Are all models wrong? Fundamental limits in distribution-free empirical model falsification

Manuel M. Müller MM2559@CAM.AC.UK
University of Cambridge

Yuetian Luo YUETIAN@UCHICAGO.EDU
University of Chicago

Rina Foygel Barber RINA@UCHICAGO.EDU
University of Chicago

Editors: Nika Haghtalab and Ankur Moitra

Abstract

In statistics and machine learning, when we train a fitted model on available data, we typically want to ensure that we are searching within a model class that contains at least one accurate model—that is, we would like to ensure an upper bound on the *model class risk* (the lowest possible risk that can be attained by any model in the class). However, it is also of interest to establish lower bounds on the model class risk, for instance so that we can determine whether our fitted model is at least approximately optimal within the class, or, so that we can decide whether the model class is unsuitable for the particular task at hand. Particularly in the setting of interpolation learning where machine learning models are trained to reach zero error on the training data, we might ask if, at the very least, a positive lower bound on the model class risk is possible—or are we unable to detect that "all models are wrong"? In this work, we answer these questions in a distribution-free setting by establishing a model-agnostic, fundamental hardness result for the problem of constructing a lower bound on the best test error achievable over a model class, and examine its implications on specific model classes such as tree-based methods and linear regression.

Keywords: Distribution-free Inference, Risk Bounds, Interpolation Learning

1. Introduction

A central goal in machine learning and statistics is to learn a rule that captures properties of interest of an unknown, data-generating distribution. For instance, in a regression setting, we might seek to identify a function whose output can accurately predict the response given an observed covariate. Typically in such a setting, we do not allow the fitted function to take any arbitrary form, but rather require it to be chosen from a (parametric or non-parametric) family of functions, a so-called *model class*, $\mathcal{F}$. Such restrictions to certain model classes can have many different motivations, such as encoding prior information, aiding interpretability, or allowing computationally efficient selection of a function. Indeed, any specific choice of analysis method restricts our model class in some way. This restriction may be explicit—for instance, with linear regression, we clearly limit ourselves to linear functions. But such a restriction can also be more implicit, such as through the choice of a specific neural network architecture, which is known to restrict the set of functions that can be represented, or the employed method used to eventually select a rule from the model class (DeVore et al., 2021; Petrova and Wojtaszczyk, 2023).

Choosing a model class $\mathcal{F}$ involves trading off between multiple considerations. A model class $\mathcal{F}$ that is too constrained might mean that there is no good model $f \in \mathcal{F}$ (all models in the class

are too simple to capture the true distribution). On the other hand, a model class $\mathcal{F}$ that is too complex may lead to challenges in selecting a good model $f \in \mathcal{F}$, if the available sample size is too small. In statistical learning theory, this is often referred to as the *approximation–estimation* tradeoff (Shalev-Shwartz and Ben-David, 2014). Moreover, an overly rich model class $\mathcal{F}$ can also lead to computational constraints, where searching within $\mathcal{F}$ is too costly, leading to an approximation–estimation–computation tradeoff (Bottou and Bousquet, 2007).

With this tradeoff in mind, we may ask the following question: after using our observed data to select a model $\hat{f} \in \mathcal{F}$, can we determine whether this selected model $\hat{f}$ is, at least approximately, optimal within this model class $\mathcal{F}$? This question has deep practical importance—if we are dissatisfied with $\hat{f}$'s accuracy (say, for predicting future data), then if we determine that $\hat{f}$ is nearly optimal within $\mathcal{F}$, we will need to consider switching to a richer class of models than $\mathcal{F}$; on the other hand, if $\hat{f}$ is far from optimal within $\mathcal{F}$, then we might simply need to gather more data in order to find a better model in the same class.

To make this question more precise, let $R_P(f)$ denote the risk of a model relative to the (unknown) data distribution P—for example, the misclassification probability, in the context of a classification problem. Then we would like to determine whether the quantity $R_P(\hat{f}) - \inf_{f\in\mathcal{F}} R_P(f)$ (sometimes referred to as the *estimation error* (Devroye et al., 1996)) is approximately zero, without relying on untestable assumptions on P. In order to do so, let us examine the feasibility of estimating both quantities: $R_P(\hat{f})$, the risk of our fitted model, and $\inf_{f\in\mathcal{F}} R_P(f)$, which we will refer to as the *model class risk*. Indeed, the difficulty of estimation is very different for these two quantities:

- Estimating $R_P(\hat{f})$ has a simple solution: by training $\hat{f}$ on only part of the available data, we may simply use the remaining data as a holdout set to provide an estimate of $R_P(\hat{f})$.

- Estimating $\inf_{f\in\mathcal{F}} R_P(f)$, on the other hand, is potentially quite challenging. Particularly in an overparameterized regime where a sample of size n can be interpolated by the model class, minimizing the empirical risk does not necessarily provide a reliable estimate of the true model class risk $\inf_{f\in\mathcal{F}} R_P(f)$.

While the former question is hence easy to address with a holdout set, in this paper we investigate the latter question. Is it possible to provide nontrivial bounds on the model class risk $\inf_{f\in\mathcal{F}} R_P(f)$, without placing assumptions on P—that is, in a distribution-free setting?

Outline. The remainder of this paper is organized as follows. Section 2 defines the questions of interest in more detail, and provides background and definitions. In Section 3, we present our results for two regimes, where the model class $\mathcal{F}$ is "low-complexity" or "high-complexity", while Section 4 examines examples lying in an interesting in-between regime. We conclude with a discussion in Section 5, including connections to related literature. Most proofs are deferred to the Appendix.

2. Problem formulation and background

In this section, we introduce some notation to formalize our questions, and give an overview of our contributions on the problem of inference on the model class risk.

2.1. Setting and notation

Let P denote the unknown distribution on the data space $\mathcal{Z}$, and let $\ell : \mathcal{F} \times \mathcal{Z} \to [0, \infty)$ denote a loss function, where $\mathcal{F}$ is the model class of interest. As a canonical example, in the setting of a binary classification problem, we might have $\mathcal{Z} = \mathbb{R}^d \times \{0, 1\}$, where a data point $Z = (X, Y)$ consists of features $X \in \mathbb{R}^d$ and a binary response $Y \in \{0, 1\}$, and we may then use the zero/one loss, which is defined as $\ell(f, Z) = \ell(f, (X, Y)) = \mathbb{1}\{f(X) \neq Y\}$.

Given a choice of loss function ℓ, we define the risk of a particular model f as

$$R_P(f) = \mathbb{E}_P[\ell(f, Z)],$$

and define the model class risk as $R_P(\mathcal{F}) := \inf_{f \in \mathcal{F}} R_P(f)$. Given a data set $\mathcal{D}_n = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$, the empirical risk is defined as

$$\hat{R}(f, \mathcal{D}_n) = \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i),$$

and we also write $\hat{R}(\mathcal{F}, \mathcal{D}_n) := \inf_{f \in \mathcal{F}} \hat{R}(f, \mathcal{D}_n)$ to denote the lowest risk achieved on the sample over all possible models $f \in \mathcal{F}$.

In the regression setting with data points $Z_i = (X_i, Y_i)$, it is common to say that $\mathcal{F}$ can *interpolate* the observed data $\mathcal{D}_n$ if we can find some $f \in \mathcal{F}$ with $f(X_i) = Y_i$ for all $i \in [n] := \{1, \ldots, n\}$; in this work, we will use this term more broadly and will say that $\mathcal{F}$ interpolates the data whenever $\hat{R}(\mathcal{F}, \mathcal{D}_n) = 0$, even though our work is not restricted to regression problems.

2.2. Key questions: lower bounds and upper bounds

As highlighted above, we are interested in performing distribution-free inference on the model class risk $R_P(\mathcal{F})$, which we formalize as follows.

Definition 1 *Fix $\alpha \in (0, 1)$, $n \geq 1$, and model class $\mathcal{F}$. We say that $\hat{L}_\alpha(\mathcal{F}, \cdot) : \mathcal{Z}^n \to [0, \infty]$ is a valid distribution-free lower bound on the model class risk of $\mathcal{F}$ if*

$$\mathbb{P}_P\Big\{R_P(\mathcal{F}) \geq \hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n)\Big\} \geq 1 - \alpha \quad \text{for all distributions } P \text{ on } \mathcal{Z}, \tag{1}$$

and similarly, $\hat{U}_\alpha(\mathcal{F}, \cdot) : \mathcal{Z}^n \to [0, \infty]$ is a valid distribution-free upper bound on the model class risk of $\mathcal{F}$ if

$$\mathbb{P}_P\Big\{R_P(\mathcal{F}) \leq \hat{U}_\alpha(\mathcal{F}, \mathcal{D}_n)\Big\} \geq 1 - \alpha \quad \text{for all distributions } P \text{ on } \mathcal{Z},$$

where $\mathbb{P}_P\{\cdot\}$ denotes that the probability is computed with respect to $\mathcal{D}_n \sim P^n$, i.e., a sample of size n drawn i.i.d. from P.[1]

1. We may also allow randomization in our lower and upper bounds. Formally, for the lower bound (and similarly for the upper bound), we define $\hat{L}_\alpha(\mathcal{F}, \cdot, \cdot) : \mathcal{Z}^n \times [0, 1] \to [0, \infty]$, where $\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n, \xi)$ now depends also on a random seed ξ, and the probability in (1) is now computed with respect to both $\mathcal{D}_n \sim P^n$ and $\xi \sim \text{Unif}[0, 1]$. All our results hold for both deterministic and randomized bounds, but for simplicity we suppress the random seed ξ in our notation.

While the questions of obtaining lower and upper bounds appear symmetric in the above definition, in fact they are substantially different. This is because the target of inference is an infimum, $R_P(\mathcal{F}) = \inf_{f\in\mathcal{F}} R_P(f)$. Therefore, a valid upper bound on $R_P(\mathcal{F})$ can be obtained by simply bounding $R_P(f)$ for any *single* choice of $f \in \mathcal{F}$. For example, a common strategy would be to partition the available data $\mathcal{D}_n$ into two independent subsets, $\mathcal{D}_{n/2} = (Z_1, \ldots, Z_{\lceil n/2 \rceil})$ and $\mathcal{D}'_{n/2} = (Z_{\lceil n/2 \rceil + 1}, \ldots, Z_n)$, then train a model $\hat{f} \in \mathcal{F}$ using $\mathcal{D}_{n/2}$, and finally use $\mathcal{D}'_{n/2}$ as a hold-out set to provide an unbiased estimate $\hat{R}(\hat{f}, \mathcal{D}'_{n/2})$ of the risk $R_P(\hat{f})$ (and we can also construct an upper bound on $R_P(\hat{f})$, via concentration arguments).

On the other hand, a lower bound on $R_P(\mathcal{F})$ is valid only if it bounds $R_P(f)$ simultaneously for *all* $f \in \mathcal{F}$, which is therefore a much more challenging task. The central aim of this paper is to explore the problem of constructing valid distribution-free lower bounds for $R_P(\mathcal{F})$. Of course, we may simply take $\hat{L}_\alpha(\mathcal{F}, \cdot) \equiv 0$ to ensure validity, but this is not an informative lower bound. If we allow randomization in our answer, moreover, we can even provide a trivial solution without always returning zero: since a valid lower bound can have error α according to Definition 1, we may return a trivial lower bound

$$\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) = \begin{cases} 0, & \text{with probability } 1-\alpha, \\ \infty, & \text{with probability } \alpha. \end{cases} \tag{2}$$

In other words, any meaningful answer to this question must have $\mathbb{P}_P\{\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0\}$ substantially larger than α; we will therefore refer to any lower bound with $\mathbb{P}_P\{\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0\} \leq \alpha + \mathrm{o}(1)$ as *trivial*. Thus, the central question of this work is to determine whether it is possible to construct a lower bound on $R_P(\mathcal{F})$ that is always valid, and is also nontrivial, to ensure that it is more informative than the meaningless solution constructed in (2).

2.3. Overview of contributions

We will now turn to the question of constructing valid distribution-free lower bounds for the model class risk. As a special case, we can ask when it is possible to verify that $R_P(\mathcal{F}) > 0$, i.e., that there is no perfect model in the class. This aim recalls George E. P. Box's often-quoted claim that, in statistics, "all models are wrong" (Box, 1976). While this well-known quotation is referring to the idea that any practical choice of the model class $\mathcal{F}$ must be wrong in terms of what it implies about the data distribution,[2] here we interpret the question in a different way: given a fixed model class $\mathcal{F}$, can we determine whether all models in the class are "wrong" in the sense that they do not lead to a zero risk?

As a starting point, consider the random variable $\hat{R}(\mathcal{F}, \mathcal{D}_n) = \inf_{f\in\mathcal{F}} \hat{R}(f, \mathcal{D}_n)$. In expectation, this quantity is an underestimate of the target, since

$$\mathbb{E}[\hat{R}(\mathcal{F}, \mathcal{D}_n)] = \mathbb{E}\left[\inf_{f\in\mathcal{F}} \hat{R}(f, \mathcal{D}_n)\right] \leq \inf_{f\in\mathcal{F}} \mathbb{E}\left[\hat{R}(f, \mathcal{D}_n)\right] = \inf_{f\in\mathcal{F}} R_P(f) = R_P(\mathcal{F}). \tag{3}$$

Therefore we might expect that the empirical model class risk $\hat{R}(\mathcal{F}, \mathcal{D}_n)$ could be a useful ingredient for constructing a distribution-free lower bound for $R_P(\mathcal{F})$. But if $\mathcal{F}$ is a high-complexity model class and can interpolate the training data, we will have $\hat{R}(\mathcal{F}, \mathcal{D}_n) = 0$. We may believe that this

2. In other words, Box (1976) uses the term "model" to refer to a class of functions $\mathcal{F}$, such as "the linear model", while in this paper we use "model" to refer to a single $f \in \mathcal{F}$.

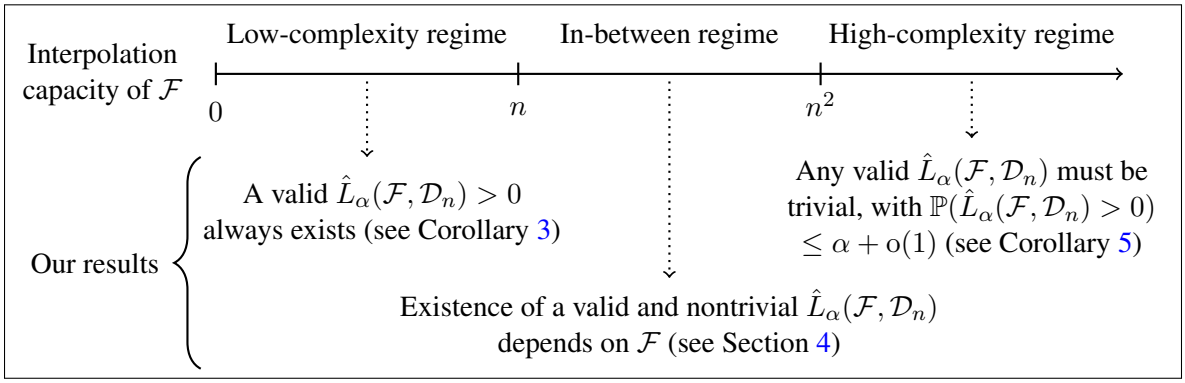

Figure 1: A schematic depiction of the role of the complexity of $\mathcal{F}$ (see Section 2.3 for discussion).

is due solely to the high complexity of the model class $\mathcal{F}$—but would it ever be possible to be confident that $R_P(\mathcal{F}) > 0$ in such a setting, or is $\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) = 0$ the only valid lower bound in this regime? More generally, is it ever possible to have a valid lower bound $\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n)$ that is higher than the empirical model class risk $\hat{R}(\mathcal{F}, \mathcal{D}_n)$?

In this paper we find that the problem can be characterized by three regimes, which are defined via the *interpolation capacity* of the model class $\mathcal{F}$—the largest sample size N for which $\hat{R}(\mathcal{F}, \mathcal{D}_N) = 0$. Of course, the definitions and results will be formalized below, but here we give an overview of our findings:

(i) **The low-complexity regime.** If $\mathcal{F}$ is not able to interpolate our data (i.e., $\hat{R}(\mathcal{F}, \mathcal{D}_n) > 0$), then we can always construct a valid distribution-free lower bound $\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0$.

(ii) **The high-complexity regime.** At the other extreme, consider a highly complex model class $\mathcal{F}$, which is able to interpolate not just our n training points but is in fact able to interpolate a data set of[3] size $\gg n^2$. Then any valid distribution-free lower bound must necessarily be trivial, with $\mathbb{P}\{\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0\} \leq \alpha + o(1)$.

(iii) **The in-between regime.** Finally, we find an interesting regime in between these two extremes. If $\mathcal{F}$ interpolates a data set of size n, but its interpolation capacity is $\mathcal{O}(n^2)$, then the question does not have a general answer: we will see specific examples of model classes with qualitatively different behavior within this regime. Counterintuitively, it is sometimes possible to verify that $R_P(\mathcal{F}) > 0$ ("all models are wrong") even when $\hat{R}(\mathcal{F}, \mathcal{D}_n) = 0$ (i.e., the empirical risk does not provide any evidence that "all models are wrong").

These three regimes are illustrated in Figure 1.

3. Information-theoretic extremes: the low-complexity and high-complexity regimes

In this section, we will develop theoretical results to determine whether it is possible to construct a meaningful and valid lower bound on $R_P(\mathcal{F})$.

3. For $a, b \geq 0$, we write $a \gg b$ in informal descriptions of our results whenever we mean that $b/a = o(1)$.

3.1. Defining interpolation capacity

As summarized in Figure 1, the low-complexity and high-complexity regimes are characterized by the *interpolation capacity* of the model class $\mathcal{F}$—informally, how large of a sample can be interpolated by some $f \in \mathcal{F}$. Before presenting our theoretical results, we begin by formalizing this measure of the complexity of $\mathcal{F}$. These definitions are closely related to complexity measures such as VC dimension and fat-shattering dimension that appear in the statistical learning theory literature (Devroye et al., 1996).

