QUANTILE RISK CONTROL: A FLEXIBLE FRAMEWORK FOR BOUNDING THE PROBABILITY OF HIGH-LOSS PREDICTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

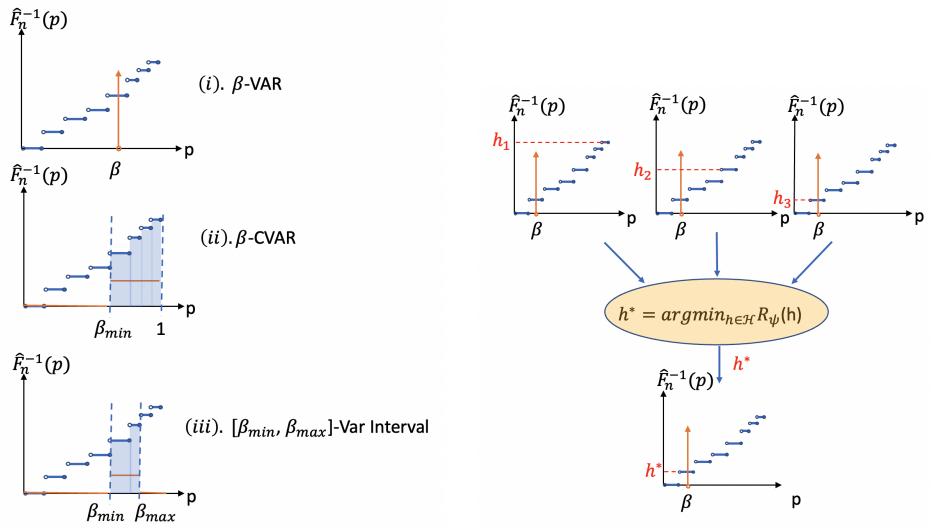
Rigorous guarantees about the performance of predictive algorithms are necessary in order to ensure their responsible use. Previous work has largely focused on bounding the expected loss of a predictor, but this is not sufficient in many risk-sensitive applications where the distribution of errors is important. In this work, we propose a flexible framework to produce a variety of bounds on quantiles of the loss distribution incurred by a predictor. Our method takes advantage of the order statistics of the observed loss values rather than relying on the sample mean alone. We show that a quantile is an informative way of quantifying predictive performance, and that our framework applies to a variety of quantile-based metrics, each targeting important subsets of the data distribution. We analyze the theoretical properties of our proposed method and demonstrate its ability to rigorously control loss quantiles on several real-world datasets.

1 INTRODUCTION

Learning-based predictive algorithms have a great opportunity for impact, particularly in domains such as healthcare, finance and government, where outcomes carry long-lasting individual and societal consequences. Predictive algorithms such as deep neural networks have the potential to automate a plethora of manually intensive tasks, saving vast amounts of time and money. Moreover, when deployed responsibly, there is great potential for a better decision process, by improving the consistency, transparency, and guarantees of the system. As just one example, a recent survey found that a majority of radiologists anticipated that AI-based solutions will lead to fewer medical errors, less time spent on each exam, and more time spent with patients (Waymel et al., 2019). In order to realize such benefits, it is crucial that predictive algorithms are rigorously yet flexibly validated prior to deployment. The validation should be *rigorous* in the sense that it produces bounds that can be trusted with high confidence. It should also be *flexible* in several ways. First, we aim to provide bounds on a variety of loss-related quantities (risk measures): the bound could apply to the mean loss, or the 90th percentile loss, or the average loss of the 20% worst cases. Furthermore, the guarantees should adapt to the difficulty of the instance: easy instances should have strong guarantees, and as the instances become harder, the guarantees weaken to reflect the underlying uncertainty. We also want to go beyond simply bounding the performance of a fixed predictor and instead choose the optimal predictor from a set of candidate hypotheses that minimizes some target risk measure.

Validating the trustworthiness and rigor of a given predictive algorithm is a very challenging task. One major obstacle is that the guarantees output by the validation procedure should hold with respect to any unknown data distribution and across a broad class of predictors including deep neural networks and complicated black-box algorithms. Recent work has built upon distribution-free uncertainty quantification to provide rigorous bounds for a single risk measure, the expected loss (Bates et al., 2021; Angelopoulos et al., 2021). However, to our knowledge there has been no work that unifies distribution-free control of a rich set of expressive risk measures into the same framework.

Our key conceptual advancement is to work with lower confidence bounds on the cumulative distribution function (CDF) of a predictor's loss distribution as a fundamental underlying representation. We demonstrate that a lower bound on the CDF can be converted to an upper bound on the quantile function. This allows our framework to seamlessly provide bounds for any risk measure that can



(a) Bounding multiple QRBMs with $\hat{F}_n^{-1}(p)$.

(b) Using the β-VaR to select the optimal h^*.

Figure 1: An overview of our quantile risk control framework. Given n validation samples $\{X_1, \ldots, X_n\}$ drawn i.i.d. from the loss distribution with CDF F induced by a predictor h, we produce a upper confidence bound $\hat{F}_n^{-1}(p)$ on the true quantile function $F^{-1}(p) \triangleq \inf\{x : F(x) \geq p\}$. Left: The same bound on the quantile function can be used to bound multiple quantile-based risk measures. Right: A bound $\hat{F}_n^{-1}$ is computed for the loss distribution of each predictor $h \in \mathcal{H}$ and the one minimizing a target risk measure $R_\psi(h) = \int_0^1 \psi(p)\hat{F}_n^{-1}(p)\, dp$ is selected.

be expressed as weighted integrals of the quantile function, known as quantile-based risk measures (QBRM) (Dowd & Blake, 2006). QBRMs are a broad class of risk measures that include expected loss, value-at-risk (VaR), conditional value-at-risk (CVaR) (Rockafellar & Uryasev, 2000), and spectral risk measures (Acerbi, 2002). Our approach inverts a nonparametric goodness-of-fit statistic to construct a lower confidence bound on the loss CDF for each candidate predictor. Furthermore, our confidence lower bounds hold simultaneously across an entire candidate set of hypotheses, and thus can be used as the basis for optimization of a target risk measure. For example, our approach can be used to choose the threshold of a deep neural network or a set of thresholds on the scores produced by a complicated black-box prediction algorithm. Figure 1 illustrates an overview of our framework.

We conduct experiments on real-world datasets where the goal is to tune the threshold of a deep neural network with respect to four representative risk measures: the expected loss, VaR, CVaR, and the VaR interval. The expected loss is the mean loss over the test distribution. The β-VaR measures the maximum loss incurred after excluding a $1 - \beta$ proportion of high-loss outliers. The β-CVaR measures the mean loss for the worst $1 - \beta$ proportion of the population. Finally, the VaR-interval can be interpreted as optimizing an uncertain loss quantile that belongs to a known range. We compare various methods for controlling these risk measures, including a novel one that is tailored to the CVaR and VaR-interval settings, and show that our approach bounds the loss CDF well across all scenarios. We also demonstrate how our framework can be used to achieve fairness by equalizing a target risk measure across groups within the population.

Our work unifies the rigorous control of quantile-based risk measures into a simple yet expressive framework grounded in lower confidence bounds on the loss CDF. Unlike previous approaches which only target a single risk measure, our approach provides a flexible and detailed understanding of how a predictive algorithm will perform after deployment. We provide a family of bounds that hold for any quantile-based risk measure, even those not optimized for prior to deployment. Practitioners using our method can thus easily and confidently respond to inquiries from regulators and other key stakeholders, thereby building trust between organizations deploying predictive algorithms and the individuals whose lives are affected by them. This rigorous and flexible validation is an important step towards the responsible use of predictive algorithms in order to benefit society.

