Appendix

Broader impact. The broader impact of the proposed framework is significant, as it extends the ability to gain trust in machine learning systems. However there are important concerns and limitations.

- **Focus on performance metrics** In this paper we propose a range of performance metrics, which extend well beyond standard metrics concerning expected loss. However, in many situations these metrics are not sufficient to capture the effects of the machine learning system. Often a number of different metrics are required to provide a clearer picture of model performance, while some effects are difficult to capture in any metric. Also, while the measures studied offer the ability to more evenly distribute a quantity across a population, they do not offer guarantees to individuals. Finally, achieving a more equal distribution of the relevant quantity (e.g., loss or income) may have negative impacts on some segments of the population.
- **Limitations** These are summarized in the Conclusion but are expanded upon here. An important assumption in this work, and in distribution-free uncertainty quantification more generally, is that the examples seen in deployment are drawn from the same distribution as those in the validation set that are used to construct the bounds. Although this is an active area of research, here we make this assumption, and the quality of the bounds produced may degrade if the assumption is violated. A second limitation is that the scope of hypotheses and predictors we can select from is limited, due to theoretical constraints: a correction must be performed based on the size of the hypothesis set. Finally, the generated bounds may not be tight, depending on the amount of available validation data and unavoidable limits of the techniques used to produce the bounds. We did some comparisons to Empirical values of the measures we obtained bounds for in the experiments; more extensive studies would be useful to elucidate the value of the bounds in practice.

Organization of the Appendix. (1) In Appendix A, we provide detailed statements and derivations of our methodology presented Section 4.1, including how to obtain bounds for those measures mentioned in Section 3; (2) in Appendix B, we introduce further societal dispersion measures, beyond those presented in Section 3 and corresponding bounds; (3) in Appendix C, we investigate the extension of our results to multi-dimensional settings; (4) lastly, in Appendix D and E, we provide more complete details and results from our experiments (Section 5).

A Derivations and proofs for bounding methods

Section A.1, we first consider how to control, or provide upper bounds on, various quantities when we are given $(\hat{F}^{\delta}_{n,L}, \hat{F}^{\delta}_{n,U})$, which are constructed by $\{X_i\}_{i=1}^n$, such that

$$\mathbb{P}(\hat{F}^{\delta,-}_{n,L} \preceq F \preceq \hat{F}^{\delta,-}_{n,U}) \geq 1 - \delta$$

where the randomness is taken over $\{X_i\}_{i=1}^n$.

Then, in Section A.2, we will show how we obtain $(\hat{F}^{\delta,-}_{n,L}, \hat{F}^{\delta,-}_{n,U})$ by extending the arguments in [32]. In addition, we show details in Section A.2.2 on how we go beyond the methods in [32] and provide a numerical optimization method for tighter bounds.

Proof of Proposition 1. We briefly describe the the proof for Proposition 1. The proof is mainly based on [32], but we include it here for completeness. Notice for any non-decreasing function $G : \mathbb{R} \to \mathbb{R}$ (not just a CDF), there exists the (general) inverse of G as $G^-(p) = \inf\{x : G(x) \geq p\}$ for any $p \in \mathbb{R}$.

Proposition 2 (Restatement of Proposition 1)**.** *For the CDF F of X, if there exists two increasing functions F_U, F_L such that $F_U \succeq F \succeq F_L$, then we have $F_L^- \succeq F^- \succeq F_U^-$.*

Proof. For any two non-decreasing function $G(p)$ and $C(p)$, by the definition of the general inverse function, $G(G^-(p)) \geq p$. If $C \succeq G$, we therefore have $C(G^-(p)) \geq G(G^-(p)) \geq p$. Applying C^- to both sides yields $C^-(C(G^-(p))) \geq C^-(p)$. But $x \geq C^- \circ C(x)$ (see e.g. Proposition 3 on p. 6

of [30]) and thus $G^-(p) \geq C^-(p)$. Plugging in F and F_U as G and C, this can yield $F^- \succeq F_U^-$. The other direcion is similar. □

A.1 Control of nonlinear functions of CDFs (Section 4.1)

A.1.1 Control for monotonic functions

Recall that we start with the simplest case where ξ is a monotonic function in the range of X. It is straightforward to have the following claim.

