1}{2e}.$$

On event E, there are $n-1$ samples at a, all with label $1/2$, and one sample at b with label 1. Let

$$S_n := \sum_{i=1}^{n} \left(\ell(g; (X_i, Y_i)) - \ell(f^\star; (X_i, Y_i))\right),$$

where

$$\ell(f; (x, y)) := -y \log f(x) - (1 - y) \log(1 - f(x)).$$

If $S_n < 0$, then g has strictly smaller empirical log loss than $f^\star$, so every log-loss ERM over $\mathcal{F}_n$ selects g.

At point a, where $Y = 1/2$, the log-loss difference is

$$\begin{aligned} d_a &:= \ell(g; (a, 1/2)) - \ell(f^\star; (a, 1/2)) \\ &= -\frac{1}{2}\log\left(\frac{1}{2} + \eta_n\right) - \frac{1}{2}\log\left(\frac{1}{2} - \eta_n\right) + \log\left(\frac{1}{2}\right) \\ &= -\frac{1}{2}\log(1 - 4\eta_n^2). \end{aligned}$$

Since $\eta_n = 1/(4\sqrt{n})$, this becomes

$$d_a = -\frac{1}{2}\log\left(1 - \frac{1}{4n}\right).$$

At point b, on the favorable label $Y = 1$, the log-loss difference is

$$\begin{aligned} d_b^+ &:= \ell(g; (b, 1)) - \ell(f^\star; (b, 1)) \\ &= -\log(3/4) + \log(1/2) \\ &= -\log(3/2). \end{aligned}$$

Therefore, on event E,

$$S_n = (n-1)d_a + d_b^+ = (n-1)\left[-\frac{1}{2}\log\left(1-\frac{1}{4n}\right)\right] - \log(3/2).$$

We now upper-bound the first term. Using

$$-\log(1-u) \le \frac{u}{1-u}, \qquad 0 \le u < 1,$$

with $u = 1/(4n)$, we obtain

$$\begin{aligned}(n-1)d_a &\le n \cdot \frac{1}{2} \cdot \frac{1/(4n)}{1-1/(4n)} \\ &= \frac{1}{8(1-1/(4n))} \\ &\le \frac{1}{6},\end{aligned}$$

where the last inequality uses $n \ge 1$. Since

$$\frac{1}{6} < \log(3/2),$$

it follows that

$$S_n < 0$$

on event E. Hence, on E, the log-loss ERM selects g.

It remains to compute the L^1 error of g. We have

$$\begin{aligned}\mathbb{E}_X|g(X) - f^\star(X)| &= \left(1-\frac{1}{n}\right)\left|\frac{1}{2}+\eta_n-\frac{1}{2}\right| + \frac{1}{n}\left|\frac{3}{4}-\frac{1}{2}\right| \\ &= \left(1-\frac{1}{n}\right)\eta_n + \frac{1}{4n}.\end{aligned}$$

In particular, for $n \ge 2$,

$$\mathbb{E}_X|g(X) - f^\star(X)| \ge \left(1-\frac{1}{n}\right)\frac{1}{4\sqrt{n}} \ge \frac{1}{8\sqrt{n}}.$$

Since $\hat{f}_{\log} = g$ on event E, we conclude that

$$\mathbb{P}\left(\mathbb{E}_X\left|\hat{f}_{\log}(X) - f^\star(X)\right| \ge \frac{1}{8\sqrt{n}}\right) \ge \mathbb{P}(E) \ge \frac{1}{2e}.$$

■

Appendix E. Proof of Theorem 4

Definition 13 *We first provide definitions for new quantities that are used throughout the proof of Theorem 4.*

$$\begin{aligned}\Delta_x :=& f_x^* - f_x \\ \overline{\Delta}_{h,x,\phi,c} :=& \overline{\big(\phi(h_x - f_x)\big)}_{[-c,c]} \\ U_x :=& \max\{(-f_x^*)\frac{-\overline{\Delta}_{h,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{h,x,\phi,c}},\ (1-f_x^*)\frac{-\overline{\Delta}_{h,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{h,x,\phi,c}}\}\end{aligned}$$

Lemma 14 *For any $x \in \mathcal{X}$, we have:*

1. $\overline{\Delta}_{f^*,x,\phi,c} = \operatorname{sign}(f_x^* - f_x)\ \big(\phi|f_x^* - f_x| \wedge c\big)$, *and* $\Delta_x\overline{\Delta}_{f^*,x,\phi,c} \geq 0$.
2. $U_x \leq \frac{1}{4}$.

Proof

1. By the definition of $\overline{\Delta}_{h,x,\phi,c}$, one can see $\overline{\Delta}_{f^*,x,\phi,c} = \overline{\big(\phi(f_x^* - f_x)\big)}_{[-c,c]} = \operatorname{sign}(f_x^* - f_x)\ \big(\phi|f_x^* - f_x| \wedge c\big)$, and $\Delta_x\overline{\Delta}_{f^*,x,\phi,c} \geq 0$ for all x.
2. Note that

$$\begin{aligned}|\overline{\Delta}_{f^*,x,\phi,c}| =& \phi|f_x^* - f_x| \wedge c \\ \leq& c \\ \leq& \frac{1}{4}. && (c \leq \tfrac{1}{4})\end{aligned}$$

If $\overline{\Delta}_{f^*,x,\phi,c} \geq 0$,

$$\begin{aligned}U_x =& \max\{(-f_x^*)\frac{-\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}},\ (1-f_x^*)\frac{-\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}\} \\ =& f_x^*\frac{\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}} \\ \leq& \overline{\Delta}_{f^*,x,\phi,c} && (\forall x,\ \Delta_x\overline{\Delta}_{f^*,x,\phi,c} \geq 0,\ 0 \leq f_x^* \leq 1) \\ \leq& \frac{1}{4}. && (|\overline{\Delta}_{f^*,x,\phi,c}| \leq \tfrac{1}{4})\end{aligned}$$

Similarly, we can show that if $\overline{\Delta}_{f^*,x,\phi,c} \leq 0$, then $U_x = (1-f_x^*)\frac{-\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}} \leq \frac{1}{4}$.

■

Lemma 15 *Let $a \in (0,1)$. Then, $\forall x \in [0,a],\ \ln(1-x) \geq \frac{-\ln(1-a)}{a} \cdot (-x)$.*

Proof Given the concavity of $\ln(1-x)$, for any $x \in [0,a]$, the function lies above the secant line connecting $(0, \ln(1-0))$ and $(a, \ln(1-a))$.

The equation of the secant line is:

$$y = \frac{\ln(1-a)}{a}x.$$

By concavity:

$$\ln(1-x) \geq \frac{\ln(1-a)}{a} x.$$

■

Lemma 16 *Let $\delta \in (0, \frac{1}{|\mathcal{F}|})$. We have,*

$$1 - |\mathcal{F}|\delta \leq \mathbb{P}\left(\forall h \in \mathcal{F},\ \phi \in [0, \overline{\phi}],\ c \in [0, \frac{1}{4}],\ \frac{1}{n} H_{\phi,c}(h, f^*) \leq \frac{1}{n} \ln(8\overline{\phi} n^2/\delta)\right)$$

Proof The plan is to fix h and show

$$1 - \delta \leq \mathbb{P}\left(\forall \phi^* \in [0, \overline{\phi}], c^* \in [0, \frac{1}{4}],\ \frac{1}{n} H_{\phi^*,c^*}(h, f^*) \leq \frac{1}{n} \ln(8\overline{\phi} n^2/\delta)\right)$$

and then take the union bound over $h \in \mathcal{F}$.

Let $\varepsilon > 0$ be a small number to be chosen later. Discretize $[0, \overline{\phi}] \times [0, \frac{1}{4}]$ as blocks of length ε by ε. The number of such blocks is $\frac{\overline{\phi}}{4\varepsilon^2}$. For any (ϕ^*, c^*), there is block, such that (ϕ^*, c^*) belongs to this block. Let U' be the uniform distribution supported on this block.

We start from the martingale

$$\frac{\mathbb{E}_{(\phi,c)\sim U'}[\exp(H_{\phi,c}(h, f^*))]}{\mathbb{E}_{(\phi,c)\sim U', \{x,y\}\sim D^n}[\exp(H_{\phi,c}(h, f^*))]}.$$

Using Markov's inequality, we have, w.p. at least $1 - \delta/(\frac{\overline{\phi}}{4\varepsilon^2})$,

$$\begin{aligned}
\ln(\mathbb{E}_{(\phi,c)\sim U'}[\exp(H_{\phi,c}(h, f^*))]) &\leq \ln(\mathbb{E}_{(\phi,c)\sim U', \{(x,y)\}\sim D^n}[\exp(H_{\phi,c}(h, f^*))]) + \ln(\frac{\overline{\phi}}{4\varepsilon^2\delta}) \\
&= \ln(\mathbb{E}_{(\phi,c)\sim U'}(\mathbb{E}_{\{(x,y)\}\sim D}[1 + (y - f^*)\overline{(\phi(h_x - f_x))}_{[-c,c]}])^n) + \ln(\frac{\overline{\phi}}{4\varepsilon^2\delta}) \\
&\qquad \text{(independence)} \\
&= \ln(\frac{\overline{\phi}}{4\varepsilon^2\delta}). \qquad (10)
\end{aligned}$$

Taking a union bound over all $\frac{\overline{\phi}}{4\varepsilon^2}$ blocks, we have with probability at least $1 - \delta$, for any U that is a uniform distribution on any block,

$$\ln(\mathbb{E}_{(\phi,c)\sim U}[\exp(H_{\phi,c}(h, f^*))]) \leq \ln(\frac{\overline{\phi}}{4\varepsilon^2\delta}).$$

We desire to lower bound the LHS of Equation (10) above as $H_{\phi^*,c^*}(h, f^*)$ plus some extra terms for any (ϕ^*, c^*) that belongs to the support of U'.