Given a model class $\mathcal{F}$ and a distribution P on $\mathcal{Z}$ (and some fixed choice of the loss function ℓ), we define

$$N(\mathcal{F}, P) = \sup\left\{n : \hat{R}(\mathcal{F}, \mathcal{D}_n) = 0 \ P\text{-almost surely}\right\}.$$

It will also be useful to define a related quantity,

$$N_+(\mathcal{F}, P) = \sup\left\{n : \mathbb{P}_P\left\{\hat{R}(\mathcal{F}, \mathcal{D}_n) = 0\right\} > 0\right\}.$$

In other words, for a sample size n, if $n \leq N(\mathcal{F}, P)$ then we will *always* observe zero empirical risk, $\hat{R}(\mathcal{F}, \mathcal{D}_n) = 0$, while if $n \leq N_+(\mathcal{F}, P)$ then we *may* observe zero empirical risk.

By definition, we have $N(\mathcal{F}, P) \leq N_+(\mathcal{F}, P)$—and in fact, in many common examples, these two measures of interpolation capacity are equal or approximately equal. For instance, in a regression setting with data points $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R} =: \mathcal{Z}$, if $\mathcal{F}$ is the class of all linear models (without an intercept), then $N(\mathcal{F}, P) = N_+(\mathcal{F}, P) = d$ for any distribution P with a density on $\mathbb{R}^d \times \mathbb{R}$. For this reason, when discussing our results, we treat these two different measures of interpolation capacity as essentially interchangeable, for the purpose of intuition.[4]

3.2. A lower bound via the empirical risk for the low-complexity regime

We will now see that the empirical risk can provide a distribution-free lower bound on the model class risk $R_P(\mathcal{F})$.

Theorem 2 *Fix $\alpha \in (0, 1)$, $n \geq 1$, and model class $\mathcal{F}$. Then*

$$\hat{L}_\alpha^{\mathrm{ERM}}(\mathcal{F}, \mathcal{D}_n) := \alpha \cdot \hat{R}(\mathcal{F}, \mathcal{D}_n)$$

is a valid distribution-free lower bound on $R_P(\mathcal{F})$.

Proof If $R_P(\mathcal{F}) = 0$, then we must have $\hat{R}(\mathcal{F}, \mathcal{D}_n) = 0$ almost surely. If instead $R_P(\mathcal{F}) > 0$, then by Markov's inequality, we have $\mathbb{P}_P\{\hat{L}_\alpha^{\mathrm{ERM}}(\mathcal{F}, \mathcal{D}_n) \leq R_P(\mathcal{F})\} = \mathbb{P}_P\{\hat{R}(\mathcal{F}, \mathcal{D}_n) \leq \alpha^{-1} R_P(\mathcal{F})\} \geq 1 - \alpha$, since $\mathbb{E}_P[\hat{R}(\mathcal{F}, \mathcal{D}_n)] \leq R_P(\mathcal{F})$ as computed in (3). ■

In general, this lower bound is far from optimal—if n is large (and so the minimum empirical risk $\hat{R}(\mathcal{F}, \mathcal{D}_n)$ exhibits strong concentration), we might hope for a lower bound that is approximately equal to $\hat{R}(\mathcal{F}, \mathcal{D}_n)$, but $\hat{L}_\alpha^{\mathrm{ERM}}(\mathcal{F}, \cdot)$ is a constant factor smaller. In the Appendix, we will establish a tighter result: in the setting of a bounded loss, we will construct a valid lower bound of the form $\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) = (1 - \mathrm{o}(1)) \cdot \hat{R}(\mathcal{F}, \mathcal{D}_n)$.

4. There are however also cases in which $N(\mathcal{F}, P) \ll N_+(\mathcal{F}, P)$—for instance, consider the classification setting from Section 2.1 with $\mathcal{F}$ containing all constant functions on $\mathbb{R}^d$ and P being a distribution of (X, Y) on $\mathbb{R}^d \times \{0, 1\}$ such that $\mathbb{P}_P(Y = 1|X)$ is in $(0, 1)$ almost surely. Then, $N(\mathcal{F}, P) = 1$, while $N_+(\mathcal{F}, P) = \infty$.

However, even though the bound constructed here is loose, it is sufficient for answering our questions about whether a nontrivial lower bound is possible: to examine the implications of this theorem, the following corollary interprets this result in terms of the interpolation capacity.

Corollary 3 *In the setting of Theorem 2, suppose* $N_+(\mathcal{F}, P) < n$. *Then*

$$\mathbb{P}_P\left\{\hat{L}_\alpha^{\mathrm{ERM}}(\mathcal{F}, \mathcal{D}_n) > 0\right\} = 1.$$

Proof If $N_+(\mathcal{F}, P) < n$, then almost surely we have $\hat{R}(\mathcal{F}, \mathcal{D}_n) > 0$ and so $\hat{L}_\alpha^{\mathrm{ERM}}(\mathcal{F}, \mathcal{D}_n) > 0$. ■

That is, whenever $\mathcal{F}$ fails to interpolate the training data, we are able to conclude that "all models are wrong", as in the low-complexity regime of Figure 1. Outside of the low-complexity regime, on the other hand, the result of Theorem 2 yields a lower bound $\hat{L}_\alpha^{\mathrm{ERM}}(\mathcal{F}, \mathcal{D}_n) = 0$—that is, the lower bound on the model class risk is valid, but meaningless. However, these results do not yet answer whether it is impossible for *any* valid lower bound to be positive—it only tells us that this particular lower bound $\hat{L}_\alpha^{\mathrm{ERM}}(\mathcal{F}, \cdot)$, computed via the empirical risk, is not informative. To resolve this remaining question, we will need to study the high-complexity regime.

3.3. Hardness results in the high-complexity regime

We now turn to the high-complexity regime, where we will see that there are fundamental limits on the ability of *any* distribution-free lower bound to provide meaningful inference for $R_P(\mathcal{F})$. Based on the results above for the low-complexity setting, we might conjecture the following:

> Is it true that, if $\hat{R}(\mathcal{F}, \mathcal{D}_n) = 0$, then we must have $\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) = 0$ as well (at least, with large probability) for any valid lower bound?

That is, if $\mathcal{F}$ can interpolate the training sample $\mathcal{D}_n$ of size n, is it *impossible* to achieve a nontrivial and valid lower bound? In fact, as previewed earlier in Figure 1, this is not exactly the case, but a weaker result holds—if $\mathcal{F}$ can interpolate a sample of size $N \gg n^2$, then indeed any valid lower bound $\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n)$ must (usually) equal zero.

Before stating our main result, we need to introduce some additional notation. We will now consider an infinite stream of data points, $Z_1, Z_2, \ldots \overset{\text{iid}}{\sim} P$. Then $\mathcal{D}_n = (Z_1, \ldots, Z_n)$ is a data set of n i.i.d. draws from P, as before, but now we will also write $\mathcal{D}_N = (Z_1, \ldots, Z_N)$ for each $N \geq n$, to denote larger data sets that contain $\mathcal{D}_n$ as a subset.

Theorem 4 *Fix* $\alpha \in (0, 1)$, $n \geq 1$, *and model class* $\mathcal{F}$. *Let* $\hat{L}_\alpha(\mathcal{F}, \cdot)$ *be a valid distribution-free lower bound on the model class risk. Then, for all* $N \geq n$,

$$\mathbb{P}_P\left\{\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > \hat{R}(\mathcal{F}, \mathcal{D}_N)\right\} \leq \alpha + \frac{n^2}{2N}.$$

The proof of this result follows the "sample–resample" strategy used for establishing certain hardness results in the distribution-free inference literature (see Angelopoulos et al. (2024)), which relies on the fact that, when sampling n times from a population of size N, the total variation distance between sampling with replacement and sampling without replacement is bounded by $\frac{n^2}{2N}$.

To interpret this theorem, let us return to the question of whether we can determine that "all models are wrong": can a valid distribution-free lower bound satisfy $\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0$, certifying that

the model class risk $R_P(\mathcal{F})$ is not zero, more often than the trivial solution in (2)? The following result shows how the interpolation capacity of $\mathcal{F}$ allows us to characterize a regime where any valid lower bound must necessarily be trivial.

Corollary 5 *In the setting of Theorem 4, for any valid distribution-free lower bound on the model class risk, it holds that*

$$\mathbb{P}_P\Big\{\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0\Big\} \leq \alpha + \frac{n^2}{2N(\mathcal{F}, P)}.$$

Proof This result follows immediately from Theorem 4: for any $N \leq N(\mathcal{F}, P)$, we have $\hat{R}(\mathcal{F}, \mathcal{D}_N) = 0$ almost surely, and so $\mathbb{P}_P\{\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0\} = \mathbb{P}_P\{\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > \hat{R}(\mathcal{F}, \mathcal{D}_N)\} \leq \alpha + \frac{n^2}{2N}$. ■

To summarize what we have seen for the low-complexity and high-complexity regimes in terms of the interpolation capacity, if $\mathcal{F}$ can interpolate data sets of size $N \gg n^2$, then $\mathbb{P}_P\{\hat{L}_\alpha(\mathcal{F}, \mathcal{D}_n) > 0\} \leq \alpha + \mathrm{o}(1)$ for every valid lower bound $\hat{L}_\alpha(\mathcal{F}, \cdot)$. Recalling the trivial solution in (2), this means that there do not exist any nontrivial valid lower bounds. On the other hand, if $\mathcal{F}$ cannot interpolate data sets of size n, then there always exists a positive valid lower bound, namely, $\hat{L}_\alpha^{\mathrm{ERM}}(\mathcal{F}, \cdot)$.

4. Examples at the boundaries: the in-between regime

Through the results of Section 3 (as summarized in Figure 1), we have seen that providing a nontrivial lower bound on the model class risk, using a sample of size n, is always possible if $\mathcal{F}$ does not interpolate n data points, but inevitably impossible if it interpolates $\gg n^2$ many data points.

In this section, we shed some light on what happens in the "in-between" regime, where the interpolation capacity $N(\mathcal{F}, P)$ lies somewhere between n and $\mathcal{O}(n^2)$. Throughout this section, we will work in a regression setting, where we observe data points $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R} =: \mathcal{Z}$, the model class $\mathcal{F}$ is some set of functions $f : \mathbb{R}^d \to \mathbb{R}$, and ℓ is the squared loss, $\ell(f, (x, y)) = (f(x) - y)^2$.

We will present two examples that exhibit different behavior in terms of how interpolation capacity relates to the problem of inference on $R_P(\mathcal{F})$:

- For one example (piecewise constant functions), we will see that a nontrivial lower bound is possible whenever $N(\mathcal{F}, P) = \mathcal{O}(n^2)$—and in particular, we may have $N(\mathcal{F}, P) \gg n$.
- For the second example (linear functions), we will see that it is impossible to construct a nontrivial lower bound whenever $N(\mathcal{F}, P) \gg n$.

In particular, the contrast between these two examples implies that there is no universal phase transition for the problem of constructing a nontrivial lower bound: while a nontrivial lower bound is possible for any $\mathcal{F}$ in the low-complexity regime, and impossible for any $\mathcal{F}$ in the high-complexity regime, its existence in the in-between regime depends on the nature of $\mathcal{F}$.

4.1. Example: piecewise constant regression

Define the model class of *piecewise constant functions* with $\leq m$ components as

$$\mathcal{F}_{\mathrm{pwc}}^{(m)} = \Big\{f : \mathbb{R}^d \to \mathbb{R} \ : \ \big|\{f(x) : x \in \mathbb{R}^d\}\big| \leq m\Big\}.$$

Equivalently, a function f lies in $\mathcal{F}^{(m)}_{\text{pwc}}$ if it can be expressed as

$$f(x) = \sum_{i=1}^{m} y_i \cdot \mathbb{1}_{x \in A_i},$$

for some $y_1, \ldots, y_m \in \mathbb{R}$ and some partition $\mathbb{R}^d = A_1 \cup \cdots \cup A_m$. For example, methods such as regression trees or random forests will produce fitted models of this form.

For any distribution P on $\mathbb{R}^d \times \mathbb{R}$ such that its marginal P_X is nonatomic (i.e., $\mathbb{P}_P\{X = x\} = 0$ for all $x \in \mathbb{R}^d$), we can calculate the interpolation complexity of this model class as

$$N_+(\mathcal{F}^{(m)}_{\text{pwc}}, P) \geq N(\mathcal{F}^{(m)}_{\text{pwc}}, P) \geq m,$$

since, for any $(X_1, Y_1), \ldots, (X_m, Y_m)$ with distinct values $X_1, \ldots, X_m$, we can construct a function $f \in \mathcal{F}^{(m)}_{\text{pwc}}$ with $f(X_i) = Y_i$ for each $i \in [m]$. In particular, the lower bound

$$\hat{L}^{\text{ERM}}_{\alpha}(\mathcal{F}^{(m)}_{\text{pwc}}, \mathcal{D}_n) = \alpha \cdot \hat{R}(\mathcal{F}^{(m)}_{\text{pwc}}, \mathcal{D}_n)$$

constructed in Theorem 2 is a valid but meaningless lower bound for any sample size $n \leq m$, since it is zero almost surely. However, we will now show that a better result can be obtained by constructing a different valid lower bound.

Theorem 6 *Fix $\alpha \in (0, 1)$, and let $\alpha_0 + \alpha_1 = \alpha$, with $\alpha_0, \alpha_1 > 0$. Fix any $n \geq 1$ and $m \geq 1$ with*

$$m \leq \frac{n(n-1)}{2\log(1/\alpha_0)}.$$

Then

$$\hat{L}^{\text{pwc}}_{\alpha}(\mathcal{F}^{(m)}_{\text{pwc}}, \mathcal{D}_n) := \alpha_1 \cdot \hat{R}(\mathcal{F}^{(n-1)}_{\text{pwc}}, \mathcal{D}_n)$$

is a valid distribution-free lower bound on $R_P(\mathcal{F}^{(m)}_{\text{pwc}})$.

The intuition for proving the above theorem is the following: consider any $f \in \mathcal{F}^{(m)}_{\text{pwc}}$, which takes (at most) m distinct values in its output. If m is not too large, then with high probability, at most $n - 1$ of these values will actually be observed in a random sample of size n—that is, at least one observed value will be repeated. Consequently, on the sample $\mathcal{D}_n$, the empirical risk of f can also be achieved by some function in $\mathcal{F}^{(n-1)}_{\text{pwc}}$, which is a much less complex model class.

How should we interpret this result? Consider a setting where P_X and P_Y are both nonatomic. In this case, $\hat{R}(\mathcal{F}^{(n-1)}_{\text{pwc}}, \mathcal{D}_n) > 0$ must hold almost surely, since we cannot interpolate n unique Y values with any function $f \in \mathcal{F}^{(n-1)}_{\text{pwc}}$. In particular, since we can take $m \propto n^2$ in this theorem, this result verifies that the fundamental hardness result provided in Theorem 4 is tight: for this particular model class, although $N(\mathcal{F}^{(m)}_{\text{pwc}}, P) \geq m \propto n^2$, it is nonetheless possible to provide a nontrivial valid lower bound, since $\hat{L}^{\text{pwc}}_{\alpha}(\mathcal{F}^{(m)}_{\text{pwc}}, \mathcal{D}_n) > 0$ almost surely.[5]

5. We will also prove a stronger version of this result in the Appendix, that establishes a more precise connection between the scaling of m and the behavior of a valid lower bound.

4.2. Example: linear models

For our second example, define the class of all linear predictors,

$$\mathcal{F}_{\text{lin}}^{(d)} = \{x \mapsto x^\top \beta : \beta \in \mathbb{R}^d\}.$$

For this class, for any distribution P with a density on $\mathbb{R}^d \times \mathbb{R}$, we have

$$N_+(\mathcal{F}_{\text{lin}}^{(d)}, P) = N(\mathcal{F}_{\text{lin}}^{(d)}, P) = d,$$

since any d data points in generic position can be interpolated by some $f \in \mathcal{F}_{\text{lin}}^{(d)}$. In particular, the valid lower bound $\hat{L}_\alpha^{\text{ERM}}(\mathcal{F}_{\text{lin}}^{(d)}, \cdot)$, constructed in Theorem 2, provides a nontrivial lower bound for the model class risk if and only if $d < n$. However, might there be some other valid lower bound that is nontrivial even for $d \geq n$?

Our main result in this section will show that, in terms of finding a nontrivial lower bound, the performance of $\hat{L}_\alpha^{\text{ERM}}(\mathcal{F}_{\text{lin}}^{(d)}, \cdot)$ is essentially the best that we could hope for, over a wide class of distributions. That is, when $d \gg n$, it is impossible to construct a nontrivial lower bound.

We first need to set up some notation and definitions. To begin, let us define a class of noise-free linear distributions on $\mathbb{R}^d \times \mathbb{R}$,

$$\mathcal{Q}_{\text{lin}} := \Big\{P \ : \ \text{for some } \beta \in \mathbb{R}^d, Y = X^\top \beta \text{ holds } P\text{-almost surely}\Big\},$$

and the class of n-fold product distributions from this class,

$$\mathcal{Q}_{\text{lin}}^{(n)} = \{P^n : P \in \mathcal{Q}_{\text{lin}}\}.$$

Let $\tilde{\mathcal{Q}}_{\text{lin}}^{(n)}$ denote the family of mixtures of distributions in $\mathcal{Q}_{\text{lin}}^{(n)}$, which can be seen as the convex hull of $\mathcal{Q}_{\text{lin}}^{(n)}$. Finally, for any distribution P on $\mathbb{R}^d \times \mathbb{R}$, define

$$\lambda_{n,d}(P) := \inf_{Q \in \tilde{\mathcal{Q}}_{\text{lin}}^{(n)}} \mathrm{d_{TV}}\big(P^n, Q\big).$$

The following theorem provides a hardness result that bounds our ability to provide a nontrivial valid lower bound on the model class risk.