Claim 1. *If we have $\hat{F}^{\delta,-}_{n,L} \preceq F \preceq \hat{F}^{\delta,-}_{n,U}$ with probability at least $1-\delta$ for some $\delta \in (0,1)$, if ξ is an increasing function, then*

$$\xi(\hat{F}^{\delta,-}_{n,L}) \succeq \xi(\hat{F}^{-}) \succeq \xi(\hat{F}^{\delta,-}_{n,U})$$

with probability at least $1-\delta$. Similarly, if ξ is a decreasing function, then $\xi(\hat{F}^{\delta,-}_{n,L}) \preceq \xi(\hat{F}^{-}) \preceq \xi(\hat{F}^{\delta,-}_{n,U})$ with probability at least $1-\delta$.

We show how this could be applied to provide bounds for Gini coefficient and Atkinson index by controlling the numerator and denominator separately as integrals of monotonic functions of F^-.

Example 1 (Gini coefficient)**.** *If given a $(1-\delta)$-CBP $(\hat{F}^{\delta}_{n,L}, \hat{F}^{\delta}_{n,U})$ and $\hat{F}^{\delta}_{n,L} \succeq 0$*[1]*, we can provide the following bound for the Gini coefficient. Notice that*

$$\mathcal{G}(X) = \frac{\int_0^1 (2p-1)F^-(p)dp}{\int_0^1 F^-(p)dp} = \frac{\int_0^1 2pF^-(p)dp}{\int_0^1 F^-(p)dp} - 1.$$

Given $F^-(p) \geq 0$ (since we only consider non-negative losses, i.e. X is always non-negative), we know

$$\mathcal{G}(X) \leq \frac{\int_0^1 2p\hat{F}^{\delta,-}_{n,L}(p)dp}{\int_0^1 \hat{F}^{\delta,-}_{n,U}(p)dp} - 1,$$

with probability at least $1-\delta$.

Example 2 (Atkinson index)**.** *First, we present the complete version of Atkinson index. Namely,*

$$\mathcal{A}(\varepsilon, X) := \begin{cases} 1 - \frac{\left(\int_0^1 (F^-(p))^{1-\varepsilon} dp\right)^{\frac{1}{1-\varepsilon}}}{\int_0^1 F^-(p)dp}, & \text{if } \varepsilon \geq 0,\ \varepsilon \neq 1; \\ 1 - \frac{\exp(\int_0^1 \ln(F^-(p))dp)}{\int_0^1 F^-(p)dp}, & \text{if } \varepsilon = 1. \end{cases}$$

Notice that for $\varepsilon \geq 0$, $(\cdot)^{1-\varepsilon}$ and $\ln(\cdot)$ are increasing functions, thus, for Atkinson index and a $(1-\delta)$-CBP $(\hat{F}^{\delta}_{n,L}, \hat{F}^{\delta}_{n,U})$, if $\hat{F}^{\delta}_{n,L} \succeq 0$, let us define $\mathcal{A}^{\delta}_U(\varepsilon, X) := 1 - \frac{\left(\int_0^1 (\hat{F}^{\delta,-}_{n,U}(p))^{1-\varepsilon} dp\right)^{\frac{1}{1-\varepsilon}}}{\int_0^1 \hat{F}^{\delta,-}_{n,L}(p)dp}$, if $\varepsilon \geq 0, \varepsilon \neq 1$; $1 - \frac{\exp(\int_0^1 \ln(\hat{F}^{\delta,-}_{n,U}(p))dp)}{\int_0^1 \hat{F}^{\delta,-}_{n,L}(p)dp}$, if $\varepsilon = 1$. Then, with probability at least $1-\delta$, $\mathcal{A}^{\delta}_U(\varepsilon, X)$ is an upper bound for $\mathcal{A}(\varepsilon, X)$ for all $\varepsilon \in [0,1)$.