2 BACKGROUND & RELATED WORK

Distribution-free uncertainty quantification. The line of work most closely related to ours pertains to distribution-free uncertainty quantification. Such work originated with conformal prediction (Vovk et al., 2005; Shafer & Vovk, 2008), which seeks to produce *coverage* guarantees: *prediction sets* or *prediction intervals* that are guaranteed to contain the target output with high probability. Conformal prediction offers a lightweight algorithm that can be applied on top of complex predictors to obtain provable coverage guarantees, even when the distribution of the training data differs from the validation data. Recent work has generalized conformal prediction to bounding the expected loss with respect to more complex loss functions (Bates et al., 2021; Angelopoulos et al., 2021). Like these works, we also are interested in provable distribution-free guarantees for general loss functions and output spaces. While previous work focuses on bounding the expected loss, part of our contribution lies in bounding the loss for quantiles of the data distribution.

Empirical processes. Several methods consider the relationship between the empirical and true CDF. The Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (Dvoretzky et al., 1956) generates CDF-based confidence bounds and can provide a confidence band. Later variants (Massart, 1990; Naaman, 2021) provide tighter bounds or generalize to wider settings, including hypothesis testing: whether given samples are drawn from a specific CDF F. The confidence bound provided by the DKW inequality is closely related to the Kolmogorov–Smirnov (KS) test (Massey, 1951) – a classic test of goodness of fit. However, the KS test imposes uniform weight on the test range and therefore has low power for detecting deviations at the tail of a distribution. Other variance-weighted tests of goodness of fit address this, e.g., Anderson & Darling (1952); Cramér (1928); Donoho & Jin (2004). Nevertheless, these tests still fail to handle the extreme tails. To our knowledge, none of these techniques have been used as in our work, to produce bounds on quantiles of the loss distribution incurred by a predictor and for model selection.

Risk metrics. A number of approaches have been used to estimate confidence intervals for estimated risk measures. Our focus is on a broad class of quantile-based risk meausres (QBRM's), which includes VaR, CVaR, and spectral risk measures. Value-at-Risk (VaR) has long been used to express the risk associated with financial positions. The value-at-risk (VaR) at the β level (β-VaR) of an investment represents the maximum loss that the investment will incur with probability β. VaR is useful for financial organizations as it represents the amount of capital that must be reserved in order to maintain solvency with the desired probability. Conditional value-at-risk (CVaR) (Rockafellar & Uryasev, 2000) represents the mean loss incurred in the tail of the distribution. The trade-offs between these risk measures are nuanced, with CVaR being easier to optimize and VaR being applicable to a wider class of loss functions (Wang, 2021). The work of Dowd (2010) inspires us; it presents a general approach for estimating confidence intervals for any type of quantile-based risk measure (QBRM). Here we construct confidence bounds on the loss CDF in order to bound such measures for the general problem of predictor selection.

Generalization properties of risk-sensitive loss functions were studied in (Lee et al., 2020) where confidence bounds are given for a large class of OCE risk measures, including CVaR and many of the quantile-based risk measures studied here. In the context of reinforcement learning, Chandak et al. (2021) propose an off-policy method for obtaining confidence bounds of similar risk measures.

3 PRELIMINARIES

In this section, we introduce our problem formulation and briefly discuss previous approaches that rigorously bound the expected loss. We then review quantile-based risk measures (QBRMs), an expressive class of risk measures that includes VaR, CVaR, and spectral risk measures.

3.1 PROBLEM SETUP

We assume the presence of a black-box predictor $h : \mathcal{Z} \to \hat{\mathcal{Y}}$ that maps from an input space $\mathcal{Z}$ to a space of predictions $\hat{\mathcal{Y}}$. We also assume a loss function $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$ that quantifies the quality of a prediction $\hat{Y}$ with respect to the target output Y. Let (Z, Y) be drawn from an unknown data distribution $\mathcal{D}$ over $\mathcal{Z} \times \mathcal{Y}$ and define the random variable $X \triangleq \ell(h(Z), Y)$ to be the loss induced by h on $\mathcal{D}$. Recall that the cumulative distribution function (CDF) of a random variable X is defined

as $F(x) \triangleq P(X \leq x)$. Our goal is to produce rigorous upper bounds on the risk $R(F)$ for a rich class of risk measures $R \in \mathcal{R}$, given a set of validation loss samples $X_{1:n} = \{X_1, \ldots, X_n\}$.

3.2 Distribution-Free Control of Expected Loss

Previous work has investigated distribution-free rigorous control of the expected loss, defined as $R(F) \triangleq \mathbb{E}[X]$, where $X \sim F$. These methods typically involve constructing a confidence region based on the sample mean $\bar{X} \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i$ within which the true expected loss will lie with high probability. For example, suppose that $X \in [0, 1]$ and let $\mu \triangleq \mathbb{E}[X]$. Hoeffding's inequality states that $P(\bar{X} - \mu \leq -\varepsilon) \leq \exp(-2n\varepsilon^2)$. This may be inverted to provide an upper confidence bound on the expected loss: $P(R(F) \leq \hat{R}^+(X_{1:n})) \geq 1 - \delta$, where $\hat{R}^+(X_{1:n}) = \bar{X} + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$. More sophisticated concentration inequalities, such as the Hoeffding-Bentkus inequality (Bates et al., 2021), have also been used to control the expected loss. Methods in this class differ primarily in how they achieve simultaneous control of multiple predictors: RCPS (Bates et al., 2021) does so by assuming monotonicity of the loss with respect to a one-dimensional space of predictors and Learn then Test (LTT) (Angelopoulos et al., 2021) applies a Bonferroni correction to δ in order to achieve family-wise error rate control (Lehmann & Romano, 2008, Sec. 9.1).

3.3 Quantile-based Risk Measures

Unlike previous work that considers only expected loss, we consider a class of risk measures known quantile-based risk measures (QBRM) (Dowd & Blake, 2006). Recall that the quantile function is defined as $F^{-1}(p) \triangleq \inf\{x : F(x) \geq p\}$.

Definition 1 (Quantile-based Risk Measure)**.** *Let $\psi(p)$ be a weighting function such that $\psi(p) \geq 0$ and $\int_0^1 \psi(p)\, dp = 1$. The quantile-based risk measure defined by ψ is*

$$R_\psi(F) \triangleq \int_0^1 \psi(p) F^{-1}(p)\, dp. \tag{1}$$

Several representative QBRMs are shown in Table 1. The expected loss targets the overall expectation across the data distribution, whereas the VaR targets the maximum loss incurred by the majority of the population excluding outliers. CVaR on the other hand targets the worst-off/highest loss tail of the distribution. Finally, the VaR-Interval is valuable when the precise cutoff for outliers is not known. Other instances of QBRMs, such as spectral risk measures (Acerbi, 2002), can easily be handled by our framework, but we do not explicitly consider them here.

Risk Measure	Weighting Function $\psi(p)$
Expected loss	1
β-VaR	$\delta_\beta(p)$
β-CVaR	$\psi(p) = \begin{cases} \frac{1}{1-\beta}, & p \geq \beta \\ 0, & \text{otherwise} \end{cases}$
$[\beta_{\text{min}}, \beta_{\text{max}}]$-VaR-Interval	$\psi(p) = \begin{cases} \frac{1}{\beta_{\text{max}} - \beta_{\text{min}}}, & \beta_{\text{min}} \leq p \leq \beta_{\text{max}} \\ 0, & \text{otherwise} \end{cases}$

Table 1: Several quantile-based risk measures and their corresponding weight functions. The Dirac delta function centered at β is denoted by δ_β.

4 Quantile Risk Control

In this section we introduce our framework *Quantile Risk Control* (QRC) for achieving rigorous and flexible risk control of predictive algorithms. We first describe how bounding the loss CDF incurred by a predictor can be used to form an upper bound on any quantile-based risk measure. We then show how our approach can be used as a basis for optimization on the validation set by choosing the predictor that minimizes a target risk measure.