Theorem 7 *Fix $\alpha \in (0,1)$, $n \geq 1$, and $d \geq 1$. Let $\hat{L}_\alpha(\mathcal{F}_{\text{lin}}^{(d)}, \cdot)$ be a valid distribution-free lower bound on the model class risk. Then for any distribution P on $\mathbb{R}^d \times \mathbb{R}$,*

$$\mathbb{P}_P\left\{\hat{L}_\alpha(\mathcal{F}_{\text{lin}}^{(d)}, \mathcal{D}_n) > 0\right\} \leq \alpha + \lambda_{n,d}(P).$$

For intuition, this holds simply due to the definition of total variation distance: if $\lambda_{n,d}(P)$ is low, then the distribution P^n of the observed data is nearly indistinguishable from a convex combination of noise-free linear distributions—and consequently we cannot construct a nontrivial lower bound.

From this result, we can therefore see that any valid lower bound is essentially trivial whenever $\lambda_{n,d}(P) \approx 0$—but when will this be the case? The next result establishes that, in a high-dimensional setting where $d \geq n$, the quantity $\lambda_{n,d}(P)$ is small for a broad class of distributions P.

Proposition 8 *Let P be a distribution on $\mathbb{R}^d \times \mathbb{R}$ with a density. Let $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} P$, and define $\mathbf{X} \in \mathbb{R}^{n \times d}$ as the matrix with rows X_i. If $d \geq n$, then for any positive definite matrix $\Omega \in \mathbb{R}^{d \times d}$,*

$$\lambda_{n,d}(P) \leq \frac{1}{2} \mathbb{E}_P \left[\left| \frac{h_\Omega(\mathbf{X})}{\mathbb{E}_P[h_\Omega(\mathbf{X})]} - 1 \right| \right],$$

where $h_\Omega(\mathbf{X}) = (\det(\mathbf{X}\Omega\mathbf{X}^\top))^{-1/2}$ (and we implicitly assume that $\mathbb{E}_P[h_\Omega(\mathbf{X})] < \infty$).

Proposition 8 suggests that as long as $h_\Omega(\mathbf{X}) = (\det(\mathbf{X}\Omega\mathbf{X}^\top))^{-1/2}$ concentrates for some positive definite matrix Ω, we expect $\lambda_{n,d}(P) \approx 0$. To make this intuition more precise, consider the case $\Omega = \mathbf{I}_d$, and observe that we can write the determinant as

$$h_{\mathbf{I}_d}(\mathbf{X})^{-1} = \det(\mathbf{X}\mathbf{X}^\top)^{1/2} = \|X_1\|_2 \cdot \|\mathbf{P}^\perp_{X_1} X_2\|_2 \cdot \|\mathbf{P}^\perp_{X_1, X_2} X_3\|_2 \cdot \ldots \cdot \|\mathbf{P}^\perp_{X_1, \ldots, X_{n-1}} X_n\|_2,$$

where for each i, the notation $\mathbf{P}^\perp_{X_1, \ldots, X_i}$ denotes projection to the orthogonal complement of the span of $X_1, \ldots, X_i \in \mathbb{R}^d$. Thus, as long as we expect concentration of the norms of the X_i's, and also expect that the X_i's are not likely to be strongly collinear, we can expect $h_{\mathbf{I}_d}(\mathbf{X})$ to have low variance (and so $\lambda_{n,d}(P)$ will be small).

To make this more concrete, the next result derives an explicit upper bound on $\lambda_{n,d}(P)$ for the special case of a Gaussian distribution.[6]

Corollary 9 *Let P be any distribution on $\mathbb{R}^d \times \mathbb{R}$ with a density, such that the marginal distribution of X is $P_X = \mathcal{N}(0, \Sigma)$ for some positive definite $\Sigma \in \mathbb{R}^{d \times d}$. Then, if $d \geq n + 2$,*

$$\lambda_{n,d}(P) \leq \frac{1}{2} \sqrt{\frac{n}{d - n - 1}}.$$

In particular, we have $\lambda_{n,d}(P) \approx 0$ for any $d \gg n$ in the Gaussian setting, meaning that any valid lower bound on the model class risk of $\mathcal{F}^{(d)}_{\text{lin}}$ is essentially trivial—in particular, this holds for, say, $d \propto n^{1+a}$ for any $a > 0$, meaning that we may still have $d \ll n^2$.

This result therefore offers another example for the "in-between" regime of Figure 1, which complements the findings of Section 4.1 for our first example: it verifies that the low-complexity regime results of Theorem 2 and Corollary 3 are essentially tight, in the sense that if $N(\mathcal{F}, P) \gg n$, then we cannot provide a universal guarantee that a nontrivial lower bound is always possible for any $\mathcal{F}$. A related question is also addressed by Kong and Valiant (2019, Theorem 3), who show that even when restricted to the case where P is known to be Gaussian but nothing is to be assumed about the covariance matrix, it is impossible to distinguish between $R_P(\mathcal{F}^{(d)}_{\text{lin}}) > 0$ and $R_P(\mathcal{F}^{(d)}_{\text{lin}}) = 0$ when given access to only $n = \mathcal{O}(d)$ data points.

5. Discussion

In this paper, we have discussed a central question in learning theory: when can we show that a model class is "wrong", in the sense that even the best instance of a model in the class cannot achieve zero test error? We do so by considering the more general task of lower bounding the model class risk empirically and in a distribution-free manner. Focusing on the case of highly flexible model

6. We show in the Appendix how this type of bound can be extended to more general distributions.

classes capable of often perfectly interpolating the data, we identify a regime of "mild" interpolation (where, although $\mathcal{F}$ interpolates the data $\mathcal{D}_n$, it is not able to interpolate $\gg n^2$ many data points), a setting in which we may still empirically be able to perform nontrivial inference on $R_P(\mathcal{F})$. In contrast, in the high-complexity setting (where $\mathcal{F}$ can interpolate $\gg n^2$ many data points), we may say that we are in a regime of "hyper-interpolation", where nontrivial inference is no longer possible. Our examples show how these regimes are sharply characterized.

5.1. Related literature

We will next discuss connections to related areas in the statistics and machine learning literature.

Learning theory and the approximation–estimation tradeoff. As discussed in Section 1, the model class risk is closely related to the question of excess risk $R_P(\hat{f}) - \inf_{f\in\mathcal{F}} R_P(f)$ and the approximation–estimation tradeoff (see, for example, Bartlett et al. (2002); Bottou and Bousquet (2007)). Specifically, the problem of finding an upper bound on excess risk is essentially equivalent in difficulty to the problem of a lower bound for $R_P(\mathcal{F})$—and consequently, the hardness results established in our work can be interpreted as evidence that a distribution-free upper bound on excess risk is also challenging in a regime where the model class $\mathcal{F}$ is highly complex.

On the other hand, there are many exciting recent results on controlling the excess risk under only mild distributional assumptions. For instance, parametric convergence rates of the excess risk $R_P(\hat{f}) - \inf_{f\in\mathcal{F}} R_P(f)$ can be achieved when the loss satisfies certain regularity conditions and the algorithm used to select $\hat{f}$ from $\mathcal{F}$ satisfies stability properties (Klochkov and Zhivotovskiy, 2021). Mourtada and Gaïffas (2022) leveraged the idea of stability of the empirical risk minimizer and came up with a technique to construct models $\hat{f}$ (which are potentially improper, i.e., may take values not contained in $\mathcal{F}$) achieving low excess risk in logistic regression and conditional density estimation problems. These results offer a sort of converse to the hardness results established in this paper, by examining settings where nontrivial guarantees can be achieved.

Overparametrization and interpolation. In recent years, an exciting topic in machine learning research has been the *benign overfitting* phenomenon, sometimes also called *double descent*, where an increase in a model class's capacity beyond the interpolation threshold can help to reduce generalization error of the eventually fitted function (Belkin et al., 2019a). In other words, fitting a model $\hat{f}$ with $\hat{R}(\hat{f}, \mathcal{D}_n) = 0$ (meaning, implicitly, that we are searching within in a class $\mathcal{F}$ with $N(\mathcal{F}, P) \geq n$) can often lead to high accuracy, in contrast to what earlier results, based on generalization bounds from the learning theory literature, might suggest. There are a number of analyses for the generalization error upper bound for different overfitting estimators; see for example Belkin et al. (2018, 2019b); Bartlett et al. (2020). A few recent works have also provided lower bounds on the excess risk for estimators with large training errors to illustrate that interpolation or memorization is necessary for optimal generalization, for example, see Cheng et al. (2022) and references therein.

Distribution-free risk control. As discussed earlier, the problem of a distribution-free upper bound $\hat{U}_\alpha(\mathcal{F}, \cdot)$ on the model class risk is fundamentally simpler than that of a lower bound. Concretely, as described in Section 2.2, we might train a model $\hat{f} \in \mathcal{F}$ using $\mathcal{D}_{n/2}$ (the first half of the available data set $\mathcal{D}_n$), and then use the remaining data $\mathcal{D}'_{n/2}$ to provide an upper bound on $R_P(\mathcal{F})$. If the loss function ℓ takes values in a bounded range $[0, B]$, then by Hoeffding's inequality (Ho-

effding, 1962), with probability $\geq 1-\alpha$,

$$R_P(\mathcal{F}) \leq R_P(\hat{f}) \leq \hat{R}\big(\hat{f}, \mathcal{D}'_{n/2}\big) + B\sqrt{\frac{\log(1/\alpha)}{2\lfloor n/2 \rfloor}} =: \hat{U}_\alpha(\mathcal{F}, \mathcal{D}_n),$$

thus providing a distribution-free upper bound in the sense of Definition 1. Such high-probability upper confidence bounds are the starting point for the distribution-free risk control methodology of Bates et al. (2021); see also Angelopoulos et al. (2025, 2023) for related approaches.

An important existing result on distribution-free inference on a model class risk is given by Devroye et al. (1996, Theorem 8.5), which illustrates that in a distribution-free classification setting, the task of estimating the Bayes error leads to a constant minimax lower bound.

Hardness of distribution-free inference. Our work also contributes to an important line of work on impossibility or hardness results in distribution-free inference. For example, the classical work Bahadur and Savage (1956) showed that inference for the mean of a random variable is impossible without any assumption on the distribution. More recent distribution-free hardness results have been established for problems such as predictive inference with conditional coverage (Vovk, 2012; Lei and Wasserman, 2014; Barber et al., 2021), and inference on the conditional mean or median (Barber, 2020; Medarametla and Candès, 2021). Hardness results for assumption-free inference for algorithmic properties such as stability and risk have been established in Kim and Barber (2023); Luo and Barber (2024). This last result, which studies the problem of providing distribution-free inference on $\mathbb{E}[R_P(\hat{f})]$ for a model $\hat{f}$ fitted by some black-box algorithm $\mathcal{A}$, is the closest to our work in terms of the questions being investigated, but is still substantially different: while Luo and Barber (2024) study the problem of evaluating the risk of a model fitted by a specific algorithm $\mathcal{A}$, here we instead ask about the infimum risk $\inf_{f\in\mathcal{F}} R_P(f)$ regardless of whether it is possible to construct an algorithm $\mathcal{A}$ that will identify an (approximately) optimal model.

5.2. Extensions and open questions

To conclude, we will briefly mention some possible directions for extensions and further research suggested by this work. First, the questions examined in this work focus on a single model class $\mathcal{F}$, but in practice we may want to choose $\mathcal{F}$ in an adaptive way, by examining the performance of various fitted models on the data. In particular, returning to the approximation–estimation tradeoff, one approach to find a balance between these two conflicting goals is known as *structural risk minimization*. Here, the idea is to consider a nested sequence of model classes $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots$, and to choose $\mathcal{F} = \mathcal{F}_{j_n}$ based on a suitable increasing sequence j_n, $n \geq 1$, or to use complexity regularization (Bartlett et al., 2002; von Luxburg and Schölkopf, 2011). For instance, in a regression setting, $\mathcal{F}_j$ may be a sequence of increasingly more complex neural network architectures that can approximate ever more complicated functions (Barron, 1994). In this setting, our results might be useful for identifying scenarios in which it is possible, or impossible, to use empirical evidence to help navigate this tradeoff.

Second, recall that in this work we have identified an "in-between" regime: as summarized in Figure 1, if the interpolation capacity of $\mathcal{F}$ lies between n and $\mathcal{O}(n^2)$, then it may be possible or impossible to provide distribution-free inference on $R_P(\mathcal{F})$. Our two examples (in Section 4) show that these boundaries cannot be improved—no universal guarantee of nontrivial inference can be achieved outside of the low-complexity regime, and no universal hardness result can be shown outside of the high-complexity regime. However, an important open question remains: might there be

an alternative notion of model class complexity (i.e., different from the interpolation capacity), with which it would be possible to identify a universal phase transition directly from the low-complexity to high-complexity regime?

Acknowledgments

R.F.B. was partially supported by the National Science Foundation via grant DMS-2023109, and by the Office of Naval Research via grant N00014-24-1-2544. We thank John Duchi for bringing helpful references to our attention.

References

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2023.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Raghu R. Bahadur and Leonard J. Savage. The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, 27(4):1115–1122, 1956.

Rina Foygel Barber. Is distribution-free inference possible for binary regression? *Electronic Journal of Statistics*, 14:3487–3524, 2020.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Andrew R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14:115–133, 1994.

Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Mikhail Belkin, Daniel J. Hsu, and Partha Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31, 2018.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019b.

Anna Ben-Hamou, Stéphane Boucheron, and Mesrob I Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23(1):249–287, 2017.

Léon Bottou and Olivier Bousquet. The Tradeoffs of Large Scale Learning. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities: A nonasymptotic theory of independence, 2013.

George E.P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356): 791–799, 1976.

Chen Cheng, John Duchi, and Rohith Kuditipudi. Memorize to generalize: on the necessity of interpolation in high dimensional linear regression. In *Conference on Learning Theory*, pages 5528–5560. PMLR, 2022.

Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 1996.

N.R. Goodman. The distribution of the determinant of a complex wishart distributed matrix. *The Annals of Mathematical Statistics*, 34(1):178–180, 1963.

Louis Gordon. A stochastic approach to the gamma function. *The American Mathematical Monthly*, 101(9):858–865, 1994.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Institute of Statistics, Mimeo Series No. 326*, 1962.

Byol Kim and Rina Foygel Barber. Black-box tests for algorithmic stability. *Information and Inference: A Journal of the IMA*, 12(4):2690–2719, 2023.

Yegor Klochkov and Nikita Zhivotovskiy. Stability and Deviation Optimal Risk Bounds with Convergence Rate $O(1/n)$. In *Advances in Neural Information Processing Systems*, volume 34, pages 5065–5076. Curran Associates, Inc., 2021.

Weihao Kong and Gregory Valiant. Estimating learnability in the sublinear data regime. *arXiv preprint arXiv:1805.01626*, 2019.

Yonghoon Lee and Rina Barber. Distribution-free inference for regression: discrete, continuous, and in between. *Advances in Neural Information Processing Systems*, 34:7448–7459, 2021.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

Yuetian Luo and Rina Foygel Barber. The limits of assumption-free tests for algorithm performance. *arXiv preprint arXiv:2402.07388*, 2024.

Colin McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, volume 16, pages 195–248. Springer, 1998.

Dhruv Medarametla and Emmanuel Candès. Distribution-free conditional median inference. *Electronic Journal of Statistics*, 15(2):4625–4658, 2021.

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

A.G. Munford. A note on the uniformity assumption in the birthday problem. *The American Statistician*, 31(3):119–119, 1977.

Guergana Petrova and Przemyslaw Wojtaszczyk. Limitations on approximation by deep and shallow neural networks. *Journal of Machine Learning Research*, 24(353):1–38, 2023.

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

Adriaan J Stam. Distance between sampling with and without replacement. *Statistica Neerlandica*, 32(2):81–91, 1978.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

Ulrike von Luxburg and Bernhard Schölkopf. Statistical learning theory: Models, concepts, and results. In Dov M. Gabbay, Stephan Hartmann, and John Woods, editors, *Inductive Logic*, volume 10 of *Handbook of the History of Logic*, pages 651–706. North-Holland, 2011.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25, pages 475–490. PMLR, 2012.

Appendix A. Proofs and extensions for results in Section 3

A.1. A tighter lower bound for the low-complexity regime

In this section, we present a stronger version of Theorem 2, for the setting of a bounded loss.

Theorem A.1 *Fix $\alpha \in (0,1)$, $n \geq 1$, and model class $\mathcal{F}$. Assume that the loss function ℓ is bounded, taking values in $[0,B]$ for some $B > 0$. Define $\Delta_n \in (0,1]$ as the unique solution of*[7]

$$-\Delta_n - \log(1-\Delta_n) = \frac{B\log(1/\alpha)}{n\hat{R}(\mathcal{F},\mathcal{D}_n)}. \tag{A.1}$$

Then

$$\hat{L}_\alpha^{\mathrm{ERM},B}(\mathcal{F},\mathcal{D}_n) := (1-\Delta_n)\hat{R}(\mathcal{F},\mathcal{D}_n)$$

is a valid distribution-free lower bound on $R_P(\mathcal{F})$.

As for the lower bound constructed in Theorem 2, we again have $\hat{L}_\alpha^{\mathrm{ERM},B}(\mathcal{F},\mathcal{D}_n) > 0$ if and only if $\hat{R}(\mathcal{F},\mathcal{D}_n) > 0$ (because this event is equivalent to $\Delta_n < 1$). But the theorem can be interpreted in a more quantitative way: since $-\Delta_n - \log(1-\Delta_n) \geq \Delta_n^2/2$, by (A.1) we have

$$\hat{L}_\alpha^{\mathrm{ERM},B}(\mathcal{F},\mathcal{D}_n) \geq \left(1 - \sqrt{\frac{2B\log(1/\alpha)}{n\hat{R}(\mathcal{F},\mathcal{D}_n)}}\right)\hat{R}(\mathcal{F},\mathcal{D}_n) = \hat{R}(\mathcal{F},\mathcal{D}_n) - \sqrt{\hat{R}(\mathcal{F},\mathcal{D}_n)\cdot\frac{2B\log(1/\alpha)}{n}}. \tag{A.2}$$

In other words, our lower bound is of the form $(1-\mathrm{o}(1))\cdot\hat{R}(\mathcal{F},\mathcal{D}_n)$.

