Data-Adaptive Tradeoffs among Multiple Risks in Distribution-Free Prediction

Drew T. Nguyen^{*1}, Reese Pathak^{*2}, Anastasios N. Angelopoulos², Stephen Bates³, and Michael I. Jordan¹

> ¹Department of Statistics, UC Berkeley ²Department of EECS, UC Berkeley ³Department of EECS, MIT

> > March 2024

Abstract

Decision-making pipelines are generally characterized by tradeoffs among various risk functions. It is often desirable to manage such tradeoffs in a data-adaptive manner. As we demonstrate, if this is done naively, state-of-the art uncertainty quantification methods can lead to significant violations of putative risk guarantees. To address this issue, we develop methods that permit valid control of risk when threshold and tradeoff parameters are chosen adaptively. Our methodology supports monotone and nearly-monotone risks, but otherwise makes no distributional assumptions. To illustrate the benefits of our approach, we carry out numerical experiments on synthetic data and the large-scale vision dataset MS-COCO.

1 Introduction

In modern machine learning, a focus on complex prediction models and autonomous decision-making is typical, reflecting the engineering focus of its practitioners. However, guaranteeing that the quality of these predictions (or decisions) are within desired tolerances requires good uncertainty quantification (UQ), a classically statistical issue. A burgeoning literature on conformal prediction proposes a solution for this problem, based on treating these complex predictors as unknown black boxes [1].

Many such "black box" conformal methods, as applied in supervised learning, are able to use an auxiliary calibration dataset $\{(X_i, Y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ to produce a prediction set which is a subset of the label space \mathcal{Y} , for a guarantee of the form

$$\mathbb{P}\left[R(\mathcal{C}(X_{n+1}), Y_{n+1}) \le \alpha\right] \ge 1 - \delta,$$

where the risk function $R: 2^{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ measures the quality of a prediction set in containing the true label. For example, in a K-class, multi-label classification setting, the prediction set $\mathcal{C}(X)$ could correspond to possible classes for input X, and the risk may be a *false negative rate*: the proportion of positive classes in the labels Y that are missed by the elements of the prediction set $\mathcal{C}(X)$. In practice, these predictions often constitute the final decision made by an autonomous system, and an appropriate risk measures the consequences of incorrect decisions.

In this work, we attempt to address a major shortcoming of existing UQ methods as described above: they typically assume that the data analyst has selected the tolerance level α in advance of observing any data, and has already determined a risk R to control. Practitioners, on the other hand, often select tolerance levels in a data-dependent way, invalidating the guarantees of UQ methods and hindering their applicability. This holds especially for complex machine learning systems, which are expensive to train and tune.

^{*}Equal contribution. Contact: drew.t.nguyen@berkeley.edu, pathakr@berkeley.edu



Figure 1: Histograms for 20K realizations of the risk gap α – FNR under a data-dependent choice of α , given observations of n calibration points. Each histogram represents different ways to set α . The solid outline has total area $\delta = 0.1$.

Despite this, one may naïvely hope that the guarantees hold approximately well, so as to be useful in an engineering setting. That is, a system builder may wish to proceed without considering the data-dependence in the choice of a tolerance level α , hoping that this will not lead to undesirable violations in practice. But is this actually true? In general, no. In Figure 1, we present the results of an experimental case study, discussed further in Section 1.1, in which we measure the risk gap—the difference between α and the false negative rate (FNR) of a deployed multi-label classifier—when applying methods from the papers [2, 3] with a data-dependent choice of α and with the exceeding α with probability greater than $\delta = 0.1$, which runs counter to the conservative finite-sample guarantees of [2] that hold for fixed α . In fact, the exceedance occurs with empirical probability 0.14 when n = 300 and n = 500. The high-level problem here is clear: selecting α after the observation of data can lead to a notable lack of control.

The present work addresses this problem by providing risk guarantees that account for data-dependent, post hoc choices of α . Figure 1 also depicts the risk gap for a method we propose called *restricted risk resampling* (RRR), which controls the FNR even after α is chosen to optimize a tradeoff. Though the risk gap may seem large,¹ this reflects the flexibility of RRR—it allows the analyst to revise their choice of α in any way based on the calibration data while still retaining a guarantee of low risk. This is particularly useful when data is scarce, as in medical diagnostics, and a classifier is employed at multiple different sensitivity levels. The underlying mathematical tool here is a simultaneity guarantee, and one of our main contributions is to develop methods that possess a general simultaneity property, by establishing new theoretical results regarding the uniform convergence of monotone functions.

Section 2 is the core of this work. It states the key results in the form of uniform confidence bounds, including a functional analog of an inequality of [4]. We adapt these results, as corollaries, into several risk-control procedures, valid for monotone losses and risks: (1) control via a nonasymptotic upper bound, (2) risk resampling, a bootstrap-based procedure which is asymptotically exact, (3) restricted risk resampling, a refinement of risk resampling which optionally ignores large choices of α , and (4) extensions to certain classes of non-monotone functions.

The remainder of the paper supports these core results. Section 3 contains experiments using the results of Section 2. The theoretical underpinnings of Section 2, regarding empirical process theory for monotonically-indexed function classes, are presented in Section 4, with careful proofs deferred to Appendix B. We discuss and conclude in Section 5.



Figure 2: Prediction sets on an example from MS-COCO using the classifier C_t of Section 1.1, as the threshold t is varied. Also plotted is the FNR and FPR averaged over 60K held-out images, which are unobserved in practice; additionally, a rug plot of the scheme $t = \hat{t}$ optimized as in equation (1.1), based on simulation draws of size n = 500.

1.1 Case study: risk tradeoffs and risk control on MS-COCO

We provide a basic description of the experimental setup presented in Figure 1, with complete details deferred to Appendix A.2. The 2014 MS COCO dataset [5] consists of images $X \in \mathcal{X}$ depicting everyday scenes in which any number of K = 80 common objects may be present (e.g., dog, train, chair). The label of each image $X \in \mathcal{X}$ is a vector $Y \in \{0, 1\}^K$, corresponding to the 80 classes which may be present in the image. The task of predicting such a Y is called *multi-label* classification. We now outline how we used this dataset in the experiment presented in Figure 1.

To create a predictor, we trained a neural network that outputs scores, $f(X) \in [0,1]^K$, such that a higher score $[f(X)]_k$ on the *k*th component roughly corresponds to a higher chance that $Y_k = 1$. The final classification is then performed with a threshold classifier, implemented as $C_t : \mathcal{X} \to \{0,1\}^K$ with *k*-th component

$$[\mathcal{C}_t(X)]_k = 1\{[f(X)]_k > (1-t)\}.$$

For a classifier C_t , and a labeled image (X, Y), define the false negative proportion as $\ell(t; (X, Y)) = \#\{k : [C_t(X)]_k = 0\}/\#\{k : [Y]_k = 1\}$, which is the number of false negatives over true positives, using the threshold t. The false positive proportion $q(t, (X, Y)) = \#\{k : [C_t(X)]_k = 1\}/\#\{k : [Y]_k = 0\}$ is likewise the number of false positives over true negatives, using the threshold t.

After training, we observe n additional calibration data points $\{(X_i, Y_i)\}_{i=1}^n$. In this example, we wish to use them to choose the threshold t to *trade off* the false negative rate (FNR) and the false positive rate (FPR), denoted as L(t) and Q(t), respectively:

$$L(t) = \mathbb{E}[\ell(t, (X, Y))], \quad Q(t) = \mathbb{E}[q(t, (X, Y))].$$

In real problems, any scheme used to choose the threshold tends to be based on data visualization and intuitive consideration of the problem domain. For concreteness we make a stylized choice here, one that could be plausibly implemented by a practitioner. Specifically, given the n observations, we take \hat{t} to trade off the empirical risks evenly:

$$\hat{t} = \underset{t \in [0,1]}{\operatorname{argmin}} \sum_{i=1}^{n} \ell(t; (X_i, Y_i)) + q(t; (X_i, Y_i)).$$

As context for this choice of \hat{t} , see Figure 2 for an illustration of the tradeoffs inherent in this problem.

¹When n = 500, the risk gap of RRR is larger than that of the previous method by about 0.01 on average.

We would also like to use the same n calibration data points to *control* these risks. For example, we say the FNR is controlled below α with probability at least $1 - \delta$ if

$$\mathbb{P}\Big(L(\hat{t}) \le \alpha\Big) \ge 1 - \delta.$$

In previous work [2] such control is achieved via the following approach: choose the threshold with a special scheme \hat{t} such that the upper bound of [6] is below α for all $t \geq \hat{t}$. In general, however, such an approach is inadequate, because the control level α can be data-dependent or random. In particular, in the current example our choice of \hat{t} is determined by an estimated risk-reward tradeoff, as is typical in applications.

We now need to plug in some quantity for α satisfying equation (1.1), which simply amounts to using the tightest upper bound available for the random variable $L(\hat{t})$. Here we will consider two alternatives. First, we can ignore the data dependence in α and use the tight bound of [6], which is valid for fixed α . Call this value α_1 . Second, we can apply the bound that comes from the RRR procedure of Section 2.4, specifically by plugging the chosen threshold \hat{t} into the right-hand side of Equation (3)—call this α_2 .

Figure 1 plots the histograms for the random variables $\alpha_1 - L(\hat{t})$ in front and $\alpha_2 - L(\hat{t})$ behind. The lack of control shown by the histogram in front, and the difficulty of formalizing the choice of tradeoffs in practice, motivates our proposal of the RRR method, which is valid even without an explicit scheme such as Equation (1.1).

1.2 Prior work

Our work belongs to the general body of work in distribution-free, frequentist uncertainty quantification, as exemplified by conformal prediction [1, 7]. While our work is in the general vein of conformal prediction, it is also distinct in its focus on general risk functions, as in earlier work on risk-controlling prediction sets [2]. As in that work, we augment the classical framework of conformal prediction via multiple-testing-based arguments and concentration results. The multiple-testing aspect of our work draws on a long tradition of methodology for simultaneous inference in various problem domains, from Scheffé's method for inference on all contrasts in linear regression [8] to more modern problems [9–12]. We also draw from empirical process theory, where uniform versions of concentration results provide tools for simultaneous inference in general settings [13, 14].

Our focus is hypothesis testing. Recall that in the standard Neyman-Pearson paradigm the methodological goal is to maximize power subject to type-I error control at pre-specified level α . It has been noted that this paradigm is unjustified from a decision-theoretic standpoint [15–17]. Instead, in the words of Lehmann and Romano, α should be chosen "in relation to the attainable power" [15]. Attempts to find principled ways to set α in hypothesis testing go back to the 1950s [18]. There are also more direct connections between distribution-free inference and decision-making with more explicit utility maximization [19–21].

There has been recent interest in the control of tradeoffs between multiple risks [22–24], notably the trading off of set size and coverage in conformal prediction [25, 26]. This work generally falls within the Neyman-Pearson framework of constrained optimization and is distinct conceptually from our work.

Uniform bounds have been studied recently in distribution-free novelty detection problems [27, 28]. The bounds obtained in that work focus on empirical cumulative distribution functions, which correspond to binary-valued losses in our context. We also note the work of [29] who share our focus on tradeoffs in distribution-free uncertainty quantification via uniform bounds, but again, their scope is limited to binary losses. Further results on the binary setting can be found in the work of [30–34].

2 Risk Control with Uniform Bounds

In this section, we focus on uniform convergence results for monotone losses, and show how these results can be adapted for risk control. In Section 2.1, we set notation and set up our risk-control problem. We then present two uniform convergence results, the first of which (Section 2.2) is a nonasymptotic concentration inequality, and the second (Section 2.3) a functional central limit theorem. We then present corollaries of these results that demonstrate how they can be used to establish risk control in practice. The proofs of the main results are deferred to Section 4, with technical arguments deferred to Appendix B.