Note that

$$\mathbb{E}_{(\phi,c)\sim U'}[\exp(H_{\phi,c}(h, f^*))] = \mathbb{E}_{(\phi,c)\sim U'}\left(\prod_{(x,y)}\left(1 + (y - f_x^*)\overline{(\phi(h_x - f_x^*))}_{[-c,c]}\right)\right).$$

Note that if $|\phi^* - \phi| \le \varepsilon$ and $|c^* - c| \le \varepsilon$, then using 1-Lipschitzness of $F_1(\phi) = 1 + (y - f_x^*)\overline{\left(\phi(h_x - f_x^*)\right)}_{[-c,c]}$, and 1-Lipschitzness of $F_2(c) = 1 + (y - f_x^*)\overline{\left(\phi(h_x - f_x^*)\right)}_{[-c,c]}$,

$$\begin{aligned}
&1 + (y - f_x^*)\overline{\left(\phi(h_x - f_x^*)\right)}_{[-c,c]} \\
&\ge 1 + (y - f_x^*)\overline{\left(\phi^*(h_x - f_x^*)\right)}_{[-c,c]} - \varepsilon \\
&\ge 1 + (y - f_x^*)\overline{\left(\phi^*(h_x - f_x^*)\right)}_{[-c^*,c^*]} - 2\varepsilon \\
&= (1 + (y - f_x^*)\overline{\left(\phi^*(h_x - f_x^*)\right)}_{[-c^*,c^*]}) \cdot (1 - \frac{2\varepsilon}{1 + (y - f_x^*)\overline{\left(\phi^*(h_x - f_x^*)\right)}_{[-c^*,c^*]}}) \\
&\ge (1 + (y - f_x^*)\overline{\left(\phi^*(h_x - f_x^*)\right)}_{[-c^*,c^*]}) \cdot (1 - \frac{8}{3}\varepsilon). \qquad (c^* \le \tfrac{1}{4})
\end{aligned}$$

Thus,

$$\begin{aligned}
\ln(\mathbb{E}_{(\phi,c)\sim U}[\exp(H_{\phi,c}(h, f^*))]) &\ge \sum_{(x,y)} \ln\left(1 + (y - f_x^*)\overline{\left(\phi^*(h_x - f_x^*)\right)}_{[-c^*,c^*]}\right) + n\ln(1 - \frac{8}{3}\varepsilon) \\
&\ge \sum_{(x,y)} \ln\left(1 + (y - f_x^*)\overline{\left(\phi^*(h_x - f_x^*)\right)}_{[-c^*,c^*]}\right) - n\varepsilon.
\end{aligned}$$

(Lemma 15; $\varepsilon \le \frac{1}{8}$)

This implies that

$$\frac{1}{n} H_{\phi^*,c^*}(h, f^*) \le \varepsilon + \frac{1}{n}\ln(\frac{\overline{\phi}}{4\varepsilon^2\delta}).$$

Choosing $\varepsilon = \frac{1}{4n}$, the RHS of above inequality can be upper bounded as:

$$\varepsilon + \frac{1}{n}\ln(\frac{\overline{\phi}}{4\varepsilon^2\delta}) \le \frac{1}{n}\ln(8\overline{\phi}n^2/\delta),$$

concluding the proof. ■

Lemma 17 *Let $\delta \in (0, \frac{1}{|\mathcal{F}|})$. Then,*

$$\begin{aligned}
1 - |\mathcal{F}|\delta \le \mathbb{P}\Bigg(&\forall f \in \mathcal{F},\ \phi \in [0, \overline{\phi}],\ c \in [0, \frac{1}{4}], \\
&-\frac{1}{n} H_{\phi,c}(f^*, f) \le \mathbb{E}_x\left[-\frac{\Delta_x \overline{\Delta}_{f^*,x,\phi,c}}{1 + \Delta_x \overline{\Delta}_{f^*,x,\phi,c}} + \frac{4}{3} \cdot \sigma_x^2 \overline{\Delta}^2_{f^*,x,\phi,c}\right] + \frac{1}{n}\ln(24\overline{\phi}n^2/\delta)\Bigg).
\end{aligned}$$

Proof The plan is to fix f and show

$$\begin{aligned}
1 - \delta \le \mathbb{P}\Bigg(&\forall \phi^* \in [0, \overline{\phi}], c^* \in [0, \frac{1}{4}],\ -\frac{1}{n} H_{\phi^*,c^*}(f^*, f) \\
&\le \mathbb{E}_x\left[-\frac{\Delta_x \overline{\Delta}_{f^*,x,\phi^*,c^*}}{1 + \Delta_x \overline{\Delta}_{f^*,x,\phi^*,c^*}} + \frac{4}{3} \cdot \sigma_x^2 \overline{\Delta}^2_{f^*,x,\phi^*,c^*}\right] + \frac{1}{n}\ln(24\overline{\phi}n^2/\delta)\Bigg)
\end{aligned}$$

and then take the union bound over $f \in \mathcal{F}$.

Let $\varepsilon > 0$ be a small number to be chosen later. Discretize $[0, \overline{\phi}] \times [0, \frac{1}{4}]$ as blocks of length ε by ε. The number of such blocks is $\frac{\overline{\phi}}{4\varepsilon^2}$. For any (ϕ^*, c^*), there is block, such that (ϕ^*, c^*) belongs to this block. Let U' be the uniform distribution supported on this block.

We start from the martingale

$$\frac{\mathbb{E}_{(\phi,c)\sim U'}[\exp(-H_{\phi,c}(f^*,f))]}{\mathbb{E}_{(\phi,c)\sim U',\{x,y\}\sim D}[\exp(-H_{\phi,c}(f^*,f))]}.$$

Using Markov's inequality, we have, w.p. at least $1-\delta/(\frac{\overline{\phi}}{4\varepsilon^2})$,

$$\begin{aligned}
&\ln(\mathbb{E}_{(\phi,c)\sim U'}[\exp(-H_{\phi,c}(f^*,f))]) \\
\le&\ln(\mathbb{E}_{(\phi,c)\sim U',\{(x,y)\}\sim D^n}[\exp(-H_{\phi,c}(f^*,f))]) + \ln(\frac{\overline{\phi}}{4\varepsilon^2}/\delta) \\
\le& n\,\mathbb{E}_{(\phi,c)\sim U'}\,\mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}} + \frac{1}{(1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c})^3(1-U_x)}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi,c}\right] + \ln(\frac{\overline{\phi}}{4\varepsilon^2}/\delta)
\end{aligned} \tag{11}$$

where the last inequality is by Lemma 18.

Taking a union bound over all $\frac{\overline{\phi}}{4\varepsilon^2}$ blocks, we have with probability at least $1-\delta$, for any U that is a uniform distribution on any block,

$$\begin{aligned}
&\ln(\mathbb{E}_{(\phi,c)\sim U}[\exp(-H_{\phi,c}(f^*,f))]) \\
\le& n\,\mathbb{E}_{(\phi,c)\sim U'}\,\mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}} + \frac{1}{(1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c})^3(1-U_x)}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi,c}\right] + \ln(\frac{\overline{\phi}}{4\varepsilon^2}/\delta).
\end{aligned}$$

We upper bound the RHS of (11) as follows:

$$\begin{aligned}
&\mathbb{E}_{(\phi,c)\sim U'}\,\mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}} + \frac{1}{(1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c})^3(1-U_x)}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi,c}\right] \\
\le&\,\mathbb{E}_{(\phi,c)\sim U'}\,\mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}} + \frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi,c}\right] \text{ (By Lemma 14: } \Delta_x\overline{\Delta}_{g,x,c}\ge 0,\ U_x\le\tfrac{1}{4})
\end{aligned}$$

One can see that $\overline{\Delta}_{f^*,x,\phi,c} = \overline{(\phi(f_x^*-f_x))}_{[-c,c]}$ is 1-Lipschitz in ϕ and 1-Lipschitz in c, i.e., $F_1(\phi)=\overline{\Delta}_{f^*,x,\phi,c}$ is 1-Lipschitz, and $F_2(c)=\overline{\Delta}_{f^*,x,\phi,c}$ is 1-Lipschitz. Further, $F_1^2(\phi)$ and $F_2^2(c)$ are 1-Lipschitz since $\overline{\Delta}_{f^*,x,\phi,c}\le c\le\frac{1}{4}$. In addition, since for $x\in[0,\frac{1}{4}]$, $|\frac{\mathrm{d}}{\mathrm{d}x}\frac{x}{1+x}| = \frac{1}{(1+x)^2}\le 1$, $\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}$ is 1-Lipschitz in ϕ and 1-Lipschitz in c.

Note that if $|\phi^*-\phi|\le\varepsilon$ and $|c^*-c|\le\varepsilon$, then using Lipschitzness arguments above, as well as $\sigma_x^2\le\frac{1}{4},\ \forall x$, we have

$$\begin{aligned}
-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}} + \frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi,c} &\le -\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi^*,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi^*,c}} + \frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi^*,c} + 2\varepsilon \\
&\le -\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi^*,c^*}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi^*,c^*}} + \frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi^*,c^*} + 4\varepsilon.
\end{aligned}$$

This implies that,

$$\begin{aligned}
&\mathbb{E}_{(\phi,c)\sim U'}\,\mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}} + \frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi,c}\right] \\
\le&\,\mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi^*,c^*}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi^*,c^*}} + \frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi^*,c^*}\right] + 4\varepsilon.
\end{aligned}$$

For the LHS of (11),

$$\mathbb{E}_{(\phi,c)\sim U'}[\exp(-H_{\phi,c}(f^*,f))] = \mathbb{E}_{(\phi,c)\sim U'}\left(\prod_{(x,y)}\frac{1}{\left(1+(y-f_x)\overline{(\phi(f_x^*-f_x))}_{[-c,c]}\right)}\right).$$

Note that if $|\phi^*-\phi|\le\varepsilon$ and $|c^*-c|\le\varepsilon$, then using 1-Lipschitzness of $F_3(\phi)=1+(y-f_x)\overline{(\phi(f_x^*-f_x))}_{[-c,c]}$, and 1-Lipschitzness of $F_4(c)=1+(y-f_x)\overline{(\phi(f_x^*-f_x))}_{[-c,c]}$,

$$\begin{aligned}
&1+(y-f_x)\overline{(\phi(f_x^*-f_x))}_{[-c,c]}\\
&\le 1+(y-f_x)\overline{(\phi^*(f_x^*-f_x))}_{[-c,c]}+\varepsilon\\
&\le 1+(y-f_x)\overline{(\phi^*(f_x^*-f_x))}_{[-c^*,c^*]}+2\varepsilon\\
&= (1+(y-f_x)\overline{(\phi^*(f_x^*-f_x))}_{[-c^*,c^*]})\cdot(1+\frac{2\varepsilon}{1+(y-f_x)\overline{(\phi^*(f_x^*-f_x))}_{[-c^*,c^*]}})\\
&\le (1+(y-f_x)\overline{(\phi^*(f_x^*-f_x))}_{[-c^*,c^*]})\cdot(1+\frac{8}{3}\varepsilon) && (c^*\le\tfrac{1}{4})
\end{aligned}$$