In fact, this type of result can be applied to the setting of an unbounded loss, as well, via a truncation step. For any $f \in \mathcal{F}$, define its truncated empirical risk as $\hat{R}(f,\mathcal{D}_n;B) = \frac{1}{n}\sum_{i=1}^n \min\{\ell(f,Z_i),B\}$, and define $\hat{R}(\mathcal{F},\mathcal{D}_n;B) = \inf_{f\in\mathcal{F}}\hat{R}(f,\mathcal{D}_n;B)$.

Theorem A.2 *Fix $\alpha \in (0,1)$, $n \geq 1$, and model class $\mathcal{F}$. Fix any $B > 0$, and define $\Delta_{n,B} \in (0,1]$ as the unique solution of*

$$-\Delta_{n,B} - \log(1-\Delta_{n,B}) = \frac{B\log(1/\alpha)}{n\hat{R}(\mathcal{F},\mathcal{D}_n;B)}.$$

Then

$$\hat{L}_\alpha^{\mathrm{ERM-trunc},B}(\mathcal{F},\mathcal{D}_n) := (1-\Delta_{n,B})\hat{R}(\mathcal{F},\mathcal{D}_n;B)$$

is a valid distribution-free lower bound on $R_P(\mathcal{F})$.

To help interpret this lower bound, we can observe that as in (A.2) above, it holds that

$$\begin{aligned}\hat{L}_\alpha^{\mathrm{ERM-trunc},B}(\mathcal{F},\mathcal{D}_n) &\geq \hat{R}(\mathcal{F},\mathcal{D}_n;B) - \sqrt{\hat{R}(\mathcal{F},\mathcal{D}_n;B)\cdot\frac{2B\log(1/\alpha)}{n}} \\ &\geq \hat{R}(\mathcal{F},\mathcal{D}_n;B) - \sqrt{\frac{2B^2\log(1/\alpha)}{n}},\end{aligned} \tag{A.3}$$

where the last step holds simply because $\hat{R}(\mathcal{F},\mathcal{D}_n;B) \leq B$ by construction.

The proofs of these results rely on the following lemma.

7. Note that, since $\Delta \mapsto -\Delta - \log(1-\Delta)$ is strictly increasing on $[0,1)$, with $\lim_{\Delta\searrow 0}\{-\Delta-\log(1-\Delta)\} = 0$ and $\lim_{\Delta\nearrow 1}\{-\Delta-\log(1-\Delta)\} = +\infty$, this is well-defined for any value $\hat{R}(\mathcal{F},\mathcal{D}_n) \geq 0$. In particular if $\hat{R}(\mathcal{F},\mathcal{D}_n) = 0$, then this equation is solved by $\Delta_n = 1$.

Lemma A.3 *Let $Z_1, \dots, Z_n$ be independent random variables taking values in $[0,1]$. Denote $S := \sum_{i=1}^n Z_i$ and $\mu := \mathbb{E}[S]$. Fix $\alpha \in (0,1)$ and let $\Delta \in (0,1]$ be the unique solution to*

$$-\Delta - \log(1-\Delta) = \log(1/\alpha)/S.$$

Then

$$\mathbb{P}\big((1-\Delta)S \leq \mu\big) \geq 1 - \alpha.$$

Proof of Lemma A.3. First we consider the degenerate case where $\mu = 0$, i.e., $Z_i = 0$ almost surely for each i. Then, almost surely, $S = 0$ and $\Delta = 1$, and the result holds trivially. From this point on, then, we assume $\mu > 0$ (note that we might still have $S = 0$ with positive probability).

By a multiplicative Chernoff bound (McDiarmid, 1998, Theorem 2.3(b)), we have

$$\mathbb{P}\Big\{S \leq (1+\delta)\mu\Big\} \geq 1 - \exp\big(-h(\delta)\mu\big)$$

for $\delta > 0$, where $h(\delta) = (1+\delta)\log(1+\delta) - \delta$. Since h is continuous and strictly increasing on $[0,\infty)$, with $h(0) = 0$ and $h(\delta) \to \infty$ as $\delta \to \infty$, there exists a unique $\delta_\mu > 0$ such that $h(\delta_\mu) = \log(1/\alpha)/\mu$, so that we have

$$\mathbb{P}\Big\{S \leq (1+\delta_\mu)\mu\Big\} \geq 1 - \alpha.$$

To complete the proof, we therefore only need to verify that

$$S \leq (1+\delta_\mu)\mu \implies (1-\Delta)S \leq \mu.$$

First, define

$$\Delta' = \frac{\delta_\mu}{1+\delta_\mu} \in (0,1).$$

Then $1 + \delta_\mu = 1/(1-\Delta')$, and so

$$S \leq (1+\delta_\mu)\mu \implies (1-\Delta')S \leq \mu.$$

Our last step is to check that, on the event $(1-\Delta')S \leq \mu$, we have $\Delta' \leq \Delta$. We have

$$\frac{\log(1/\alpha)}{\mu} = h(\delta_\mu) = (1+\delta_\mu)\log(1+\delta_\mu) - \delta_\mu = \frac{1}{1-\Delta'}\log\left(\frac{1}{1-\Delta'}\right) - \frac{\Delta'}{1-\Delta'},$$

so rearranging terms,

$$-\Delta' - \log(1-\Delta') = (1-\Delta') \cdot \frac{\log(1/\alpha)}{\mu} \leq \frac{\log(1/\alpha)}{S} = -\Delta - \log(1-\Delta),$$

where the inequality holds on the event that $(1-\Delta')S \leq \mu$, and the last step holds by definition of Δ. Finally, since $u \mapsto -u - \log(1-u)$ is a strictly increasing function on $[0,1)$, this implies $\Delta' \leq \Delta$, which completes the proof. ■

We are now in a position to state the proofs of Theorems A.1 and A.2.

Proof of Theorem A.1. This result is simply a special case of Theorem A.2 (since, when the loss ℓ takes values in $[0, B]$, we have $\hat{R}(\mathcal{F}, \mathcal{D}_n; B) = \hat{R}(\mathcal{F}, \mathcal{D}_n)$). ■

Proof of Theorem A.2. For any $\varepsilon > 0$, we can find some $f_\varepsilon \in \mathcal{F}$ such that $R_P(\mathcal{F}) \leq R_P(f_\varepsilon) \leq R_P(\mathcal{F}) + \varepsilon$. Noting that $\frac{\hat{R}(f_\varepsilon, \mathcal{D}_n; B)}{B}$ is an average of n i.i.d. terms, which each lie in $[0, 1]$, we can apply Lemma A.3, by taking Δ_ε to be the unique solution to

$$-\Delta_\varepsilon - \log(1 - \Delta_\varepsilon) = \frac{B \log(1/\alpha)}{n\hat{R}(f_\varepsilon, \mathcal{D}_n; B)}$$

to see that

$$1 - \alpha \leq \mathbb{P}\bigg\{(1 - \Delta_\varepsilon)\hat{R}(f_\varepsilon, \mathcal{D}_n; B) \leq \mathbb{E}[\hat{R}(f_\varepsilon, \mathcal{D}_n; B)]\bigg\} \\ \leq \mathbb{P}\bigg\{(1 - \Delta_\varepsilon)\hat{R}(\mathcal{F}, \mathcal{D}_n; B) \leq \mathbb{E}[\hat{R}(f_\varepsilon, \mathcal{D}_n; B)]\bigg\},$$

where the last step holds since $\hat{R}(f_\varepsilon, \mathcal{D}_n; B) \geq \hat{R}(\mathcal{F}, \mathcal{D}_n; B)$. And, again using the fact that $\hat{R}(f_\varepsilon, \mathcal{D}_n; B) \geq \hat{R}(\mathcal{F}, \mathcal{D}_n; B)$, we have

$$-\Delta_\varepsilon - \log(1 - \Delta_\varepsilon) = \frac{B \log(1/\alpha)}{n\hat{R}(f_\varepsilon, \mathcal{D}_n; B)} \leq \frac{B \log(1/\alpha)}{n\hat{R}(\mathcal{F}, \mathcal{D}_n; B)} = -\Delta_{n,B} - \log(1 - \Delta_{n,B})$$

by definition of $\Delta_{n,B}$. Since $u \mapsto -u - \log(1-u)$ is strictly increasing on $[0, 1)$, it follows that we must have $\Delta_\varepsilon \leq \Delta_{n,B}$. We conclude

$$1 - \alpha \leq \mathbb{P}\bigg\{(1 - \Delta_{n,B})\hat{R}(\mathcal{F}, \mathcal{D}_n; B) \leq \mathbb{E}[\hat{R}(f_\varepsilon, \mathcal{D}_n; B)]\bigg\}.$$

Next, we also have

$$\mathbb{E}[\hat{R}(f_\varepsilon, \mathcal{D}_n; B)] \leq \mathbb{E}[\hat{R}(f_\varepsilon, \mathcal{D}_n)] = R_P(f_\varepsilon) \leq R_P(\mathcal{F}) + \varepsilon,$$

where the first step holds since $\hat{R}(f, \mathcal{D}_n; B) \leq \hat{R}(f, \mathcal{D}_n)$ for any $f \in \mathcal{F}$ by definition of the truncated empirical risk, and the last step holds by definition of f_ε. Therefore,

$$\mathbb{P}\bigg\{(1 - \Delta_{n,B})\hat{R}(\mathcal{F}, \mathcal{D}_n; B) \leq R_P(\mathcal{F}) + \varepsilon\bigg\} \geq 1 - \alpha,$$

and since $\varepsilon > 0$ can be chosen to be arbitrarily small, the result follows. ■

A.2. Proof of Theorem 4 (the hardness result)

For any fixed values $z_1, \ldots, z_N \in \mathcal{Z}$, let $Q = \frac{1}{N}\sum_{i=1}^N \delta_{z_i}$ denote their empirical distribution (where δ_z is the point mass at z). Letting $I_1, \ldots, I_n \overset{\text{iid}}{\sim} \text{Unif}([N])$, we then see that $(z_{I_1}, \ldots, z_{I_n})$ is a data set that contains n i.i.d. draws from Q. Since $\hat{L}_\alpha(\mathcal{F}, \cdot)$ is a valid distribution-free lower bound, it

must therefore be valid as a lower bound for $R_Q(\mathcal{F})$ when applied to data sampled from Q—and therefore,

$$\mathbb{P}\left\{\hat{L}_\alpha(\mathcal{F}, (z_{I_1}, \ldots, z_{I_n})) \leq \hat{R}(\mathcal{F}, (z_1, \ldots, z_N))\right\} \geq 1 - \alpha,$$

by noting that $R_Q(\mathcal{F}) = \hat{R}(\mathcal{F}, (z_1, \ldots, z_N))$ by definition of Q as an empirical distribution. Next, consider indices $J_1, \ldots, J_n \in [N]$ sampled uniformly *without* replacement. By a total variation distance bound on the difference between sampling with and without replacement (e.g., Stam (1978); see also Angelopoulos et al. (2024, Lemma 4.15)), we therefore have

$$\mathbb{P}\left\{\hat{L}_\alpha(\mathcal{F}, (z_{J_1}, \ldots, z_{J_n})) \leq \hat{R}(\mathcal{F}, (z_1, \ldots, z_N))\right\} \geq 1 - \alpha - \frac{n^2}{2N}.$$

Since this is true for any fixed sequence of values $z_1, \ldots, z_N \in \mathcal{Z}$, it is also true for randomly sampled values, $Z_1, \ldots, Z_N \overset{\text{iid}}{\sim} P$—that is, we have

$$\mathbb{P}\left\{\hat{L}_\alpha(\mathcal{F}, (Z_{J_1}, \ldots, Z_{J_n})) \leq \hat{R}(\mathcal{F}, (Z_1, \ldots, Z_N))\right\} \geq 1 - \alpha - \frac{n^2}{2N},$$

where now the probability is computed with respect to both Z_i's and J_i's. But since the data points are i.i.d., by symmetry it is equivalent to write

$$\mathbb{P}\left\{\hat{L}_\alpha(\mathcal{F}, (Z_1, \ldots, Z_n)) \leq \hat{R}(\mathcal{F}, (Z_1, \ldots, Z_N))\right\} \geq 1 - \alpha - \frac{n^2}{2N},$$

which proves the desired claim.

Appendix B. Proofs and extensions for results in Section 4.1

B.1. Proof of Theorem 6 (piecewise constant functions)

Fix any $\varepsilon > 0$, and choose some $f_\varepsilon \in \mathcal{F}^{(m)}_{\text{pwc}}$ with $R_P(f_\varepsilon) \leq \inf_{f \in \mathcal{F}^{(m)}_{\text{pwc}}} R_P(f) + \varepsilon = R_P(\mathcal{F}^{(m)}_{\text{pwc}}) + \varepsilon$. By definition of the model class, we can express f_ε as

$$f_\varepsilon(x) = \sum_{j=1}^{m} y_j \cdot \mathbb{1}_{x \in A_j},$$

for some $y_1, \ldots, y_m \in \mathbb{R}$ and some partition $\mathbb{R}^d = A_1 \cup \cdots \cup A_m$. Now define

$$I(\mathcal{D}_n) = \left\{j \in [m] : \sum_{i=1}^{n} \mathbb{1}_{X_i \in A_j} > 0\right\} \subseteq [m],$$

which indexes those sets A_j that were observed at least once in the data set. Define a function

$$\hat{f}_\varepsilon(x) = \sum_{j \in I(\mathcal{D}_n)} y_j \cdot \mathbb{1}_{x \in A_j} + y_{\min I(\mathcal{D}_n)} \cdot \mathbb{1}\left\{x \in \bigcup_{j \in [m] \setminus I(\mathcal{D}_n)} A_j\right\}.$$

We can observe that $\hat{f}_\varepsilon \in \mathcal{F}^{(|I(\mathcal{D}_n)|)}_{\text{pwc}}$, since the function's output always lies in $\{y_j : j \in I(\mathcal{D}_n)\}$, and moreover, $\hat{f}_\varepsilon(X_i) = f_\varepsilon(X_i)$ for all $i \in [n]$, by construction. On the event that $|I(\mathcal{D}_n)| \leq n-1$, we then have $\mathcal{F}^{(n-1)}_{\text{pwc}} \supseteq \mathcal{F}^{(|I(\mathcal{D}_n)|)}_{\text{pwc}} \ni \hat{f}_\varepsilon$. Thus

$$\text{If } |I(\mathcal{D}_n)| \leq n-1 \text{ then } \hat{R}(f_\varepsilon, \mathcal{D}_n) = \hat{R}(\hat{f}_\varepsilon, \mathcal{D}_n) \geq \hat{R}(\mathcal{F}^{(n-1)}_{\text{pwc}}, \mathcal{D}_n) = \alpha_1^{-1} \cdot \hat{L}^{\text{pwc}}_\alpha(\mathcal{F}^{(m)}_{\text{pwc}}, \mathcal{D}_n).$$

Therefore,

$$\mathbb{P}_P\left\{\hat{L}_\alpha^{\text{pwc}}(\mathcal{F}_{\text{pwc}}^{(m)}, \mathcal{D}_n) \le R_P(f_\varepsilon)\right\} \\ \ge \mathbb{P}_P\left\{\hat{R}(f_\varepsilon, \mathcal{D}_n) \le \alpha_1^{-1} R_P(f_\varepsilon) \text{ and } |I(\mathcal{D}_n)| \le n-1\right\} \ge 1-\alpha_0-\alpha_1 = 1-\alpha,$$

where the second step holds since $\hat{R}(f_\varepsilon, \mathcal{D}_n) \le \alpha_1^{-1} R_P(f_\varepsilon)$ with probability $\ge 1-\alpha_1$ by Markov's inequality, and $\mathbb{P}_P\{|I(\mathcal{D}_n)| \le n-1\} \ge 1 - e^{-\frac{n(n-1)}{2m}} \ge 1-\alpha_0$ by Lemma B.1 below together with our assumption on m. Recalling the definition of f_ε, we have therefore proved that

$$\mathbb{P}_P\left\{\hat{L}_\alpha^{\text{pwc}}(\mathcal{F}_{\text{pwc}}^{(m)}, \mathcal{D}_n) \le R_P(\mathcal{F}_{\text{pwc}}^{(m)}) + \varepsilon\right\} \ge 1-\alpha.$$

Since $\varepsilon > 0$ can be taken to be arbitrarily small, this completes the proof of the theorem.

Lemma B.1 *Consider a discrete distribution P_Y with support $\{y_1, \dots, y_m\}$, and let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} P_Y$. Let*

$$I = \sum_{j=1}^{m} \mathbb{1}\left\{\sum_{i=1}^{n} \mathbb{1}_{Y_i = y_j} > 0\right\}$$

denote the number of unique values observed in the sample. Then

$$\mathbb{P}_{P_Y}\{I \le n-1\} \ge 1 - e^{-\frac{n(n-1)}{2m}}.$$

Proving bounds on this quantity I is related to the *occupancy problem* in probability theory (e.g., see Ben-Hamou et al. (2017)); here the bound we need is slightly different from the type of results in the existing literature.