As mentioned in Remark 1, instead of calculating bounds separately for each ε, simple post-processing enables us to efficiently issue a family of bounds.

Example 3 (CVaR fairness-risk measures and beyond)**.** *Recall that for $\alpha \in (0,1)$,*

$$\mathcal{D}_{CV,\alpha}(T(F_g)) = \min_{\rho \in \mathbb{R}} \left\{ \rho + \frac{1}{1-\alpha} \cdot \mathbb{E}_{g \sim \mathcal{P}_{ldx}}[T(F_g) - \rho]_+ \right\} - \mathbb{E}_{g \sim \mathcal{P}_{ldx}}[T(F_g)].$$

The function $[T(F_g) - \rho]_+$ is an increasing function when ρ is fixed and its further composition with the expectation operation is still increasing. If we have $(T^{\delta}_L(F_g), T^{\delta}_U(F_g))$ such that $T^{\delta}_L(F_g) \leq$

[1] This can be easily achieved by taking truncation over 0. Also, the construction of $\hat{F}^{\delta}_{n,L}$ in Section A.2 always satisies this constraint.

$T(F_g) \le T_U^\delta(F_g)$[2] *for all* g *with probability at least* $1-\delta$*, then we have*

$$\mathcal{D}_{CV,\alpha}(T(F_g)) \le \min_{\rho\in\mathbb{R}} \left\{ \rho + \frac{1}{1-\alpha} \cdot \mathbb{E}_{g\sim\mathcal{P}_{Idx}}[T_U^\delta(F_g) - \rho]_+ \right\} - \mathbb{E}_{g\sim\mathcal{P}_{Idx}}[T_L^\delta(F_g)],$$

and the first term of RHS can be minimized easily since it is a convex function of ρ*.*

A.1.2 Control for absolute and polynomial functions

Recall that if $s_L \le s \le s_U$, then

$$s_L \mathbf{1}\{s_L \ge 0\} - s_U \mathbf{1}\{s_U \le 0\} \le |s| \le \max\{|s_U|, |s_L|\}.$$

More generally, for any polynomial function $\phi(s) = \sum_{k=0} \alpha_k s^k$. Notice if k is odd, s^k is monotonic w.r.t. s and we can bound

$$\begin{aligned}\phi(s) \le & \sum_{\{k \text{ is odd},\ \alpha_k \ge 0\}} \alpha_k s_U^k + \sum_{\{k \text{ is odd},\ \alpha_k < 0\}} \alpha_k s_L^k \\ & + \sum_{\{k \text{ is even},\ \alpha_k \ge 0\}} \alpha_k \max\{|s_L|^k, |s_U|^k\} + \sum_{\{k \text{ is even},\ \alpha_k < 0\}} \alpha_k (s_L \mathbf{1}\{s_L \ge 0\} - s_U \mathbf{1}\{s_U \le 0\})^k.\end{aligned}$$

So, for $\phi(F^-)$, we can plug in $\hat{F}_{n,L}^{\delta,-}$ and $\hat{F}_{n,U}^{\delta,-}$ to replace s_U and s_L to obtain an upper bound with probability at least $(1-\delta)$. The derivation for the lower bound is similar. We summarize our results as the following proposition.

Proposition 3. *If given a* $(1-\delta)$*-CBP* $(\hat{F}_{n,L}^\delta, \hat{F}_{n,U}^\delta)$*,*

$$\hat{F}_{n,U}^{\delta,-} \mathbf{1}\{\hat{F}_{n,U}^{\delta,-} \ge 0\} - \hat{F}_{n,L}^{\delta,-} \mathbf{1}\{\hat{F}_{n,L}^{\delta,-} \le 0\} \preceq |F^-| \preceq \max\{|\hat{F}_{n,L}^{\delta,-}|, |\hat{F}_{n,}^{\delta,-}|\}.$$