4.1 CDF LOWER BOUNDS ARE RISK UPPER BOUNDS

We begin by establishing that a lower bound G on the true CDF F can be used to bound $R_\psi(F)$ for any QBRM weighting function. For CDFs F and G, let $F \succeq G$ denote $F(x) \geq G(x)$ for all $x \in \mathbb{R}$.

Theorem 1. *Let F and G be CDFs such that $F \succeq G$. Then $R_\psi(F) \leq R_\psi(G)$ for any weighting function $\psi(p)$ as defined in Definition 1.*

However, as we only have access to a finite sample $X_{1:n}$, it is useful to consider lower confidence bounds (LCB) on the CDF. We call $\hat{G}$ a $(1-\delta)$-CDF-LCB if for any F, $P(F \succeq \hat{G}(X_{1:n})) \geq 1-\delta$, where $X_{1:n} \sim^{iid} F$. We consider the task of producing such lower confidence bounds in Section 5.

Corollary 1. *Suppose that $\hat{G}(X_{1:n})$ is a $(1-\delta)$-CDF-LCB. Then $P(R_\psi(F) \leq R_\psi(\hat{G}(X_{1:n}))) \geq 1-\delta$.*

4.2 SELECTING A PREDICTOR TO MINIMIZE A TARGET RISK MEASURE

We now suppose that we have a finite set of predictors $\mathcal{H}$ defined by a set of functions applied to the output of a complex algorithm such as a deep neural network. Let ϕ be an arbitrary black box function and let $t_1, \ldots, t_m$ be a set of functions, e.g. different thresholding operations. Then the set of predictors is $\mathcal{H} = \{t_1 \circ \phi, \ldots, t_m \circ \phi\}$. The loss r.v. for a random predictor $h \in \mathcal{H}$ is denoted by $X^h \sim F^h$ and the corresponding validation loss samples are $X^h_{1:n}$.

Theorem 2. *Suppose that $\hat{G}$ is a $(1-\delta/|\mathcal{H}|)$-CDF-LCB. Then $P(\forall h \in \mathcal{H} : F^h \succeq \hat{G}(X^h_{1:n})) \geq 1-\delta$.*

Theorem 2 implies that we can simultaneously produce a lower confidence bound on each predictor's loss function with probability at least $1-\delta$, given that $\hat{G}$ is a $(1-\delta/|\mathcal{H}|)$-CDF-LCB. From the set of predictors $\mathcal{H}$, our framework selects h^*_ψ with respect to a target weighting function $\psi(p)$ as

$$h^*_\psi \triangleq \arg\min_{h \in \mathcal{H}} R_\psi(\hat{G}(X^h_{1:n})). \tag{2}$$

By Corollary 1, $P(R_{\psi'}(F^{h^*_\psi}) \leq R_{\psi'}(\hat{G}(X^{h^*_\psi}_{1:n}))) \geq 1-\delta$, including $\psi' \neq \psi$. This means that we are able to bound any quantile-based risk metric, even if it was not targeted during optimization.

5 CONSTRUCTING LOWER CONFIDENCE BOUNDS ON THE LOSS CDF

In this section, we discuss how to produce distribution-free lower confidence bounds on a CDF F given only a finite set of samples drawn i.i.d. from F.

5.1 INVERTING GOODNESS-OF-FIT (GOF) TEST STATISTICS

Our technique centers around producing a set of lower confidence bounds on the uniform order statistics, and using these to bound F. Let $U_1, \ldots, U_n \sim^{iid} \mathrm{U}(0,1)$ and let $U_{(1)} \leq \ldots \leq U_{(n)}$ denote the corresponding order statistics. Consider a one-sided minimum goodness-of-fit (GoF) statistic of the following form:

$$S \triangleq \min_{1 \leq i \leq n} s_i(U_{(i)}), \tag{3}$$

where $s_1, \ldots, s_n : [0,1] \to \mathbb{R}$ are right continuous monotone nondecreasing functions. This statistic can be used to determine a set of constraints on a CDF as follows.

Theorem 3. *Let $s_\alpha = \inf_r\{r : P(S \geq r) \geq 1-\alpha\}$, where $\alpha \in (0,1)$ and S is defined above. Let $X_1, \ldots, X_n \sim^{iid} F$, where F is an arbitrary CDF and let $X_{(1)} \leq \ldots \leq X_{(n)}$ be the corresponding order statistics. Then $P(\forall i : F(X_{(i)}) \geq s_i^{-1}(s_\alpha)) \geq 1-\alpha$, where $\forall s \in \mathbb{R}$, $s_i^{-1}(s) \triangleq \inf_u\{u : s_i(u) \geq s\}$ is the generalized inverse of s_i.*

5.2 CONSERVATIVE CDF COMPLETION

We now show how to use constraints generated by Theorem 3 to form a lower bound on F via the so-called conservative completion of the CDF (Learned-Miller & Thomas, 2020). This defines the "lowest lower bound" of F given constraints.

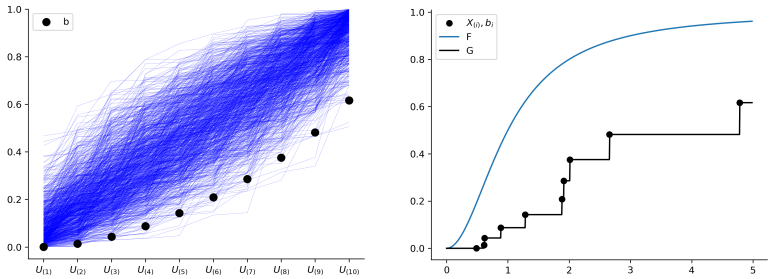


Figure 2: Constructing a distribution-free CDF lower bound by bounding uniform order statistics. Left: Berk-Jones with $\alpha = 0.05$ on uniform order statistics with $n = 10$. Shown in blue are 1000 realizations of $U_{(1)}, \ldots, U_{(n)}$. Right: The corresponding $\hat{G}_\alpha^{BJ}$ constructed using samples from a log-logistic distribution. The true CDF F is shown in blue.

Theorem 4. *Let F be an arbitrary CDF satisfying $F(x_i) \geq b_i$ for all $i \in \{1, \ldots, n\}$, where $x_1 \leq \ldots \leq x_n$ and $0 \leq b_1 \leq \ldots \leq b_n < 1$. Let us denote $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and $\mathbf{b} = (b_1, b_2, \cdots, b_n)$ and let $x^+ \in \mathbb{R} \cup \infty$ be an upper bound on $X \sim F$, i.e. $F(x^+) = 1$. Then $F \succeq G_{(\mathbf{x},\mathbf{b})}$, where $G_{(\mathbf{x},\mathbf{b})}(x)$ is defined as 0 if $x < x_1$; b_i if $x_i \leq x < x_{i+1}$ for $i \in \{1, 2, \cdots, n-1\}$; b_n if $x_n \leq x < x^+$; and 1 if $x^+ \leq x$.*

Corollary 2. *Let S be a test statistic defined as equation 3 and recall $s_\alpha = \inf_r\{r : Pr(S \geq r) \geq 1 - \alpha\}$. Let $X_1, \ldots, X_n \sim^{iid} F$, where F is an arbitrary CDF and let $X_{(1)} \leq \ldots \leq X_{(n)}$ be the corresponding order statistics. Then with probability at least $1 - \alpha$, $F \succeq G_{(X_{(1:n)}, s_{1:n}^{-1}(s_\alpha))}$, where $X_{(1:n)} \triangleq (X_{(1)}, \ldots, X_{(n)})$ and $s_{1:n}^{-1}(s_\alpha) \triangleq (s_1^{-1}(s_\alpha), \ldots, s_n^{-1}(s_\alpha))$.*

Thus we now have a way for constructing a lower confidence bound on F given the order statistics of a finite sample from F that is defined by a set of functions $s_1, \ldots, s_n$. In other words, $\hat{G}(X_{1:n}) \triangleq G_{(X_{(1:n)}, s_{1:n}^{-1}(s_\alpha))}$ is a $(1-\alpha)$-CDF-LCB and can be used to select a predictor to minimize a risk measure as in Section 4.2. The critical value s_α can be computed efficiently by bisection using software packages, e.g. (Moscovich, 2020), that compute the one-sided non-crossing probability. An illustration of the process can be found in Figure 2.