Proof of Lemma B.1. If $m < n$, the result is immediate. Suppose from this point on then that $m \ge n$. Since I takes values in the set $[n]$, we have $\mathbb{P}_{P_Y}\{I \le n-1\} = 1 - \mathbb{P}_{P_Y}\{I = n\}$. In order to have $I = n$ we must have all observations $Y_1, \dots, Y_n$ distinct, meaning that

$$\mathbb{P}_{P_Y}\{I = n\} = \sum_{\text{distinct } i_1,\dots,i_n \in [m]} \prod_{j=1}^{n} p_{i_j} = n! \cdot \sum_{\substack{S \subseteq [m] \\ |S| = n}} \prod_{i \in S} p_i = n! \cdot f(p),$$

where $p_j := \mathbb{P}_{P_Y}(Y_1 = y_j)$, and where for a probability vector $p = (p_1, \dots, p_m)$ we define $f(p) = \sum_{\substack{S \subseteq [m] \\ |S| = n}} \prod_{i \in S} p_i$. The function f attains its maximum at $p = (\frac{1}{m}, \dots, \frac{1}{m})$ (Munford, 1977), and therefore,

$$\mathbb{P}_{P_Y}\{I = n\} = n! \cdot f(p) \\ \le n! \cdot f\left(\left(\frac{1}{m}, \dots, \frac{1}{m}\right)\right) = n! \cdot \binom{m}{n}\left(\frac{1}{m}\right)^n = \frac{m(m-1)\dots(m-n+1)}{m^n} \\ = \left(1 - \frac{1}{m}\right) \cdot \dots \cdot \left(1 - \frac{n-1}{m}\right) \le e^{-1/m} \cdot \dots \cdot e^{-(n-1)/m} = e^{-\frac{n(n-1)}{2m}}.$$

■

B.2. Extensions for the piecewise constant example

Next, we state and prove several extensions of Theorem 6, to give a more precise characterization of how the effective complexity m of the model class $\mathcal{F}_{\text{pwc}}^{(m)}$ influences the behavior of the lower bound.

Theorem B.2 *Fix $\alpha \in (0,1)$, $n \geq 1$, and $m \geq 1$. Let $\alpha_0 + \alpha_1 = \alpha$, with $\alpha_0, \alpha_1 > 0$. Then*

$$\hat{L}_\alpha^{\text{pwc},r}(\mathcal{F}_{\text{pwc}}^{(m)}, \mathcal{D}_n) := \alpha_1 \cdot \hat{R}(\mathcal{F}_{\text{pwc}}^{(n-r)}, \mathcal{D}_n)$$

is a valid distribution-free lower bound on $R_P(\mathcal{F}_{\text{pwc}}^{(m)})$, where

$$r := \left\lceil \left(\frac{n(n-1)}{n+2m} - 2\sqrt{\frac{n(n-1)}{n+2m}\log(1/\alpha_0)} \right)_+ \right\rceil.$$

To interpret this result, observe that if P_X and P_Y are nonatomic, then we would typically expect to have

$$\hat{R}(\mathcal{F}_{\text{pwc}}^{(n-r)}, \mathcal{D}_n) \propto \frac{r}{n},$$

for the squared loss. This is because the data contains n i.i.d. Y values, but the model class $\mathcal{F}_{\text{pwc}}^{(n-r)}$ only allows for functions f that return $\leq n-r$ distinct Y values, and so informally, we expect that r/n of the variance of Y remains unexplained by any $f \in \mathcal{F}_{\text{pwc}}^{(n-r)}$. For $m \geq n$, as long as $m \leq cn^2$ for some appropriately chosen constant c, we have $r \propto n^2/m$—and therefore, we expect that the lower bound will scale as

$$\hat{L}_\alpha^{\text{pwc},r}(\mathcal{F}_{\text{pwc}}^{(m)}, \mathcal{D}_n) \propto \frac{n}{m}.$$

In particular, if $m = \mathcal{O}(n)$, then $\hat{L}_\alpha^{\text{pwc},r}(\mathcal{F}_{\text{pwc}}^{(m)}, \mathcal{D}_n)$ is likely bounded away from zero—a very informative lower bound.

We are now ready to prove this extension.

Proof of Theorem B.2. Following an identical argument as in the proof of Theorem 6, and defining f_ε as in that proof, we have

$$\begin{aligned}\mathbb{P}_P\left\{\hat{L}_\alpha^{\text{pwc},r}(\mathcal{F}_{\text{pwc}}^{(m)}, \mathcal{D}_n) \leq R_P(f_\varepsilon)\right\} \\ \geq \mathbb{P}_P\left\{\hat{R}(f_\varepsilon, \mathcal{D}_n) \leq \alpha_1^{-1} R_P(f_\varepsilon) \text{ and } |I(\mathcal{D}_n)| \leq n-r\right\} \geq 1-\alpha_0-\alpha_1 = 1-\alpha,\end{aligned}$$

as long as we can show that

$$\mathbb{P}_P\{|I(\mathcal{D}_n)| \leq n-r\} \geq 1-\alpha_0,$$

which is established in Lemma B.3 below. Therefore,

$$\mathbb{P}_P\left\{\hat{L}_\alpha^{\text{pwc},r}(\mathcal{F}_{\text{pwc}}^{(m)}, \mathcal{D}_n) \leq R_P(\mathcal{F}_{\text{pwc}}^{(m)}) + \varepsilon\right\} \geq 1-\alpha,$$

and since $\varepsilon > 0$ can be taken to be arbitrarily small, this completes the proof. ■

Lemma B.3 *Consider a discrete distribution P_Y with support $\{y_1, \dots, y_m\}$, and let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} P_Y$. Let*

$$I = \sum_{j=1}^{m} \mathbb{1}\left\{ \sum_{i=1}^{n} \mathbb{1}_{Y_i = y_j} > 0 \right\}$$

denote the number of unique values observed in the sample. Then for any $\alpha_0 \in (0,1)$,

$$\mathbb{P}_{P_Y}\{I \le n - r\} \ge 1 - \alpha_0,$$

where

$$r := \left\lceil \left(\frac{n(n-1)}{n+2m} - 2\sqrt{\frac{n(n-1)}{n+2m}\log(1/\alpha_0)} \right)_+ \right\rceil.$$

Proof of Lemma B.3. First, for any $k \ge 1$ and any $Y = (Y_1, \dots, Y_k) \in \{y_1, \dots, y_m\}^k$, define

$$r(Y) = \sum_{j=1}^{m} (C_j(Y) - 1)_+,$$

where $C_j(Y) = \sum_{i=1}^{n} \mathbb{1}_{Y_i = y_j}$ counts the number of times that the value y_j was observed in the vector Y. Fixing any $k \ge 2$, and defining $Y_{-i} = (Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_k) \in \{y_1, \dots, y_m\}^{k-1}$, note that $C_j(Y) = C_j(Y_{-i})$ for all i with $Y_i \neq y_j$. Therefore, for any i, j with $Y_i = y_j$,

$$r(Y) - r(Y_{-i}) = (C_j(Y) - 1)_+ - (C_j(Y_{-i}) - 1)_+ = \mathbb{1}_{C_j(Y) \ge 2}.$$

In particular, this implies

$$0 \le r(Y) - r(Y_{-i}) \le 1$$

for all $i \in [k]$, and also,

$$\sum_{i=1}^{k} \big(r(Y) - r(Y_{-i})\big) = \sum_{j=1}^{m} C_j(Y) \cdot \mathbb{1}_{C_j(Y) \ge 2} \le \sum_{j=1}^{m} 2(C_j(Y) - 1)_+ = 2r(Y).$$

This means that r is a $(2,0)$-strongly-self-bounding function, in the terminology of Boucheron et al. (2013, Section 6.11). Applying Boucheron et al. (2013, Theorem 6.20), then, for $Y = (Y_1, \dots, Y_n)$ where $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} P_Y$,

$$\mathbb{P}_{P_Y}\{r(Y) \ge \mathbb{E}_{P_Y}[r(Y)] - t\} \ge 1 - e^{-t^2/4\mathbb{E}_{P_Y}[r(Y)]}$$

for all $t > 0$, i.e., $r(Y)$ has a subgaussian left tail. Choosing $t = 2\sqrt{\mathbb{E}_{P_Y}[r(Y)]\log(1/\alpha_0)}$,

$$\mathbb{P}_{P_Y}\left\{ r(Y) \ge \mathbb{E}_{P_Y}[r(Y)] - 2\sqrt{\mathbb{E}_{P_Y}[r(Y)]\log(1/\alpha_0)} \right\} \ge 1 - \alpha_0. \tag{B.1}$$

Next we calculate $\mathbb{E}_{P_Y}[r(Y)]$. First, $C_j \sim \text{Binom}(n, p_j)$, where $p_j = \mathbb{P}_{P_Y}\{Y_1 = y_j\}$. Thus

$$\mathbb{E}_{P_Y}[(C_j - 1)_+] = \mathbb{E}_{P_Y}[C_j - 1 + \mathbb{1}_{C_j = 0}] = np_j - 1 + (1 - p_j)^n,$$

and so

$$\mathbb{E}_{P_Y}[r(Y)] = \sum_{j=1}^{m} \left(np_j - 1 + (1-p_j)^n\right) = n - m + \sum_{j=1}^{m}(1-p_j)^n.$$

Now let $p := (p_1, \ldots, p_m)$, which lies in the probability simplex. Since $p \mapsto \sum_{j=1}^{m}(1-p_j)^n$ is convex, this means that $\sum_{j=1}^{m}(1-p_j)^n$ is minimized at $p = (\frac{1}{m}, \ldots, \frac{1}{m})$, i.e.,

$$\mathbb{E}_{P_Y}[r(Y)] \geq n - m + m\left(1 - \frac{1}{m}\right)^n \geq \frac{n(n-1)}{n+2m},$$

where the last step holds by Lee and Barber (2021, Lemma 4).

Combining this lower bound on $\mathbb{E}_{P_Y}[r(Y)]$ with the calculation (B.1), we therefore have

$$\mathbb{P}_{P_Y}\left\{r(Y) \geq \left(\frac{n(n-1)}{n+2m} - 2\sqrt{\frac{n(n-1)}{n+2m} \cdot \log(1/\alpha_0)}\right)_+\right\} \geq 1 - \alpha_0,$$

since $t \mapsto (t - 2\sqrt{t\log(1/\alpha_0)})_+$ is a nondecreasing function on $t \geq 0$, and $r(Y)$ is nonnegative. By definition of r, and using the fact that $r(Y)$ is integer-valued, we therefore have $\mathbb{P}_{P_Y}\{r(Y) \geq r\} \geq 1 - \alpha_0$.

As the last step, we need to relate the target quantity I to the newly defined $r(Y)$. Note that $\sum_{j=1}^{m} C_j = n$ by construction. We therefore have

$$I = \sum_{j=1}^{m} \mathbb{1}_{C_j > 0} = \sum_{j=1}^{m} (C_j - (C_j - 1)_+) = n - r(Y),$$

which completes the proof. ■

Theorem B.2 offers a more informative lower bound than our original construction, in Theorem 6, both of which are valid without placing a bound on the loss. However, by combining the proof ideas from Theorem A.2 (which uses a truncated loss) and Theorem B.2, we can obtain the following potentially even more informative lower bound.

Theorem B.4 *Fix $\alpha \in (0,1)$, $n \geq 1$, and $m \geq 1$. Let $\alpha_0 + \alpha_1 = \alpha$, with $\alpha_0, \alpha_1 > 0$. Fix any $B > 0$, and define $\Delta_{n,B} \in (0,1]$ as the unique solution of*

$$-\Delta_{n,B} - \log(1 - \Delta_{n,B}) = \frac{B\log(1/\alpha_1)}{n\hat{R}(\mathcal{F}_{\text{pwc}}^{(n-r)}, \mathcal{D}_n; B)},$$

where r is defined as in Theorem B.2, and where the truncated empirical risk $\hat{R}(\cdot, \mathcal{D}_n; B)$ is defined as in Theorem A.2. Then

$$\hat{L}_\alpha^{\text{pwc}-\text{trunc},r,B}(\mathcal{F}_{\text{pwc}}^{(m)}, \mathcal{D}_n) := (1 - \Delta_{n,B})\hat{R}(\mathcal{F}_{\text{pwc}}^{(n-r)}, \mathcal{D}_n; B)$$

is a valid distribution-free lower bound on $R_P(\mathcal{F}_{\text{pwc}}^{(m)})$.

Proof of Theorem B.4. By the same argument and using the same notation as in the proofs of Theorems 6 and B.2, we have $\mathbb{P}_P\{|I(\mathcal{D}_n)| \leq n-r\} \geq 1-\alpha_0$. Let $\Delta'_{n,B}$ be the unique solution to

$$-\Delta'_{n,B} - \log(1-\Delta'_{n,B}) = \frac{B\log(1/\alpha_1)}{n\hat{R}(f_\varepsilon, \mathcal{D}_n; B)}.$$

Therefore, on the event $\{\hat{f}_\varepsilon \in \mathcal{F}_{\text{pwc}}^{(n-r)}\} \supseteq \{|I(\mathcal{D}_n)| \leq n-r\}$, we have $\hat{R}(f_\varepsilon, \mathcal{D}_n; B) = \hat{R}(\hat{f}_\varepsilon, \mathcal{D}_n; B) \geq \hat{R}(\mathcal{F}_{\text{pwc}}^{(n-r)}, \mathcal{D}_n; B)$ as in the proof of Theorem B.2, and therefore

$$-\Delta_{n,B} - \log(1-\Delta_{n,B}) = \frac{B\log(1/\alpha_1)}{n\hat{R}(\mathcal{F}_{\text{pwc}}^{(n-r)}, \mathcal{D}_n; B)} \geq \frac{B\log(1/\alpha_1)}{n\hat{R}(f_\varepsilon, \mathcal{D}_n; B)} = -\Delta'_{n,B} - \log(1-\Delta'_{n,B}).$$

Thus, on the event that $|I(\mathcal{D}_n)| \leq n-r$, we have $\Delta'_{n,B} \leq \Delta_{n,B}$ and so

$$\hat{L}_\alpha^{\text{pwc}-\text{trunc},r,B}(\mathcal{F}_{\text{pwc}}^{(m)}, \mathcal{D}_n) \leq (1-\Delta_{n,B})\hat{R}(f_\varepsilon, \mathcal{D}_n; B) \leq (1-\Delta'_{n,B})\hat{R}(f_\varepsilon, \mathcal{D}_n; B).$$

By Lemma A.3, $\mathbb{P}_P\big\{(1-\Delta'_{n,B})\hat{R}(f_\varepsilon, \mathcal{D}_n; B) \leq \mathbb{E}[\hat{R}(f_\varepsilon, \mathcal{D}_n; B)]\big\} \geq 1-\alpha_1$, and so

$$\begin{aligned}
&\mathbb{P}_P\left\{\hat{L}_\alpha^{\text{pwc}-\text{trunc},r,B}(\mathcal{F}_{\text{pwc}}^{(m)}, \mathcal{D}_n) \leq \mathbb{E}[\hat{R}(f_\varepsilon, \mathcal{D}_n; B)]\right\} \\
&\qquad \geq \mathbb{P}_P\big\{(1-\Delta'_{n,B})\hat{R}(f_\varepsilon, \mathcal{D}_n; B) \leq \mathbb{E}[\hat{R}(f_\varepsilon, \mathcal{D}_n; B)] \text{ and } |I(\mathcal{D}_n)| \leq n-r\big\} \geq 1-\alpha.
\end{aligned}$$

As in the proof of Theorem A.2, we have $\mathbb{E}[\hat{R}(f_\varepsilon, \mathcal{D}_n; B)] \leq R_P(\mathcal{F}_{\text{pwc}}^{(m)}) + \varepsilon$, so it therefore holds that $\mathbb{P}_P\left\{\hat{L}_\alpha^{\text{pwc}-\text{trunc},r,B}(\mathcal{F}_{\text{pwc}}^{(m)}, \mathcal{D}_n) \leq R_P(\mathcal{F}_{\text{pwc}}^{(m)}) + \varepsilon\right\} \geq 1-\alpha$, and since $\varepsilon > 0$ can be taken to be arbitrarily small, this completes the proof. ∎

Appendix C. Proofs and extensions for results in Section 4.2

C.1. Proofs for results on the linear model example

In this section, we prove all results stated in Section 4.2 for the linear model example.

Proof of Theorem 7. First, for any $Q \in \mathcal{Q}_{\text{lin}}$, we have $R_Q(\mathcal{F}_{\text{lin}}^{(d)}) = 0$ by definition. Therefore, by distribution-free validity of $\hat{L}_\alpha(\mathcal{F}_{\text{lin}}^{(d)}, \cdot)$, we must have

$$\mathbb{P}_{\mathcal{D}_n \sim Q^n}\left\{\hat{L}_\alpha(\mathcal{F}_{\text{lin}}^{(d)}, \mathcal{D}_n) = 0\right\} \geq 1-\alpha.$$

Since this is true for every $Q^n \in \mathcal{Q}_{\text{lin}}^{(n)}$, it must therefore also hold for any distribution in the convex hull of this class of distributions, i.e.,

$$\mathbb{P}_{\mathcal{D}_n \sim Q}\left\{\hat{L}_\alpha(\mathcal{F}_{\text{lin}}^{(d)}, \mathcal{D}_n) = 0\right\} \geq 1-\alpha \text{ for all } Q \in \tilde{\mathcal{Q}}_{\text{lin}}^{(n)}.$$

By definition of total variation distance, therefore,

$$\mathbb{P}_{\mathcal{D}_n \sim P^n}\left\{\hat{L}_\alpha(\mathcal{F}_{\text{lin}}^{(d)}, \mathcal{D}_n) = 0\right\} \geq 1-\alpha-\mathrm{d}_{\text{TV}}(P^n, Q) \text{ for all } Q \in \tilde{\mathcal{Q}}_{\text{lin}}^{(n)},$$

which completes the proof. ■

Proof of Proposition 8. Since $\tilde{\mathcal{Q}}_{\text{lin}}^{(n)}$ is defined by taking mixtures of noiseless linear models, the quantity $\lambda_{n,d}(P)$ is invariant to taking linear transformations of the features X—that is, for any invertible $A \in \mathbb{R}^{d\times d}$, if P' is the distribution of (AX, Y) when we draw $(X, Y) \sim P$, then $\lambda_{n,d}(P') = \lambda_{n,d}(P)$. Therefore, without loss of generality, from this point on we can assume $\Omega = \mathbf{I}_d$, and we will write $h(\mathbf{x}) = h_{\mathbf{I}_d}(\mathbf{x}) = (\det(\mathbf{x}\mathbf{x}^\top))^{-1/2}$.