These results are then extended to be more practically useful by relaxing the requirements of uniformity (Section 2.4) and monotonicity (Section 2.5). Proofs for Section 2.4 are provided in Appendix B.1.

2.1 Setting

Consider an input space \mathcal{X} , an output space \mathcal{Y}' , and a label space \mathcal{Y} . Consider also a family of predictors, $\mathcal{C}_t \colon \mathcal{X} \to \mathcal{Y}'$, indexed by a real-valued parameter $t \in [0, 1]$. Furthermore, let $\ell \colon \mathcal{Y}' \times \mathcal{Y} \to [0, 1]$ denote a bounded loss function. We impose the condition that the loss is *monotone* in the parameter t, meaning the following implication holds:

if
$$t \leq s$$
, then $\ell(\mathcal{C}_t(x), y) \geq \ell(\mathcal{C}_s(x), y)$, for all $x \in \mathcal{X}, y \in \mathcal{Y}$.

Finally, suppose we are given a dataset $\mathcal{D}_n := (X_1, Y_1), \ldots, (X_n, Y_n)$, which is drawn i.i.d. from a probability distribution P. Before deployment of the predictor \mathcal{C}_t , the user chooses $t = \hat{t}$ based on this dataset, for example by considering risk tradeoffs. Define the population and empirical risks as

$$L(t) := \mathbb{E}_{(X,Y)\sim P} \left[\ell(\mathcal{C}_t(X), Y) \right] \quad \text{and} \quad \hat{L}_n(t) = \frac{1}{n} \sum_{i=1}^n \ell(\mathcal{C}_t(X_i), Y_i)$$

respectively. Our goal is to compute a bound α such that $\mathbb{P}(L(\hat{t}) \leq \alpha) \geq 1 - \delta$, which means that the risk of $\mathcal{C}_{\hat{t}}$ is controlled below α with probability at least $1 - \delta$.

A wide range of schemes may be used to choose \hat{t} in practice, including schemes that are difficult to formalize, and thus we prefer not to target any specific scheme. Instead, our approach to risk control is based on upper confidence bounds that are *uniform*.

Risk control with uniform bounds. For some range $\mathcal{T} \subset [0,1]$, compute a upper confidence bound $\hat{L}_n^+(t)$ such that

 $L(t) \leq \hat{L}_n^+(t)$ simultaneously for all $t \in \mathcal{T}$,

with probability at least $1 - \delta$. Then for any $\hat{t} \in \mathcal{T}$, the risk of $\mathcal{C}_{\hat{t}}$ is controlled below $\alpha = \hat{L}_n^+(\hat{t})$.

2.2 Finite-sample result for monotone losses

We first present a finite-sample uniform concentration bound based on a novel extension of the Dvoretsky-Kiefer-Wolfowitz (DKW) inequality [35] to monotone losses. To state it, we recall some notation. Associated with the population and empirical risks are the following rescaled and centered processes and associated suprema:

$$\mathbb{G}_n(t) := \sqrt{n}(L_n(t) - L(t)), \quad D_n^+ := \sup_{t} \mathbb{G}_n(t), \quad \text{and} \quad D_n^- := \sup_{t} -\mathbb{G}_n(t).$$

We then have the following nonasymptotic concentration inequality.

Theorem 1. For every $\lambda > 0$, we have

$$\mathbb{P}\{D_n^- > \lambda\} \vee \mathbb{P}\{D_n^+ > \lambda\} \le e \exp(-2\lambda^2).$$

See Section 4.2.1 for a proof of this theorem. It can be viewed as a functional analog of the inequalities of [4], in particular his equation (1.1).

Note that the dependence on λ is optimal, but the prefactor e appearing in Theorem 1 is known to be improvable (to 1) in the special case that ℓ is a indicator function, as shown by [35]. Before proceeding, we record an immediate consequence of Theorem 1: an upper confidence bound (which can be used for risk control), and a lower confidence bound.

Corollary 2.1 (Nonasymptotic confidence bounds). For every sample size $n \ge 1$ and any $\delta \in (0, 1)$ it holds that

$$L(t) \leq \hat{L}_n(t) + \sqrt{\frac{\log(e/\delta)}{2n}}$$
 simultaneously for all t,

with probability at least $1 - \delta$. Similarly, we have

$$L(t) \ge \hat{L}_n(t) - \sqrt{\frac{\log(e/\delta)}{2n}}$$
 simultaneously for all t,

with probability at least $1 - \delta$.

We note that this bound can be quite conservative for practical problems as it does not account for the variance in the process \mathbb{G}_n , which can be substantially smaller than the worst case. In the extreme case of zero variance, meaning that the loss $\ell(C_t(X), Y)$ is a non-random function of t, then $D_n^+ = D_n^- = 0$ deterministically, which Theorem 1 does not account for.

This issue also occurs for bounds based on standard concentration results such as Hoeffding's inequality. In practice, it is often more appealing to use confidence bounds based on an asymptotic normal approximation which do adapt to the variance, as we describe next.

2.3 Asymptotic result for monotone losses

Motivated by the lack of instance-adaptivity of the nonasymptotic concentration inequality, we now present a tighter uniform confidence band. It hinges on the following functional central limit theorem, proven in Section 4, using the monotonicity assumption on the losses.

Theorem 2. The rescaled, centered process \mathbb{G}_n converges in distribution to a centered Gaussian process \mathbb{G} .

The next lemma is technical and is proven in Appendix B.1. It essentially shows that to estimate the process \mathbb{G} , we can sample with replacement from the data—in other words, a bootstrap approach suffices, asymptotically. To state it, we need to introduce some notation relating to the bootstrap.

We use the notation $\mathcal{D}_n^{\star} := \{(X_i^{\star}, Y_i^{\star})\}_{i=1}^n$ to denote *n* samples, drawn with replacement, from the original dataset \mathcal{D}_n . We define the *bootstrap empirical risk* as follows:

$$\hat{L}_n^{\star}(t) := \frac{1}{n} \sum_{i=1}^n \ell(\mathcal{C}_t(X_i^{\star}), Y_i^{\star})$$

and the associated rescaled and centered process as $\mathbb{G}_n^{\star}(t) = \sqrt{n} (\hat{L}_n^{\star}(t) - \hat{L}_n(t))$. Denote $D_{n,\star}^{\pm} = \sup_t \pm \mathbb{G}_n^{\star}(t)$ to be the supremum of this process, where the sign \pm denotes either + or -, and $D_{n,\star} = D_{n,\star}^+ \vee D_{n,\star}^-$.

Lemma 1. If \mathbb{G}_n converges in distribution to a Gaussian process \mathbb{G} , then the conditional distribution of the process $\mathbb{G}_n^{\star} \mid \mathcal{D}_n$ converges to the distribution of \mathbb{G} in probability; also, the conditional distribution of the random variable $D_{n,\star}^{\pm} \mid \mathcal{D}_n$ converges to the distribution of $\sup_t \pm \mathbb{G}(t)$ in probability.

The following corollary is an immediate consequence of Theorem 2 and Lemma 1, and the first result in the corollary, which is an upper confidence bound, can be used directly for risk control. We refer to the overall functional bootstrap procedure as *risk resampling* (RR).

Corollary 2.2 (Confidence bounds via risk resampling). Fix $\delta \in (0,1)$. If \hat{q} satisfies $\mathbb{P}(D_{n,\star}^- > \hat{q} \mid \mathcal{D}_n) \leq \delta$, as $n \to \infty$, we have with probability at least $1 - \delta$,

$$L(t) \leq \hat{L}_n(t) + \frac{\hat{q}}{\sqrt{n}}$$
 simultaneously for all t.

Alternately, if \hat{q} satisfies $\mathbb{P}(D_{n,\star}^+ > \hat{q} \mid \mathcal{D}_n) \leq \delta$, then with probability at least $1 - (\delta + o(1))$, as $n \to \infty$, we have with probability at least $1 - \delta$,

$$L(t) \ge \hat{L}_n(t) - \frac{\hat{q}}{\sqrt{n}}$$
 simultaneously for all t,

and if \hat{q} satisfies $\mathbb{P}(D_{n,\star} > \hat{q} \mid \mathcal{D}_n) \leq \delta$, then with probability at least $1 - (\delta + o(1))$, as $n \to \infty$, we have with probability at least $1 - \delta$,

$$|L(t) - \hat{L}_n(t)| \le \frac{\hat{q}}{\sqrt{n}}$$
 simultaneously for all t.

When the quantity \hat{q} denotes the conditional $1 - \delta$ quantile, $\inf_q \{q : \mathbb{P}(D_{n,\star}^{\pm} > q \mid \mathcal{D}_n) \leq \delta)\}$, it can be computed exactly, but only in principle; this is generally infeasible as it requires enumeration over all possible $\binom{2n-1}{n}$ realizations of the bootstrap dataset \mathcal{D}_n^{\star} . In practice (and also in our experiments, as presented in Section 3) we suggest using Monte Carlo to approximate \hat{q} via a sample quantile \hat{q}_{boot} ; note that this is just the usual, basic bootstrap procedure [36, 37].

As the Monte Carlo error cannot be ignored, it is of interest to ask how many replicates B are needed. At a minimum, B should be chosen large enough that \hat{q}_{boot} is stable, conditional on \mathcal{D}_n . A more principled rule of thumb could be to take B large enough so that $|\hat{q} - \hat{q}_{\text{boot}}| < 0.01 \hat{q}_{\text{boot}}$ with high probability conditional on \mathcal{D}_n , based on a DKW confidence band.²

As shown in Section 3, we find that the resulting bootstrap quantile \hat{q} is much smaller than the factor $\sqrt{\log(e/\delta)}$ guaranteed by the nonasymptotic inequality; indeed, Lemma 1 says it is asymptotically the best possible. Another advantage to the bootstrap procedure from Lemma 1 is that it can automatically extend to a different choice of the index set of the parameter t, namely a subset $\mathcal{T} \subset [0, 1]$. We leverage this in the next section.

2.4 Extension to localized, simultaneous risk control

The bootstrap procedure of Section 2.3, which mimics the data distribution using resamples from the empirical distribution, is flexible in that it allows for certain refinements. In this section, we illustrate one possible refinement which is inspired by [38].

Recalling our setting as outlined in Section 2.1, suppose that before deployment of a predictive algorithm C_t , the user only wishes to choose the parameter t within a subset of [0, 1]. For example, in Section 1.1, where L(t) represents the FNR risk in image classification, the user may wish to restrict to the sublevel set

$$\mathcal{T}_r := \left\{ t \in [0,1] : L(t) \le r \right\}.$$

Setting r = 0.1, say, would encode a belief that a good algorithm cannot have FNR greater than this.

Unsurprisingly, it is wasteful to use the uniform bounds of Section 2.2 and 2.3, which account for coverage violations in all of [0, 1], and not just \mathcal{T}_r which may be significantly smaller. Note that the set \mathcal{T}_r is unknown, so in practice the user can only restrict themselves to a data-dependent set such as

$$\widehat{\mathcal{T}}_r := \left\{ t \in [0,1] : \widehat{L}_n(t) \le r \right\}$$

The approach in this section assumes the user has done this. We give bounds that are valid simultaneously for all $t \in \hat{\mathcal{T}}_r$, rather than all $t \in [0, 1]$. We find empirically that they much tighter than the previous bounds.