Moreover, for any polynomial function $\phi(s) = \sum_{k=0} \alpha_k s^k$*, we have*

$$\begin{aligned}\phi(F^-) \preceq & \sum_{\{k \text{ is odd},\ \alpha_k \ge 0\}} \alpha_k (\hat{F}_{n,L}^{\delta,-})^k + \sum_{\{k \text{ is odd},\ \alpha_k < 0\}} \alpha_k (\hat{F}_{n,U}^{\delta,-})^k \\ & + \sum_{\{k \text{ is even},\ \alpha_k \ge 0\}} \alpha_k \max\{|\hat{F}_{n,U}^{\delta,-}|^k, |\hat{F}_{n,L}^{\delta,-}|^k\} \\ & + \sum_{\{k \text{ is even},\ \alpha_k < 0\}} \alpha_k (\hat{F}_{n,U}^{\delta,-} \mathbf{1}\{\hat{F}_{n,U}^{\delta,-} \ge 0\} - \hat{F}_{n,L}^{\delta,-} \mathbf{1}\{\hat{F}_{n,L}^{\delta,-} \le 0\})^k.\end{aligned}$$

Example 4. *If we have* $(T_L^\delta(F_g), T_U^\delta(F_g))$ *such that* $T_L^\delta(F_g) \le T(F_g) \le T_U^\delta(F_g)$ *holds for all* g *we consider, then we can provide high probability upper bounds for*

$$\xi(T(F_{g_1}) - T(F_{g_2}))$$

for any polynomial functions or the absolute function ξ*. For example, with probability at least* $1-\delta$

$$|T(F_{g_1}) - T(F_{g_2})| \le \max\{|T_U^\delta(F_{g_1}) - T_L^\delta(F_{g_2})|, |T_L^\delta(F_{g_1}) - T_U^\delta(F_{g_2})|\}.$$

We will further show in Appendix B how our results are applied to specific examples.

A.1.3 Control for a general function

To handle general non-linearity, we need to introduce the class of functions of bounded variation on a certain interval, which is a very rich class that includes all the functions that are continuously differentiable or Lipchitz continuous on that interval.

Definition 4 (Functions of bounded total variation [28])**.** *Define the set of paritions on* $[a,b]$ *as*

$$\Pi = \{\pi = (x_0, x_1, \cdots, x_{n_\pi}) \mid \pi \text{ is a partition of } [a,b] \text{ satisfying } x_i \le x_{i+1} \text{ for all } 0 \le i \le n_\pi - 1\}.$$

[2] $T(F_g)$ here is one of the functionals in the form we studied, so that we can provide upper and lower bounds for it.

Then, the total variation of a continuous real-valued function ξ, defined on $[a,b] \subset \mathbb{R}$ is defined as

$$V_a^b(\xi) := \sup_{\pi \in \Pi} \sum_{i=0}^{n_\pi} |\xi(x_{i+1}) - \xi(x_i)|$$

where Π is the set of all partitions, and we say a function ξ is of bounded variation, i.e. $\xi \in BV([a,b])$ iff $V_a^b(\xi) < \infty$.

Recall that $X \geq 0$ in our cases, then, for $\xi(F^-)$, we can have the following bound.

Theorem 2 (A restatement & formal version of Theorem 1)**.** *For a $(1-\delta)$-CBP $(\hat{F}^\delta_{n,L}, \hat{F}^\delta_{n,U})$, for any $p \in [0,1]$ such that the total variation of ξ is finite on $[0, \hat{F}^{\delta,-}_{n,L}(p)]$, then*

$$\xi(F^-(p)) \leq V_0^{\hat{F}^{\delta,-}_{n,L}(p)}(\xi) - V_0^{\hat{F}^{\delta,-}_{n,U}(p)}(\xi) + \xi(\hat{F}^{\delta,-}_{n,U}(p)).$$

Moreover, if ξ is continuously differentiable on $[0, \hat{F}^{\delta,-}_{n,L}(p)]$, we can express $V_0^s(\xi)$ as $\int_0^x |\frac{d\xi}{ds}(s)| ds$ for any $x \in [0, \hat{F}^{\delta,-}_{n,L}(p)]$.