First, we define two continuous distributions on $(\mathbf{X}, \beta) \in \mathbb{R}^{n\times d} \times \mathbb{R}^d$. Let $f(x, y)$ denote the density of the distribution P on $\mathbb{R}^d \times \mathbb{R}$. Now, fix any constant $c > 0$, and define $\tilde{P}$ as the distribution with density

$$g_{\tilde{P}}(\mathbf{x}, b) = \prod_{i=1}^n f(x_i, x_i^\top b) \cdot (2\pi c)^{-\frac{d-n}{2}} e^{-\|b\|_2^2/(2c)} \cdot \left[\sqrt{\det(\mathbf{x}\mathbf{x}^\top)} \cdot e^{\|\mathbf{P}_\mathbf{x} b\|_2^2/(2c)}\right],$$

where $\mathbf{P}_\mathbf{x} = \mathbf{x}^\top(\mathbf{x}\mathbf{x}^\top)^{-1}\mathbf{x} \in \mathbb{R}^{d\times d}$ is the projection matrix to the row span of $\mathbf{x}$. (Note that we are implicitly assuming $\mathbf{x}\mathbf{x}^\top \in \mathbb{R}^{n\times n}$ is invertible—since we are defining a density, it is sufficient that this holds for almost every $\mathbf{x} \in \mathbb{R}^{n\times d}$.) Next, define $\tilde{Q}$ as the distribution with density

$$g_{\tilde{Q}}(\mathbf{x}, b) = \prod_{i=1}^n f(x_i, x_i^\top b) \cdot (2\pi c)^{-\frac{d-n}{2}} e^{-\|b\|_2^2/(2c)} \cdot C^{-1},$$

where

$$C = \mathbb{E}_{\tilde{P}}\left[(\det(\mathbf{X}\mathbf{X}^\top))^{-1/2} \exp\left\{-\frac{1}{2c}\|\mathbf{P}_\mathbf{X}\beta\|_2^2\right\}\right] \le \mathbb{E}_{\tilde{P}}\left[h(\mathbf{X})\right], \tag{C.1}$$

where the inequality holds by the definition of $h(\cdot)$. (We will verify below that these functions are indeed well-defined densities, i.e., that C is finite and positive, and that $g_{\tilde{P}}$ and $g_{\tilde{Q}}$ each integrate to 1.) Next, let $\tilde{P}_{(\mathbf{X},\mathbf{Y})}$ denote the distribution of $(\mathbf{X}, \mathbf{Y}) := (\mathbf{X}, \mathbf{X}\beta) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$ induced by drawing $(\mathbf{X}, \beta) \sim \tilde{P}$, and define $\tilde{Q}_{(\mathbf{X},\mathbf{Y})}$ analogously.

The intuition of the remainder of the proof is the following. First, in Step 1, we will verify that $\tilde{P}$ and $\tilde{Q}$ are close in total variation distance—that is, we will bound $\mathrm{d_{TV}}(\tilde{P}, \tilde{Q})$, which therefore induces a bound on $\mathrm{d_{TV}}(\tilde{P}_{(\mathbf{X},\mathbf{Y})}, \tilde{Q}_{(\mathbf{X},\mathbf{Y})})$, by construction of $\tilde{P}_{(\mathbf{X},\mathbf{Y})}, \tilde{Q}_{(\mathbf{X},\mathbf{Y})}$ from $\tilde{P}, \tilde{Q}$. Next, in Step 2, we will verify that

$$\tilde{P}_{(\mathbf{X},\mathbf{Y})} = P^n, \text{ and } \tilde{Q}_{(\mathbf{X},\mathbf{Y})} \in \tilde{\mathcal{Q}}_{\text{lin}}^{(n)}, \tag{C.2}$$

and consequently, this will mean that $\lambda_{n,d}(P) \le \mathrm{d_{TV}}(\tilde{P}_{(\mathbf{X},\mathbf{Y})}, \tilde{Q}_{(\mathbf{X},\mathbf{Y})})$. Finally, in Step 3, we will take a limit as $c \to \infty$ to complete the proof.

Step 1: bounding the total variation distance. We calculate

$$\begin{aligned}
\mathrm{d_{TV}}(\tilde{P},\tilde{Q}) &= \mathbb{E}_{\tilde{P}}\left[\left(1-\frac{g_{\tilde{Q}}(\mathbf{X},\beta)}{g_{\tilde{P}}(\mathbf{X},\beta)}\right)_+\right] \\
&= \mathbb{E}_{\tilde{P}}\left[\left(1-\frac{C^{-1}}{\sqrt{\det(\mathbf{X}\mathbf{X}^\top)}\cdot e^{\|\mathbf{P}_\mathbf{X}\beta\|_2^2/(2c)}}\right)_+\right] \\
&\le \mathbb{E}_{\tilde{P}}\left[\left(1-\frac{h(\mathbf{X})}{\mathbb{E}_{\tilde{P}}[h(\mathbf{X})]}\cdot e^{-\|\mathbf{P}_\mathbf{X}\beta\|_2^2/(2c)}\right)_+\right] \quad \text{by definition of } h(\cdot) \text{ and by (C.1)} \\
&\le \mathbb{E}_{\tilde{P}}\left[e^{-\|\mathbf{P}_\mathbf{X}\beta\|_2^2/(2c)}\cdot\left(1-\frac{h(\mathbf{X})}{\mathbb{E}_{\tilde{P}}[h(\mathbf{X})]}\right)_+\right] + \mathbb{E}_{\tilde{P}}\left[1-e^{-\|\mathbf{P}_\mathbf{X}\beta\|_2^2/(2c)}\right] \\
&\le \mathbb{E}_{\tilde{P}}\left[\left(1-\frac{h(\mathbf{X})}{\mathbb{E}_{\tilde{P}}[h(\mathbf{X})]}\right)_+\right] + \mathbb{E}_{\tilde{P}}\left[1-e^{-\|\mathbf{P}_\mathbf{X}\beta\|_2^2/(2c)}\right] \\
&= \frac{1}{2}\mathbb{E}_{\tilde{P}}\left[\left|\frac{h(\mathbf{X})}{\mathbb{E}_{\tilde{P}}[h(\mathbf{X})]}-1\right|\right] + \mathbb{E}_{\tilde{P}}\left[1-e^{-\|\mathbf{P}_\mathbf{X}\beta\|_2^2/(2c)}\right],
\end{aligned}$$

where the last step holds since for any random variable A with $\mathbb{E}[A]=1$, it holds by symmetry that $\mathbb{E}[(1-A)_+]=\mathbb{E}[(A-1)_+]=\frac{1}{2}\mathbb{E}[|A-1|]$.

Next, we have that $\tilde{P}_{(\mathbf{X},\mathbf{Y})}$-almost surely

$$\mathbf{P}_\mathbf{X}\beta = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\beta = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y},$$

by definition of $\mathbf{Y}=\mathbf{X}\beta$. From (C.2) (which we will verify below), we know that $P^n=\tilde{P}_{(\mathbf{X},\mathbf{Y})}$, and therefore,

$$\mathrm{d_{TV}}(\tilde{P},\tilde{Q}) \le \frac{1}{2}\mathbb{E}_P\left[\left|\frac{h(\mathbf{X})}{\mathbb{E}_P[h(\mathbf{X})]}-1\right|\right] + \mathbb{E}_P\left[1-e^{-\|\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y}\|_2^2/(2c)}\right]. \tag{C.3}$$

Step 2: proving (C.2). We now need to verify (C.2) (and, along the way, to check that $g_{\tilde{P}}$ and $g_{\tilde{Q}}$ are well-defined densities). We begin by considering a joint distribution $\tilde{P}_*$ on $(\mathbf{X},\mathbf{Y},\beta)\in\mathbb{R}^{n\times d}\times\mathbb{R}^n\times\mathbb{R}^d$ generated as follows: first sample $(\mathbf{X},\mathbf{Y})\sim P^n$, then sample

$$\beta \mid (\mathbf{X},\mathbf{Y}) \sim \mathcal{N}\left(\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y}, c\mathbf{P}_\mathbf{X}^\perp\right),$$

where $\mathbf{P}_\mathbf{X}^\perp=\mathbf{I}_d-\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}=\mathbf{I}_d-\mathbf{P}_\mathbf{X}$ denotes the projection matrix onto the orthogonal complement of the row span of $\mathbf{X}$. (Since the distribution P has a density, note that $\mathbf{X}$ has rank n, almost surely, and so $\mathbf{X}\mathbf{X}^\top$ is invertible, almost surely.) Now consider the joint distribution of $(\mathbf{X},\beta)$ under $\tilde{P}_*$ (i.e., we are marginalizing out $\mathbf{Y}$), by first calculating the conditional density of $\beta\mid\mathbf{X}$. Let $U\in\mathbb{R}^{d\times(d-n)}$ be an orthonormal matrix satisfying $UU^\top=\mathbf{P}_\mathbf{X}^\perp$. Define

$$\gamma = \begin{pmatrix}\mathbf{X}\\ U^\top\end{pmatrix}\cdot\beta.$$

Then by construction,

$$\gamma \mid (\mathbf{X}, \mathbf{Y}) \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{X} \\ U^\top \end{pmatrix} \cdot \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y}, \begin{pmatrix} \mathbf{X} \\ U^\top \end{pmatrix} \cdot c\mathbf{P}_{\mathbf{X}}^{\perp} \cdot \begin{pmatrix} \mathbf{X} \\ U^\top \end{pmatrix}^\top\right)$$
$$= \mathcal{N}\left(\begin{pmatrix} \mathbf{Y} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & c\mathbf{I}_{d-n} \end{pmatrix}\right).$$

Then we can calculate that the distribution of γ conditional on $\mathbf{X}$ has the following density on $\mathbb{R}^d$ (note that we are now marginalizing over $\mathbf{Y}$):

$$g_{\gamma|\mathbf{X}}(z \mid \mathbf{X}) = \prod_{i=1}^{n} f_{Y|X}(z_i \mid X_i) \cdot (2\pi c)^{-\frac{d-n}{2}} e^{-\|(z_{n+1},\ldots,z_d)\|_2^2/(2c)}.$$

(Here $f_{Y|X}$ denotes the conditional density of Y given X under the joint distribution $(X, Y) \sim P$.) Moreover, when we condition on $\mathbf{X}$, since β is simply a linear transformation of γ, we therefore have

$$g_{\beta|\mathbf{X}}(b \mid \mathbf{X}) = \left|\det\left(\begin{pmatrix} \mathbf{X} \\ U^\top \end{pmatrix}\right)\right| \cdot g_{\gamma|\mathbf{X}}\left(\begin{pmatrix} \mathbf{X} \\ U^\top \end{pmatrix} \cdot b\right).$$

We calculate

$$\left|\det\left(\begin{pmatrix} \mathbf{X} \\ U^\top \end{pmatrix}\right)\right| = \sqrt{\det\left(\begin{pmatrix} \mathbf{X} \\ U^\top \end{pmatrix}\begin{pmatrix} \mathbf{X} \\ U^\top \end{pmatrix}^\top\right)}$$
$$= \sqrt{\det\begin{pmatrix} \mathbf{X}\mathbf{X}^\top & \mathbf{X}U \\ U^\top\mathbf{X}^\top & U^\top U \end{pmatrix}} = \sqrt{\det\begin{pmatrix} \mathbf{X}\mathbf{X}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-n} \end{pmatrix}} = \sqrt{\det(\mathbf{X}\mathbf{X}^\top)}.$$

And,

$$g_{\gamma|\mathbf{X}}\left(\begin{pmatrix} \mathbf{X} \\ U^\top \end{pmatrix} \cdot b\right) = \prod_{i=1}^{n} f_{Y|X}(X_i^\top b \mid X_i) \cdot (2\pi c)^{-\frac{d-n}{2}} e^{-\|U^\top b\|_2^2/(2c)}$$
$$= \prod_{i=1}^{n} f_{Y|X}(X_i^\top b \mid X_i) \cdot (2\pi c)^{-\frac{d-n}{2}} e^{-\|\mathbf{P}_{\mathbf{X}}^{\perp} b\|_2^2/(2c)},$$

since $UU^\top = \mathbf{P}_{\mathbf{X}}^{\perp}$. Therefore,

$$g_{\beta|\mathbf{X}}(b \mid \mathbf{X}) = \sqrt{\det(\mathbf{X}\mathbf{X}^\top)} \cdot \prod_{i=1}^{n} f_{Y|X}(X_i^\top b \mid X_i) \cdot (2\pi c)^{-\frac{d-n}{2}} e^{-\|\mathbf{P}_{\mathbf{X}}^{\perp} b\|_2^2/(2c)}$$
$$= \prod_{i=1}^{n} f_{Y|X}(X_i^\top b \mid X_i) \cdot (2\pi c)^{-\frac{d-n}{2}} e^{-\|b\|_2^2/(2c)} \cdot \left[\sqrt{\det(\mathbf{X}\mathbf{X}^\top)} \cdot e^{\|\mathbf{P}_{\mathbf{X}} b\|_2^2/(2c)}\right].$$

Writing f_X as the marginal density of X under the joint distribution $(X, Y) \sim P$, we then see that the density of $(\mathbf{X}, \beta)$ under $\tilde{P}_*$ (i.e., after marginalizing out $\mathbf{Y}$) is given by

$$\begin{aligned}
&\prod_{i=1}^n f_X(x_i) \cdot g_{\beta|\mathbf{X}}(b \mid \mathbf{x}) \\
&\quad = \prod_{i=1}^n f_X(x_i) \cdot \prod_{i=1}^n f_{Y|X}(x_i^\top b \mid x_i) \cdot (2\pi c)^{-\frac{d-n}{2}} e^{-\|b\|_2^2/(2c)} \cdot \left[\sqrt{\det(\mathbf{x}\mathbf{x}^\top)} \cdot e^{\|\mathbf{P}_\mathbf{x} b\|_2^2/(2c)}\right] \\
&\quad = g_{\tilde{P}}(\mathbf{x}, b).
\end{aligned}$$

In particular, this verifies that $g_{\tilde{P}}$ is a well-defined density, and also verifies that $\tilde{P}$ is the marginal distribution of $(\mathbf{X}, \beta)$, when $(\mathbf{X}, \mathbf{Y}, \beta) \sim \tilde{P}_*$. Moreover, under the distribution $\tilde{P}_*$, we have $(\mathbf{X}, \mathbf{Y}) \sim P^n$ by construction, and we also have $\mathbf{Y} = \mathbf{X}\beta$ almost surely and therefore $(\mathbf{X}, \mathbf{X}\beta) \sim P^n$ also holds. This proves the first part of (C.2).

Next consider $\tilde{Q}$. First, we verify that C is finite, since

$$C \leq \mathbb{E}_{\tilde{P}}[h(\mathbf{X})] = \mathbb{E}_P[h(\mathbf{X})] < \infty,$$

where the first step holds by (C.1), the second holds by the first part of (C.2), and the third holds by assumption in the proposition. Moreover, $C > 0$ since it is the expected value of a random variable that is positive almost surely. Next, note that

$$g_{\tilde{Q}}(\mathbf{x}, b) = g_{\tilde{P}}(\mathbf{x}, b) \cdot C^{-1} \cdot \frac{1}{\sqrt{\det(\mathbf{x}\mathbf{x}^\top)} \cdot e^{\|\mathbf{P}_\mathbf{x} b\|_2^2/(2c)}},$$

by definition of $g_{\tilde{P}}, g_{\tilde{Q}}$. Plugging in the definition of C, we have

$$g_{\tilde{Q}}(\mathbf{x}, b) = g_{\tilde{P}}(\mathbf{x}, b) \cdot \frac{(\det(\mathbf{x}\mathbf{x}^\top))^{-1/2} \exp\left\{-\frac{1}{2c}\|\mathbf{P}_\mathbf{x} b\|_2^2\right\}}{\mathbb{E}_{\tilde{P}}\left[(\det(\mathbf{X}\mathbf{X}^\top))^{-1/2} \exp\left\{-\frac{1}{2c}\|\mathbf{P}_\mathbf{X} b\|_2^2\right\}\right]},$$

which means that $g_{\tilde{Q}}$ must also integrate to 1 (i.e., it is a well-defined density). Next we need to verify that $\tilde{Q}_{(\mathbf{X},\mathbf{Y})} \in \tilde{\mathcal{Q}}_{\text{lin}}^{(n)}$. Define $\tilde{Q}_\beta$ as the distribution of $(\mathbf{X}, \mathbf{Y})$ conditional on β, under the joint distribution $\tilde{Q}$ on $(\mathbf{X}, \beta)$, when we define $\mathbf{Y} = \mathbf{X}\beta$. Clearly, $\tilde{Q}_{(\mathbf{X},\mathbf{Y})}$ can be expressed as a mixture of such distributions. Moreover, the data points (X_i, Y_i) are i.i.d. conditional on β (since by examining the density $g_{\tilde{Q}}$ we can see that it factors over data points i), and satisfy $Y_i = X_i^\top \beta$ almost surely. Therefore, $\tilde{Q}_\beta \in \mathcal{Q}_{\text{lin}}^{(n)}$, which completes the proof of (C.2).