5.3 Relationship to Standard Goodness-of-fit Tests

Several standard tests of uniformity may be viewed as instances of the minimum-type statistic S defined in equation 3. For example, the well-known one-sided Kolmogorov-Smirnov (KS) statistic with the uniform as the null hypothesis can be expressed as

$$D_n^+ \triangleq \max_{1 \leq i \leq n} \left(\frac{i}{n} - U_{(i)} \right) = -\min_{1 \leq i \leq n} \left(U_{(i)} - \frac{i}{n} \right), \tag{4}$$

which can be viewed as $D_n^+ = -S$, where $s_i(u) = u - \frac{i}{n}$. The Berk-Jones (BJ) statistic achieves greater tail sensitivity by taking advantage of the fact that the marginal distribution of $U_{(i)}$ is $\mathrm{Beta}(i, n-i+1)$, which has lower variance in the tails. The one-sided Berk-Jones statistic is defined as

$$M_n^+ \triangleq \min_{1 \leq i \leq n} I_{U_{(i)}}(i, n-i+1), \tag{5}$$

where $I_x(a, b)$ is the CDF of $\mathrm{Beta}(a, b)$ evaluated at x. This can be recognized as S where $s_i(u) = I_u(i, n-i+1)$, the inverse of which can be efficiently computed using standard numerical software packages. We let $\hat{G}_\alpha^{KS}$ and $\hat{G}_\alpha^{BJ}$ denote the $(1-\alpha)$-CDF-LCB defined by applying Corollary 2 to the KS and BJ statistics, respectively.

5.4 Novel Truncated Berk-Jones

In this section we introduce a novel truncated Berk-Jones variant that targets a one-sided range of quantiles $[\beta_{\text{min}}, 1)$. The key idea is to drop lower order statistics that do not affect the bound on $F^{-1}(\beta_{\text{min}})$. We define the truncated one-sided Berk-Jones as $M_{n,k}^+ \triangleq \min_{k \leq i \leq n} I_{U_{(i)}}(i, n-i+1)$,

which can be realized by using $s_i(u) = I_u(i, n-i+1)$ for $k \leq i \leq n$ and $s_i(u) = 1$ otherwise. For a given β_{min}, define $k^*(\beta_{\text{min}}) \triangleq \min\{k : s_k^{-1}(s_\alpha^k) \geq \beta_{\text{min}}\}$, where s_α^k is the critical value of $M_{n,k}^+$. Bisection search can be used to compute k^* and thus the inversion of $M_{n,k^*(\beta_{\text{min}})}^+$ provides a CDF lower bound targeted at quantiles β_{min} and above.

5.5 Relationship to Standard Quantile Upper Confidence Bounds

Standard distribution-free one-sided confidence bounds for quantiles involve selecting the minimal order statistic that bounds the desired quantile level with high probability. The bound discussed in Section G.2.2 of (Meeker et al., 2017) is equivalent to a modified Berk-Jones statistic where only a single order statistic is used. An analogous approach can be taken using the DKW inequality (see Appendix B.1 for details). We call these methods "Order Stats" and "DKW," respectively. They can be adapted to handle a VaR interval by simultaneously bounding multiple quantile levels, e.g. using a grid over the interval $[\beta_{\text{min}}, \beta_{\text{max}}]$, via Bonferroni correction.

6 Experiments

The goal of our experiments is to test the strength of the previously described bounding methods using real world datasets. A good bound will be close to the true CDF and give a low loss guarantee, and will also lead to a favorable actual (i.e. test) loss distribution according to the bounded metric. We work with a variety of loss functions in these experiments, demonstrating the flexibility of our framework. Discussion of our results is in Section 7 and additional results are presented in Appendix B.6.

In all experiments, the guarantees produced by the bounding methods are used to find a threshold on the model output that minimizes one of our risk measures: Expected-value (**Mean**); β-VaR (**VaR**); $[\beta_{\text{min}}, \beta_{\text{max}}]$-VaR-Interval (**Interval**); or **CVaR**. In general, each method is applied on top of a pre-trained model that produces a per-class score, i.e. $W = \phi(Z)$, where $W \in \mathbb{R}^K$ and K is the number of classes. The prediction set is formed by selecting all classes with a score at least some value t: $\hat{Y} = \{k : W_k \geq t\}$. Therefore a predictor is described by its corresponding threshold, and our algorithms choose the predictor that minimizes the target metric according to the bound produced by a given method. Results are measured on the test data using the predictor with the best guarantee for a particular target metric. See Appendix B.7 for more details on the predictor space.

Our comparisons are made using the point-wise methods Learn Then Test-Hoeffding Bentkus (Ltt-HB), DKW, and Order Statistics (Order Stats) and goodness of fit-based methods Kolmogorov-Smirnov (KS), Berk-Jones (BJ), and Truncated Berk-Jones (Truncated-BJ), all as described in previous sections.

6.1 COCO Experiments on all Risk Measures

For our main experiment we compare the performance of various bounding methods across the full canonical set of metrics (Mean, VaR, Interval, and CVaR). We perform this experiment using MS-COCO (Lin et al., 2014), a large object detection dataset with 80 different categories where an image may contain one or more classes of objects and thus may have one or more ground truth labels.

Performance is measured using a balanced accuracy metric, which evenly combines sensitivity and specificity measures, where sensitivity encourages larger prediction set sizes and specificity encourages smaller set sizes (see Appendix B.2 for an exact description). When targeting a VaR-Interval or the CVaR, the loss value is computed as the area under the inverse loss CDF in the range of interest. Mean and VaR are computed as expected. Other experiment details can be found in Appendix B.3.

All results for these experiments are reported in Table 2 with additional plots in Figure 3 (Appendix B.6.1). Table 2 includes the target metric used to choose a predictor, bounding method, guarantee returned on the validation set, loss quantity observed on the test set, and the percent of trials where the guarantee was violated. Berk-Jones gives the tightest bound on the Mean, while the point-wise Order Statistics method outperforms the GoF-based methods on bounding a single VaR metric. When bounding a VaR-Interval or the CVaR, which is the focus of our work, Berk-Jones (especially its truncated version) achieves a lower loss guarantee and actual loss.

In order to highlight the trade-offs inherent in choosing a target metric, we run further experiments on COCO to measure the guaranteed and actual loss given by the Kolmogorov-Smirnov and Berk-Jones bounds when targeting each of the metrics individually. Presentation of these results is left to Table 5 and Figure 4 in Appendix B.6.1. Instead of comparing the bounding methods, we compare the target metrics with the same bounding method. Our results illustrate that achieving a tighter bound in general is as important as choosing the right target risk metric.

Metric	Method	Guarantee	Actual	Violation %
Mean	LttHB	0.097 ± 0.004	0.048 ± 0.001	0.0 ± 0.0
	KS	0.142 ± 0.003	0.048 ± 0.001	0.0 ± 0.0
	BJ	**0.088 ± 0.003**	0.048 ± 0.001	0.0 ± 0.0
VaR	LttHB	0.968 ± 0.036	0.154 ± 0.010	0.0 ± 0.0
	DKW	0.287 ± 0.023	0.150 ± 0.020	0.0 ± 0.0
	OrderStats	**0.166 ± 0.010**	**0.132 ± 0.004**	0.003 ± 0.055
	KS	0.287 ± 0.023	0.150 ± 0.020	0.0 ± 0.0
	BJ	0.179 ± 0.010	0.134 ± 0.006	0.001 ± 0.032
Interval	DKW	0.744 ± 0.005	0.146 ± 0.012	0.0 ± 0.0
	OrderStats	0.189 ± 0.011	0.138 ± 0.007	0.001 ± 0.032
	KS	0.581 ± 0.008	0.144 ± 0.013	0.0 ± 0.0
	BJ	0.189 ± 0.011	0.138 ± 0.007	0.001 ± 0.032
	Truncated-BJ	**0.183 ± 0.010**	**0.137 ± 0.006**	0.001 ± 0.032
CVaR	DKW	1.0 ± 0.0	0.500 ± 0.000	0.0 ± 0.0
	OrderStats	0.454 ± 0.015	0.208 ± 0.011	0.0 ± 0.0
	KS	0.971 ± 0.002	0.228 ± 0.022	0.0 ± 0.0
	BJ	0.453 ± 0.015	0.208 ± 0.011	0.0 ± 0.0
	Truncated-BJ	**0.419 ± 0.015**	**0.207 ± 0.010**	0.0 ± 0.0

Table 2: Results for experiments with COCO Object Detection and all canonical metrics. Best results in bold, except for actual mean where all methods produced the same loss.