Proof. By the property of functions of bounded total variation [28], if ξ is of bounded total variation on $[0, \hat{F}^{\delta,-}_{n,L}(p)]$, then, we have that: for any $x \in [0, \hat{F}^{\delta,-}_{n,L}(p)]$

$$\xi(x) = V_0^x(\xi) - (V_0^x(\xi) - \xi(x))$$

where both $f_1(x) := V_0^x(\xi)$ and $f_2(x) := V_0^x(\xi) - \xi(x)$ are increasing functions. Moreover,

$$V_0^x(\xi) = \int_0^x \left|\frac{d\xi}{ds}(s)\right| ds$$

if ξ is continuously differentiable.

Thus, by taking advantage of the monotonicity, we have

$$\xi(F^-(p)) \leq V_0^{\hat{F}^{\delta,-}_{n,L}(p)}(\xi) - V_0^{\hat{F}^{\delta,-}_{n,U}(p)}(\xi) + \xi(\hat{F}^{\delta,-}_{n,U}(p)).$$

So, if ξ is of bounded variation on the range of X, then

$$\xi(F^-) \preceq V_0^{\hat{F}^{\delta,-}_{n,L}}(\xi) - V_0^{\hat{F}^{\delta,-}_{n,U}}(\xi) + \xi(\hat{F}^{\delta,-}_{n,U}) = f_1(\hat{F}^{\delta,-}_{n,L}) - f_2(\hat{F}^{\delta,-}_{n,U}).$$

□

A.2 Methods to obtain confidence two-sided bounds for CDFs (Section 4.3)

We provide details for two-sided bounds and our numerical methods in the following.

A.2.1 The reduction approach to constructing upper bounds of CDFs (Section 4.3.1)

We here provide the proof of Lemma 1.

Lemma 2 (A restatement & formal version of Lemma 1)**.** *For $0 \leq L_1 \leq L_2 \cdots \leq L_n \leq 1$, since $\mathbb{P}(\forall i : F(X_{(i)}) \geq L_i) \geq \mathbb{P}(\forall i : U_{(i)} \geq L_i)$ by [32], if we further have $\mathbb{P}(\forall i : U_{(i)} \geq L_i) \geq 1-\delta$, then we have*

$$\mathbb{P}(\forall i : \lim_{\epsilon \to 0^+} F(X_{(i)} - \epsilon) \leq 1 - L_{n-i+1}) \geq 1 - \delta.$$

Furthermore, let $R(x)$ be defined as

$$R(x) = \begin{cases} 1 - L_n, & \text{for } x < X_{(1)} \\ 1 - L_{n-1}, & \text{for } X_{(1)} \leq x < X_{(2)} \\ \dots & \\ 1 - L_1, & \text{for } X_{(n-1)} \leq x < X_{(n)} \\ 1, & \text{for } X_{(n)} \leq x. \end{cases}$$

Then, $F \preceq R$.

Proof. Notice that for given order statistics $\{X_{(i)}\}_{i=1}^n$, let $\mathbb{P}_{\{X_{(i)}\}_{i=1}^n}$ denote the probability taken over the randomness of $\{X_{(i)}\}_{i=1}^n$, and $\mathbb{P}_X$ denote the probability taken over the randomness of X, which is an independent random variable drawn from F. Let us denote $B = -X$, and $B_{(i)}$ as the i-th order statistic for samples $\{-X_i\}_{i=1}^n$. It is easy to see that $B_{(n-i+1)} = -X_{(i)}$. We also denote $\mathbb{P}_B$ as the probability taken over the randomness of B, and F_B as the CDF of B.