Step 3: combining everything. Combining our calculations so far, we have shown that

$$\begin{aligned}
\lambda_{n,d}(P) &\leq \mathrm{d_{TV}}(\tilde{P}_{(\mathbf{X},\mathbf{Y})}, \tilde{Q}_{(\mathbf{X},\mathbf{Y})}) \leq \mathrm{d_{TV}}(\tilde{P}, \tilde{Q}) \\
&\leq \frac{1}{2}\mathbb{E}_P\left[\left|\frac{h(\mathbf{X})}{\mathbb{E}_P[h(\mathbf{X})]} - 1\right|\right] + \mathbb{E}_P\left[1 - e^{-\|\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y}\|_2^2/(2c)}\right],
\end{aligned}$$

where the first step holds by (C.2), the second step holds by construction of $\tilde{P}_{(\mathbf{X},\mathbf{Y})}, \tilde{Q}_{(\mathbf{X},\mathbf{Y})}$ from $\tilde{P}, \tilde{Q}$, and the third step holds by (C.3). Moreover, since this is true for any $c > 0$, and since for any

random variable $A \geq 0$ it holds that $\lim_{c\to\infty} \mathbb{E}[e^{-A/c}] = 1$ by the dominated convergence theorem, we therefore have

$$\lambda_{n,d}(P) \leq \frac{1}{2}\mathbb{E}_P\left[\left|\frac{h(\mathbf{X})}{\mathbb{E}_P[h(\mathbf{X})]} - 1\right|\right],$$

as desired. ■

Proof of Corollary 9. By Proposition 8 (applied with $\Omega = \Sigma^{-1}$), we have

$$\begin{aligned}
\lambda_{n,d}(P) &\leq \frac{1}{2}\mathbb{E}_P\left[\left|\frac{h_\Omega(\mathbf{X})}{\mathbb{E}_P[h_\Omega(\mathbf{X})]} - 1\right|\right] \\
&\leq \frac{1}{\sqrt{2}}\left\{\mathbb{E}_P\left[\log\left(\frac{\mathbb{E}_P[h_\Omega(\mathbf{X})]}{h_\Omega(\mathbf{X})}\right)\right]\right\}^{1/2} \\
&= \frac{1}{\sqrt{2}}\left\{\log\left(\mathbb{E}_P\left[(\det(\mathbf{X}\Sigma^{-1}\mathbf{X}^\top))^{-1/2}\right]\right) - \mathbb{E}_P\left[\log\left((\det(\mathbf{X}\Sigma^{-1}\mathbf{X}^\top))^{-1/2}\right)\right]\right\}^{1/2},
\end{aligned}$$

where the second step holds by Pinsker's inequality (to be more concrete, we are using the fact that, for a random variable $Z \geq 0$ with $\mathbb{E}[Z] = 1$, $\frac{1}{2}\mathbb{E}[|Z-1|] \leq \sqrt{\frac{1}{2}\mathbb{E}[\log(1/Z)]}$). Next we compute each of these expected values. First, observe that $\mathbf{X}\Sigma^{-1}\mathbf{X}^\top \sim W_n(\mathbf{I}_n, d)$ (a Wishart distribution). Goodman (1963) proves that the determinant of a random matrix drawn from the Wishart distribution $W_n(\mathbf{I}_n, d)$ is distributed as the product of independent random variables with a χ^2-distribution, i.e.,

$$\det(\mathbf{X}\Sigma^{-1}\mathbf{X}^\top) \overset{\mathrm{d}}{=} G_d \cdot G_{d-1} \cdot \ldots \cdot G_{d-n+1},$$

where $G_k \sim \chi^2_k$ for each k, and the G_k's are mutually independent. It is straightforward to check that, since $G_k \sim \chi^2_k$,

$$\mathbb{E}[G_k^{-1/2}] = \int_0^\infty \frac{1}{\sqrt{x}}\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}\,\mathrm{d}x = \frac{1}{\sqrt{2}}\frac{\Gamma(\frac{k-1}{2})}{\Gamma(\frac{k}{2})},$$

and

$$\begin{aligned}
\mathbb{E}\big[\log G_k\big] - \log 2 = \mathbb{E}\big[\log(G_k/2)\big] &= \int_0^\infty \frac{\log(x/2)}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}\,\mathrm{d}x \\
&= \int_0^\infty \frac{\log(t)}{\Gamma(k/2)}t^{k/2-1}e^{-t}\,\mathrm{d}t = \psi\Big(\frac{k}{2}\Big),
\end{aligned}$$

where the final equality follows from the integral representation of the digamma function $\psi(u) := \frac{\mathsf{d}}{\mathsf{d}u}\log\Gamma(u) = \int_0^\infty \log(x)x^{u-1}e^{-x}\,\mathsf{d}x/\Gamma(u)$, $u>0$ (Gordon, 1994). Therefore,

$$\begin{aligned}
&\log\mathbb{E}_P[(\det(\mathbf{X}\Sigma^{-1}\mathbf{X}^\top))^{-1/2}] - \mathbb{E}_P\left[\log(\det(\mathbf{X}\Sigma^{-1}\mathbf{X}^\top)^{-1/2})\right] \\
&= \log\mathbb{E}[(G_d\cdot\ldots\cdot G_{d-n+1})^{-1/2}] + \frac{1}{2}\mathbb{E}\left[\log(G_d\cdot\ldots\cdot G_{d-n+1})\right] \\
&= \log\prod_{k=d-n+1}^{d}\mathbb{E}[G_k^{-1/2}] + \frac{1}{2}\sum_{k=d-n+1}^{d}\mathbb{E}[\log G_k] \\
&= \sum_{k=d-n+1}^{d}\left\{\log\left(\frac{1}{\sqrt{2}}\frac{\Gamma(\frac{k-1}{2})}{\Gamma(\frac{k}{2})}\right) + \frac{1}{2}\psi\Big(\frac{k}{2}\Big) + \frac{\log 2}{2}\right\} \\
&= \sum_{j=1}^{n}\left\{\log\left(\frac{\Gamma(\frac{d-1}{2}+\frac{1-j}{2})}{\Gamma(\frac{d}{2}+\frac{1-j}{2})}\right) + \frac{1}{2}\psi\Big(\frac{d}{2}+\frac{1-j}{2}\Big)\right\}.
\end{aligned}$$

Next, we bound each term in the sum. Since $\psi(u) = \frac{\mathsf{d}}{\mathsf{d}u}\log\Gamma(u)$ and $d \geq n+2$, we have

$$\begin{aligned}
&\log\left(\Gamma\left(\frac{d-1}{2}+\frac{1-j}{2}\right)\right) - \log\left(\Gamma\left(\frac{d}{2}+\frac{1-j}{2}\right)\right) + \frac{1}{2}\psi\Big(\frac{d}{2}+\frac{1-j}{2}\Big) \\
&= \frac{1}{2}\left[\psi\Big(\frac{d}{2}+\frac{1-j}{2}\Big) - \psi\Big(\frac{d-c}{2}+\frac{1-j}{2}\Big)\right] \quad \text{for some } c\in[0,1], \text{ by Taylor's theorem} \\
&\leq \frac{1}{2}\left[\psi\Big(\frac{d}{2}+\frac{1-j}{2}\Big) - \psi\Big(\frac{d-1}{2}+\frac{1-j}{2}\Big)\right] \quad \text{since } \psi(u) \text{ is an increasing function on } u>0 \\
&\leq \frac{1}{4}\left[\psi\Big(\frac{d}{2}+\frac{1-j}{2}\Big) - \psi\Big(\frac{d-2}{2}+\frac{1-j}{2}\Big)\right] \quad \text{since } \psi(u) \text{ is a concave function on } u>0 \\
&= \frac{1}{4}\left[\frac{1}{\frac{d-2}{2}+\frac{1-j}{2}}\right] = \frac{1}{2(d-1-j)},
\end{aligned}$$

where we use the fact that $\psi(u+1) = \psi(u) + \frac{1}{u}$ holds for all $u>0$. Therefore, combining all our calculations,

$$\begin{aligned}
2\big(\lambda_{n,d}(P)\big)^2 \leq \log\mathbb{E}_P[(\det(\mathbf{X}\Sigma^{-1}\mathbf{X}^\top))^{-1/2}] + \frac{1}{2}\mathbb{E}_P\left[\log(\det(\mathbf{X}\Sigma^{-1}\mathbf{X}^\top))\right] \\
\leq \sum_{j=1}^{n}\frac{1}{2(d-1-j)} \leq \frac{n}{2(d-1-n)},
\end{aligned}$$

which completes the proof. ■

C.2. Extensions for the linear model example: generalizing to other distributions

We extend the results of Section 4.2, to prove a bound on $\lambda_{n,d}(P)$ for a broader family of distributions.

Proposition C.1 *Let P, Q be distributions on $\mathbb{R}^d \times \mathbb{R}$, such that Q is absolutely continuous with respect to P. Assume $\frac{\mathrm{d}Q}{\mathrm{d}P}(x,y) \le \varepsilon^{-1}$ for P-almost every (x,y). Then, for all $n \ge 1$*

$$\lambda_{n,d}(Q) \le \lambda_{\lceil 2n/\varepsilon \rceil, d}(P) + e^{-n/4}.$$

For instance, if P is a distribution with a Gaussian marginal P_X, then we know that $\lambda_{\lceil 2n/\varepsilon \rceil, d}(P) \approx 0$ (as long as $d \gg n/\varepsilon$), by Corollary 9. This means that $\lambda_{n,d}(Q)$ will also be small for any distribution Q with sufficiently light tails, and therefore, Theorem 7 shows that a nontrivial lower bound is not possible for the model class risk $R_Q(\mathcal{F}_{\mathrm{lin}}^{(d)})$.

In order to prove this result, first we need a lemma on rejection sampling.

Lemma C.2 *Let P, Q be distributions on $\mathcal{Z}$, such that Q is absolutely continuous with respect to P. Assume $\frac{\mathrm{d}Q}{\mathrm{d}P}(z) \le \varepsilon^{-1}$ for P-almost every z.*

Let $\tilde{P}$ denote the distribution on $(Z, B) \in \mathcal{Z} \times \{0,1\}$ constructed by sampling $Z \sim P$, then sampling $B \mid Z \sim \mathrm{Bernoulli}(\varepsilon \cdot \frac{\mathrm{d}Q}{\mathrm{d}P}(Z))$. Then, for any $N \ge n \ge 1$, and for any function $f : \mathcal{Z}^n \to [0,1]$,

$$0 \le \mathbb{E}_{Q^n}[f] - \mathbb{E}_{\tilde{P}^N}\left[\sum_{1 \le i_1 < \cdots < i_n \le N} f(Z_{i_1}, \ldots, Z_{i_n}) \cdot \frac{\mathbb{1}_{B_{i_1} = \cdots = B_{i_n} = 1}}{1 \vee \binom{\sum_i B_i}{n}}\right] \le \mathbb{P}\{\mathrm{Binom}(N, \varepsilon) < n\}.$$

To understand this lemma, we can interpret it as a result about rejection sampling. To draw one sample from the distribution of $f(Z_1, \ldots, Z_n)$ under Q^n, we do the following:

- We draw N samples from P, given by $Z_1, \ldots, Z_N$, and draw the Bernoulli random variables $B_1, \ldots, B_N$ to perform rejection sampling.
- The "accepted" draws—that is, all Z_i for which $B_i = 1$—can be viewed as a random sample drawn i.i.d. from Q. If $\sum_{i=1}^N B_i \ge n$, then, we can choose a random subset of n of these accepted draws (indices $1 \le i_1 < \cdots < i_n \le N$), and evaluate $f(Z_{i_1}, \ldots, Z_{i_n})$. This is a draw from the target distribution.
- However, on the event that $\sum_{i=1}^N B_i < n$ (which is highly unlikely if we choose N to be sufficiently large), we do not have sufficiently many accepted samples; instead we might simply return 0 since no estimate is available.

With this intuition in place, the proof is straightforward.

Proof of Lemma C.2. First we condition on $B_1, \ldots, B_N$. On the event that $\sum_{i=1}^N B_i \ge n$, by the argument described above, for any indices $1 \le i_1 < \cdots < i_n \le N$ such that $B_{i_1} = \cdots = B_{i_n} = 1$, it holds that $f_{i_1,\ldots,i_n} := f(Z_{i_1}, \ldots, Z_{i_n})$ is therefore a draw from the target distribution—that is,

$$\mathbb{E}_{\tilde{P}^N}\left[f_{i_1,\ldots,i_n} \mid B_{i_1} = \cdots = B_{i_n} = 1\right] = \mathbb{E}_{Q^n}[f].$$

Then

$$\mathbb{E}_{\tilde{P}^N}\left[f_{i_1,\ldots,i_n} \cdot \mathbb{1}_{B_{i_1} = \cdots = B_{i_n} = 1} \mid B_1, \ldots, B_N\right] = \mathbb{E}_{Q^n}[f] \cdot \mathbb{1}_{B_{i_1} = \cdots = B_{i_n} = 1},$$

and so averaging over all possible collections of indices, we then have

$$\frac{\mathbb{E}_{\tilde{P}^N}\left[\sum_{1 \le i_1 < \cdots < i_n \le N} f_{i_1,\ldots,i_n} \cdot \mathbb{1}_{B_{i_1} = \cdots = B_{i_n} = 1} \,\middle|\, B_1, \ldots, B_N\right]}{\sum_{1 \le i_1 < \cdots < i_n \le N} \mathbb{1}_{B_{i_1} = \cdots = B_{i_n} = 1}} = \mathbb{E}_{Q^n}[f],$$

on the event that the sum in the denominator is positive. Note that this denominator is equal to $\binom{\sum_i B_i}{n}$, which is nonzero if and only if $\sum_i B_i \geq n$, so equivalently we have

$$\frac{\mathbb{E}_{\tilde{P}^N}\left[\sum_{1\leq i_1<\cdots<i_n\leq N} f_{i_1,\ldots,i_n}\cdot \mathbb{1}_{B_{i_1}=\cdots=B_{i_n}=1}\,\middle|\, B_1,\ldots,B_N\right]}{\binom{\sum_i B_i}{n}} = \mathbb{E}_{Q^n}[f],$$

on the event that $\sum_i B_i \geq n$. To cover both cases, then, we have

$$\frac{\mathbb{E}_{\tilde{P}^N}\left[\sum_{1\leq i_1<\cdots<i_n\leq N} f_{i_1,\ldots,i_n}\cdot \mathbb{1}_{B_{i_1}=\cdots=B_{i_n}=1}\,\middle|\, B_1,\ldots,B_N\right]}{1\vee\binom{\sum_i B_i}{n}} = \mathbb{E}_{Q^n}[f]\cdot \mathbb{1}\left\{\sum_i B_i \geq n\right\}.$$

Marginalizing over the B_i's, then,

$$\mathbb{E}_{\tilde{P}^N}\left[\frac{\sum_{1\leq i_1<\cdots<i_n\leq N} f_{i_1,\ldots,i_n}\cdot \mathbb{1}_{B_{i_1}=\cdots=B_{i_n}=1}}{1\vee\binom{\sum_i B_i}{n}}\right] = \mathbb{E}_{Q^n}[f]\cdot \mathbb{P}_{\tilde{P}^N}\left\{\sum_i B_i \geq n\right\}.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{Q^n}[f] - \mathbb{E}_{\tilde{P}^N}\left[\frac{\sum_{1\leq i_1<\cdots<i_n\leq N} f_{i_1,\ldots,i_n}\cdot \mathbb{1}_{B_{i_1}=\cdots=B_{i_n}=1}}{1\vee\binom{\sum_i B_i}{n}}\right] \\
&\qquad = \mathbb{E}_{Q^n}[f]\cdot \mathbb{P}_{\tilde{P}^N}\left\{\sum_i B_i < n\right\} = \mathbb{E}_{Q^n}[f]\cdot \mathbb{P}\{\text{Binom}(N,\varepsilon) < n\} \\
&\qquad\qquad \in \Big[0, \mathbb{P}\{\text{Binom}(N,\varepsilon) < n\}\Big].
\end{aligned}$$

■

With this lemma in place, we are now ready to prove the extension.

Proof of Proposition C.1. First we will prove the result in the case where $\frac{\mathsf{d}Q}{\mathsf{d}P}(x,y) > 0$ for all $(x,y) \in \mathbb{R}^d \times \mathbb{R}$.

Fix any $N \geq n$ and any $\delta > 0$. By the definition of $\lambda_{N,d}(P)$, there exists a distribution $P_* \in \tilde{\mathcal{Q}}_{\text{lin}}^{(N)}$, such that

$$\mathrm{d_{TV}}(P^N, P_*) \leq \lambda_{N,d}(P) + \delta.$$

By definition of $\tilde{\mathcal{Q}}_{\text{lin}}^{(N)}$, the distribution P_* is equal to a mixture of distributions in $\mathcal{Q}_{\text{lin}}^{(N)}$, i.e., there exists a probability measure ν on the space of distributions $\mathcal{Q}_{\text{lin}}$, such that sampling from P_* is the same as sampling from $(P_0)^N$ after sampling $P_0 \sim \nu$.

For any $P_0 \sim \nu$, define a corresponding distribution $Q(P_0)$ given by

$$\frac{\mathsf{d}Q(P_0)}{\mathsf{d}P_0}(x,y) = \frac{\frac{\mathsf{d}Q}{\mathsf{d}P}(x,y)}{\mathbb{E}_{(X,Y)\sim P_0}\left[\frac{\mathsf{d}Q}{\mathsf{d}P}(X,Y)\right]}.$$

Since $P_0 \in \mathcal{Q}_{\text{lin}}$ and $Q(P_0)$ is absolutely continuous with respect to P_0, it follows that $Q(P_0) \in \mathcal{Q}_{\text{lin}}$ as well (i.e., since there exists some $\beta \in \mathbb{R}^d$ such that $Y = X^\top\beta$ holds P_0-almost surely, we also

have that $Y = X^\top \beta$ holds $Q(P_0)$-almost surely). Consequently, we can define $Q_* \in \tilde{\mathcal{Q}}_{\text{lin}}^{(n)}$ as the mixture obtained by sampling $P_0 \sim \nu$ and then returning $(Q(P_0))^n$.