Thus,

$$\begin{aligned}
\ln\left(\mathbb{E}_{(\phi,c)\sim U'}[\exp(-H_{\phi,c}(f^*,f))]\right) &= \mathbb{E}_{(\phi,c)\sim U'}\left(\prod_{(x,y)}\frac{1}{\left(1+(y-f_x)\overline{(\phi(f_x^*-f_x))}_{[-c,c]}\right)}\right)\\
&\ge \sum_{(x,y)}-\ln\left(1+(y-f_x)\overline{(\phi^*(f_x^*-f_x))}_{[-c^*,c^*]}\right)-n\ln(1+\frac{8}{3}\varepsilon)\\
&\ge \sum_{(x,y)}-\ln\left(1+(y-f_x)\overline{(\phi^*(f_x^*-f_x))}_{[-c^*,c^*]}\right)-n\frac{8}{3}\varepsilon\\
&&(\ln(1+x)\le x)
\end{aligned}$$

Combining the bounds for the LHS and RHS of (11),

$$\begin{aligned}
&\sum_{(x,y)}-\ln\left(1+(y-f_x)\overline{(\phi^*(f_x^*-f_x))}_{[-c^*,c^*]}\right)-n\frac{8}{3}\varepsilon\\
&\le n\,\mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi^*,c^*}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi^*,c^*}}+\frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi^*,c^*}\right]+4n\varepsilon+\ln(\frac{\overline{\phi}}{4\varepsilon^2}/\delta).
\end{aligned}$$

This implies that

$$-\frac{1}{n}H_{\phi^*,c^*}(f^*,f)\le\frac{20}{3}\varepsilon+\mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi^*,c^*}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi^*,c^*}}+\frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi^*,c^*}\right]+\frac{1}{n}\ln(\frac{\overline{\phi}}{4\varepsilon^2}/\delta).$$

Choosing $\varepsilon=\frac{1}{4n}$,

$$\begin{aligned}
-\frac{1}{n}H_{\phi^*,c^*}(f^*,f)&\le\frac{20}{3}\varepsilon+\mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi^*,c^*}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi^*,c^*}}+\frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi^*,c^*}\right]+\frac{1}{n}\ln(\frac{\overline{\phi}}{4\varepsilon^2}/\delta)\\
&\le\mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi^*,c^*}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi^*,c^*}}+\frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi^*,c^*}\right]+\frac{1}{n}\ln(24\overline{\phi}n^2/\delta).
\end{aligned}$$

■

Lemma 18 *Recall the definition of the loss function $H_{\phi,c}$ and $U_x = \max\{(-f_x^*)\frac{-\overline{\Delta}_{h,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{h,x,\phi,c}},\ (1-f_x^*)\frac{-\overline{\Delta}_{h,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{h,x,\phi,c}}\}$. Let V be a distribution of (ϕ,c) supported on a subset of $[0,\overline{\phi}]\times[0,\frac{1}{4}]$. Then for any $h, f\in\mathcal{F}$, we have*

$$\begin{aligned}&\ln(\mathbb{E}_{(\phi,c)\sim V,\{(x,y)\}\sim D^n}[\exp(-H_{\phi,c}(h,f))])\\ \le& n\,\mathbb{E}_{(\phi,c)\sim V}\,\mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{h,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{h,x,\phi,c}}+\frac{1}{(1+\Delta_x\overline{\Delta}_{h,x,\phi,c})^3(1-U_x)}\cdot\sigma_x^2\overline{\Delta}^2_{h,x,\phi,c}\right].\end{aligned}$$

Proof Let $\eta := y - f^*$, then $\forall x\in\mathcal{X},\ \mathbb{E}[\eta\mid x]=0$ and $\mathbb{E}[\eta^2\mid x]=\sigma_x^2$. We have

$$\begin{aligned}&(\mathbb{E}_{(\phi,c)\sim V,\{(x,y)\}\sim D^n}[\exp(-H_{\phi,c}(h,f))])^{\frac{1}{n}}\\ &=\mathbb{E}_{(\phi,c)\sim V,\{(x,y)\}\sim D}[(\frac{1}{1+(y-f_x)\overline{\big(\phi(h_x-f_x)\big)}_{[-c,c]}})]\\ &=\mathbb{E}_{(\phi,c)\sim V,\{(x,y)\}\sim D}[(\frac{1}{1+(f_x^*+\eta-f_x)\overline{\big(\phi(h_x-f_x)\big)}_{[-c,c]}})]\\ &=\mathbb{E}_{(\phi,c)\sim V,\{(x,y)\}\sim D}[(\frac{1}{1+\Delta_x\overline{\Delta}_{h,x,\phi,c}+\eta\overline{\Delta}_{h,x,\phi,c}})]\\ &=\mathbb{E}_{(\phi,c)\sim V}\,\mathbb{E}_x\left[\frac{1}{1+\Delta_x\overline{\Delta}_{h,x,\phi,c}}\,\mathbb{E}_\eta[\frac{1}{1+\eta\overline{\Delta}_{h,x,\phi,c}\cdot(1+\Delta_x\overline{\Delta}_{h,x,\phi,c})^{-1}}]\right].\end{aligned}$$

Using the fact that $\frac{1}{1+x}=1-x+\frac{x^2}{1+x}$ with $x=\eta\overline{\Delta}_{h,x,\phi,c}\cdot(1+\Delta_x\overline{\Delta}_{h,x,\phi,c})^{-1}$, we have

$$\mathbb{E}_\eta[\frac{1}{1+\eta\overline{\Delta}_{h,x,\phi,c}\cdot(1+\Delta_x\overline{\Delta}_{h,x,\phi,c})^{-1}}]=1+\mathbb{E}_\eta[\frac{\eta^2\overline{\Delta}^2_{h,x,\phi,c}}{(1+\Delta_x\overline{\Delta}_{h,x,\phi,c})^2}\cdot\frac{1}{1+\eta\overline{\Delta}_{h,x,\phi,c}\cdot(1+\Delta_x\overline{\Delta}_{h,x,\phi,c})^{-1}}].$$

If $\overline{\Delta}_{h,x,\phi,c}\ge 0$, then the RHS $\le 1+\frac{\sigma_x^2\overline{\Delta}^2_{h,x,\phi,c}}{(1+\Delta_x\overline{\Delta}_{h,x,\phi,c})^2}\cdot\frac{1}{1+(-f^*)\overline{\Delta}_{h,x,\phi,c}\cdot(1+\Delta_x\overline{\Delta}_{h,x,\phi,c})^{-1}}$. Else if $\overline{\Delta}_{h,x,\phi,c}<0$, then the RHS $\le 1+\frac{\sigma_x^2\overline{\Delta}^2_{h,x,\phi,c}}{(1+\Delta_x\overline{\Delta}_{h,x,\phi,c})^2}\cdot\frac{1}{1+(1-f^*)\overline{\Delta}_{h,x,\phi,c}\cdot(1+\Delta_x\overline{\Delta}_{h,x,\phi,c})^{-1}}$.

Thus, with $U_x=\max\{(-f_x^*)\frac{-\overline{\Delta}_{h,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{h,x,\phi,c}},\ (1-f_x^*)\frac{-\overline{\Delta}_{h,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{h,x,\phi,c}}\}$,

$$\begin{aligned}&\frac{1}{n}\ln(\mathbb{E}_{(\phi,c)\sim V,\{(x,y)\}\sim D^n}[\exp(-H_{\phi,c}(h,f))])\\ \le&\ln\mathbb{E}_{(\phi,c)\sim V}\,\mathbb{E}_x\left[\frac{1}{1+\Delta_x\overline{\Delta}_{h,x,\phi,c}}\left(1+\sigma_x^2\overline{\Delta}^2_{h,x,\phi,c}\cdot\frac{1}{(1+\Delta_x\overline{\Delta}_{h,x,\phi,c})^2(1-U_x)}\right)\right]\\ \le&\,\mathbb{E}_{(\phi,c)\sim V}\,\mathbb{E}_x\left[\frac{1}{1+\Delta_x\overline{\Delta}_{h,x,\phi,c}}\left(1+\sigma_x^2\overline{\Delta}^2_{h,x,\phi,c}\cdot\frac{1}{(1+\Delta_x\overline{\Delta}_{h,x,\phi,c})^2(1-U_x)}\right)-1\right]\\ &\qquad\qquad(\ln x\le x-1)\\ =&\,\mathbb{E}_{(\phi,c)\sim V}\,\mathbb{E}_x\left[\frac{1}{1+\Delta_x\overline{\Delta}_{h,x,\phi,c}}\left(\sigma_x^2\overline{\Delta}^2_{h,x,\phi,c}\cdot\frac{1}{(1+\Delta_x\overline{\Delta}_{h,x,\phi,c})^2(1-U_x)}-\Delta_x\overline{\Delta}_{h,x,\phi,c}\right)\right],\end{aligned}$$

completing the proof. ■

Theorem 19 (Restatement of Theorem 4) *Recall that*

$$L_n(f) := \max_{h\in\mathcal{F}} \max_{\phi\in[0,\overline{\phi}]} \max_{c\in[0,\frac{1}{4}]} \frac{1}{n} \sum_{(x,y)\in D_n} \ln\left(1 + (y - f_x)\overline{(\phi(h_x - f_x))}_{[-c,c]}\right)$$

With probability at least $1-\delta$, $\forall f \in \mathcal{F}$,

$$\mathbb{E}_x |f_x - f_x^*|$$
$$\leq \sqrt{\frac{25}{12}\mathbb{E}\,\sigma_x^2 \cdot \left(\frac{2}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) + (L_n(f) - L_n(f^*))\right)} + \frac{6}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) + \frac{5}{2}(L_n(f) - L_n(f^*))$$

Proof Define the events

$$A_1 := 3\{\forall h \in \mathcal{F},\ \phi \in [0,\overline{\phi}],\ c \in [0,\frac{1}{4}],\ \frac{1}{n}H_{\phi,c}(h,f^*) \leq \frac{1}{n}\ln\left(\frac{16|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) 3\}$$

$$A_2 := \Bigg\{\forall f \in \mathcal{F},\ \phi \in [0,\overline{\phi}],\ c \in [0,\frac{1}{4}],$$
$$-\frac{1}{n}H_{\phi,c}(f^*,f) \leq \mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}} + \frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi,c}\right] + \frac{1}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right)\Bigg\}$$

$A := A_1 \cap A_2$.