$$\begin{aligned}
\mathbb{P}_{\{X_{(i)}\}_{i=1}^n}(\forall i : \lim_{\epsilon \to 0^+} F(X_{(i)} - \epsilon) \leq 1 - L_{n-i+1}) =& \mathbb{P}_{\{X_{(i)}\}_{i=1}^n}(\forall i : \mathbb{P}_X(X \geq X_{(i)}) > L_{n-i+1}) \\
=& \mathbb{P}_{\{X_{(i)}\}_{i=1}^n}(\forall i : \mathbb{P}_X(-X \leq -X_{(i)}) > L_{n-i+1}) \\
=& \mathbb{P}_{\{X_{(i)}\}_{i=1}^n}(\forall i : \mathbb{P}_B(B \leq B_{(n-i+1)}) > L_{n-i+1}) \\
=& \mathbb{P}(\forall i : F_B \circ F_B^-(U_{(n-i+1)}) > L_{n-i+1}) \\
\geq& \mathbb{P}(\forall i : U_{(n-i+1)} > L_{n-i+1}).
\end{aligned}$$

where we use the fact that $F_B^-(U_{(n-i+1)})$ is of the same distribution as $B_{(n-i+1)}$ and the last inequality follows from Proposition 1, eq. 24 on p.5 of [30].

Notice that $\mathbb{P}(\forall i : U_{(n-i+1)} > L_{n-i+1}) = \mathbb{P}(\forall i : U_{(n-i+1)} \geq L_{n-i+1})$, and according to [32] and our assumption, $\mathbb{P}(\forall i : F(X_{(i)}) \geq L_i) \geq \mathbb{P}(\forall i : U_{(i)} \geq L_i) \geq 1 - \delta$.

The conservative construction of R satisfies $R \succeq F$ straightforwardly if $\forall i : \lim_{\epsilon \to 0^+} F(X_{(i)} - \epsilon) \leq 1 - L_{n-i+1})$ holds. Thus, we know $R \succeq F$ with probability at least $1 - \delta$. Our proof is complete. □

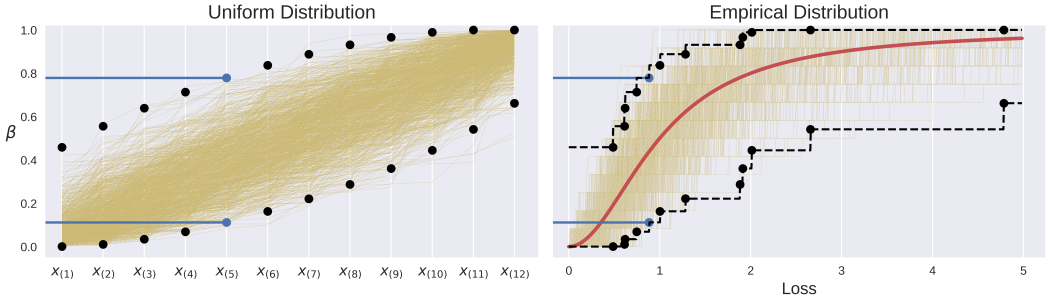

Figure 4: Example illustrating the construction of distribution-free CDF lower and upper bounds by bounding order statistics. On the left, order statistics are drawn from a uniform distribution. On the right, samples are drawn from a real loss distribution, and the corresponding Berk-Jones CDF lower and upper bound are shown in black. Our distribution-free method gives bound $b_i^{(l)}$ and $b_i^{(u)}$ on each sorted order statistic such that the bound depends only on i, as illustrated in the plots for $i = 5$ (shown in blue). On the left, 1000 realizations of $x_{(1)}, \ldots, x_{(n)}$ are shown in yellow. On the right, 1000 empirical CDFs are shown in yellow, and the true CDF F is shown in red.

A.2.2 Details of numerical optimization method (Section 4.3.2)

Now, we introduce the details of our numerical optimization method. Recall that one drawback of the QBRM bounding approach is that it is not weight function aware: when controlling $\int_0^1 \psi(p) F^-(p) dp$ for a non-negative weight function ψ, the procedure ignores the structure of ψ, as it first obtains $\hat{F}_{n,L}^{\delta}$, then provides an upper bound $\int_0^1 \psi(p) \hat{F}_{n,L}^{\delta,-}(p) dp$.

Our numerical approach can overcome that drawback and can also easily be applied to handle mixtures of multiple functionals. The bounds obtained by our method are significantly tighter than those provided by methods in [32] in the regime of small data size. Notice that the small data size regime is the one people care about because when the data size is large, all the bounds we discussed will converge to the same value, and the gap between different bounds will shrink to 0 as the data size grows.