6.2 Exploring VaR Interval Bounds in Other Datasets

We performed additional experiments on other datasets—multi-modal and natural language processing classification tasks—to study interval-based quantile bounds derived by the various point-wise and uniform methods. We use a multi-modal CLIP Radford et al. (2021) embedding model in order to perform zero-shot image classification on the CIFAR-100 dataset using the method described in Radford et al. (2021). We furthermore analyze the performance of these methods in a natural language processing setting using the Go Emotions dataset Demszky et al. (2020). Details for these experiments are presented in Appendix B.5. Results for these experiments are reported in Table 3 and Appendix B.6.2. Truncated Berk Jones gives the best loss guarantees, although all bounds achieve similar test loss.

6.3 Fairness Experiment with UCI Nursery Dataset

It has been noted that one shortcoming of distribution-free guarantees for controlling expected loss is that they apply only marginally across the data distribution (Angelopoulos et al., 2021). However, work in algorithmic fairness has aimed at achieving coverage guarantees across groups (Romano et al., 2019; 2020), while Williamson & Menon (2019) proposed CVaR as a notion of fairness that can be viewed as the worst-case mean loss for subgroups of a particular size. However, our approach goes a step further than these. First and most importantly, our framework allows for bounding any interval of the VaR, not just a single point or a range that always extends to 1.0. Next, it is applicable to a broader range of loss functions than handled by previous work.

For our final experiment we consider a fairness problem, and show how our framework can be applied to generate a predictor that optimizes CVaR, or worst-case outcomes, individually with respect to each group according to some protected attribute. We employ the UCI Nursery dataset

Dataset	Method	Guarantee	Actual	Violation %
CiFAR-100	DKW	0.315 ± 0.016	0.142 ± 0.012	0.0 ± 0.0
	OrderStats	0.228 ± 0.019	0.143 ± 0.012	0.0 ± 0.0
	KS	0.254 ± 0.016	**0.140 ± 0.011**	0.0 ± 0.0
	BJ	0.225 ± 0.019	0.143 ± 0.012	0.0 ± 0.0
	Truncated-BJ	**0.220 ± 0.019**	0.142 ± 0.011	0.0 ± 0.0
Go Emotions	DKW	0.252 ± 0.013	0.119 ± 0.014	0.0 ± 0.0
	OrderStats	0.146 ± 0.012	0.111 ± 0.010	0.0 ± 0.0
	KS	0.191 ± 0.013	0.116 ± 0.014	0.0 ± 0.0
	BJ	0.144 ± 0.012	0.110 ± 0.010	0.0 ± 0.0
	Truncated-BJ	**0.140 ± 0.012**	**0.108 ± 0.009**	0.0 ± 0.0

Table 3: Results for zero-shot image classification and textual emotion recognition.

(Dua & Graff, 2017), a multi-class classification problem where categorical features about families are used to assign a predicted rank to an applicant $Y \in \{1, ..., 5\}$ ranging from "not recommended" ($Y = 1$) to "special priority" ($Y = 5$). We measure a class-dependent weighted accuracy loss. Exact specifications of the loss function, dataset, and other experiments details can be found in Appendix B.4. Results for the fair classification experiment are reported in Table 4 (Appendix B.4) and Figure 6 (Appendix B.6.2). Applying the Truncated Berk-Jones bound individually to each group leads to the best CVaR among all methods for both groups.

7 Discussion

Our experimental results (see Section 6 and Appendix B.6.1 and B.6.2) show that in order to bound a VaR-Interval or the CVaR, GoF-based methods, especially Berk-Jones and its Truncated variant, give tighter bounds than running point-wise methods such as LTT or RCPS over a grid of β values. On the other hand, point-wise bounds are tighter for targeted weight functions like that used to calculate β-VaR.

While methods that focus on a single VaR value can give a tighter bound when the target metric is known beforehand, a key advantage of Quantile Risk Control comes from the flexibility to change the target metric post hoc, within some range, and still maintain a valid bound. Note that there are some ways to convert between these classes (e.g., Mean to Quantile with Markov's Inequality; CVaR is directly an upper bound on VaR), but these are weak as shown by our experiments.

Table 5 and Figure 4 in Appendix B.6.1 show results for using different target metrics (Mean, VaR, Interval, and CVaR) to produce guarantees and test loss distributions on all four metrics. The bounds produced by Berk-Jones give low loss guarantees and produce a favorable loss distribution even when the target metric in deployment does not match that used to select the predictor. This makes QRC well-suited to scenarios in which a stakeholder may want to issue a guarantee or several guarantees from a single predictor.

A natural extension of this work is to consider returning multiple predictors with our algorithm. In this case, while the target metric may not be known beforehand, we could produce multiple predictors that optimally bound specific points in the range, instead of a single predictor that gives the best bound on average over the range.

8 Conclusion

In this work we propose Quantile Risk Control, a flexible framework for producing a range of bounds on quantiles of a predictor's loss distribution. In general, our proposed framework allows an organization to tailor a bound to its specific loss function, and to the particular quantile of the data distribution. It can then produce a rigorous bound using one of the methods, including the novel and effective truncated variant introduced here. Our method also opens the door for optimizing functionals of the loss distribution by working with the CDF of the loss distribution itself. Further work could also extend fairness applications of our framework, for instance bounding the difference in loss guarantees across groups.

ETHICS STATEMENT

Some care needs to be taken to ensure that the guarantees under our framework are clear, in that they do not apply to every individual in the test distribution and only hold if test examples are drawn from the same distribution as validation examples.

REPRODUCIBILITY STATEMENT

All of our code will be available to the public on github. For all experiments, datasets and splits are specified, and we use publicly available pre-trained models for feature extraction when possible.

REFERENCES

Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, July 2002. ISSN 03784266. doi: 10.1016/S0378-4266(02)00281-9. URL https://linkinghub.elsevier.com/retrieve/pii/S0378426602002819.

Theodore W Anderson and Donald A Darling. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics*, pp. 193–212, 1952.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control. *arXiv:2110.01052 [cs, stat]*, November 2021. URL http://arxiv.org/abs/2110.01052. arXiv: 2110.01052.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-Free, Risk-Controlling Prediction Sets. *arXiv:2101.02703 [cs, stat]*, August 2021. URL http://arxiv.org/abs/2101.02703. arXiv: 2101.02703.

Yash Chandak, Scott Niekum, Bruno da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S. Thomas. Universal Off-Policy Evaluation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27475–27490. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/e71e5cd119bbc5797164fb0cd7fd94a4-Abstract.html.

Harald Cramér. On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74, 1928.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions, 2020. URL https://arxiv.org/abs/2005.00547.

David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

Kevin Dowd. Using Order Statistics to Estimate Confidence Intervals for Quantile-Based Risk Measures. *Journal of Derivatives*, 17(3):9–14,4, 2010. ISSN 10741240. URL https://www.proquest.com/docview/220481200/abstract/B5705588A9544594PQ/1. Num Pages: 7 Place: New York, United Kingdom Publisher: Pageant Media.