By Lemma 16, $\mathbb{P}(A_1) \geq 1 - \frac{\delta}{2}$; by Lemma 17, $\mathbb{P}(A_2) \geq 1 - \frac{\delta}{2}$. Taking a union bound, one can see that

$$\mathbb{P}(A) \geq 1 - \delta.$$

The subsequent reasoning conditions on A. $\forall f \in \mathcal{F}$, we have

$$L_n(f^*) - L_n(f)$$
$$= \max_{h\in\mathcal{F},\phi'\in[0,\overline{\phi}],c'\in[0,\frac{1}{4}]} \min_{h'\in\mathcal{F},\phi\in[0,\overline{\phi}],c\in[0,\frac{1}{4}]} \frac{1}{n}H_{\phi',c'}(h,f^*) - \frac{1}{n}H_{\phi,c}(h',f) \qquad \text{(definition of } L\text{)}$$
$$\leq \max_{h\in\mathcal{F},\phi'\in[0,\overline{\phi}],c'\in[0,\frac{1}{4}]} \min_{\phi\in[0,\overline{\phi}],c\in[0,\frac{1}{4}]} \frac{1}{n}H_{\phi',c'}(h,f^*) - \frac{1}{n}H_{\phi,c}(f^*,f) \qquad (f^* \in \mathcal{F})$$
$$\leq \max_{h\in\mathcal{F},\phi'\in[0,\overline{\phi}],c'\in[0,\frac{1}{4}]} \min_{\phi\in[0,\overline{\phi}],c\in[0,\frac{1}{4}]} \frac{1}{n}\ln\left(\frac{16|\mathcal{F}|\overline{\phi}n^2}{\delta}\right)$$
$$+ \mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}} + \frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi,c}\right] + \frac{1}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right)$$
(definition of A_1, A_2)
$$= \min_{\phi\in[0,\overline{\phi}],c\in[0,\frac{1}{4}]} \frac{1}{n}\ln\left(\frac{16|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) + \mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}} + \frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi,c}\right] + \frac{1}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right)$$
$$\leq \min_{\phi\in[0,\overline{\phi}],c\in[0,\frac{1}{4}]} \mathbb{E}_x\left[-\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}} + \frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi,c}\right] + \frac{2}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right).$$

That is,

$$\max_{\phi\in[0,\overline{\phi}]}\max_{c\in[0,\frac{1}{4}]}\underbrace{\mathbb{E}_x\left[\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}-\frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi,c}\right]}_{=:\,\mathrm{LHS}}\le\frac{2}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right)+(L_n(f)-L_n(f^*)).$$

Recall that $\Delta_x=f_x^*-f_x$ and $\overline{\Delta}_{f^*,x,\phi,c}=\overline{\left(\phi(f_x^*-f_x)\right)}_{[-c,c]}$. By Lemma 14, $\Delta_x\overline{\Delta}_{f^*,x,\phi,c}\ge 0$ for all x and $U_x\le\frac{1}{4}$. Therefore,

$$\begin{aligned}
\mathrm{LHS}&=\mathbb{E}_x\left[\frac{\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}{1+\Delta_x\overline{\Delta}_{f^*,x,\phi,c}}-\frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi,c}\right]\\
&\ge\mathbb{E}_x\left[\frac{4}{5}\Delta_x\overline{\Delta}_{f^*,x,\phi,c}-\frac{4}{3}\cdot\sigma_x^2\overline{\Delta}^2_{f^*,x,\phi,c}\right]\qquad(\Delta_x\overline{\Delta}_{f^*,x,\phi,c}\ge 0,\ |\overline{\Delta}_{f^*,x,\phi,c}|\le\tfrac{1}{4},|\Delta_x|\le 1)\\
&=\mathbb{E}_x\left[\frac{4}{5}|\Delta_x|\left(\phi|f^*-f|\wedge c\right)-\frac{4}{3}\cdot\sigma_x^2\left(\phi|f^*-f|\wedge c\right)^2\right]\\
&\qquad\qquad(\overline{\Delta}_{f^*,x,\phi,c}=\mathrm{sign}(f^*-f)\left(\phi|f^*-f|\wedge c\right))\\
&=\frac{4}{5}\,\mathbb{E}_x\left[|\Delta_x|^2\left(\phi\wedge\frac{c}{|\Delta_x|}\right)\left[1-\frac{5}{3}\cdot\sigma_x^2\left(\phi\wedge\frac{c}{|\Delta_x|}\right)\right]\right].
\end{aligned}$$

We want to set c and ϕ such that

$$\mathbb{E}\,\frac{1}{2}|\Delta_x|^2\left(\phi\wedge\frac{c}{|\Delta_x|}\right)\ge\mathbb{E}\,\frac{5}{3}\cdot|\Delta_x|^2\sigma_x^2\left(\phi\wedge\frac{c}{|\Delta_x|}\right)^2,\tag{12}$$

which will give us the inequality of

$$\frac{4}{5}\,\mathbb{E}\,\frac{1}{2}|\Delta_x|^2\left(\phi\wedge\frac{c}{|\Delta_x|}\right)\le\frac{2}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right)+(L_n(f)-L_n(f^*)).\tag{13}$$

We choose ϕ such that $\phi=\frac{c}{\Delta^*}$ for some Δ^* to be chosen later, we can see that $\phi\wedge\frac{c}{|\Delta_x|}=c\left(\frac{1}{\Delta^*}\wedge\frac{1}{|\Delta_x|}\right)$. Using this, the above inequality (12) becomes:

$$\mathbb{E}\,\frac{1}{2}|\Delta_x|^2c\left(\frac{1}{\Delta^*}\wedge\frac{1}{|\Delta_x|}\right)\ge\mathbb{E}\,\frac{5}{3}\cdot|\Delta_x|^2c^2\sigma_x^2\left(\frac{1}{\Delta^*}\wedge\frac{1}{|\Delta_x|}\right)^2.$$

We choose $c:=c_0\wedge\frac{1}{4}$, where

$$c_0=\frac{\mathbb{E}\,\frac{1}{2}|\Delta_x|^2\left(\frac{1}{\Delta^*}\wedge\frac{1}{|\Delta_x|}\right)}{\mathbb{E}\,\frac{5}{3}\cdot|\Delta_x|^2\sigma_x^2\left(\frac{1}{\Delta^*}\wedge\frac{1}{|\Delta_x|}\right)^2}$$

- If $c_0 \leq \frac{1}{4}$, then $c = c_0$. Plugging this into (13) along with the fact $\phi \wedge \frac{c}{|\Delta_x|} = c\left(\frac{1}{\Delta^*} \wedge \frac{1}{|\Delta_x|}\right)$, we have

$$\begin{aligned}
&\frac{4}{5}\left[\mathbb{E}\,\frac{1}{2}|\Delta_x|^2 c\left(\frac{1}{\Delta^*} \wedge \frac{1}{|\Delta_x|}\right)\right]^2 \\
&\leq \mathbb{E}\,\frac{5}{3}\cdot|\Delta_x|^2\sigma_x^2\left(\frac{1}{\Delta^*} \wedge \frac{1}{|\Delta_x|}\right)^2 \cdot \left(\frac{2}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) + (L_n(f) - L_n(f^*))\right) \\
&\leq \mathbb{E}\,\frac{5}{3}\sigma_x^2 \cdot \left(\frac{2}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) + (L_n(f) - L_n(f^*))\right) \\
\implies &\left[\mathbb{E}\,|\Delta_x|^2\left(\frac{1}{\Delta^*} \wedge \frac{1}{|\Delta_x|}\right)\right]^2 \\
&\leq \frac{25}{12}\,\mathbb{E}\,\sigma_x^2 \cdot \left(\frac{2}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) + (L_n(f) - L_n(f^*))\right).
\end{aligned}$$

  We could lower bound the LHS above by picking out the region with $|\Delta_x| \geq \Delta^*$ to arrive at:

$$\mathbb{E}\,\mathbb{1}\left\{|\Delta_x| \geq \Delta^*\right\}|\Delta_x| \leq \sqrt{\frac{25}{12}\,\mathbb{E}\,\sigma_x^2 \cdot \left(\frac{2}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) + (L_n(f) - L_n(f^*))\right)}$$

- If $c_0 > \frac{1}{4}$, then $c = \frac{1}{4}$. $c = \frac{1}{4} < c_0$ implies that (12) is true.

  Plugging $c = \frac{1}{4}$ into (13) along with the fact $\phi \wedge \frac{c}{|\Delta_x|} = c\left(\frac{1}{\Delta^*} \wedge \frac{1}{|\Delta_x|}\right)$, we have

$$\frac{2}{5}\,\mathbb{E}\,|\Delta_x|^2\left(\frac{1}{\Delta^*} \wedge \frac{1}{|\Delta_x|}\right) \leq \frac{2}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) + (L_n(f) - L_n(f^*)).$$

  We could lower bound the LHS above by picking out the region with $|\Delta_x| \geq \Delta^*$ to arrive at:

$$\begin{aligned}
&\frac{2}{5}\,\mathbb{E}\,\mathbb{1}\left\{|\Delta_x| \geq \Delta^*\right\}|\Delta_x| \leq \frac{2}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) + (L_n(f) - L_n(f^*)) \\
\implies &\mathbb{E}\,\mathbb{1}\left\{|\Delta_x| \geq \Delta^*\right\}|\Delta_x| \leq 5\frac{1}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) + \frac{5}{2}(L_n(f) - L_n(f^*))
\end{aligned}$$

In either case, we have:

$$\begin{aligned}
\mathbb{E}\,\mathbb{1}\left\{|\Delta_x| \geq \Delta^*\right\}|\Delta_x| \leq &\sqrt{\frac{25}{12}\,\mathbb{E}\,\sigma_x^2 \cdot \left(\frac{2}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) + (L_n(f) - L_n(f^*))\right)} \\
&+ \frac{5}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) + \frac{5}{2}(L_n(f) - L_n(f^*)).
\end{aligned}$$

We choose $\Delta^* = \frac{1}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right)$, which gives us,

$$\mathbb{E}\,\mathbb{1}\left\{|\Delta_x| < \Delta^*\right\}|\Delta_x| \le \frac{1}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right).$$

Altogether, we have,

$$\mathbb{E}_x\,|\Delta_x|$$
$$\le \sqrt{\frac{25}{12}\,\mathbb{E}\,\sigma_x^2 \cdot \left(\frac{2}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) + (L_n(f) - L_n(f^*))\right)} + \frac{6}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right) + \frac{5}{2}(L_n(f) - L_n(f^*)).$$

We verify the choice ϕ is valid as follows.

$$\begin{aligned}
\phi = \frac{c}{\Delta^*} &\le \frac{1}{4\Delta^*} && (c \le \tfrac{1}{4})\\
&= \frac{1}{4\frac{1}{n}\ln\left(\frac{48|\mathcal{F}|\overline{\phi}n^2}{\delta}\right)}\\
&= \frac{1}{4\frac{1}{n}\ln\left(\frac{12|\mathcal{F}|n^3}{\delta}\right)} && (\overline{\phi} = \tfrac{n}{4})\\
&\le \overline{\phi}
\end{aligned}$$

which validates that $\phi \in [0, \overline{\phi}]$.