To develop this approach, we need to extend the notation of the previous section. Define the bootstrap suprema based on the rescaled and normalized bootstrap process $\mathbb{G}_n^{\star}(t) = \sqrt{n}(\hat{L}_n^{\star}(t) - \hat{L}_n(t))$ on sets $\mathcal{T} \subset [0, 1]$:

$$D_{n,\star}^+(\mathcal{T}) = \sup_{t \in \mathcal{T}} \mathbb{G}_n^{\star}(t), \quad D_{n,\star}^-(\mathcal{T}) = \sup_{t \in \mathcal{T}} -\mathbb{G}_n^{\star}(t), \quad \text{and} \quad D_{n,\star}(\mathcal{T}) = D_{n,\star}^+(\mathcal{T}) \lor D_{n,\star}^-(\mathcal{T}).$$

²Specifically, let $F(x) = \mathbb{P}(D_{n,\star}^{\pm} \leq x \mid \mathcal{D}_n)$. Based on *B* bootstrap replicates, the DKW inequality gives a confidence band $[C^-(x), C^+(x)]$, and if $\hat{q}^{\pm} = \inf_q \{q: 1 - C^{\pm}(q) \leq \delta\}$, then $[\hat{q}^+, \hat{q}^-]$ contains both \hat{q}_{boot} and \hat{q} with high probability; *B* could be chosen until $\hat{q}^- - \hat{q}^+$ seems small.

Our approach begins by fixing a level r and two tolerance parameters $\delta_{\text{glob}}, \delta_{\text{loc}} \in [0, 1]$. We then proceed in three steps:

1. Global estimation: Select a δ_{glob} -bootstrap quantile \hat{q}_{glob} the two-sided supremum risk over [0, 1] satisfying

$$\mathbb{P}\left\{D_{n,\star}([0,1]) > \hat{q}_{\text{glob}} \mid \mathcal{D}_n\right\} \leq \delta_{\text{glob}}.$$

2. Localization: Form an adjusted empirical sublevel set:

$$\widehat{\mathcal{T}}_r^+ := \left\{ t \in [0,1] : \widehat{L}_n(t) \le r + 2\frac{\widehat{q}_{\text{glob}}}{\sqrt{n}} \right\},\$$

containing the original empirical sublevel set $\widehat{\mathcal{T}}_r$.

3. Local estimation: Select a δ_{loc} -bootstrap quantile \hat{q}_{loc} of the one-sided supremum risk over $\hat{\mathcal{T}}_r^+$ satisfying

$$\mathbb{P}\left\{D_{n,\star}^{-}(\widehat{\mathcal{T}}_{r}^{+}) > \hat{q}_{\mathrm{loc}} \mid \mathcal{D}_{n}\right\} \leq \delta_{\mathrm{loc}},$$

and use it to compute an upper confidence band.

The method can be understood intuitively as follows. The first step computes by how much the size of the set $\hat{\mathcal{T}}_r$ should be increased, to obtain the corrected set $\hat{\mathcal{T}}_r^+$ of the second step.³ The third step essentially carries out the bootstrap quantile estimate from the previous section, but specializing to $\hat{\mathcal{T}}_r^+$ for tighter bounds. Due to the correction, the confidence set is valid over the original, smaller set $\hat{\mathcal{T}}_r$.

The initial two-sided global estimation is key. To see why, suppose we have a confidence set that is valid for a fixed set \mathcal{T} when specializing over \mathcal{T} . Then it is valid for any *subset* of \mathcal{T} when specializing over any *superset* of \mathcal{T} , even if these sets are data dependent. Since the two-sided estimation quantifies how far \hat{L}_n is—both above and below—from the unknown mean L, we might plausibly find the fixed set \mathcal{T}_r to be "sandwiched" as $\hat{\mathcal{T}}_r \subset \mathcal{T}_r \subset \hat{\mathcal{T}}_r^+$, so that our confidence set can apply.

The following theorem gives the precise form of a uniform upper confidence bound for risk control. We refer to its computation as *restricted risk resampling* (RRR), but like risk resampling, note that it is a functional form of the bootstrap, though now combined with the localization idea of [38].

Theorem 3 (Confidence bound via restricted risk resampling). Let $r \in [0, 1]$. Fix confidence parameters $\delta_{\text{glob}}, \delta_{\text{loc}} \in [0, 1]$ and set $\delta = \delta_{\text{glob}} + \delta_{\text{loc}}$. Then we have, with probability at least $1 - (\delta + o(1))$,

$$L(t) \leq \hat{L}_n(t) + \frac{\hat{q}_{\text{loc}}}{\sqrt{n}}$$
 simultaneously for all $t \in \hat{\mathcal{T}}_r$.

See Appendix B.1 for a proof of this result.

Let us make a few remarks. First, regarding implementation: we use Monte Carlo approximation to compute the quantiles $\hat{q}_{\rm loc}, \hat{q}_{\rm glob}$, similarly to Section 2.3, and we choose a ratio of $\delta_{\rm loc}/\delta_{\rm glob} = 9$ for the confidence parameters in our experiments.

Second, note that the theorem claims validity over the observed data-dependent set $\hat{\mathcal{T}}_r$, rather than the population set \mathcal{T}_r . The former is arguably more useful in applications, as \mathcal{T}_r is not observed, but a guarantee involving the population set is possible with minor modifications. Specifically, if the procedure is run with the level $r' = r - \hat{q}_{glob}/\sqrt{n}$, then in addition to the conclusion of Theorem 3, we additionally have the inclusion $\hat{\mathcal{T}}_{r'} \subset \mathcal{T}_r$ with high probability, asymptotically.

Finally, regarding motivation: our goal was to spend the error budget less wastefully, when the user prefers parameters t such that $\hat{L}_n(t)$ is small. (By monotonicity, this means that t is large). We note that in

³To appreciate why such a correction is necessary, consider a fixed set $\mathcal{T} \subset [0,1]$, such as $\mathcal{T} = [0,1]$ from Section 2.3. A confidence band $\mathcal{B}(t;\mathcal{T}): [0,1] \to 2^{[0,1]}$ which is uniformly valid over \mathcal{T} —in the sense that $L(t) \in \mathcal{B}(t,\mathcal{T})$ for all $t \in \mathcal{T}$, with high probability—is typically no longer uniformly valid when a random set $\widehat{\mathcal{T}}$ is substituted for \mathcal{T} .

past work, this concern has been addressed differently, using bounds which are valid simultaneously for all t, but which have *variable width*: tighter for large t, looser for small t. The bound in this section, in contrast, is *fixed width*, but can still be tighter for large t, as it need not be valid when t is small.

One seemingly plausible approach to variable-width upper bounds in our setting is inspired by the Monte Carlo method of [27]: for some function $f(t; \gamma)$, corresponding to the width of the bound, which is decreasing in t and increasing in γ , define $\hat{L}_n^+(t) := \hat{L}_n(t) + n^{-1/2} f(t; \hat{\gamma})$, where $\hat{\gamma}$ satisfies $\mathbb{P}(\forall t, \mathbb{G}_n^*(t) \leq \hat{f}(t; \hat{\gamma}) \mid \mathcal{D}_n) \geq 1 - \delta$. Observe that a fixed-width bound corresponds to $f(t, \gamma) = \gamma$.

We do not pursue this further in the present work, but note that the present approach selects a set $\hat{\mathcal{T}}_r$, rather than a function $f(t;\gamma)$. This is advantageous when the risk L(t), and hence the set $\hat{\mathcal{T}}_r$, is more interpretable than the parameter t.

2.5 Extension to non-monotone losses

Our previous concentration inequalities and upper confidence bounds only apply to population risks arising as expectations of monotone losses. In this section, we briefly discuss two extensions that we can accommodate that involve losses which are not monotone.

2.5.1 Combinations of monotone risks

In many situations, the risk that we would like to control can be decomposed into multiple monotone components. Formally, suppose we are interested in controlling the composition of k different risks,

$$L(t) := \Psi(L_1(t), \dots, L_k(t)), \quad \text{where} \quad L_i(t) := \mathbb{E}[\ell_i(\mathcal{C}_t(X), Y)],$$

for i = 1, ..., k. We assume for simplicity that ℓ_i have range in [0, 1] and are monotone in the sense of display (2.1), but the overall function $\Psi \colon [0, 1]^k \to [0, 1]$ may possibly be non-monotone.

General approach: To obtain a upper confidence bound on L(t), we simply combine the uniform lower and upper confidence bounds for each of the components L_i , in the following two steps.

1. Develop a $1 - \delta_{n,i}$ confidence band $\widehat{\mathcal{C}}_i$ for each *i*, such that

 $L_i(t) \in \widehat{\mathcal{C}}_i(t)$ simultaneously for all $t \in \mathcal{T}$,

holds with probability at least $1 - \delta_{n,i}$.

2. Aggregate the confidence parameters and confidence sets by defining

$$\delta = \sum_{i=1}^k \delta_{n,i}, \quad \widehat{C}_{\text{low}}(t) := \inf_{\ell_i \in \widehat{C}_i(t)} \Psi(\ell_1, \dots, \ell_k), \quad \text{and} \quad \widehat{C}_{\text{up}}(t) := \sup_{\ell_i \in \widehat{C}_i(t)} \Psi(\ell_1, \dots, \ell_k).$$

Clearly, we have with probability at least $1 - \delta$ that

 $\widehat{C}_{\text{low}}(t) \leq L(t) \leq \widehat{C}_{\text{up}}(t), \text{ simultaneously for all } t \in \mathcal{T}.$

For the first step in the above approach, we can apply any of our previously described confidence bounds since the component risks $\{L_i\}$ are bounded and monotone.

Illustration for selective classification: Consider the case where $L(t) = L_1(t)/L_2(t)$, which is a special case of the above approach, having taken $L = \Psi(L_1, L_2)$ and $\Psi(\ell_1, \ell_2) = \ell_1/\ell_2$. The two risks can be define to capture a tradeoff, or may arise directly from the specification of L.

This setup arises specifically in *selective classification*, also known as classification with abstention, or classification with a reject option [39–41]. In this setting, we wish to classify covariates x as falling into one of K classes or, alternatively, we can *abstain*, for instance, if we believe we are too uncertain to commit to a point prediction. In this case, let $\mathcal{Y}' = \{1, \ldots, K\} \cup \text{abstain}$. A classifier $\mathcal{C}_t : \mathcal{X} \to \mathcal{Y}'$ can work on top of learned scores $\hat{p}(X) \in \Delta^K$ in the K-simplex as follows. Denoting $k^*(x) = \operatorname{argmax}_k \hat{p}_k(x)$, it returns the highest scoring class unless a score threshold t is not reached:

$$\mathcal{C}_t(X) = \begin{cases} k^*(X) & \hat{p}_{k^*(X)}(X) > t \\ \texttt{abstain} & \text{otherwise.} \end{cases}$$

Then the following risk L, which is the probability of misclassification given that C_t did not abstain, can be upper bounded with high probability by upper bounding the numerator and lower bounding the denominator:

$$L(t) = \mathbb{P}[\mathcal{C}_t(X) \neq Y \mid \mathcal{C}_t(X) \neq \texttt{abstain}] = \frac{\mathbb{P}[\mathcal{C}_t(X) \neq Y, \mathcal{C}_t(X) \neq \texttt{abstain}]}{\mathbb{P}[\mathcal{C}_t(X) \neq \texttt{abstain}]}$$

2.5.2 Nearly monotone risks

Now we consider the case where we have a risk $L(t) = \mathbb{E}[\ell(\mathcal{C}_t(X), Y)]$ which is the expectation of a nonmonotone loss ℓ . The following approach will allow us to provide meaningful risk control when L is "nearly" monotone.

Monotonizing the loss: Our approach is to *monotonize the loss*. Formally, we define the functions

$$\ell^{\downarrow}(\mathcal{C}_t(X), Y) := \inf_{s \le t} \ell(\mathcal{C}_s(X), Y) \text{ and } \ell^{\uparrow}(\mathcal{C}_t(X), Y) := \sup_{s \le t} \ell(\mathcal{C}_s(X), Y)$$

The functions ℓ^{\downarrow} , ℓ^{\uparrow} are essentially the running minimum and maximum, respectively, over the set $\{s \leq t\}$. We define

$$L^{\downarrow}(t) := \mathbb{E}[\ell^{\downarrow}(\mathcal{C}_t(X), Y)] \text{ and } L^{\uparrow}(t) := \mathbb{E}[\ell^{\uparrow}(\mathcal{C}_t(X), Y)].$$

Since $\ell^{\downarrow} \leq \ell \leq \ell^{\uparrow}$, we also have $L^{\downarrow} \leq L \leq L^{\uparrow}$.