Kevin Dowd and David Blake. After VaR: The Theory, Estimation, and Insurance Applications of Quantile-Based Risk Measures. *Journal of Risk <html_ent glyph="@amp;" ascii="&"/> Insurance*, 73(2):193–229, June 2006. ISSN 0022-4367, 1539-6975. doi: 10.1111/j.1539-6975.2006.00171.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1539-6975.2006.00171.x.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pp. 642–669, 1956.

Erik Learned-Miller and Philip S. Thomas. A New Confidence Interval for the Mean of a Bounded Random Variable. Technical Report arXiv:1905.06208, arXiv, November 2020. URL http://arxiv.org/abs/1905.06208. arXiv:1905.06208 [cs, math, stat] type: article.

Jaeho Lee, Sejun Park, and Jinwoo Shin. Learning bounds for risk-sensitive learning, 2020. URL https://arxiv.org/abs/2006.08138.

Erich L Lehmann and Joseph P Romano. *Testing Statistical Hypotheses*. Springer texts in statistics. Springer New York, New York, NY, 2008. ISBN 978-1-4419-3178-8.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2014. URL https://arxiv.org/abs/1405.0312.

P. Massart. The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. ISSN 00911798.

Frank J. Massey, Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951. doi: 10.1080/01621459.1951.10500769. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1951.10500769.

William Q. Meeker, Gerald J. Hahn, and Luis A. Escobar. *Statistical intervals : a guide for practitioners and researchers*. Wiley Series in Probability and Statistics. Wiley, Hoboken, New Jersey, 2017. ISBN 1-118-59495-9.

Amit Moscovich. Fast calculation of p-values for one-sided Kolmogorov-Smirnov type statistics. Technical Report arXiv:2009.04954, arXiv, September 2020. URL http://arxiv.org/abs/2009.04954. arXiv:2009.04954 [stat] type: article.

Michael Naaman. On the tight constant in the multivariate dvoretzky–kiefer–wolfowitz inequality. *Statistics & Probability Letters*, 173:109088, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture, 2020. URL https://arxiv.org/abs/2003.13630.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000. ISSN 14651211. doi: 10.21314/JOR.2000.038. URL http://www.risk.net/journal-of-risk/technical-paper/2161159/optimization-conditional-value-risk.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel J. Candès. With Malice Towards None: Assessing Uncertainty via Equalized Coverage. *arXiv:1908.05428 [cs, stat]*, August 2019.

Yaniv Romano, Stephen Bates, and Emmanuel J. Candès. Achieving Equalized Odds by Resampling Sensitive Attributes. *arXiv:2006.04292 [cs, stat]*, June 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008.

Galen R. Shorack and Jon A. Wellner. *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics, January 2009. ISBN 978-0-89871-684-9 978-0-89871-901-7. doi: 10.1137/1.9780898719017. URL http://epubs.siam.org/doi/book/10.1137/1.9780898719017.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Ruodu Wang. A tale of two risk measures: Contrasting Value-at-Risk and Expected Shortfall, August 2021. URL https://sas.uwaterloo.ca/~wang/talk/2021ARC.pdf.

Q. Waymel, S. Badr, X. Demondion, A. Cotten, and T. Jacques. Impact of the rise of artificial intelligence in radiology: What do radiologists think? *Diagnostic and Interventional Imaging*, 100(6):327–336, June 2019. ISSN 22115684. doi: 10.1016/j.diii.2019.03.015. URL https://linkinghub.elsevier.com/retrieve/pii/S2211568419300907.

Robert Williamson and Aditya Menon. Fairness risk measures. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6786–6797. PMLR, June 2019.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2019. URL https://arxiv.org/abs/1910.03771.

A PROOFS

A.1 PROOF OF THEOREM 1

Proof. Consider $G^{-1}(p)$. By the definition of the quantile function, $G(G^{-1}(p)) \geq p$. By the relationship between F and G, we therefore have $F(G^{-1}(p)) \geq G(G^{-1}(p)) \geq p$. Applying F^{-1} to both sides yields $F^{-1}(F(G^{-1}(p))) \geq F^{-1}(p)$. But $x \geq F^{-1} \circ F(x)$ (see e.g. Proposition 3 on p. 6 of (Shorack & Wellner, 2009)) and thus $G^{-1}(p) \geq F^{-1}(p)$. Then we have $R_\psi(F) \leq R_\psi(G)$ by Definition 1. □

A.2 PROOF OF COROLLARY 1

Proof. The event that $F \succeq \hat{G}(X_{1:n})$ implies that $R_\psi(F) \leq R_\psi(G)$ by Theorem 1. Since the antecedent holds with probability at least $1 - \delta$, so must the consequent. □

A.3 PROOF OF THEOREM 2

Proof. The claim follows by applying a union bound argument over $h \in \mathcal{H}$:

$$P(\exists h \in \mathcal{H} : F^h \not\succeq \hat{G}(X^h_{1:n})) \leq \sum_{h \in \mathcal{H}} P(F^h \not\succeq \hat{G}(X^h_{1:n}))) \leq \delta.$$

□

A.4 PROOF OF THEOREM 3

Proof. Given the definition of s_α and s_i, if $S \geq s_\alpha$, then we have that $U_{(i)} \geq s_i^{-1}(s_\alpha)$ for all $i \in \{1, \ldots, n\}$. By Proposition 1, eq. 24 on p.5 of (Shorack & Wellner, 2009), we have that $F \circ F^{-1}(t) \geq t$ for any $t \in [0, 1]$ and any CDF F. Therefore we have $F(F^{-1}(U_{(i)})) \geq U_{(i)} \geq s_i^{-1}(s_\alpha)$. In addition, notice X_i is distributed as $F^{-1}(U_i)$ (via the inverse transform). Moreover, F^{-1} preserves ordering (by definition of the quantile function), i.e. $F^{-1}(U_{(1)}) \leq \ldots \leq F^{-1}(U_{(n)})$. Thus, we have $X(i)$ is of the same distribution as $F^{-1}(U_{(i)})$, and we have $F(X_{(i)}) \geq s_i^{-1}(s_\alpha)$ for all i. □

A.5 PROOF OF THEOREM 4

Proof. When $x < x_1$, $F(x) \geq 0$. For any $i \in \{1, 2, \cdots, n\}$, since $F(x_i) \geq b_i$, thus, for $x \in [x_i, x_{i+1})$, $F(x) \geq F(x_i) \geq b_i$ (x_{n+1} is considered as x^+ here). Lastly, if $x \geq x^+$, $F(x) = 1$. Combined together, we have $F(x) \geq G_{(\mathbf{x},\mathbf{b})}(x)$. □

A.6 PROOF OF COROLLARY 2

Proof. Notice that by Theorem 4, if $\forall i, F(X_{(i)}) \geq s_i^{-1}(s_\alpha)$, then we have $F(x) \geq G_{(X_{(1:n)} s_{1:n}^{-1}(s_\alpha))}$, which implies

$$P\Big(F(x) \geq G_{(X_{(1:n)}, s_{1:n}^{-1}(s_\alpha))}(x)\Big) \geq Pr\Big(\forall i, F(X_{(i)}) \geq s_i^{-1}(s_\alpha)\Big).$$

Then, by Theorem 3, $P\Big(\forall i, F(X_{(i)}) \geq s_i^{-1}(s_\alpha)\Big) \geq 1 - \alpha$. □

B EXPERIMENTAL DETAILS

B.1 DKW METHOD FOR QUANTILE DISTRIBUTION-FREE UPPER CONFIDENCE BOUNDS

A distribution upper confidence bound on $F^{-1}(\beta)$ can be constructed by using a one-sided version of the DKW inequality (Dvoretzky et al., 1956; Massart, 1990). Define the empirical distribution function as $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \leq x\}$. Then for every $\varepsilon \geq \sqrt{\frac{1}{2n}\log 2}$, by the DKW inequality,

$$P\left\{\sup_{x\in\mathbb{R}} F_n(x) - F(x) > \varepsilon\right\} \leq \exp(-2n\varepsilon^2).$$

If we let $\delta = \exp(-2n\varepsilon^2)$, then we have $\varepsilon = \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$ and the above condition holds for any $\delta \leq 1/2$.