■

Appendix F. Proof of Theorem 6

Lemma 20 *Recall that*

$$L_n(f) := \max_{h\in\mathcal{F}}\max_{\phi\in[0,\overline{\phi}]}\max_{c\in[0,\frac{1}{4}]}\frac{1}{n}\sum_{(x,y)\in D_n}\ln\left(1 + (y - f_x)\overline{(\phi(h_x - f_x))}_{[-c,c]}\right)$$

L is $\frac{4}{3}n$-Lipschitz w.r.t. $\|\cdot\|_\infty$.

Proof For fixed (h, ϕ, c), define:

$$\Phi(f, h, \phi, c) := \frac{1}{n}\sum_{(x,y)\in D_n}\ln\left(1 + (y - f_x)\overline{(\phi(h_x - f_x))}_{[-c,c]}\right).$$

We first show that $\Phi(f, h, \phi, c)$ is Lipschitz in f w.r.t. $\|\cdot\|_\infty$.

Let $\varphi_1(t) := (y - t)\overline{(\phi(h_x - t))}_{[-c,c]}$ for $t \in [0, 1]$ and $\varphi_2(t) := \ln(1 + t)$ for $t \in [-1/4, 1/4]$. If $\phi(h_x - t) \in [-c, c]$, then $|\varphi_1'(t)| = \phi|(t - y) + (t - h_x)| \le \phi + c \le n$; else if $\phi(h_x - t) \notin [-c, c]$, then $|\varphi_1'(t)| = c \le \frac{1}{4}$. Hence φ_1 is n-Lipschitz. $|\varphi_2'(t)| = \frac{1}{1+t} \le \frac{4}{3}$. Therefore, for any (h, ϕ, c), $\Phi(f, h, \phi, c)$ is $\frac{4}{3}n$-Lipschitz in f w.r.t. $\|\cdot\|_\infty$:

$$\forall f, f' \in \mathcal{F},\ \ |\Phi(f, h, \phi, c) - \Phi(f', h, \phi, c)| \le \frac{4}{3}n \cdot \|f - f'\|_\infty.$$

Furthermore, $\forall f, f' \in \mathcal{F}$,

$$\begin{aligned} L_n(f) - L_n(f') &= \max_{h\in\mathcal{F},\phi\in[0,\overline{\phi}],c\in[0,\frac{1}{4}]} \Phi(f,h,\phi,c) - \max_{h\in\mathcal{F},\phi\in[0,\overline{\phi}],c\in[0,\frac{1}{4}]} \Phi(f',h,\phi,c) \\ &\leq \max_{h\in\mathcal{F},\phi\in[0,\overline{\phi}],c\in[0,\frac{1}{4}]} \Phi(f,h,\phi,c) - \Phi(f',h,\phi,c) \\ &\leq \frac{4}{3}n \cdot \|f - f'\|_\infty \end{aligned}$$

By symmetry,

$$L_n(f') - L_n(f) \leq \frac{4}{3}n \cdot \|f - f'\|_\infty.$$

Therefore, $\forall f, f' \in \mathcal{F}$,

$$|L_n(f) - L_n(f')| \leq \frac{4}{3}n \cdot \|f - f'\|_\infty.$$

■

Theorem 21 (Parametric class. Restatement of Theorem 6) *Assume the covering number of $\mathcal{F}$ satisfies Eqn. (7). Then, with probability at least $1-\delta$, the output $\hat{f}$ of Algorithm 1 satisfies:*

$$\mathbb{E}_x\left|\hat{f}_x - f^*_x\right| \leq \sqrt{\frac{25}{3}\,\mathbb{E}\,\sigma_x^2 \frac{v}{n}\ln(\frac{12(1+A)n^5}{\delta})} + 12\frac{v}{n}\ln(\frac{12(1+A)n^5}{\delta})$$

Proof

Let $\mathcal{F}_\varepsilon$ be a minimum-cardinality proper ε-cover of $\mathcal{F}$ w.r.t. the metric $\|\cdot\|_\infty$. Then, we can designate $\hat{f}^\varepsilon \in \mathcal{F}_\varepsilon$ such that $\left\|\hat{f}^\varepsilon - \hat{f}\right\|_\infty \leq \varepsilon$.

Applying Theorem 19 with $\mathcal{F}' \leftarrow \mathcal{F}_\varepsilon \cup \{f^*\}$, we have

$$\mathbb{E}_x\left|\hat{f}^\varepsilon_x - f^*_x\right| \leq \sqrt{\frac{25}{12}\,\mathbb{E}\,\sigma_x^2 \cdot \left(2\frac{L}{n} + (L^\varepsilon_n(\hat{f}^\varepsilon) - L^\varepsilon_n(f^*))\right)} + 6\frac{L}{n} + \frac{5}{2}(L^\varepsilon_n(\hat{f}^\varepsilon) - L^\varepsilon_n(f^*)),$$

where

$$\begin{aligned} L &= \ln\left(\frac{48|\mathcal{F}'|\overline{\phi}n^2}{\delta}\right) \\ L^\varepsilon_n(f) &:= \max_{h\in\mathcal{F}'}\max_{\phi\in[0,\overline{\phi}]}\max_{c\in[0,\frac{1}{4}]}\frac{1}{n}\sum_{(x,y)\in D_n}\ln\left(1 + (y-f_x)\overline{(\phi(h_x - f_x))}_{[-c,c]}\right). \end{aligned}$$

For analysis purposes, we also denote:

$$L_n(f) := \max_{h\in\mathcal{F}}\max_{\phi\in[0,\overline{\phi}]}\max_{c\in[0,\frac{1}{4}]}\frac{1}{n}\sum_{(x,y)\in D_n}\ln\left(1 + (y-f_x)\overline{(\phi(h_x - f_x))}_{[-c,c]}\right).$$

Recall that our Algorithm 1 returns $\hat{f} \in \arg\min_{f\in\mathcal{F}} L_n(f)$. We have,

$$\begin{aligned} L^\varepsilon_n(\hat{f}^\varepsilon) - L^\varepsilon_n(f^*) &= \left(L^\varepsilon_n(\hat{f}^\varepsilon) - L^\varepsilon_n(\hat{f})\right) + \left(L^\varepsilon_n(\hat{f}) - L_n(\hat{f})\right) + \left(L_n(\hat{f}) - L_n(f^*)\right) + \left(L_n(f^*) - L^\varepsilon_n(f^*)\right) \\ &\leq \frac{4}{3}n \cdot \left\|\hat{f}^\varepsilon - \hat{f}\right\|_\infty + 0 + 0 + \left(L_n(f^*) - L^\varepsilon_n(f^*)\right), \end{aligned} \tag{14}$$

where the first term is by Lemma 20, the second term is by the definition of L^ε_n and L_n, and the third term is by the definition of $\hat{f}$.

We bound $L_n(f^*) - L_n^\varepsilon(f^*)$ as follows. Let h^* be such that

$$L_n(f^*) = \max_{\phi\in[0,\overline{\phi}]} \max_{c\in[0,\frac{1}{4}]} \frac{1}{n} \sum_{(x,y)\in D_n} \ln\left(1 + (y - f_x^*)\overline{\left(\phi(h_x^* - f_x^*)\right)}_{[-c,c]}\right).$$

Let $h_\varepsilon^* \in \mathcal{F}_\varepsilon$ be such that $\|h^* - h_\varepsilon^*\|_\infty \le \varepsilon$. Then,

$$\begin{aligned}
&L_n(f^*) - L_n^\varepsilon(f^*) \\
&\le \max_{\phi\in[0,\overline{\phi}]} \max_{c\in[0,\frac{1}{4}]} \frac{1}{n} \sum_{(x,y)\in D_n} \ln\left(1 + (y - f_x^*)\overline{\left(\phi(h_x^* - f_x^*)\right)}_{[-c,c]}\right) \\
&\quad - \max_{\phi\in[0,\overline{\phi}]} \max_{c\in[0,\frac{1}{4}]} \frac{1}{n} \sum_{(x,y)\in D_n} \ln\left(1 + (y - f_x^*)\overline{\left(\phi(h_{\varepsilon,x}^* - f_x^*)\right)}_{[-c,c]}\right).
\end{aligned}$$

Note that for any (ϕ, c),

$$\begin{aligned}
&\left|(y - f^*)\overline{\left(\phi\cdot(h^* - f^*)\right)}_{[-c,c]} - (y - f^*)\overline{\left(\phi\cdot(h_\varepsilon^* - f^*)\right)}_{[-c,c]}\right| \\
&= |(y - f^*)| \cdot \left|\overline{\left(\phi\cdot(h^* - f^*)\right)}_{[-c,c]} - \overline{\left(\phi\cdot(h_\varepsilon^* - f^*)\right)}_{[-c,c]}\right| \\
&\le |(y - f^*)| \cdot \frac{n}{4}\varepsilon \qquad (\phi \in [0, \tfrac{n}{4}])
\end{aligned}$$

Using $t \mapsto \ln(1+t)$ is $\frac{4}{3}$-Lipschitz for $t \in [-\frac{1}{4}, \frac{1}{4}]$,

$$\begin{aligned}
L_n(f^*) - L_n^\varepsilon(f^*) &\le \frac{4}{3}\cdot\frac{n}{4}\varepsilon \cdot \frac{1}{n}\sum_{(x,y)} |(y - f_x^*)| \\
&\le \frac{1}{3} n\varepsilon.
\end{aligned}$$

Plugging back into Eqn. (14),

$$L_n^\varepsilon(\hat{f}^\varepsilon) - L_n^\varepsilon(f^*) \le \frac{4}{3} n \cdot \left\|\hat{f}^\varepsilon - \hat{f}\right\|_\infty + \left(L_n(f^*) - L_n^\varepsilon(f^*)\right) \le 2n\varepsilon \tag{15}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_x\,|\hat{f}_x - f_x^*| \\
&\le \mathbb{E}_x\,|\hat{f}_x - \hat{f}_x^\varepsilon| + \mathbb{E}_x\,|\hat{f}_x^\varepsilon - f_x^*| && \text{(Triangle inequality)} \\
&\le \varepsilon + \sqrt{\frac{25}{12}\,\mathbb{E}\,\sigma_x^2 \cdot \left(2\frac{L}{n} + (L_n^\varepsilon(\hat{f}^\varepsilon) - L_n^\varepsilon(f^*))\right)} + 6\frac{L}{n} + \frac{5}{2}(L_n^\varepsilon(\hat{f}^\varepsilon) - L_n^\varepsilon(f^*)) && \text{(Theorem 19)} \\
&\le \varepsilon + \sqrt{\frac{25}{12}\,\mathbb{E}\,\sigma_x^2 \cdot \left(2\frac{L}{n} + 2n\varepsilon\right)} + 6\frac{L}{n} + \frac{5}{2}(2n\varepsilon) && \text{(Eqn. (15))} \\
&\le \sqrt{\frac{25}{12}\,\mathbb{E}\,\sigma_x^2(\frac{2}{n}\ln(\frac{(1+(A/\varepsilon)^v)48\overline{\phi}n^2}{\delta}) + 2n\varepsilon)} + 6\frac{1}{n}\ln(\frac{(1+(A/\varepsilon)^v)48\overline{\phi}n^2}{\delta}) + 6n\varepsilon,
\end{aligned}$$

where the last inequality is because the covering number of $\mathcal{F}$ satisfies Eqn. (7), i.e., for every $\varepsilon > 0$, $N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \le (\frac{A}{\varepsilon})^v$.