If the loss ℓ is bounded, then the functions $L^{\downarrow}, L^{\uparrow}$ satisfy the monotonicity assumptions needed to develop our lower and upper confidence bounds. In particular, we define the monotonized empirical risks,

$$\hat{L}_n^{\downarrow}(t) := \frac{1}{n} \sum_{i=1}^n \ell^{\downarrow}(\mathcal{C}_t(X), Y) \quad \text{and} \quad \hat{L}_n^{\uparrow}(t) := \frac{1}{n} \sum_{i=1}^n \ell^{\uparrow}(\mathcal{C}_t(X), Y).$$

Then, using $\hat{L}_n^{\downarrow}, \hat{L}_n^{\uparrow}$, we can develop, respectively, simultaneous lower and upper confidence bounds on the risk L using the inequalities developed in the previous sections.

Batch-and-monotonize: One concern that we may have with the approach developed above is that even when L is close to monotone, the loss ℓ may be far from monotone, resulting in a larger than desired gap between the population risks L and the monotonized variants $L^{\downarrow}, L^{\uparrow}$. A way to address this is to *batch* the samples, and then monotonize on these individual batches.

Specifically, using k data points at a time can get tighter bounds. Assuming for simplicity that n is divisible by k, we define for $j \in \{0, 1, ..., n/k - 1\}$ the dataset and loss

$$Z_j = \{ (X_{kj+i}, Y_{kj+i}) \}_{i=1}^k \quad \text{and} \quad \ell_k(t, Z_j) = \frac{1}{k} \sum_{i=1}^k \ell(\mathcal{C}_t(X_{kj+i}), Y_{kj+i}).$$

We can then monotonize ℓ_k as described in the previous paragraph, and use this to develop simultaneous lower and upper confidence bounds on the population risk. At first glance this may appear to be lossy because there are only n/k data points Z_j ; however, values of ℓ_k can be expected to have lower variance than ℓ , so variance-aware methods such as Corollary 2.2 and Theorem 3 will adapt.

3 Experiments and Examples

In this section, we will demonstrate the performance of the upper bounds of Section 2: the nonasymptotic bound (NASM),⁴ risk resampling (RR),⁵ and restricted risk resampling (RRR).⁶. Each method was run with confidence parameter $\delta = 0.1$; the RR and RRR methods are run with 1000 bootstrap resamples, and the RRR method was run with risk tolerance r = 0.1 and global and local parameters $\delta_{glob} = 0.01$, $\delta_{loc} = 0.09$.

These methods each amount to different ways to compute a uniform upper bound, which we denote generically as $\hat{L}^+(t)$. We investigate two settings: a fully simulated setting, as well as the MS COCO setting from the Introduction.

Since these bounds are *fixed-width* bounds of the form $\hat{L}^+(t) = \hat{L}_n(t) + \hat{q}/\sqrt{n}$, they must satisfy

$$\mathbb{P}\left(\sup_{t} L(t) - \hat{L}_n(t) \le \frac{\hat{q}}{\sqrt{n}}\right) \ge 1 - \delta.$$

Hence in each setting we provide a quantile plot of \hat{q}/\sqrt{n} against the true $1 - \delta$ quantile of $D_n = \sup_t L(t) - \hat{L}_n(t)$, which constitutes the best possible fixed-width bound.

We also display miscoverage metrics. Call the quantity $\mathbb{P}(\exists t \text{ s.t. } L(t) > \hat{L}^+(t))$ the anywhere miscoverage probability. Additionally, for the selected set $\hat{S} = \{t : \hat{L}_n(t) \leq 0.1\}$, call the quantity $\mathbb{P}(\exists t \in \hat{S} \text{ s.t. } L(t) > \hat{L}^+(t))$ the selected set miscoverage probability; we plot these two quantities in bar charts.

Though the resampling-based bounds are asymptotically exact, their use may seem unreasonable if they are very wide. Hence, on the MS COCO example, we model the behavior of an analyst who trades off two risks. Choosing the parameter \hat{t} as a function of the data, call $\mathbb{E}[\hat{L}^+(\hat{t}) - L(\hat{t})]$ the *average conservatism*; we also plot this in a bar chart. For details on the specific function \hat{t} , see Appendix A.3.

In each setting, we consider multiple different loss functions, and plot the results for each method applied on each loss. In addition, we include a fourth method,⁷, referred to as "pointwise," which is not uniformly valid, but only pointwise valid in the sense that $\mathbb{P}(L(t) \leq \hat{L}^+(t)) \geq 1 - \delta$ for every t and finite n. This method, due to [6], gives remarkably tight estimates of means of bounded random variables, so we display it as a benchmark against our methods which are uniformly valid.

Lastly, we note that these probabilities and expectations cannot be computed exactly on finite data, so in the MS COCO example we must compute surrogates based on splitting our datasets in halves into a holdout set \mathcal{H} and a sampling set \mathcal{S} ; refer to Section A.2 for details. Additionally, the suprema and miscoverage quantities are computed for t in a grid; for the Gaussian example, it is a grid on [-3, 3] with size 1000, and for MS COCO it is a grid on [0, 1] of size 500.

Replication code can be found at github.com/drewtnguyen/risk-tradeoffs-experiments.

3.1 Simulated data

To define a completely synthetic monotone loss function, consider empirical CDFs on batches of data. Let Z_1, \ldots, Z_n be i.i.d., where each Z_i is a batch of five equi-correlated Gaussians, having covariance with diagonal values 1 and off-diagonals $\rho \in [-1, 1]$:

$$Z_i = (X_{i1}, \ldots, X_{i5}) \sim \mathcal{N}_5(0, \rho \mathbb{1} + (1-\rho)\mathbb{I}).$$

Define the loss $\ell(t, Z_i)$ as

$$\ell(t, Z_i) = \sum_{j=1}^{5} 1\{X_{ij} \le t\}$$

⁴Corollary 2.1, right-hand side of Equation (2.1).

⁵Corollary 2.2, right-hand side of Equation (2.2).

⁶Theorem 3, right-hand side of Equation (3).

⁷Theorem A.1 Equation (A.1).



Figure 3: Simulated data: Log-log plot of true quantile of D_n and its median bootstrap estimate (and 90/10% quantiles) computed from 3K Monte Carlo runs.



Figure 4: Simulated data: Anywhere miscoverage based on 20K Monte Carlo runs.

Evidently the risk is just the standard normal CDF $L(t) = \Phi(t)$, even though varying ρ constitutes different loss distributions (we choose $\rho \in \{-0.2, 0.2, 0.6\}$ in the experiments). Note that if $\rho = 0$, uniform upper bounds on L could be obtained by standard arguments for CDFs.

The quantile plot of Figure 3 shows the nonasymptotic upper bound, as well as convergence of the estimated bootstrap quantile (which can be predicted from Lemma 1). The plots show that the nonasymptotic bound is, as predicted, valid for all sample sizes, but is very conservative; RR and RRR work for moderate sample sizes; and the pointwise bound is not uniformly valid.

Note that the convergence of the bootstrap appears to be slower for small ρ , which is when the underlying process is closest to its mean. More generally, constant pre-factors in the convergence rate may depend on the problem setting.

3.2 MS COCO

We revisit multi-label classification on the MS COCO data set that was discussed in the Introduction. The results are presented in Figures 6-9, for four different multi-label classification risks of interest: FNR, FPR, FDR, and SetSize.

The FNR and FPR were defined in the Introduction; the false discovery rate (FDR) is the expectation of the number of false positives over selected classes, while SetSize is the expectation of the normalized number of selected classes. For precise definitions and illustrations of these risks, see Appendix A.3.



Figure 5: Simulated data: Selected set miscoverage based on 20K Monte Carlo runs.



Figure 6: MS COCO data: Log-log plot of true quantile of D_n and its median bootstrap estimate (and 90/10% quantiles) computed from 3K Monte Carlo runs.

We can interpret control of the FDR as a guarantee that the selected classes are mostly true positives on average. Note that it is not a monotone risk, so we monotonize it as described in Section 2.5.

The figures are qualitatively similar to those of the previous section. Again, the nonasymptotic bound is extremely conservative and the pointwise baseline does not have the right coverage, while resampling-based methods are valid at moderate sample sizes and are quite effective. In particular, Figure 9, measuring the average tradeoff conservatism, shows that the RRR method does better than RR in terms of tightness of the bound, and compares favorably to the method that is pointwise valid.

4 Theoretical Underpinnings

This section is a short foray into the empirical process theory that underlies the main results of the paper. First, in Section 4.1 we compute the Vapnik-Chervonenkis (VC) dimension of a class of monotonicallyindexed functions. In Section 4.2, we prove the main results, Theorems 1 and 2, and sketch the proof of the technical Lemma 1 regarding the bootstrap. Careful proofs for all results can be found in Appendix B.

4.1 Monotonically-indexed function classes

Because our results hold even without reference to the risk control problem studied earlier in the paper, we adopt new notation that reflects the underlying empirical process that we are tasked with controlling.



Figure 7: MS COCO: Anywhere miscoverage based on 20K Monte Carlo runs. (Miscoverage for larger n is an artifact of having a finite population—see Appendix A.2).



Figure 8: MS COCO: Selected set miscoverage based on 20K Monte Carlo runs.



Figure 9: MS COCO: Tradeoff conservatism based on 20K Monte Carlo runs.

Let \mathcal{Z} be a probability space on which we define i.i.d. random variables Z_1, \ldots, Z_n . We begin by introducing the following notion of monotonically-indexed function classes.

Monotonically-indexed function class: We say the collection of functions $\mathcal{F} = \{f_t : \mathcal{Z} \to \mathbb{R}\}_{t \in [0,1]^K}$ is *monotonically-indexed* if

$$t \leq s$$
 implies $f_t(z) \leq f_s(z)$, for any $z \in \mathbb{Z}$.

Here, the inequality $t \leq s$ corresponds to the sequence of componentwise inequalities $t_i \leq s_i$ for $i = 1, \ldots, K$. It does not matter what the space \mathcal{Z} is over which the functions f_t are defined, as the monotonicity provides enough structure to restrict the complexity of the class.

A monotonically-indexed class \mathcal{F} should not be confused with a class \mathcal{F} of monotone functions, which is the class such that whenever $f \in \mathcal{F}$, then $z_1 \leq z_2$ implies $f(z_1) \leq f(z_2)$. This latter setting has been studied classically; see, for instance, [42].

Our first result shows that monotonically-indexed function classes have small VC subgraph dimension, as defined in [43].

Proposition 1. If \mathcal{F} is monotonically-indexed, then its VC (subgraph) dimension is at most K + 1.

See Appendix B.1 for a proof of this claim. Our proof is inspired by Lemma 9.10 in [14], which provides a proof in the case K = 1.

Finiteness of the VC dimension allows us to easily derive the following consequence, which is a functional central limit theorem (CLT) for monotonically-indexed classes. To state the result, we consider the following rescaled and centered process:

$$\mathbb{G}_n(t) := \frac{1}{\sqrt{n}} \Big(\sum_{i=1}^n (f_t(Z_i) - \mathbb{E}f_t(Z_i)) \Big).$$

Proposition 2. Let \mathcal{F} denote a monotonically-indexed function class. If the elements of \mathcal{F} are rightcontinuous and uniformly bounded, in the sense that

$$\sup_{f\in\mathcal{F}}\sup_{z\in\mathcal{Z}}|f(z)|<\infty,$$

then the process \mathbb{G}_n converges in distribution to a centered Gaussian process \mathbb{G} with covariance kernel

$$C(t,s) = \mathbb{E}[f_t(Z)f_s(Z)] - \mathbb{E}[f_t(Z)]\mathbb{E}[f_s(Z)].$$

This result is almost an immediate consequence of our Proposition 1, Theorem 19.14, and Lemma 19.15 of [43]; for details, see Appendix B.1.