Our next step is to bound $\mathrm{d_{TV}}(Q^n, Q_*)$. Fix any $A \subseteq (\mathbb{R}^d \times \mathbb{R})^n$. By Lemma C.2 (applied with $f = \mathbb{1}_A$ and $Z = (X, Y)$),

$$Q^n(A) \leq \mathbb{E}_{\tilde{P}^N}\left[\frac{\sum_{1\leq i_1<\cdots<i_n\leq N} \mathbb{1}_{((X_{i_1},Y_{i_1}),\ldots,(X_{i_n},Y_{i_n}))\in A} \cdot \mathbb{1}_{B_{i_1}=\cdots=B_{i_n}=1}}{1 \vee \binom{\sum_i B_i}{n}}\right] + \mathbb{P}\{\text{Binom}(N,\varepsilon) < n\},$$

where $\tilde{P}$ is the distribution on $(X, Y, B) \in \mathbb{R}^d \times \mathbb{R} \times \{0, 1\}$ defined as in the statement of the lemma. Next, for any $P_0 \sim \nu$, define $\tilde{P}_0$ as the distribution on $(X, Y, B) \in \mathbb{R}^d \times \mathbb{R} \times \{0, 1\}$ obtained by sampling $(X, Y) \sim P_0$, then

$$B \mid (X, Y) \sim \text{Bernoulli}\left(\varepsilon \cdot \frac{\mathsf{d}Q}{\mathsf{d}P}(X, Y)\right) = \text{Bernoulli}\left(\varepsilon_{P_0} \cdot \frac{\mathsf{d}Q(P_0)}{\mathsf{d}P_0}(X, Y)\right),$$

where $\varepsilon_{P_0} = \varepsilon \cdot \mathbb{E}_{(X,Y)\sim P_0}\left[\frac{\mathsf{d}Q}{\mathsf{d}P}(X, Y)\right]$. Again applying Lemma C.2,

$$(Q(P_0))^n(A) \geq \mathbb{E}_{(\tilde{P}_0)^N}\left[\frac{\sum_{1\leq i_1<\cdots<i_n\leq N} \mathbb{1}_{((X_{i_1},Y_{i_1}),\ldots,(X_{i_n},Y_{i_n}))\in A} \cdot \mathbb{1}_{B_{i_1}=\cdots=B_{i_n}=1}}{1 \vee \binom{\sum_i B_i}{n}}\right].$$

Therefore,

$$\begin{aligned} Q_*(A) &= \mathbb{E}_{P_0\sim\nu}[(Q(P_0))^n(A)] \\ &\geq \mathbb{E}_{P_0\sim\nu}\left[\mathbb{E}_{(\tilde{P}_0)^N}\left[\frac{\sum_{1\leq i_1<\cdots<i_n\leq N} \mathbb{1}_{((X_{i_1},Y_{i_1}),\ldots,(X_{i_n},Y_{i_n}))\in A} \cdot \mathbb{1}_{B_{i_1}=\cdots=B_{i_n}=1}}{1 \vee \binom{\sum_i B_i}{n}}\right]\right]. \end{aligned}$$

Let $\tilde{P}_*$ be the distribution on $(\mathbb{R}^d \times \mathbb{R} \times \{0, 1\})^N$ obtained as follows: sample $P_0 \sim \nu$, then return a draw from $(\tilde{P}_0)^N$. Then we equivalently have

$$Q_*(A) \geq \mathbb{E}_{\tilde{P}_*}\left[\frac{\sum_{1\leq i_1<\cdots<i_n\leq N} \mathbb{1}_{((X_{i_1},Y_{i_1}),\ldots,(X_{i_n},Y_{i_n}))\in A} \cdot \mathbb{1}_{B_{i_1}=\cdots=B_{i_n}=1}}{1 \vee \binom{\sum_i B_i}{n}}\right].$$

Since the quantity inside the expected value must always lie in $[0, 1]$, it therefore holds that

$$Q_*(A) \geq \mathbb{E}_{\tilde{P}^N}\left[\frac{\sum_{1\leq i_1<\cdots<i_n\leq N} \mathbb{1}_{((X_{i_1},Y_{i_1}),\ldots,(X_{i_n},Y_{i_n}))\in A} \cdot \mathbb{1}_{B_{i_1}=\cdots=B_{i_n}=1}}{1 \vee \binom{\sum_i B_i}{n}}\right] - \mathrm{d_{TV}}(\tilde{P}^N, \tilde{P}_*).$$

Combining everything, then,

$$Q^n(A) \leq Q_*(A) + \mathrm{d_{TV}}(\tilde{P}^N, \tilde{P}_*) + \mathbb{P}\{\text{Binom}(N,\varepsilon) < n\}.$$

Since this holds for all $A \subseteq (\mathbb{R}^d \times \mathbb{R})^n$, and since $Q_* \in \tilde{\mathcal{Q}}_{\text{lin}}^{(n)}$, we therefore have

$$\lambda_{n,d}(Q) \leq \mathrm{d_{TV}}(Q^n, Q_*) = \sup_{A\subseteq(\mathbb{R}^d\times\mathbb{R})^n} \{Q^n(A) - Q_*(A)\} \leq \mathrm{d_{TV}}(\tilde{P}^N, \tilde{P}_*) + \mathbb{P}\{\text{Binom}(N,\varepsilon) < n\}.$$

Next, by construction, for both $\tilde{P}^N$ and $\tilde{P}_*$, the conditional distribution of $(B_1, \ldots, B_N)$ given $(X_1, Y_1), \ldots, (X_N, Y_N)$ is equal to

$$\text{Bernoulli}\left(\varepsilon \cdot \frac{\mathsf{d}Q}{\mathsf{d}P}(X_1, Y_1)\right) \times \cdots \times \text{Bernoulli}\left(\varepsilon \cdot \frac{\mathsf{d}Q}{\mathsf{d}P}(X_N, Y_N)\right).$$

In other words, the total variation distance between $\tilde{P}^N$ and $\tilde{P}_*$ is solely due to the difference in marginal distributions over $((X_1, Y_1), \ldots, (X_N, Y_N))$, i.e.,

$$\mathrm{d_{TV}}(\tilde{P}^N, \tilde{P}_*) = \mathrm{d_{TV}}(P^N, P_*).$$

By our choice of P_*, we therefore have $\mathrm{d_{TV}}(\tilde{P}^N, \tilde{P}_*) = \mathrm{d_{TV}}(P^N, P_*) \leq \lambda_{N,d}(P) + \delta$, and so

$$\lambda_{n,d}(Q) \leq \lambda_{N,d}(P) + \delta + \mathbb{P}\{\text{Binom}(N, \varepsilon) < n\}.$$

Finally, taking $N = \lceil 2n/\varepsilon \rceil$, and applying a Chernoff bound (McDiarmid, 1998, Theorem 2.3(c)), we have

$$\mathbb{P}\{\text{Binom}(N, \varepsilon) < n\} \leq \exp\{-\varepsilon N/8\} \leq e^{-n/4},$$

and thus

$$\lambda_{n,d}(Q) \leq \lambda_{N,d}(P) + \delta + e^{-n/4}.$$

Since $\delta > 0$ can be taken to be arbitrarily small, this completes the proof for the case that $\frac{\mathsf{d}Q}{\mathsf{d}P}(x, y) > 0$ for all (x, y).

Next we turn to the general case. Fix some small $c \in (0, 1)$, and define $Q_c = (1 - c)Q + cP$. Then

$$\frac{\mathsf{d}Q_c}{\mathsf{d}P}(x, y) = (1 - c) \cdot \frac{\mathsf{d}Q}{\mathsf{d}P}(x, y) + c \in (0, \varepsilon^{-1}],$$

so we can apply the result of the proposition to this perturbed distribution to obtain

$$\lambda_{n,d}(Q_c) \leq \lambda_{N,d}(P) + e^{-n/4}.$$

But by definition of $\lambda_{n,d}$, we have

$$\lambda_{n,d}(Q) \leq \lambda_{n,d}(Q_c) + \mathrm{d_{TV}}(Q^n, (Q_c)^n) \leq \lambda_{n,d}(Q_c) + nc.$$

Taking $c \in (0, 1)$ to be arbitrarily small, we have completed the proof for the general case.

■

C.3. Extensions for the linear model example: a tighter bound for the low-dimensional case

In this section, we construct a more informative lower bound for $R_P(\mathcal{F}_{\text{lin}}^{(d)})$ by applying the truncated loss construction of Theorem A.2, and show that in the low-dimensional setting ($d < n$), this lower bound provides an accurate estimate of the model class risk $R_P(\mathcal{F}_{\text{lin}}^{(d)})$.

Theorem C.3 *Fix $\alpha \in (0,1)$, and $n \geq d \geq 1$. Let $\alpha_0 + \alpha_1 = \alpha$, with $\alpha_0, \alpha_1 > 0$. Define the valid lower bound $\hat{L}_\alpha^{\mathrm{ERM-trunc},B}(\mathcal{F}_{\mathrm{lin}}^{(d)}, \cdot)$ as in Theorem A.2. Then, for any data set $\mathcal{D}_n = ((X_1,Y_1),\ldots,(X_n,Y_n)) \in (\mathbb{R}^d \times \mathbb{R})^n$, if it holds that*

$$\frac{1}{n}\sum_{i=1}^n Y_i^4 \leq \gamma, \text{ and for all } b \in \mathbb{R}^d, \ \frac{1}{n}\sum_{i=1}^n (X_i^\top b)^2 \geq \lambda_0 \|b\|_2^2, \ \frac{1}{n}\sum_{i=1}^n (X_i^\top b)^4 \leq \lambda_1 \|b\|_2^4, \quad \text{(C.4)}$$

for some $\gamma, \lambda_0, \lambda_1 > 0$, then

$$\hat{L}_\alpha^{\mathrm{ERM-trunc},B}(\mathcal{F}_{\mathrm{lin}}^{(d)}, \mathcal{D}_n) \geq \frac{1}{n}\|\mathcal{P}_{\mathbf{X}}^\perp(\mathbf{Y})\|_2^2 - \sqrt{\frac{2B^2\log(1/\alpha)}{n}} - \frac{c}{B},$$

where $\mathcal{P}_{\mathbf{X}}^\perp$ denotes projection to the orthogonal complement of the span of the columns of $\mathbf{X}$, and where the constant c depends only on $\gamma, \lambda_0, \lambda_1$.

The condition (C.4) will hold with high probability under standard mild assumptions on the distribution of the data, via the usual concentration arguments (including matrix concentration arguments for the minimum eigenvalue—see, e.g., Tropp (2012)).

As an illustrative example, we can consider the setting where P follows a linear model, $Y_i = X_i^\top \beta^* + \zeta_i$, for some fixed true coefficient vector β^*, and i.i.d. noise ζ_i with mean zero and variance σ^2. In this case, we have $R_P(\mathcal{F}_{\mathrm{lin}}^{(d)}) = \sigma^2$. And, by standard arguments, when $\mathbf{X}$ has full column rank, then

$$\mathbb{E}_P\left[\frac{1}{n}\|\mathcal{P}_{\mathbf{X}}^\perp(\mathbf{Y})\|_2^2\right] = \left(1 - \frac{d}{n}\right)\sigma^2 = \left(1 - \frac{d}{n}\right) R_P(\mathcal{F}_{\mathrm{lin}}^{(d)}).$$

In particular, if $d \ll n$, and if we choose the truncation parameter B to satisfy $1 \ll B \ll \sqrt{n}$, this shows that the valid lower bound $\hat{L}_\alpha^{\mathrm{ERM-trunc},B}(\mathcal{F}_{\mathrm{lin}}^{(d)}, \mathcal{D}_n)$ is in fact an accurate estimator of the true risk $R_P(\mathcal{F}_{\mathrm{lin}}^{(d)})$.

Proof of Theorem C.3. Applying (A.3), we have

$$\hat{L}_\alpha^{\mathrm{ERM-trunc},B}(\mathcal{F}_{\mathrm{lin}}^{(d)}, \mathcal{D}_n) \geq \hat{R}(\mathcal{F}_{\mathrm{lin}}^{(d)}, \mathcal{D}_n; B) - \sqrt{\frac{2B^2\log(1/\alpha)}{n}}.$$

Consequently, from this point on, we only need to show that $\hat{R}(\mathcal{F}_{\mathrm{lin}}^{(d)}, \mathcal{D}_n; B) \geq \frac{1}{n}\|\mathcal{P}_{\mathbf{X}}^\perp(\mathbf{Y})\|_2^2 - \frac{c}{B}$ holds under the assumption (C.4).

Consider any function of the form $f(x) = x^\top \beta$, for $\beta \in \mathbb{R}^d$. First we consider the case $\|\beta\|_2^2 > \frac{4\sqrt{\gamma}}{\lambda_0}$. Define the unit vector $u = \beta/\|\beta\|_2$. Then

$$
\begin{aligned}
\hat{R}(f, \mathcal{D}_n; B) &= \frac{1}{n}\sum_{i=1}^n \min\{(Y_i - X_i^\top \beta)^2, B\} \\
&\geq \frac{1}{n}\sum_{i=1}^n \min\left\{\frac{1}{2}(X_i^\top \beta)^2 - Y_i^2, B\right\} \quad \text{since } (a+b)^2 \leq 2a^2 + 2b^2 \text{ for all } a, b \in \mathbb{R} \\
&\geq \frac{1}{n}\sum_{i=1}^n \left(\min\left\{\frac{1}{2}(X_i^\top \beta)^2, B\right\} - Y_i^2\right) \\
&\geq -\frac{1}{n}\sum_{i=1}^n Y_i^2 + \frac{1}{n}\sum_{i=1}^n \min\left\{\frac{2\sqrt{\gamma}}{\lambda_0}(X_i^\top u)^2, B\right\} \quad \text{since } \|\beta\|_2^2 > \frac{4\sqrt{\gamma}}{\lambda_0} \\
&\geq -\frac{1}{n}\sum_{i=1}^n Y_i^2 + \frac{1}{n}\sum_{i=1}^n \frac{2\sqrt{\gamma}}{\lambda_0}(X_i^\top u)^2 - \frac{1}{n}\sum_{i=1}^n \frac{2\sqrt{\gamma}}{\lambda_0}(X_i^\top u)^2 \cdot \mathbb{1}\left\{\frac{2\sqrt{\gamma}}{\lambda_0}(X_i^\top u)^2 \geq B\right\} \\
&\geq -\frac{1}{n}\sum_{i=1}^n Y_i^2 + \frac{1}{n}\sum_{i=1}^n \frac{2\sqrt{\gamma}}{\lambda_0}(X_i^\top u)^2 - \frac{1}{n}\sum_{i=1}^n \frac{1}{B}\left[\frac{2\sqrt{\gamma}}{\lambda_0}(X_i^\top u)^2\right]^2 \\
&\geq -\frac{1}{n}\sum_{i=1}^n Y_i^2 + 2\sqrt{\gamma} - \frac{4\gamma\lambda_1}{B\lambda_0^2} \quad \text{by (C.4)} \\
&\geq \frac{1}{n}\sum_{i=1}^n Y_i^2 - \frac{4\gamma\lambda_1}{B\lambda_0^2} \quad \text{since } \frac{1}{n}\sum_{i=1}^n Y_i^2 \leq \sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^4} \leq \sqrt{\gamma} \text{ by (C.4)} \\
&\geq \frac{1}{n}\|\mathcal{P}_{\mathbf{X}}^\perp(\mathbf{Y})\|_2^2 - \frac{4\gamma\lambda_1}{B\lambda_0^2} \quad \text{since } \sum_{i=1}^n Y_i^2 = \|\mathbf{Y}\|_2^2 \geq \|\mathcal{P}_{\mathbf{X}}^\perp(\mathbf{Y})\|_2^2.
\end{aligned}
$$

Next consider the case $\|\beta\|_2^2 \leq \frac{4\sqrt{\gamma}}{\lambda_0}$. Then

$$
\begin{aligned}
\hat{R}(f, \mathcal{D}_n; B) &= \frac{1}{n}\sum_{i=1}^n \min\{(Y_i - X_i^\top \beta)^2, B\} \\
&\geq \frac{1}{n}\sum_{i=1}^n (Y_i - X_i^\top \beta)^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - X_i^\top \beta)^2 \cdot \mathbb{1}\{(Y_i - X_i^\top \beta)^2 \geq B\} \\
&\geq \frac{1}{n}\|\mathcal{P}_{\mathbf{X}}^\perp(Y)\|_2^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - X_i^\top \beta)^2 \cdot \mathbb{1}\{(Y_i - X_i^\top \beta)^2 \geq B\} \\
&\geq \frac{1}{n}\|\mathcal{P}_{\mathbf{X}}^\perp(Y)\|_2^2 - \frac{1}{n}\sum_{i=1}^n \frac{1}{B}(Y_i - X_i^\top \beta)^4 \\
&\geq \frac{1}{n}\|\mathcal{P}_{\mathbf{X}}^\perp(Y)\|_2^2 - \frac{1}{n}\sum_{i=1}^n \frac{1}{B}\left(8Y_i^4 + 8(X_i^\top \beta)^4\right) \quad \text{since } (a+b)^4 \leq 8a^4 + 8b^4 \text{ for all } a, b \in \mathbb{R} \\
&\geq \frac{1}{n}\|\mathcal{P}_{\mathbf{X}}^\perp(Y)\|_2^2 - \frac{8(\gamma + \lambda_1(\frac{4\sqrt{\gamma}}{\lambda_0})^2)}{B} \quad \text{by (C.4)}.
\end{aligned}
$$

Therefore, combining both cases, we have

$$\hat{R}(\mathcal{F}_{\text{lin}}^{(d)}, \mathcal{D}_n; B) \geq \min \left\{ \frac{1}{n} \|\mathcal{P}_{\mathbf{X}}^{\perp}(Y)\|_2^2 - \frac{8(\gamma + \lambda_1(\frac{4\sqrt{\gamma}}{\lambda_0})^2)}{B}, \frac{1}{n} \|\mathcal{P}_{\mathbf{X}}^{\perp}(\mathbf{Y})\|_2^2 - \frac{4\gamma\lambda_1}{B\lambda_0^2} \right\}.$$

Choosing the constant c appropriately completes the proof. ■