Proposition 1. *Let $X_{(1)} \leq \ldots \leq X_{(n)}$ denote the order statistics of $X_1, \ldots, X_n$, where $X_i \sim F$ for all $i \in \{1, \ldots, n\}$. Let $k = \left\lceil n\left(\beta + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}\right)\right\rceil$. Then for $\beta + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}} \leq 1$ and $\delta \leq 1/2$,*

$$P(F^{-1}(\beta) \leq X_{(k)}) \geq 1 - \delta.$$

Proof. Observe that $F_n(X_{(i)}) \geq \frac{i}{n}$ for $i \in \{1, \ldots, n\}$. Thus for k defined above, we have

$$F_n(X_{(k)}) \geq \beta + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}.$$

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

$$F(x) \geq F_n(x) - \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

for all $x \in \mathbb{R}$. Thus $P(F(X_{(k)}) \geq \beta) \geq 1 - \delta \Rightarrow P(F^{-1}(\beta) \leq X_{(k)}) \geq 1 - \delta$, where the latter follows from the property that $F^{-1} \circ F(x) \leq x$ (see e.g. Proposition 3 on p. 6 of (Shorack & Wellner, 2009)). □

B.2 DEFINITIONS OF LOSS FUNCTIONS

Balanced accuracy is computed as follows, where $\hat{Y}$ is the prediction set, Y is the set of ground truth labels (which in this case will always be a single label), and K is the number of classes:

$$L(\hat{Y}, Y) = 1 - \frac{1}{2}(\text{Sens}(\hat{Y}, Y) + \text{Spec}(\hat{Y}, Y)), \text{ where}$$

$$\text{Sens}(\hat{Y}, Y) = \frac{|\hat{Y} \cap Y|}{|Y|} \text{ and } \text{Spec}(\hat{Y}, Y) = \frac{K - |Y| - |\hat{Y} \setminus Y|}{K - |Y|}.$$

B.3 COCO EXPERIMENT DETAILS

Class scores are extracted using TResNet Ridnik et al. (2020). Results are averaged over 1,000 trials, where in each trial 2,000 datapoints are randomly split into 500 validation and 1,500 test points. VaR is targeted with $\beta = 0.9$, while the Interval and CVaR target the ranges $[0.85, 0.95]$

and $[0.9, 1.0]$, respectively. KS and BJ bound the Mean by targeting the loss for the entire range $[0.0, 1.0]$. Point-wise methods (DKW, OrderStats) are used to bound intervals by calculating VaR upper bounds over a grid of β values within the range of interest using a Bonferroni correction for the grid size. For all experiments the size of this grid is optimized over $[10, 50, 100]$, and in this case is 10.

B.4 Fairness Experiment

Our sensitive attribute is a family's financial status. Applicants are separated based on this sensitive attribute, and we apply the methods to and evaluate each group ("convenient" and "inconvenient") separately. We measure a class-varying weighted accuracy loss:

$$L(\hat{Y}, Y) = (1 - \mu_Y)(1 - \text{Spec}(\hat{Y}, Y)) + \mu_Y(1 - \text{Sens}(\hat{Y}, Y))$$

where $\mu \in \mathbb{R}^5$ is a vector of class weights representing the importance of sensitivity for each test item's class label. This allows for user flexibility, where for instance a school could prefer to be more sensitive towards detecting good students. To demonstrate this, we perform our experiments with $\mu = (0.5, 0.6, 0.7, 0.8, 0.9)$.

We use $8{,}683$ samples from the UCI Nursery train split, including examples from both groups, to train a logistic regression classifier to predict a prospective student's admission ranking. In our bounding experiments, $4{,}000$ samples from the original test split are used for validation and testing, split evenly between the groups, with 500 examples from each group taken randomly for validation each trial. Point-wise methods are run with a beta grid size of 50. Since fairness is often concerned with those examples with the highest loss, we focus on bounding the CVaR in the range $[0.9, 1.0]$.

Results for the fair classification experiment are in Table 4 and Figure 6. Applying the Truncated Berk-Jones bound individually to each group leads to the best CVaR among all methods for both groups.

Group	Method	Guarantee	Actual	Violation %
Convenient	DKW	1.0 ± 0.0	0.500 ± 0.0	0.0 ± 0.0
	OrderStats	0.419 ± 0.026	0.190 ± 0.032	0.0 ± 0.0
	KS	0.966 ± 0.001	0.247 ± 0.019	0.0 ± 0.0
	BJ	0.418 ± 0.026	0.188 ± 0.031	0.0 ± 0.0
	Truncated-BJ	**0.382 ± 0.027**	**0.184 ± 0.027**	0.0 ± 0.0
Inconvenient	DKW	1.0 ± 0.0	0.500 ± 0.0	0.0 ± 0.0
	OrderStats	0.426 ± 0.027	0.204 ± 0.036	0.0 ± 0.0
	KS	0.966 ± 0.001	0.247 ± 0.019	0.0 ± 0.0
	BJ	0.424 ± 0.027	0.205 ± 0.036	0.0 ± 0.0
	Truncated-BJ	**0.389 ± 0.029**	**0.199 ± 0.034**	0.0 ± 0.0

Table 4: Results for fair classification on the Nursery dataset.

B.5 CIFAR and Go Emotions Experiment Details

The dataset features images of 100 different classes, and we use the $10{,}000$ examples available in the CIFAR-100 test split. Our feature extractor is the *ViT-B/32* version of CLIP available on Huggingface Wolf et al. (2019).

Go Emotions is a dataset of comments from Reddit where binary labels are given for 28 different fine-grained emotion categories and each comment may have one or more positive labels. $5{,}427$ examples are drawn from the test set, and our base classifier is a *bert-base-uncased* model fine-tuned on the train split, again available on Huggingface Wolf et al. (2019).

Experiments are run for $1{,}000$ trials with 500 validation examples randomly selected and the rest of the dataset used for testing. Balanced accuracy loss is bounded in the VaR-Interval $[0.6, 0.9]$ in both cases. For zero-shot CLIP, point-wise methods are run over a grid of 50 β values, while for Go Emotions the grid size is 100.

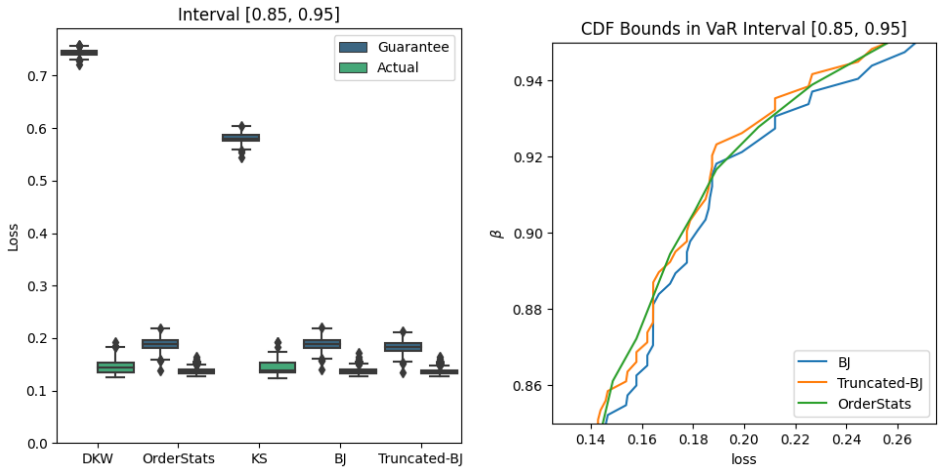


Figure 3: COCO experiment results for bounding the VaR interval $[0.85, 0.95]$. On the left is a box plot of the guarantees given and loss incurred by each method. Stronger methods return loss guarantees closer to observed test loss. On the right is a plot of the CDF bounds in the VaR interval returned by Order Stats, Berk-Jones, and Truncated Berk-Jones methods. Both plots reflect that on average the Truncated Berk-Jones bound achieves a lower loss over the interval.