Comparing general monotone functions to indicators: One interesting interpretation of Proposition 2 arises in the case K = 1. Let \mathcal{F} be a monotonically-indexed function class; let \mathbb{G} and C denote the limit process and kernel associated with \mathbb{G}_n as guaranteed by Proposition 2. Additionally, suppose for simplicity that $F(t) = \mathbb{E}f_t(Z)$ also satisfies F(0) = 0 and F(1) = 1 (the general case can be reduced this case by translation and rescaling). Then, we consider a hypothetical process where we sample Z'_i in an i.i.d. fashion according to the CDF F. We consider the following process:

$$\mathbb{G}'_{n}(t) := \frac{1}{\sqrt{n}} \Big(\sum_{i=1}^{n} (\mathbf{1}\{Z'_{i} \le t\} - F(t)) \Big).$$

By Proposition 2 we have that \mathbb{G}'_n converges to a centered Gaussian process with covariance kernel $C'(t,s) = F(t \wedge s) - F(t)F(s)$. Interestingly, we also see that the limit processes \mathbb{G}, \mathbb{G}' and their covariance kernels C, C' are related via

 $C'(t,s) = C(t,s) + \Delta(t,s), \text{ where } \Delta(t,s) = \mathbb{E}[f(t \wedge s, Z) - f_t(Z)f_s(Z)].$

It is straightforward to check that Δ is a positive semidefinite kernel on $[0,1] \times [0,1]$, and we obtain

$$\mathbb{G}' = \mathbb{G} + \mathbb{W},$$

where \mathbb{W} is a centered Gaussian process on [0, 1] with the covariance Δ . Asymptotically, this relation shows that the empirical process associated with a monotone function is stochastically no larger than that of a corresponding CDF.

4.2 Proofs of main results

We now present the proofs of our main results, Theorems 1 and 2, and sketch the proof of Lemma 1 on validity of the bootstrap.

4.2.1 Proof of Theorem 1

In Appendix B.3, we prove the following result for function classes which are monotonically-indexed by a single parameter, which can be seen as a generalization of the DKW inequality.

Theorem 4. Let $\mathcal{F} = (f_t)_{t \in [0,1]}$ denote a monotonically-indexed function class, with K = 1. Then, the rescaled and centered process

$$\mathbb{G}_n(t) := \frac{1}{\sqrt{n}} \Big(\sum_{i=1}^n (f_t(Z_i) - \mathbb{E}f_t(Z_i)) \Big)$$

satisfies

$$\mathbb{P}\Big\{\sup_{t} \mathbb{G}_{n}(t) > x\Big\} \vee \mathbb{P}\Big\{\sup_{t} -\mathbb{G}_{n}(t) > x\Big\} \le e \exp(-2x^{2}),$$

for all x > 0.

Theorem 1 now follows by taking $\lambda = x$, and identifying

$$Z_i = (X_i, Y_i)$$
 and $f_t(Z_i) = \ell(\mathcal{C}_t(X_i), Y_i)$, for $i = 1, \dots, n$.

Note that the monotonicity assumption in Section 2.1 (decreasing) is the reverse of the one for monotonicallyindexed classes (increasing).

4.2.2 Proof of Theorem 2

We first identify $Z_i = (X_i, Y_i)$ and $f_t(Z_i) = \ell(\mathcal{C}_t(X_i), Y_i)$, as in the proof of Theorem 1. Then Theorem 2 follows from Proposition 2, because for any z, right continuity in t can be assumed without loss of generality by redefining $f_t(z)$ at the discontinuity points, which are countably many due to monotonicity, and uniform boundedness holds because $0 \le f_t(z) \le 1$.

4.2.3 Proof sketch of Lemma 1

let $\mathcal{D}_n = Z_1, \ldots, Z_n$ be i.i.d, and let the centered, rescaled bootstrap distribution be

$$\mathbb{G}_n^{\star}(t) = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n (M_{n,i} - 1) f_t(Z_i) \right)$$

where $M_n \sim \text{Multinomial}(n, 1/n, \dots, 1/n)$.

The functional CLT in Theorem 2 holds if and only if the bootstrap distribution is accurate, in the sense that the conditional law of $\mathbb{G}_n^* \mid \mathcal{D}_n$ converges to the distribution of \mathbb{G} , in probability. The conclusion follows by a version of the continuous mapping theorem that holds for bootstrap distributions. See Appendix B.1 for details.

5 Discussion

We have presented an approach to distribution-free predictive inference that allows post hoc optimization of bounded, monotone risk functions. Unlike most existing work, our bounds are *uniform*, enabling exploratory revision of levels after seeing the data.

An alternative would be to set α ahead of time via data splitting. Neither approach is better than the other in general [44]. But in the machine-learning problems that motivate our work, where uncertainty quantification may be one component of an overall complex engineering system, the property of uniform bounds seems particularly useful. It allows the choice of t to be a complex function of other components of the system.

Our best performing methods, based on the bootstrap, have asymptotic validity but do not have provable finite-sample validity. This is a familiar issue—the inequality of Theorem 1 is akin to a Hoeffding inequality, whereas the bootstrap is akin to a central limit theorem. This begs the question: it possible to derive an analog of a Bernstein inequality—a tight, variance-aware, finite-sample bound? Such a result was recently demonstrated for binary losses [45], and we defer further investigation to figure work.

Another natural question concerns the extension of the present tools to confidence bounds for truly non-monotone risks, rather than near monotone risks or combinations of them. The method of Learn Then Test [3] achieves finite-sample risk control in the binary setting by gridding the space of parameters into ppoints and performing tests at each point to assess whether the risk is below a level α , with multiplicity correction. That work, however, does not estimate the underlying dependencies between the tests, a task at which bootstrap methods succeed asymptotically.

References

- A. N. Angelopoulos and S. Bates. "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification." (Dec. 7, 2022), (visited on 10/06/2023), preprint.
- [2] S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. Jordan, "Distribution-free, Risk-controlling Prediction Sets," *Journal of the ACM*, vol. 68, no. 6, 43:1–43:34, Sep. 30, 2021.
- [3] A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei. "Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control." (Sep. 29, 2022), (visited on 10/06/2023), preprint.
- [4] V. Bentkus, "On Hoeffding's inequalities," The Annals of Probability, vol. 32, no. 2, Apr. 1, 2004.
- [5] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. "Microsoft COCO: Common Objects in Context." (Feb. 20, 2015), (visited on 11/09/2023), preprint.
- [6] I. Waudby-Smith and A. Ramdas, "Estimating means of bounded random variables by betting," Journal of the Royal Statistical Society Series B: Statistical Methodology, qkad009, Feb. 16, 2023.
- [7] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Cham: Springer International Publishing, 2022.
- [8] H. Scheffé, "A method for judging all contrasts in the analysis of variance," *Biometrika*, vol. 40, no. 1-2, pp. 87–110, 1953.
- [9] T. Dickhaus, Simultaneous Statistical Inference: With Applications in the Life Sciences. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.
- [10] J. J. Goeman and A. Solari, "Multiple Testing for Exploratory Research," *Statistical Science*, vol. 26, no. 4, pp. 584–597, Nov. 2011.
- [11] E. Katsevich and A. Ramdas, "Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings," *The Annals of Statistics*, vol. 48, no. 6, Dec. 1, 2020.
- [12] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao, "Valid post-selection inference," The Annals of Statistics, vol. 41, no. 2, pp. 802–837, Apr. 2013.

- [13] G. R. Shorack and J. A. Wellner, Empirical Processes with Applications to Statistics (Classics in Applied Mathematics). Society for Industrial and Applied Mathematics, Jan. 2009, 991 pp.
- [14] M. R. Kosorok, Introduction to Empirical Processes and Semiparametric Inference (Springer Series in Statistics). New York, NY: Springer New York, 2008.
- [15] "Uniformly Most Powerful Tests," in *Testing Statistical Hypotheses*, ser. Springer Texts in Statistics,
 E. L. Lehmann and J. P. Romano, Eds., New York, NY: Springer, 2005, pp. 56–109.
- [16] A. Tetenov, "An economic theory of statistical testing," The IFS, Sep. 27, 2016.
- [17] P. Grünwald. "Beyond Neyman-Pearson." (Feb. 15, 2023), (visited on 10/06/2023), preprint.
- [18] E. L. Lehmann, "Significance Level and Power," The Annals of Mathematical Statistics, vol. 29, no. 4, pp. 1167–1176, Dec. 1958.
- [19] V. Vovk and C. Bendtsen, "Conformal predictive decision making," in Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications, PMLR, Jun. 7, 2018, pp. 52– 62.
- [20] E. Straitouri, L. Wang, N. Okati, and M. G. Rodriguez, "Improving Expert Predictions with Conformal Prediction," in *Proceedings of the 40th International Conference on Machine Learning*, PMLR, Jul. 3, 2023, pp. 32633–32653.
- [21] J. Lekeufack, A. N. Angelopoulos, A. Bajcsy, M. I. Jordan, and J. Malik. "Conformal Decision Theory: Safe Autonomous Decisions from Imperfect Predictions." (Oct. 9, 2023), (visited on 11/08/2023), preprint.
- [22] B. Laufer-Goldshtein, A. Fisch, R. Barzilay, and T. Jaakkola. "Efficiently Controlling Multiple Risks with Pareto Testing." (Oct. 14, 2022), (visited on 11/08/2023), preprint.
- [23] Z. Lin, S. Trivedi, C. Xiao, and J. Sun, "Fast online value-maximizing prediction sets with conformal cost control," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23, vol. 202, Honolulu, Hawaii, USA: JMLR.org, Sep. 30, 2023, pp. 21182–21203.
- [24] J. Teneggi, M. Tivnan, J. W. Stayman, and J. Sulam. "How to Trust Your Diffusion Model: A Convex Optimization Approach to Conformal Risk Control." (Jun. 13, 2023), (visited on 10/06/2023), preprint.
- [25] M. Sadinle, J. Lei, and L. Wasserman, "Least Ambiguous Set-Valued Classifiers with Bounded Error Levels," *Journal of the American Statistical Association*, vol. 114, no. 525, pp. 223–234, Jan. 2, 2019.
- [26] G. S. Dhillon, G. Deligiannidis, and T. Rainforth. "On the Expected Size of Conformal Prediction Sets." (Jun. 12, 2023), (visited on 11/11/2023), preprint.
- [27] S. Bates, E. Candès, L. Lei, Y. Romano, and M. Sesia, "Testing for outliers with conformal p-values," *The Annals of Statistics*, vol. 51, no. 1, pp. 149–178, 2023.
- [28] U. Gazin, G. Blanchard, and E. Roquain, "Transductive conformal inference with adaptive scores," arXiv preprint arXiv:2310.18108, 2023.
- [29] S. Sarkar and A. K. Kuchibhotla. "Post-selection Inference for Conformal Prediction: Trading off Coverage for Precision." (Jun. 30, 2023), (visited on 11/11/2023), preprint.
- [30] V. Vovk. "Conditional validity of inductive conformal predictors." (Sep. 24, 2012), (visited on 11/09/2023), preprint.
- [31] S. Park, S. Li, I. Lee, and O. Bastani, "Pac confidence predictions for deep neural network classifiers," arXiv preprint arXiv:2011.00716, 2020.
- [32] M. Bian and R. F. Barber, "Training-conditional coverage for distribution-free predictive inference," *Electronic Journal of Statistics*, vol. 17, no. 2, pp. 2044–2066, Jan. 2023.
- [33] R. Liang and R. F. Barber, "Algorithmic stability implies training-conditional coverage for distribution-free prediction methods," arXiv preprint arXiv:2311.04295, 2023.
- [34] N. Amann, H. Leeb, and L. Steinberger, "Assumption-lean conditional predictive inference via the jackknife and the jackknife+," arXiv preprint arXiv:2312.14596, 2023.