Choosing $\varepsilon = \frac{1}{n^2}$ gives:

$$\mathbb{E}_x\,|\hat{f}_x - f_x^*| \le \sqrt{\frac{25}{3}\,\mathbb{E}\,\sigma_x^2\frac{v}{n}\ln(\frac{12(1+A)n^5}{\delta})} + 12\frac{v}{n}\ln(\frac{12(1+A)n^5}{\delta}).$$

■

Appendix G. Proof of Corollary 7

Corollary 22 (Linear class. Restatement of Corollary 7) *Let $\mathcal{F}$ be a linear function class in d-dimensional space: $\mathcal{F} = \{x \mapsto x^\top\theta + \frac{1}{2} : \|\theta\|_2 \leq \frac{1}{2}\}$ and the instance space $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. Then, with probability at least $1-\delta$, the output $\hat{f}$ of Algorithm 1 satisfies:*

$$\mathbb{E}_x\,|\hat{f}_x - f^*_x| \leq \sqrt{\frac{25}{3}\,\mathbb{E}\,\sigma_x^2\frac{d}{n}\ln(\frac{36n^5}{\delta})} + 12\frac{d}{n}\ln(\frac{36n^5}{\delta})$$

Proof To make use of Theorem 6, we just need to show that the L_∞ covering number of this class, $N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)$, grows polynomially in $1/\varepsilon$. The covering number of the linear class is not new (e.g., Exercise 20.3 of Lattimore and Szepesvári (2018)), we include the proof for completeness.

Denote by $\mathcal{W}$ the parameter space: $\mathcal{W} := \{\theta : \|\theta\|_2 \leq \frac{1}{2}\}$.

Let f_u and f_v be two functions in $\mathcal{F}$. We have,

$$\begin{aligned}
\|f_u - f_v\|_\infty &= \sup_{x\in\mathcal{X}} |x^\top u - x^\top v| \\
&= \sup_{x\in\mathcal{X}} |x^\top(u-v)| \\
&\leq \sup_{x\in\mathcal{X}} \|u-v\|_2\|x\|_2 && \text{(Cauchy-Schwarz)} \\
&\leq \|u-v\|_2 && (\|x\|_2 \leq 1)
\end{aligned}$$

This implies that an ε-cover of the parameter space $\mathcal{W}$ induces an ε-cover of the function class $\mathcal{F}$. Therefore, we can bound the covering number of the function class by the covering number of the parameter space:

$$N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \leq N(\varepsilon, \mathcal{W}, \|\cdot\|_2) \tag{16}$$

The problem is now reduced to finding the covering number of the parameter space $\mathcal{W}$, which is a ball of radius $\frac{1}{2}$ in a d-dimensional Euclidean space. This is a standard geometric result. The number of ε-balls to cover a ball of radius B is bounded by:

$$N(\varepsilon, \mathcal{W}, \|\cdot\|_2) \leq \left(\frac{2B}{\varepsilon} + 1\right)^d = \left(\frac{1}{\varepsilon} + 1\right)^d \leq \left(\frac{2}{\varepsilon}\right)^d \text{ for } \varepsilon \leq 1$$

Combining with Eqn. (16), we arrive at:

$$N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \leq \left(\frac{2}{\varepsilon}\right)^d$$

Applying Theorem 6 with $v = d$ and $A = 2$ concludes the proof. ■

Appendix H. Proof of Corollary 8

Proof We verify that $\mathcal{F}$ has polynomial covering number in $\|\cdot\|_\infty$, and then apply Theorem 6.

For $\theta \in \mathbb{R}^d$, write

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

The sigmoid function satisfies

$$\sigma'(z) = \sigma(z)(1-\sigma(z)) \le \frac{1}{4} \qquad \text{for all } z \in \mathbb{R}.$$

Therefore, for any θ, θ' with $\|\theta\|_2, \|\theta'\|_2 \le B$,

$$\begin{aligned} \|f_\theta - f_{\theta'}\|_\infty &= \sup_{x \in \mathcal{X}} \left|\sigma(\theta^\top x) - \sigma((\theta')^\top x)\right| \\ &\le \frac{1}{4} \sup_{\|x\|_2 \le R} \left|x^\top(\theta - \theta')\right| \\ &\le \frac{R}{4}\|\theta - \theta'\|_2. \end{aligned}$$

Thus, if $\{\theta_1, \dots, \theta_N\}$ is a $(4\varepsilon/R)$-cover of the Euclidean ball

$$\Theta_B := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le B\}$$

in $\|\cdot\|_2$, then $\{f_{\theta_1}, \dots, f_{\theta_N}\}$ is an ε-cover of $\mathcal{F}$ in $\|\cdot\|_\infty$. Hence, for $R > 0$,

$$N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \le N\left(\frac{4\varepsilon}{R}, \Theta_B, \|\cdot\|_2\right).$$

By the standard covering-number bound for Euclidean balls,

$$N(r, \Theta_B, \|\cdot\|_2) \le \left(1 + \frac{2B}{r}\right)^d.$$

Taking $r = 4\varepsilon/R$, we obtain

$$N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \le \left(1 + \frac{BR}{2\varepsilon}\right)^d.$$

For $0 < \varepsilon \le 1$,

$$1 + \frac{BR}{2\varepsilon} \le \frac{1 + BR/2}{\varepsilon} \le \frac{1 + BR}{\varepsilon}.$$

Therefore,

$$N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \le \left(\frac{1 + BR}{\varepsilon}\right)^d, \qquad 0 < \varepsilon \le 1.$$

Thus $\mathcal{F}$ satisfies the polynomial $\|\cdot\|_\infty$-entropy condition with

$$v = d, \qquad A = 1 + BR.$$

Applying Theorem 6 with these parameters gives

$$\mathbb{E}_x\,|\hat{f}_x - f_x^*| \le C_1\sqrt{\mathbb{E}[\sigma_x^2]\,\frac{d}{n}\log\left(\frac{C_0(2+BR)n^5}{\delta}\right)} + C_2\frac{d}{n}\log\left(\frac{C_0(2+BR)n^5}{\delta}\right),$$

after absorbing universal constants into C_0, C_1, C_2. This proves the claim. ∎

Appendix I. A Cautionary Lipschitz-Class Example

This appendix provides evidence that extending Theorem 6 to rich nonparametric classes is nontrivial. The issue is related to known phenomena in optimistic-rate theory: Srebro et al. (2010) show that,

for scale-sensitive classes, fast rates can fail even for smooth and strongly convex losses, with $n^{-1/2}$ excess-risk rates being unavoidable in general. Our construction below is different in nature, but points in the same direction: for a Lipschitz class, the empirical betting loss can prefer a function $f_0 \neq f^\star$ over the true regression function $f^\star$ with constant probability.

We interpret this negative result as evidence that the betting-loss ERM may behave differently in rich nonparametric classes than in finite or parametric classes. Intuitively, the multiple max operations that define the betting loss can substantially change the geometry of the optimization in sufficiently rich classes like Lipschitz, making betting loss behave as if it were minimizing an L^1 (median-seeking) criterion. As a result, the predictor is pulled toward the conditional median rather than the conditional mean f^*. We therefore leave fast second-order rates for general nonparametric classes as open problems.

Theorem 23 *Consider the following problem instance. Let $\mathcal{X} = [0,1]$, and $\mathcal{F} \subset \{\mathcal{X} \to [0,1]\}$ be the Lipschitz function class with Lispchitz parameter $L = 101$, i.e., $\forall f \in \mathcal{F},\ \forall x, x' \in \mathcal{X},\ \left|f(x) - f(x')\right| \leq L \cdot \left|x - x'\right|$. Let $\mathcal{D}_X$, the marginal distribution of X, be the uniform distribution on $\mathcal{X}$, and $\forall x \in \mathcal{X},\ \mathcal{D}_{Y|X=x}$ be the Bernoulli distribution with parameter x.*

For this problem instance, the following holds:

there exists $f_0 \in \mathcal{F},\ f_0 \neq f^*$, constants $N > 0$ and $c_0, c_1 > 0$, such that $\forall n \geq N$,

$$\mathbb{P}(L_n(f^*) - L_n(f_0) > c_1) \geq c_0 \ .$$

Proof In this instance, the true regression function is $f^*_x = \mathbb{E}[Y \mid X = x] = x$. Note that f^* is 1-Lipschitz, so $f^* \in \mathcal{F}$.

We construct f_0 as follows:

$$f_0(x) = \begin{cases} 0 & x \in [0, 1/4] \\ 2x - 1/2 & x \in [1/4, 3/4] \\ 1 & x \in [3/4, 1] \end{cases}$$

Recall that $\forall f \in \mathcal{F}$,

$$H_{\phi,c}(h, f) := \sum_{(x,y)\in D_n} \ln\left(1 + (y - f_x)\overline{\left(\phi(h_x - f_x)\right)}_{[-c,c]}\right)$$

$$L_n(f) := \max_{h\in\mathcal{F}} \max_{\phi\in[0,\bar{\phi}]} \max_{c\in[0,\frac{1}{4}]} \frac{1}{n} H_{\phi,c}(h, f)$$

We may overload the notation $L_n(f)$ to $L_n(f, h, \phi, c)$, to denote the dependence on h, ϕ, c.