B.6 Additional Results

B.6.1 Additional COCO Results

In order to understand the trade-offs that exist in targeting different metrics, we perform an experiment where each metric is used separately to optimize over the predictor space, and the performance of each chosen predictor is measured on all metrics.

Results for these experiments are reported in Table 5 and Figure 4.

	Bound	Target	Mean	VaR	Interval	CVaR
Guarantee	KS	Mean	**0.142 ± 0.003**	0.346 ± 0.083	0.595 ± 0.009	0.975 ± 0.003
		VaR	0.156 ± 0.016	**0.287 ± 0.023**	0.587 ± 0.011	0.974 ± 0.003
		Interval	0.158 ± 0.010	0.310 ± 0.043	**0.581 ± 0.008**	0.973 ± 0.003
		CVaR	0.177 ± 0.023	0.312 ± 0.033	0.592 ± 0.014	**0.971 ± 0.002**
	BJ	Mean	**0.088 ± 0.003**	0.203 ± 0.014	0.211 ± 0.011	0.472 ± 0.017
		VaR	0.101 ± 0.006	**0.179 ± 0.010**	0.191 ± 0.011	0.460 ± 0.017
		Interval	0.100 ± 0.007	0.181 ± 0.011	**0.189 ± 0.011**	0.458 ± 0.017
		CVaR	0.107 ± 0.014	0.190 ± 0.016	0.197 ± 0.015	**0.453 ± 0.015**
Actual	KS	Mean	**0.048 ± 0.001**	0.154 ± 0.010	0.154 ± 0.009	0.222 ± 0.010
		VaR	0.061 ± 0.016	0.150 ± 0.020	0.151 ± 0.019	0.213 ± 0.015
		Interval	0.064 ± 0.009	**0.141 ± 0.014**	**0.144 ± 0.013**	**0.203 ± 0.008**
		CVaR	0.082 ± 0.023	0.175 ± 0.031	0.176 ± 0.030	0.228 ± 0.022
	BJ	Mean	**0.048 ± 0.001**	0.149 ± 0.010	0.150 ± 0.008	0.218 ± 0.008
		VaR	0.061 ± 0.004	**0.134 ± 0.006**	**0.138 ± 0.006**	**0.198 ± 0.005**
		Interval	0.061 ± 0.005	**0.134 ± 0.007**	**0.138 ± 0.007**	0.199 ± 0.005
		CVaR	0.068 ± 0.013	0.148 ± 0.019	0.150 ± 0.018	0.208 ± 0.011

Table 5: COCO results illustrating trade-offs between targeting different metrics in predictor selection. Bold results are best for a given split, metric, and bound.

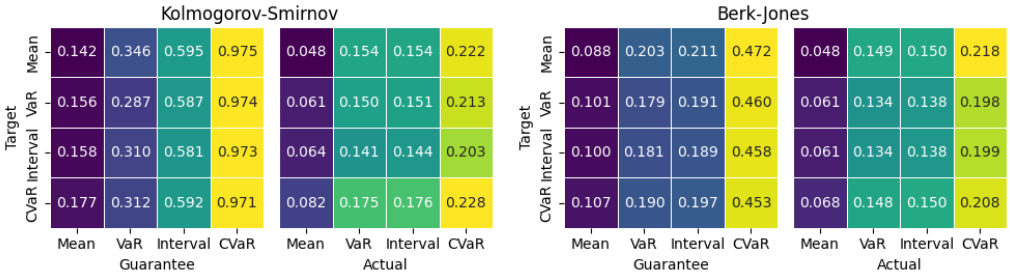


Figure 4: Results for COCO tradeoff experiments. Darker boxes correspond to lower loss values. The best guarantees always come from bounding a metric directly, although the tightest bounds do not always lead to the lowest empirical loss.

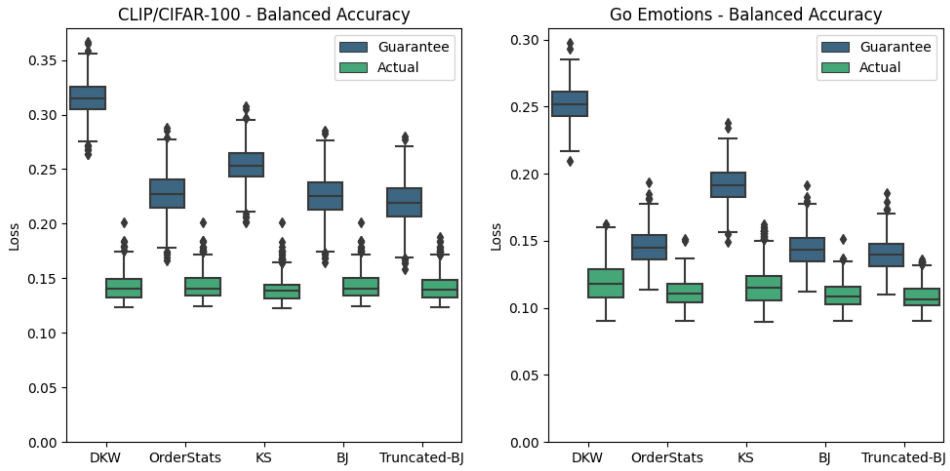


Figure 5: Results for zero-shot image classification and textual emotion recognition.

B.6.2 Additional Plots for Zero-shot CLIP, NLP, and Fairness Experiments

See Figure 5 for box plots of the results of our experiments bounding the VaR Interval for CIFAR-100 and Go Emotions datasets. Figure 6 plots the bounds in the Nursery CVaR experiment.

B.7 Predictor Space

For each experiment trial and random data split, we test a finite sample of 500 linearly spaced thresholds (or predictors) on the validation set. For computational efficiency, upper and lower bounds are set as the maximum and minimum values of the scores over the entire dataset, although this could easily be replaced by the maximum and minimum value of the scores on the validation set alone. Results are measured on the test data using the predictor with the best guarantee for a particular target metric.

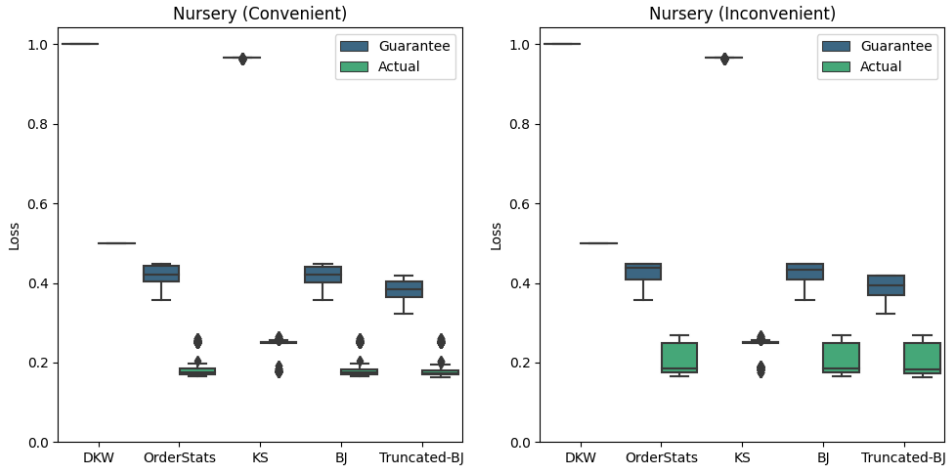


Figure 6: Plots of guaranteed and actual CVaR loss on the Nursery dataset.