- [35] P. Massart, "The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality," The Annals of Probability, vol. 18, no. 3, pp. 1269–1283, 1990.
- [36] R. J. Tibshirani and B. Efron, An Introduction to the Bootstrap. New York: Chapman and Hall/CRC, May 14, 1994, 456 pp.
- [37] A. C. Davison and D. V. Hinkley, Bootstrap Methods and Their Application (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge: Cambridge University Press, 1997.
- [38] T. Zrnic and W. Fithian. "Locally Simultaneous Inference." (May 7, 2023), (visited on 11/09/2023), preprint.
- [39] C.-K. Chow, "An optimum character recognition system using decision functions," IRE Transactions on Electronic Computers, no. 4, pp. 247–254, 1957.
- [40] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on information theory*, vol. 16, no. 1, pp. 41–46, 1970.
- [41] C. Cortes, G. DeSalvo, and M. Mohri, "Learning with rejection," in Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27, Springer, 2016, pp. 67–82.
- [42] J. Dehardt, "Generalizations of the Glivenko-Cantelli Theorem," The Annals of Mathematical Statistics, vol. 42, no. 6, pp. 2050–2055, 1971.
- [43] A. W. van der Vaart, Asymptotic statistics (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press, Cambridge, 1998, vol. 3, pp. xvi+443.
- [44] A. K. Kuchibhotla, J. E. Kolassa, and T. A. Kuffner, "Post-Selection Inference," Annual Review of Statistics and Its Application, vol. 9, no. 1, pp. 505–527, 2022.
- [45] D. Bartl and S. Mendelson. "On a variance dependent dvoretzky-kiefer-wolfowitz inequality." (2023).
- [46] T. Ridnik, H. Lawen, A. Noy, E. B. Baruch, G. Sharir, and I. Friedman. "TResNet: High Performance GPU-Dedicated Architecture." (Aug. 27, 2020), (visited on 11/09/2023), preprint.
- [47] V. Bentkus, N. Kalosha, and M. van Zuijlen, "On domination of tail probabilities of (super)martingales: Explicit bounds," *Liet. Mat. Rink.*, vol. 46, no. 1, pp. 3–54, 2006.
- [48] I. Pinelis, "Fractional sums and integrals of r-concave tails and applications to comparison probability inequalities," Advances in stochastic inequalities (Atlanta, GA, 1997), vol. 234, pp. 149–168, 1999.
- [49] D. Panchenko, "Symmetrization approach to concentration inequalities for empirical processes," The Annals of Probability, vol. 31, no. 4, pp. 2068–2081, 2003.
- [50] "Univariate Monotone Convex and Related Orders," in *Stochastic Orders*, ser. Springer Series in Statistics, M. Shaked and J. G. Shanthikumar, Eds., New York, NY: Springer, 2007, pp. 181–232.

A Experimental details

A.1 A concentration inequality

The following bound is drawn from the work of [6]:

$$\hat{L}^+(t) = \inf\left\{p \ge 0 : \max_{i=1,\dots,n} \mathcal{K}_i(p;t) > \frac{1}{\delta}\right\}.$$

where \mathcal{K} is referred to as a capital process in *i*, defined in terms of further quantities:

$$\mathcal{K}_{i}(p;t) = \prod_{j=1}^{i} \left\{ 1 - \lambda_{j}(t)(\ell_{j}(t) - p) \right\}, \text{ where}$$
$$\hat{\mu}_{i}(t) = \frac{1/2 + \sum_{j=1}^{i} \ell_{j}(t)}{1+i}, \ \hat{\sigma}_{i}^{2}(t) = \frac{1/4 + \sum_{j=1}^{i} (\ell_{j}(t) - \hat{\mu}_{j}(t))^{2}}{1+i}, \ \lambda_{i}(t) = \min\left\{ 1, \sqrt{\frac{2\log(1/\delta)}{n\hat{\sigma}_{i-1}^{2}(t)}} \right\}$$

Theorem A.1 (Based on Theorem 3 of [6]). For any parameter t and any finite sample size n,

$$\mathbb{P}(L(t) \le \hat{L}^+(t)) \ge 1 - \delta.$$

As L(t) is simply the mean of a bounded random variable, the same conclusion can hold for simpler upper bounds $\hat{L}^+(t)$, such as the Hoeffding bound, but typically Waudby-Smith and Ramdas' bound seems tighter than any other bound with proven finite-sample validity.

In the present setting, however, this bound is only pointwise valid, not uniform; that is, it is not simultaneously valid for all $t \in [0, 1]$, or any other subset $\mathcal{T} \subset [0, 1]$ which is not a singleton.

A.2 More details for MS COCO

In this section, we provide additional detail regarding how we computed the quantities discussed in Section 1.1 and Section 3 for the MS COCO dataset.

Splitting the MS-COCO dataset. The 2014 MS COCO dataset [5] consists of about 200K labeled images, of which ~120K are designated as either training or validation images. The label of each image $X \in \mathcal{X}$ is a vector $Y \in \{0, 1\}^{80}$, corresponding to which of 80 classes are present in the image. We separated the 120K train/val images into three splits.

Split 1: Training a classifer. Half of the images went to a split for training a TResnet model [46] for 29 epochs, which computes logits given X; we obtained a model of scores $f : \mathcal{X} \to [0, 1]^K$ by converting the raw logits using a sigmoid activation, and a classifier C_t via equation (1.1). The model was cached for every epoch.

Split 2: Choosing the best epoch. A second split of 1K images was used to choose from the epochs the best performing classifier in terms of mean average precision (mAP), namely epoch 5. We obtain a final classifier C_t .

Split 3: Calculating performance metrics. Denote the remaining third split of ~60K images from the train/val set as \mathcal{D}_{COCO} . The MS COCO image/label pairs (X, Y) can be thought of as samples from some joint distribution of MS COCO-type images. But we do not know this distribution and cannot sample from it; in particular, we do not know ground truth risks, such as the FNR, from which to exactly calculate performance metrics such as miscoverage.

(Some notation: Let $\hat{L}(t; \mathcal{D})$ denote an empirical risk calculated using data \mathcal{D} , and similarly let $\hat{L}^+(t; \mathcal{D})$ denote some upper bound. Let \mathcal{D}_n^* denote a sample of size n with replacement from \mathcal{D} , and let true_n be an iid sample of size n from the true distribution of MS COCO images.)

To compute a surrogate quantity, on each simulation we randomly split $\mathcal{D}_{\text{COCO}}$ into two halves, a holdout set \mathcal{H} and a sampling set \mathcal{S} ; we picked n points with replacement \mathcal{S}_n^* ; using these n points, we computed \hat{t} from optimizing a trade-off (see Section A.3), where t takes values on an even grid of 500 points in [0, 1]; and then we evaluate performance metrics by treating $\hat{L}(t; \mathcal{H})$ as ground truth.

For example, we approximate anywhere miscoverage, which is

$$\mathbb{P}\left(L(t) > \hat{L}^+(t; \mathtt{true}_n) \text{ for all } t \in [0, 1]\right),$$

by the surrogate

$$\mathbb{P}\left(\hat{L}_n(t,\mathcal{H}) > \hat{L}^+(t;\mathcal{S}_n^*) \text{ for all } t \in [0,1] \mid \mathcal{D}_{\text{COCO}}\right)$$

so the probability is taken over the split of $\mathcal{D}_{\text{COCO}}$ as well as the sample of size n from \mathcal{S} . The function $\hat{L}_n(\cdot;\mathcal{H})$ serves as a surrogate for the true mean $L(\cdot)$, computed using a holdout set, and the sample of n points \mathcal{S}_n^* serves as a surrogate for sampling n points iid from the true distribution of MS COCO-type images.



Figure 10: Four different risks on the MS COCO distribution, and an empirical estimate of them based on n = 300 data points.

A.3 MS COCO losses and risks

Consider the classifier from Split 2 of Appendix A.2. The following are multi-label classification losses used for the MS COCO dataset:

$$\ell_{\rm FNP}(\mathcal{C}_t(X), Y) = \frac{\sum_{k=1}^{K} 1\{[\mathcal{C}_t(X)]_k = 0, [Y]_k = 1\}}{1 \vee \sum_{k=1}^{K} 1\{[Y]_k = 1\}}$$
$$\ell_{\rm FPP}(\mathcal{C}_t(X), Y) = \frac{\sum_{k=1}^{K} 1\{[\mathcal{C}_t(X)]_k = 1, [Y]_k = 0\}}{1 \vee \sum_{k=1}^{K} 1\{[Y]_k = 0\}}$$
$$\ell_{\rm FDP}(\mathcal{C}_t(X), Y) = \frac{\sum_{k=1}^{K} 1\{[\mathcal{C}_t(X)]_k = 1, [Y]_k = 0\}}{1 \vee \sum_{k=1}^{K} 1\{[\mathcal{C}_t(X)]_k = 1\}}$$
$$\ell_{\rm SetSize}(\mathcal{C}_t(X), Y) = \frac{\sum_{k=1}^{K} 1\{[\mathcal{C}_t(X)]_k = 1\}}{K},$$

and the resulting risks are as follows:

 $FNR(t) = \mathbb{E}[\ell_{FNP}(\mathcal{C}_t(X), Y)]$ $FPR(t) = \mathbb{E}[\ell_{FPP}(\mathcal{C}_t(X), Y)]$ $FDR(t) = \mathbb{E}[\ell_{FDP}(\mathcal{C}_t(X), Y)]$ $SetSize(t) = \mathbb{E}[\ell_{SetSize}(\mathcal{C}_t(X), Y)].$

An illustration of these risks is displayed in Figure 10. Technically, the ground truth plotted in blue is actually computed based on a holdout dataset \mathcal{H} , and the empirical estimate from sampling n = 300 points from a disjoint dataset \mathcal{S} ; see Appendix A.2 for details.

Risk tradeoffs. All the risks are provably monotone, except for FDR, which appears to be nearly monotone. Since FNR is the only one which is decreasing, it makes sense to trade all the others off of FNR.

In particular, let $\hat{t} = \operatorname{argmin}_{t \in \widehat{\mathcal{T}}_r} \varphi(\hat{L}_n(t), \hat{Q}_n(t))$, where \hat{L}_n , \hat{Q}_n represent given empirical risks, and φ aggregates them in some way. The constraint $\widehat{\mathcal{T}}_r = \{t : \hat{L}_n(t) \leq r\}$ models the idea that L_n too large would be intolerable; we set r = 0.1 in the experiments. Usually, this constraint was non-binding.

We changed φ, \hat{L} , and \hat{Q} depending on the experiment. Specifically, Figure 9 computes the tradeoff conservatism $\mathbb{E}[\hat{L}^+(\hat{t}) - L(\hat{t})]$ for different choices of φ, \hat{L} , and \hat{Q} , depending on the risk being controlled:

- FNR control: \hat{L} is empirical FNR, \hat{Q} is empirical FPR, and $\varphi(\ell, q) = \ell + q$.
- FPR control: \hat{L} is empirical FPR, \hat{Q} is empirical FNR, and $\varphi(\ell, q) = \ell + q$.
- FDR control: \hat{L} is empirical FDR, \hat{Q} is empirical FNR, and $\varphi(\ell, q) = -\text{dist}((\ell, q), \texttt{line}((1, 0), (0, 1)))$.
- SetSize control: \hat{L} is empirical SetSize, \hat{Q} is empirical FNR, and $\varphi(\ell, q) = -\text{dist}((\ell, q), \texttt{line}((1, 0), (0, 1)))$.