Note that on any datapoint (x, y),

$$\max_{h,\phi,c}(y - f_{0,x})\overline{\left(\phi(h_x - f_{0,x})\right)}_{[-c,c]} \leq |y - f_{0,x}| \cdot c \leq |y - f_{0,x}| \cdot \frac{1}{4}$$

Thus,

$$\begin{aligned} L_n(f_0) &\leq \frac{1}{n} \sum_{(x,y)\in D_n} \max_{h,\phi,c} \ln\left(1 + (y - f_{0,x})\overline{\left(\phi(h_x - f_{0,x})\right)}_{[-c,c]}\right) \\ &\leq \frac{1}{n} \sum_{(x,y)\in D_n} \ln\left(1 + |y - f_{0,x}| \cdot \frac{1}{4}\right) := \ell_{f_0} \end{aligned}$$

As $n \to \infty$, by the weak law of large numbers, the sample average converges in probability to its true expectation:

$$\ell_{f_0} = \frac{1}{n} \sum_{(x,y)\in D_n} \ln\left(1 + |y - f_{0,x}| \cdot \frac{1}{4}\right) \xrightarrow{p} \mathbb{E} \ln\left(1 + |y - f_{0,x}| \cdot \frac{1}{4}\right) = 0.0625$$

By the definition of convergence in probability, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\ell_{f_0} - 0.0625| < \varepsilon) = 1,$$

which implies that there exists N_1, such that $\forall n \geq N_1$,

$$\begin{aligned} & \mathbb{P}(|\ell_{f_0} - 0.0625| < 0.0005) \geq 0.9 \\ \implies & \mathbb{P}(\ell_{f_0} < 0.0630) \geq 0.9 \\ \implies & \mathbb{P}(L_n(f_0) < 0.0630) \geq 0.9 \end{aligned} \tag{17}$$

Next, we turn to lower bounding $L_n(f^*)$ for large enough n.

Recall that in our problem, X follows a uniform distribution on $[0, 1]$. In a sample drawn from this distribution, let a "gap" be the distance between two adjacent samples, i.e., $X_{(j)} - X_{(j-1)}$ for any $j \in [1, n]$ as well as $X_{(1)}$, where $X_{(j)}$ is the increasingly sorted sample.

We call it a "good" gap if it is $\geq \frac{0.0202}{n}$. By Lemma 24, if $n \geq 10^4$, then with probability at least 0.9, the number of "good" gaps is at least $0.97n$. Suppose the event that "the number of 'good' gaps is at least $0.97n$" happens.

We choose (h^*, ϕ^*, c^*) (not necessarily a maximizer) such that:

- $h^*_x - f^*_x = \text{sign}(y - f^*_x) \cdot \frac{1}{n}$ for only those data points next to a good gap. Indeed, such h^* is $(L = 101)$-Lipschitz: let $x_{(j)}$ be an observed datapoint next to a good gap, $x_{(j^-)}$ be the immediate previous datapoint that is next to a good gap, then $x_{(j)} - x_{(j^-)} \geq \frac{0.0202}{n}$, hence,

$$\begin{aligned} \left|h^*_{x_{(j)}} - h^*_{x_{(j^-)}}\right| &= \left|(h^*_{x_{(j)}} - f^*_{x_{(j)}}) - (h^*_{x_{(j^-)}} - f^*_{x_{(j^-)}}) + f^*_{x_{(j)}} - f^*_{x_{(j^-)}}\right| \\ &= \left|\text{sign}(y_{(j)} - f^*_{x_{(j)}}) \cdot \frac{1}{n} - \text{sign}(y_{(j^-)} - f^*_{x_{(j^-)}}) \cdot \frac{1}{n} + x_{(j)} - x_{(j^-)}\right| \\ & \qquad\qquad (f^*_x = x) \\ &\leq \frac{2}{n} + (x_{(j)} - x_{(j^-)}) \qquad \text{(Triangle inequality; } x_{(j)} - x_{(j^-)} \geq 0) \\ &\leq (\frac{2}{0.0202} + 1)(x_{(j)} - x_{(j^-)}) \qquad (x_{(j)} - x_{(j^-)} \geq \tfrac{0.0202}{n}) \\ &\leq 101 \cdot (x_{(j)} - x_{(j^-)}) \end{aligned}$$

  which means such h^* exists in $\mathcal{F}$.
- $c^* = \frac{1}{4}$.
- $\phi^* = \frac{n}{4}$.

Thus, $L_n(f^*) \geq L_n(f^*, h^*, \phi^*, c^*)$.

Let $z := |y - f^*_x|$ and sort z_i in increasing order as $z_{(i)}$. We can see that

$$L_n(f^*, h^*, \phi^*, c^*) \geq \frac{1}{n}\left[\sum_{i=1}^{0.97n} \ln\left(1 + z_{(i)} \cdot \frac{1}{4}\right) + \sum_{i=0.97n}^{n} \ln\left(1 - z_{(i)} \cdot \frac{1}{4}\right)\right] := \ell_{f^*} \tag{18}$$

As $n \to \infty$, by the weak law of large numbers,

$$\begin{aligned}\ell_{f^*} =& \frac{1}{n}\left[\sum_{i=1}^{0.97n} \ln\left(1 + z_{(i)} \cdot \frac{1}{4}\right) + \sum_{i=0.97n}^{n} \ln\left(1 - z_{(i)} \cdot \frac{1}{4}\right)\right] \\ \xrightarrow{p} & \int_0^{q_{0.97}} \ln\left(1 + z \cdot \frac{1}{4}\right) f_Z(z)\,\mathrm{d}z + \int_{q_{0.97}}^{1} \ln\left(1 - z \cdot \frac{1}{4}\right) f_Z(z)\,\mathrm{d}z, \end{aligned} \tag{19}$$

where $q_{0.97}$ is the 97th percentile of the distribution of Z.

We calculate the CDF and PDF of Z:

First note that

$$\begin{aligned}\mathbb{P}(Z \le z \mid X = x) &= \mathbb{P}(|Y - X| \le z \mid X = x) \\ &= (1 - x)\,\mathbb{1}\{x \le z\} + x\,\mathbb{1}\{1 - x \le z\}\end{aligned}$$

The reason is: if $Z = x$ then $Z \le z$ iff $x \le z$; if $Z = 1 - x$ then $Z \le z$ iff $1 - x \le z$. Thus for $z \in [0, 1]$,

$$\begin{aligned} F_Z(z) &= \mathbb{P}(Z \le z) \\ &= \mathbb{E}[\mathbb{P}(Z \le z) \mid X] && \text{(law of total probability)} \\ &= \int_0^1 [(1 - x)\,\mathbb{1}\{x \le z\} + x\,\mathbb{1}\{1 - x \le z\}]\,\mathrm{d}x \\ &= \int_0^1 [(1 - x)\,\mathbb{1}\{x \le z\} + x\,\mathbb{1}\{1 - z \le x\}]\,\mathrm{d}x \\ &= \int_0^z (1 - x)\,\mathrm{d}x + \int_{1-z}^1 x\,\mathrm{d}x \\ &= 2\int_0^z (1 - x)\,\mathrm{d}x && \text{(symmetry)} \\ &= 2(z - \frac{z^2}{2}) \\ &= 2z - z^2 \end{aligned}$$

Hence,

$$\begin{aligned} F_Z(z) &= 2z - z^2,\ z \in [0, 1] \\ f_Z(z) &= 2(1 - z),\ z \in [0, 1] \end{aligned}$$

Setting $F_Z(z) = 0.97$, the solution that lies within $[0, 1]$ is:

$$q_{0.97} = 1 - \sqrt{1 - 0.97} \approx 0.8268$$

We have,

$$\begin{aligned} &\int_0^{q_{0.97}} \ln\left(1 + z \cdot \frac{1}{4}\right) f_Z(z)\,\mathrm{d}z + \int_{q_{0.97}}^{1} \ln\left(1 - z \cdot \frac{1}{4}\right) f_Z(z)\,\mathrm{d}z \\ =& \int_0^{q_{0.97}} \ln\left(1 + z \cdot \frac{1}{4}\right) 2(1 - z)\,\mathrm{d}z + \int_{q_{0.97}}^{1} \ln\left(1 - z \cdot \frac{1}{4}\right) 2(1 - z)\,\mathrm{d}z \\ =& 0.0726 - 0.0075 \\ =& 0.0651 \end{aligned} \tag{20}$$

Putting together the above calculations (Eqns. (19) to (20)), we have, as $n \to \infty$,

$$\ell_{f^*} = \frac{1}{n}\left[\sum_{i=1}^{0.97n} \ln\left(1 + z_{(i)} \cdot \frac{1}{4}\right) + \sum_{i=0.97n}^{n} \ln\left(1 - z_{(i)} \cdot \frac{1}{4}\right)\right] \xrightarrow{p} 0.0651$$

By the definition of convergence in probability, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\ell_{f^*} - 0.0651| < \varepsilon) = 1,$$

which implies that there exists N_2, such that $\forall n \geq N_2$,

$$\begin{aligned} &\mathbb{P}(|\ell_{f^*} - 0.0651| < 0.0005) \geq 0.9 \\ \implies &\mathbb{P}(\ell_{f^*} > 0.0646) \geq 0.9 \end{aligned} \tag{21}$$

Combining $L_n(f^*) \geq L_n(f^*, h^*, \phi^*, c^*)$, Eqns. (18) (21) and taking a union bound with the event "the number of 'good' gaps is at least $0.97n$", which, as we mentioned and by Lemma 24, happens with probability at least 0.9, we have, $\forall n \geq \max\{10^4, N_2\}$,

$$\mathbb{P}(L_n(f^*) > 0.0646) \geq 0.8 \tag{22}$$

Combining Equations (17) and (22) with a union bound, we get $\forall n \geq \max\{10^4, N_1, N_2\}$,

$$\mathbb{P}(L_n(f_0) < 0.0630 < 0.0646 < L_n(f^*)) \geq 0.7$$

■

Lemma 24 *Suppose $n \geq 10^4$. In a sample set drawn i.i.d. from a uniform distribution on $[0, 1]$, let a "gap" be the distance between two adjacent samples.i.e., $X_{(j)} - X_{(j-1)}$ for any $j \in [1, n]$ as well as $X_{(1)}$ and where $X_{(j)}$ is the increasingly sorted sample.*

We call a cap "good" if it is $\geq \frac{k}{n}$, where $k = -\ln 0.98 = 0.0202$.

With probability at least 0.9, the number of "good" gaps is at least $0.97n$.

Proof In a sample drawn from the distribution on $[0, 1]$, by symmetry, all gaps i.e., $X_{(j)} - X_{(j-1)}$ for any $j \in [1, n]$ as well as $X_{(1)}$ where $X_{(j)}$ is the sorted sample, have the same probability distribution.

To study the properties of these gaps, we will find the distribution of the first gap, $X_{(1)}$, which is the simplest to compute.