Here, dist $((\ell, q), \texttt{line}(A, B))$ refers to the distance between the point $(\ell, q) \in \mathbb{R}^2$ and the closest point that lies on the line between A, B. Maximizing this distance finds the "elbow" on an ROC type curve (L(t), Q(t)).

B Proofs

B.1 Proofs of Lemmas and Propositions

To prove Lemma 1 we will draw upon two textbook theorems, written in the notation of Section 4. In particular, let $\mathcal{D}_n = Z_1, \ldots, Z_n$ be i.i.d. with common distribution P, and define

$$\mathbb{G}_n^{\star}(t) = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n (M_{n,i} - 1) f_t(Z_i) \right),$$

where $M_n \sim \text{Multinomial}(n, 1/n, \dots, 1/n)$.

The first is a central limit theorem for bootstrap processes.

Theorem B.1 (Theorem 2.6 (i, ii) from [14]). A function class \mathcal{F} is P-Donsker if and only if

 $\mathbb{G}_n^{\star} \to \mathbb{G}$

in distribution (conditionally on \mathcal{D}_n , in probability), and also \mathbb{G}_n^{\star} is asymptotically measurable.

Next, following the notation of the textbook [14], let the Banach spaces $\mathbb{D} = \ell^{\infty}(\mathcal{F})$ and $\mathbb{E} = \mathbb{R}$ have the uniform norm. We have the following continuous mapping theorem for bootstrap processes.

Theorem B.2 (Theorem 10.8 from [14]). Let $g : \mathbb{D} \to \mathbb{E}$ be continuous, and assume that the map taking $M_n \mapsto h(\mathbb{G}_n^*)$ is measurable for every bounded, continuous $h : \mathbb{D} \to \mathbb{R}$. Then if \mathbb{G}_n^* converges in distribution to a tight process \mathbb{G} (conditionally on \mathcal{D}_n , in probability), then $g(\mathbb{G}_n^*)$ converges in distribution to $g(\mathbb{G})$ (conditionally on \mathcal{D}_n , in probability).

Proof of Lemma 1. First, a minor note. This lemma was stated in the notation of Section 2; it remains true as stated using the notation of Section 4, and it is easy to translate between the two.

Now the fact that \mathbb{G}_n converges in distribution to a Gaussian process \mathbb{G} is the definition of \mathcal{F} being a *P*-Donsker class. Hence we may use Theorem B.1 to claim that \mathbb{G}_n^* converges in distribution to the Gaussian process \mathbb{G} (conditionally on \mathcal{D}_n , in probability). Now we may plug this into Theorem B.2 using $g(F) = \sup_{t \in [0,1]} \pm F(t)$, after the continuity and measurability conditions are checked. But the continuity follows from the continuity of the uniform norm, and the measurability certainly holds, so the conclusion of Theorem B.2 holds, which gives the result. \Box

Proof of Proposition 1. Consider a set of distinct points $P = \{p_1, \ldots, p_{K+1}\} \in \mathbb{Z} \times \mathbb{R}$, and let $S_t = \{(z, s) : f_t(z) < s\}$ denote a subgraph. Essentially, we perform the classical calculation of the VC dimension of half-intervals in \mathbb{R}^K .

Suppose w.l.o.g that for every $k, p_k \in S_t$ for some t, and define $t_{j,k}^* = \inf\{r : \exists t \text{ s.t. } t_j = r, p_k \in S_t\}$, the smallest t_j such that S_t picks out p_k . Define $t_j^{**} = \max_{k \in [K]} t_{j,k}^*$. We now consider two cases.

If there is k' such that $t_{j,k'}^* < t_j^{**}$ for all j, then whenever $P \setminus \{p_{k'}\} \subset S_t$, then $t_j \ge t_j^{**}$ for each j. This implies that $t_j > t_{j,k'}^*$ for all j, so by monotonicity, $P \subset S_t$, so S_t cannot pick out the configuration $P \setminus \{p_{k'}\}$.

If instead for every k, we have $t_{j,k}^* = t_j^{**}$ for some j, then since $j \in \{1, \ldots, K\}$ and $k \in \{1, \ldots, K+1\}$, there must exist k', k'' such that $t_{j,k'}^* = t_{j,k''}^{**} = t_j^{**}$. In this case, the two-point set $\{p_{k'}, p_{k''}\}$ cannot be shattered, because if there is t satisfying $p_{k'} \in S_t$, $p_{k''} \notin S_t$, then

$$t_j^{**} = t_{j,k'}^* \le t_j < t_{j,k''}^* = t_j^{**},$$

which is a contradiction.

Proof of Proposition 2. From Theorem 19.14 and Lemma 19.15 of [43], if a "suitably measurable" function class has finite VC dimension and uniformly bounded, then it is P-Donsker. The definition of P-Donsker is precisely the convergence given in the result.

Suitable measurability is a technical condition that [43] does not define, but he does note that it is sufficient that there is a countable collection of functions \mathcal{G} in which for each $f_t(\cdot)$ we can find a sequence $g^1, g^2, \ldots : \mathcal{Z} \to \mathbb{R}$ satisfying $\lim_{m\to\infty} g^m(z) \to f_t(z)$ at each z. By the right continuity of f_t , it suffices to take \mathcal{G} equal to the subset of \mathcal{F} where t is rational.

B.2 Proof of Theorem 3

The following proof uses the same core idea as that of Theorem 1 of [38], except that we extend their original, nonasymptotic examples to the present asymptotic setting of the bootstrap. Also, we use the notation of Section 2 rather than their general notation, which enables a less abstract proof.

In addition to the notation of Section 2, we set additional notation for the proofs and two lemmas. Let the realized set of selected t be denoted as

$$\widehat{\mathcal{T}}_r = \{t : \widehat{L}_n(t) \le r\},\$$

let the population sublevel set be denoted as

$$\mathcal{T}_r = \{t : \hat{L}_n(t) \le r\},\$$

let the predictive set of selected t, with known mean, be denoted as

$$\mathcal{T}_r^+ = \left\{ t : L(t) \le r + \frac{\hat{q}_{\text{glob}}}{\sqrt{n}} \right\},\,$$

and let the predictive set of selected t, with estimated mean, be denoted as

$$\widehat{\mathcal{T}}_r^+ = \left\{ t : \widehat{L}_n(t) \le r + 2\frac{\widehat{q}_{\text{glob}}}{\sqrt{n}} \right\}.$$

Though we do not directly use this fact, the predictive sets have the property that when an independent copy of \hat{L}_n is observed, and hence an independent copy of the realized set $\hat{\mathcal{T}}_r$, the predictive sets contain the new realized set with high probability.

For each t, let the one-sided and two-sided confidence sets for L(t) be

$$\mathcal{B}_1(t,\mathcal{T}_r^+) = \left\{ y : y \le \hat{L}_n(t) + \frac{\hat{q}_{\text{loc}}(\mathcal{T}_r^+)}{\sqrt{n}} \right\}$$

and

$$\mathcal{B}_2(t, [0, 1]) = \left\{ y : |y - \hat{L}_n(t)| \le \frac{\hat{q}_{\text{glob}}}{\sqrt{n}} \right\}$$

They are confidence bands when interpreted as functions of t.

Finally, let \mathbb{G} refer to the Gaussian process of Theorem 2; it is the process to which the empirical process $\sqrt{n}(\hat{L}_n(t) - L(t))$ is convergent.

Now we establish a lemma on the convergence of bootstrap quantiles.

Lemma 2. If $\hat{q}_{\text{glob}} = \inf_q \{q : \mathbb{P}(D_{n,\star} > q \mid \mathcal{D}_n) \leq \delta_{\text{glob}})\}$ is the conditional $1 - \delta_{\text{glob}}$ quantile of $D_{n,\star}$, and for any \mathcal{T} , $\hat{q}_{\text{loc}}(\mathcal{T}) = \inf_q \{q : \mathbb{P}(D_{n,\star}^-(\mathcal{T}) > q \mid \mathcal{D}_n) \leq \delta_{\text{loc}})\}$ is the conditional $1 - \delta_{\text{loc}}$ quantile of $D_{n,\star}^-(\mathcal{T})$, then

$$\hat{q}_{\mathrm{glob}} \xrightarrow{p} q_{\mathrm{glob}}$$

and

$$\hat{q}_{\mathrm{loc}}(\mathcal{T}_r^+) \xrightarrow{p} q_{\mathrm{loc}}(\mathcal{T}_r),$$

where $q_{\text{glob}} = \inf_{q} \{q : \mathbb{P}(\sup_{t \in [0,1]} |\mathbb{G}(t)| > q) \le \delta_{\text{loc}})\}$ is the limiting global quantile, and $q_{\text{loc}}(\mathcal{T}_r) = \inf_{q} \{q : \mathbb{P}(\sup_{t \in \mathcal{T}_r} \mathbb{G}(t) > q) \le \delta_{\text{loc}})\}$ is the limiting one-sided quantile.

Proof. First, equation (2) holds as an immediate consequence of Lemma 1. Next, note that $D_{n,\star}^-(\mathcal{T}_r^+)$ can be expressed as

$$D_{n,\star}^-(\mathcal{T}_r^+) = \sup_{t \in [0,1]} \mathbb{G}_n^\star(t) \cdot 1\{L(t) \le r + \frac{q_{\text{glob}}}{\sqrt{n}}\}$$

By Lemma 1, the conditional law of \mathbb{G}_n^* converges to that of \mathbb{G} in probability, and the process $1\{L(t) \leq r + \frac{\hat{q}_{\text{glob}}}{\sqrt{n}}\}$ converges to the constant function $1\{L(t) \leq r\} = 1\{t \in \mathcal{T}_r\}$ in probability, so by Slutsky's lemma, the conditional law of

$$\mathbb{G}_n^{\star}(t) \cdot 1\{L(t) \le r + \frac{\hat{q}_{\text{glob}}}{\sqrt{n}}\}$$

converges to the law of

$$\mathbb{G}(t) \cdot 1\{t \in \mathcal{T}_r\}$$

in probability. Finally, by a similar application of Theorem B.2 as in the proof of Lemma 1, the conditional law of $D_{n,\star}^-(\mathcal{T}_r^+)$ converges to that of $\sup_{t\in\mathcal{T}_r} \mathbb{G}(t)$ in probability. This directly implies equation (2).

Second, we establish the key lemma of the proof, which mirrors Lemma 1 of [38].

Lemma 3. For any n, the inequality

$$\mathbb{P}(L(t) \in \mathcal{B}_1(t, \widehat{\mathcal{T}}_r^+) \text{ for all } t \in \widehat{\mathcal{T}}_r) \ge \mathbb{P}\Big(\{L(t) \in \mathcal{B}_1(t, \mathcal{T}_r^+) \text{ for all } t \in \mathcal{T}_r^+\}$$

AND $\{L(t) \in \mathcal{B}_2(t, [0, 1]) \text{ for all } t \in [0, 1]\}\Big)$

holds.

The two events in the right-hand side have been bracketed for clarity. Note that these events depend on the sample size n, but this dependence has been suppressed.

Proof. Because $\mathcal{B}_1(t,\mathcal{T}_1) \subset \mathcal{B}_1(t,\mathcal{T}_2)$ when $\mathcal{T}_1 \subset \mathcal{T}_2$, it is sufficient to show the inclusion

$$\widehat{\mathcal{T}}_r \subset \mathcal{T}_r^+ \subset \widehat{\mathcal{T}}_r^+$$

on the event $L(t) \in \mathcal{B}_2(t, [0, 1])$ for all $t \in [0, 1]$.

On this event, the first inclusion holds because $L(t) \leq \hat{L}_n(t) + \hat{q}_{glob}/\sqrt{n}$ for all $t \in [0, 1]$, giving the chain of implications

$$\hat{L}_n(t) \le r \Rightarrow \hat{L}_n(t) + \hat{q}_{\text{glob}}/\sqrt{n} \le r + \hat{q}_{\text{glob}}/\sqrt{n} \Rightarrow L(t) \le r + \hat{q}_{\text{glob}}/\sqrt{n}.$$

The second inclusion holds because $\hat{L}_n(t) \leq \hat{L}_n(t) + \hat{q}_{glob}/\sqrt{n}$ for all $t \in [0, 1]$, giving the chain of implications

$$L(t) \le r + \hat{q}_{\text{glob}} / \sqrt{n} \Rightarrow L(t) + \hat{q}_{\text{glob}} / \sqrt{n} \le r + 2\hat{q}_{\text{glob}} / \sqrt{n} \Rightarrow \hat{L}_n(t) \le r + 2\hat{q}_{\text{glob}} / \sqrt{n}.$$

Proof of Theorem 3. We wish to show that the event

$$\left\{ L(t) \le \hat{L}_n(t) + \frac{\hat{q}_{\text{loc}}(\widehat{\mathcal{T}}_r^+)}{\sqrt{n}}, \text{ simultaneously for all } t \in \widehat{\mathcal{T}}_r \right\}$$

occurs with probability at least $1 - (\delta + o(1))$. Observe that we may re-express this event as the event

$$\left\{ L(t) \in \mathcal{B}_1(t, \widehat{\mathcal{T}}_r^+), \text{ for all } t \in \widehat{\mathcal{T}}_r \right\}.$$

Applying Lemma 3, we can lower bound the probability of this event in terms of two events with easier-to-handle sets \mathcal{T} :

$$\mathbb{P}(L(t) \in \mathcal{B}_1(t, \widehat{\mathcal{T}}_r^+) \text{ for all } t \in \widehat{\mathcal{T}}_r) \ge \mathbb{P}\Big(\{L(t) \in \mathcal{B}_1(t, \mathcal{T}_r^+) \text{ for all } t \in \mathcal{T}_r^+\}$$

AND $\{L(t) \in \mathcal{B}_2(t, [0, 1]) \text{ for all } t \in [0, 1]\}\Big)$

Rewrite the right-hand side as

$$\mathbb{P}\Big(\{\sqrt{n}(L(t) - \hat{L}_n(t)) \le \hat{q}_{\text{loc}}(\mathcal{T}_r^+) \text{ for all } t \in \mathcal{T}_r^+\} \text{ AND } \{\sqrt{n}|L(t) - \hat{L}_n(t)| \le \hat{q}_{\text{glob}} \text{ for all } t \in [0,1]\}\Big),$$

which can be further re-expressed as

$$\mathbb{P}\Big(\{\sup_{t\in[0,1]} \big(\sqrt{n}(L(t)-\hat{L}_n(t)\big)\cdot 1\{L(t)\leq r+\frac{\hat{q}_{\text{glob}}}{\sqrt{n}}\}\leq \hat{q}_{\text{loc}}(\mathcal{T}_r^+)\} \text{ AND } \{\sup_{t\in[0,1]}\sqrt{n}|L(t)-\hat{L}_n(t)|\leq \hat{q}_{\text{glob}}\}\Big)$$

and by the union bound, this is greater than

$$1 - \mathbb{P}\Big(\sup_{t \in [0,1]} \left(\sqrt{n}(L(t) - \hat{L}_n(t)) \cdot 1\{L(t) \le r + \frac{\hat{q}_{\text{glob}}}{\sqrt{n}}\} > \hat{q}_{\text{loc}}(\mathcal{T}_r^+)\Big) - \mathbb{P}\Big(\sup_{t \in [0,1]} \sqrt{n}|L(t) - \hat{L}_n(t)| > \hat{q}_{\text{glob}}\Big).$$

Now applying the functional central limit theorem (Theorem 2), Lemma 2, and Slutsky's lemma, this equals

$$1 - \mathbb{P}\Big(\sup_{t \in [0,1]} \mathbb{G}(t) \cdot 1\{t \in \mathcal{T}_r\} \le q_{\text{loc}}(\mathcal{T}_r)\Big) - \mathbb{P}\Big(\sup_{t \in [0,1]} |\mathbb{G}(t)| \le q_{\text{glob}}\Big) - o(1),$$

where we have assumed each \hat{q} and q represent exact quantiles as in Lemma 2; this is fine to assume because this quantity lower bounds the case where they are not exact quantiles.

Importantly, \mathcal{T}_r is not a random set, but is deterministic, so because $q_{\text{loc}}(\mathcal{T}_r)$ and q_{glob} are quantiles, definitionally we have that the previous display is lower bounded by

$$1 - \delta_{\text{loc}} - \delta_{\text{glob}} - o(1) = 1 - (\delta + o(1)),$$

as claimed.

B.3 Proof of Theorem 4

To prove this theorem we need two lemmas.

Lemma 4. Let $(Y(t))_{t \in \mathbb{R}}$ be a real-valued stochastic process with bounded sample paths. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex, non-decreasing function. Let $V = \sup_t \mathbb{E}[Y(t) \mid Z]$ and $V^* = \sup_t Y(t)$. Then

$$\mathbb{E}[\varphi(V)] \le \mathbb{E}[\varphi(V^*)].$$

Proof of Lemma 4. First, note that for any function $f : \mathbb{R} \to \mathbb{R}$ with $\sup_t f(t) < \infty$, then if $\varphi : \mathbb{R} \to \mathbb{R}$ is monotone increasing and continuous,

$$\varphi(\sup_{t} f(t)) = \sup_{t} \varphi(f(t)).$$

The direction $\varphi(\sup_t f(t)) \ge \sup_t \varphi(f(t))$ follows by using monotonicity. The direction $\varphi(\sup_t f(t)) \le \sup_t \varphi(f(t))$ follows from continuity: $\varphi(\tilde{f}) \le \sup_t \varphi(f(t))$ for any \tilde{f} in the range of f, so take an increasing sequence $\tilde{f}_1, \tilde{f}_2, \ldots$ converging to $\sup_t f(t)$.

We get the chain of inequalities, noting that φ must be continuous due to convexity:

$$\mathbb{E}[\varphi(\sup_{t} \mathbb{E}[Y(t) \mid Z])] = \mathbb{E}[\sup_{t} \varphi(\mathbb{E}[Y(t) \mid Z])] \leq \mathbb{E}[\sup_{t} \mathbb{E}[\varphi(Y(t)) \mid Z]]$$

$$\leq \mathbb{E}[\mathbb{E}[\sup_{t} \varphi(Y(t)) \mid Z]] = \mathbb{E}[\sup_{t} \varphi(Y(t))] = \mathbb{E}[\varphi(\sup_{t} Y(t))],$$

where the first inequality is Jensen's.

The next lemma concerns what we call the "Bentkus transform," defined in [47]. For any function $S : \mathbb{R} \to \mathbb{R}$, define the log-concave hull S° as the smallest function such that $S \leq S^{\circ}$ and $x \to -\log S^{\circ}(x)$ is a convex function. If S is a survival function, define its Bentkus transform as

$$\mathcal{B}[S](x) = \inf_{r < x} \frac{\mathbb{E}[(X - r)_{+}]}{(x - r)_{+}} = \inf_{r < x} \frac{1}{(x - r)_{+}} \int_{r}^{\infty} S(y) \, dy$$

where $X \sim 1 - S$.

Lemma 5. For a survival function S, for all $x \in \mathbb{R}$,

$$S(x) \stackrel{(i)}{\leq} \mathcal{B}[S](x) \stackrel{(ii)}{\leq} eS^{\circ}(x).$$

Proof. Inequality (i) can be shown by Markov's inequality, applied to the random variable $(X - r)_+$. Inequality (ii) is proved in more generality as Lemma 4.2 by [4], or alternately Lemma 1.1 [47].

[4] attributes this lemma to [48], and a special case to Kemperman, citing Ch. 25 of [13]. It seems to be well-suited for proving extremal results for random variables that are the "least averaged."

For instance, it was used to prove Theorem 1.2 of [4], which showed that binary random variables are, in some sense, more stochastic than variables bounded in [0, 1]. Meanwhile, Ch. 25 of [13] demonstrates a DKW inequality for independent but not identically distributed random variables by showing that the iid case is the most stochastic. Evidently, these results are related to ours.

Proof of Theorem 4. Let $h(t, Z) := f_t(Z)$, and we extend its domain to $t \in \mathbb{R}$ by taking h(t, Z) = 1 whenever t > 1, and h(t, Z) = 0 whenever t < 0; then

$$\sup_{t \in [0,1]} H_n(t,Z) - H(t) = \sup_{t \in \mathbb{R}} H_n(t,Z) - H(t),$$

so from now on we take suprema over \mathbb{R} , and show

$$\mathbb{P}\Big(\sup_{t\in\mathbb{R}} + \big(H_n(t,Z) - \mathbb{E}H_n(t,Z)\big) \ge x\Big) \le e\exp(-2x^2),$$

which implies the result with the + sign (the - sign is exactly similar).

The right-continuity assumption implies that, for each Z, h(t, Z) is a CDF of a random variable supported on [0, 1]; that is, it is non-decreasing, right-continuous, and h(1, Z) = 1. Then, conditionally on each Z_i , let T_i be a random variable drawn according to the CDF $h(\cdot, Z_i)$, and define

$$Y(t) = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} 1\{T_i \le t\} - H(t) \right).$$

Letting $Z = (Z_1, \ldots, Z_n)$, it follows that

$$\mathbb{E}[Y(t) \mid Z] = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} h(t, Z_i) - H(t) \right) = \sqrt{n} (H_n(t) - H(t)).$$

Let $V = \sup_t \mathbb{E}[Y(t) \mid Z]$ and $V^* = \sup_t Y(t)$. Set $\varphi_r : x \mapsto (x - r)_+$, an increasing convex function, and applying Lemma 4, we have for any x > r

$$\frac{\mathbb{E}[\varphi_r(V)]}{\varphi_r(x)} \le \frac{\mathbb{E}[\varphi_r(V^*)]}{\varphi_r(x)}.$$

Let S denote the survival function of V and S^* of V^* . Then taking infimums in r on both sides, we can write an inequality between two Bentkus transforms:

$$\mathcal{B}[S](x) \le \mathcal{B}[S^*](x).$$

Applying Lemma 5 gives

$$S(x) \le e[S^*]^{\circ}(x),$$

and finally, observe that the (one-sided) DKW inequality [35] implies $S^*(x) \leq \exp(-2x^2)$. Since the right-hand side is log-concave, in fact $[S^*]^{\circ}(x) \leq \exp(-2x^2)$. So ultimately

$$S(x) = \mathbb{P}\Big(\sup_{t \in \mathbb{R}} + \big(H_n(t, Z) - \mathbb{E}H_n(t, Z)\big) \le e \exp(-2x^2),$$

as claimed.

Inspecting the argument leading up to equation (B.3), we can extract a fact that may be of independent interest; a weaker version was also stated by [49]. It concerns the *increasing convex ordering* of random variables (see, e.g., [50], Section 4.A).

We write $A \leq_{icx} B$, read as "A is less than B in the increasing convex order," if $\mathbb{E}\varphi(A) \leq \mathbb{E}\varphi(B)$ for all non-decreasing, convex φ . Let S_A, S_B denote their survival functions.

Corollary B.1. Whenever $A \leq_{icx} B$, then $S_A(x) \leq eS_B^{\circ}(x)$.