We compute the CDF:

$$\begin{aligned} F_{X_{(1)}}(x) &= \mathbb{P}(X_{(1)} \leq x) \\ &= 1 - \mathbb{P}(X_{(1)} > x) \\ &= 1 - \mathbb{P}(\forall i \in [n],\ X_i > x) \\ &= 1 - (1 - x)^n \end{aligned}$$

Let the random variable G be a gap, that is, $G \overset{\mathrm{d}}{=} X_{(1)}$.

Recall that from the CDF of G, we have $P(G \leq t) = 1 - (1 - t)^n$. Hence $P(G \geq \frac{k}{n}) = (1 - \frac{k}{n})^n \xrightarrow{n\to\infty} e^{-k}$. Since $k = -\ln 0.98 = 0.0202$, we have w.p. 0.98, any single gap satisfies $G \geq \frac{0.0202}{n}$, i.e., w.p. 0.98, any single gap is good.

Let I_i be an indicator random variable for the i-th gap, where $i \in [n]$. $I_i = 1$ if the i-th gap is "good", i.e., $G_i \geq \frac{k}{n}$ and $I_i = 0$ otherwise. Hence, $\mathbb{P}(I_i = 1) = p = 0.98$.

Let Y be the total number of good gaps: $Y = \sum_{i=1}^{n} I_i$. By linearity of expectation,

$$\mathbb{E}[Y] = \mathbb{E}[\sum_{i=1}^{n} I_i] = 0.98n.$$

Our goal is to find an L such that $\mathbb{P}(Y \geq L) \geq 0.9$.

Note that

$$\text{Var}(Y) = \text{Var}\left(\sum_{i=1}^{n} I_i\right) = \sum_{i=1}^{n} \text{Var}(I_i) + \sum_{i \neq j} \text{Cov}(I_i, I_j).$$

We argue that $\text{Cov}(I_i, I_j) \leq 0$ for $i \neq j$. The intuitive interpretation is that, since the gaps are not independent, if one gap is very large, the others must be smaller to compensate, since their total length is fixed. This means they are negatively correlated, so $\text{Cov}(I_i, I_j) \leq 0$ for $i \neq j$. More formally, one can verify (by exchangeability) that for $i \neq j$,

$$\begin{aligned}
&\text{Cov}(I_i, I_j) = \mathbb{E}[I_i I_j] - \mathbb{E}[I_i]\,\mathbb{E}[I_j] \\
&\mathbb{E}[I_i] = \mathbb{P}(I_i = 1) = (1 - \frac{k}{n})^n \\
&\mathbb{E}[I_j] = \mathbb{P}(I_j = 1) = (1 - \frac{k}{n})^n \\
&\mathbb{E}[I_i I_j] = \mathbb{P}(I_i = 1, I_i = j) = (1 - 2\frac{k}{n})^n \\
\implies &\text{Cov}(I_i, I_j) \leq 0
\end{aligned}$$

Thus,

$$\text{Var}(Y) \leq \sum_{i=1}^{n} \text{Var}(I_i) = np(1-p) = 0.0196n.$$

Applying Chebyshev's inequality,

$$\begin{aligned}
&\mathbb{P}(Y \leq \mathbb{E}[Y] - \varepsilon) \leq \frac{\text{Var}(Y)}{\varepsilon^2} \\
\implies &\mathbb{P}(Y \leq 0.98n - \varepsilon) \leq \frac{np(1-p)}{\varepsilon^2} = \frac{0.0196n}{\varepsilon^2}
\end{aligned}$$

Solving $\frac{0.0196n}{\varepsilon^2} = 0.1$, we get $\varepsilon = \sqrt{0.196n}$. Therefore, with $n \geq 10^4$, with probability at least 0.9,

$$Y \geq 0.98n - \sqrt{0.196n} \geq 0.97n.$$

■

The implication: Theorem 23 shows that, for sufficiently large samples, with constant probability, there exists a sufficiently large gap between $L_n(f^*)$ and $L_n(f_0)$.

This theorem may support our conjecture about the inconsistency of the betting loss ERM estimator for nonparametric classes. The idea is that, if an ERM estimator is consistent, the true regression function f^* should effectively become the minimizer of the loss for large samples. However, Theorem 23 shows that this is demonstrably not true.

In more detail, let $L_n^{\min} = \min_{f \in \mathcal{F}} L_n(f)$. By definition, $L_n(\hat{f}) = L_n^{\min}$.

By Theorem 23, for all $n \geq N$,

$$\mathbb{P}(L_n(f^*) - L_n(f_0) > c_1) \geq c_0\ .$$

Since $L_n^{\min} \le L_n(f_0)$, this implies:

$$\mathbb{P}(L_n(f^*) - L_n^{\min} > c_1) \ge \mathbb{P}(L_n(f^*) - L_n(f_0) > c_1) \ge c_0 \,.$$

We conjecture that the significant difference of $L_n(f^*) - L_n(\hat{f})$ will effectively translate to a notable distance between $\hat{f}$ and f^*.

Appendix J. Comparing the Two First-Order Quantities

In this section, we first show that $\mathbb{E}_x[f^*(x) \wedge (1 - f^*(x))] \le \mathbb{E}_x[f^*(x)] \wedge \mathbb{E}_x[1 - f^*(x)]$, then we give an example where the difference between these two quantities can be arbitrarily large.

Lemma 25 *Recall that $f^* : \mathcal{X} \to [0, 1]$. We have,*

$$\mathbb{E}_x[f^*(x) \wedge (1 - f^*(x))] \le \mathbb{E}_x[f^*(x)] \wedge \mathbb{E}_x[1 - f^*(x)]$$

Proof Note that

$$\mathbb{E}_x[f^*(x) \wedge (1 - f^*(x))] \le \mathbb{E}_x[f^*(x)],$$

and

$$\mathbb{E}_x[f^*(x) \wedge (1 - f^*(x))] \le \mathbb{E}_x[1 - f^*(x)].$$

Hence,

$$\mathbb{E}_x[f^*(x) \wedge (1 - f^*(x))] \le \mathbb{E}_x[f^*(x)] \wedge \mathbb{E}_x[1 - f^*(x)].$$

■

Example 1 *Let $\varepsilon > 0$ be a small number, Y be of the distribution $\mathbb{P}(Y = \varepsilon) = \mathbb{P}(Y = 1 - \varepsilon) = \frac{1}{2}$. Then,*

$$\mathbb{E}[Y \wedge 1 - Y] = \varepsilon,$$

whereas

$$\mathbb{E}[Y] \wedge \mathbb{E}[1 - Y] = \frac{1}{2}.$$

Appendix K. Proof of Lemma 2

Proof

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathbb{E}[(Y - \mathbb{E}[Y])^2] \\
&= \mathbb{E}[Y^2 - 2Y\,\mathbb{E}[Y] + \mathbb{E}^2[Y]] \\
&\le \mathbb{E}[Y - 2Y\,\mathbb{E}[Y] + \mathbb{E}^2[Y]] && (Y \in [0, 1]) \\
&= \mathbb{E}[Y] - \mathbb{E}^2[Y] \\
&= \mathbb{E}[Y](1 - \mathbb{E}[Y])
\end{aligned}$$

One can see that the equality in the third line is attained iff Y is Bernoulli distributed. ■

Appendix L. Real-World Experiments

For experiments on the real-world datasets, we regenerated the labels to ensure realizability. All experiment settings are identical to those used in the main paper. We experiment on the Wine Quality datasets (Cortez et al., 2009), where the feature vectors are normalized via Min-Max scaling. As shown in Figure 2 and Figure 3, the betting loss consistently achieves the lowest MAE.

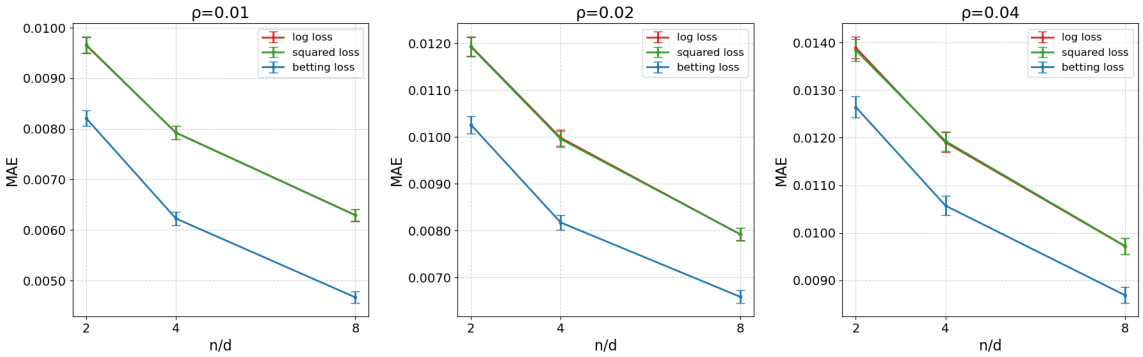

Figure 2: Comparison of average mean absolute error (MAE) obtained under log loss, squared loss and betting loss across varying n/d and ρ on the Wine Quality (Red) dataset. Error bars denote the standard error.

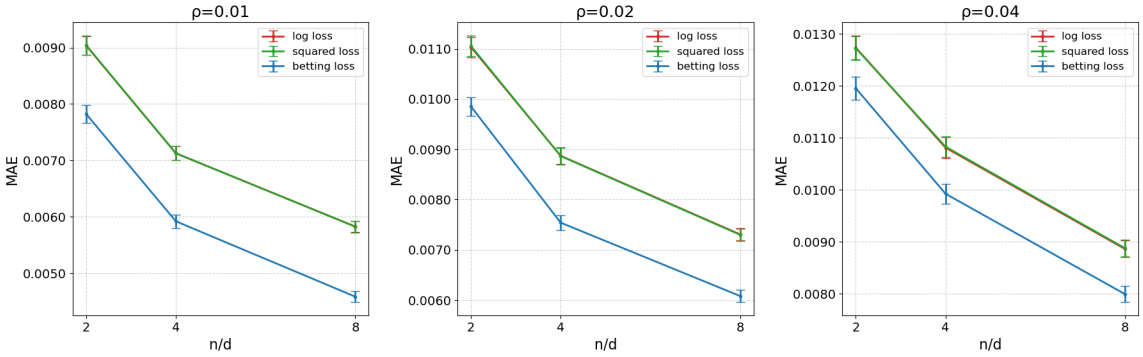

Figure 3: Comparison of average mean absolute error (MAE) obtained under log loss, squared loss and betting loss across varying n/d and ρ on the Wine Quality (White) dataset. Error bars denote the standard error.