First, by [23] and Proposition 1, eq. 24 on p.5 of [30], we have for any $0 \le L_1 \le \cdots \le L_n \le 1$,

$$\begin{aligned}\mathbb{P}\big(\forall i,\ F(X_{(i)}) \ge L_i\big) &\ge \mathbb{P}\big(\forall i,\ U_{(i)} \ge L_i\big)\\ &\ge n! \int_{L_n}^{1} dx_n \int_{L_{n-1}}^{x_n} dx_{n-1} \cdots \int_{L_1}^{x_2} dx_1,\end{aligned}$$

where the right-hand side integral is a function of $\{L_i\}_{i=1}^n$ and its partial derivatives can be exactly calculated by the package in [22]. Specifically, the package in [22] enables us to calculate

$$\upsilon(L_1, L_2, \cdots, L_n, 1) := \int_{L_n}^{1} dx_n \int_{L_{n-1}}^{x_n} dx_{n-1} \cdots \int_{L_1}^{x_2} dx_1$$

for any positive integer n. Notice that the partial derivative of $\upsilon(L_1, L_2, \cdots, L_n, 1)$ with respect to L_i is:

$$\begin{aligned}\partial_{L_i}\upsilon(L_1, L_2, \cdots, L_n, 1) = &-\int_{L_n}^{1} dx_n \int_{L_{n-1}}^{x_n} dx_{n-1} \cdots \int_{L_{i+1}}^{x_{i+2}} dx_{i+1}\\ &\cdot \int_{L_{i-1}}^{L_i} dx_{i-1} \cdots \int_{L_1}^{x_2} dx_1,\\ &= -\upsilon(L_{i+1}, \cdots, L_n, 1) \cdot \upsilon(L_1, \cdots, L_{i-1}, L_i),\end{aligned}$$

which we can also use the package in [22] to calculate the partial derivatives.

Consider providing upper or lower bounds for $\int_0^1 \psi(p)F^-(p)dp$ for non-negative weight function ψ as an example. For any $\{L_i\}_{i=1}^n$ satisfying $\mathbb{P}\big(\forall i,\ F(X_{(i)}) \ge L_i\big) \ge 1-\delta$, one can use conservative CDF completion in [32] to obtain $\hat{F}^{\delta}_{n,L}$, i.e. $\int_0^1 \psi(p)\xi(\hat{F}^{\delta,-}_{n,L}(p))dp = \sum_{i=1}^{n+1} \xi(X_{(i)}) \int_{L_{i-1}}^{L_i} \psi(p)dp$, where L_{n+1} is 1, $L_0 = 0$, and $X_{(n+1)} = \infty$ or a known upper bound for X. Then, we can formulate tightening the upper bound as an optimization problem:

$$\min_{\{L_i\}_{i=1}^n} \sum_{i=1}^{n+1} \xi(X_{(i)}) \int_{L_{i-1}}^{L_i} \psi(p)dp$$

such that

$$\mathbb{P}\big(\forall i,\ F(X_{(i)}) \ge L_i\big) \ge 1-\delta, \text{ and } 0 \le L_1 \le \cdots \le L_n \le 1.$$

Similarly, for the lower bound, we can use the CDF completion mentioned in Theorem 1, and construct $\hat{F}^{\delta}_{n,U}$, then, we can study the following lower bound for $\int_0^1 \psi(p)F^-(p)dp$,

$$\sum_{i=1}^{n} \xi(X_{(i)}) \int_{L_{n-i}}^{L_{n-i+1}} \psi(p)dp$$

where $X_{(0)} = 0$.

Parameterized model approach. Notice the above optimization problem formulation has a drawback: if more samples are drawn, i.e. n increases, then the number of parameters we need to optimize also increases. In practice, we re-parameterize $\{L_i\}_{i=1}^n$ as the following:

$$L_i(\theta) = \frac{\sum_{j=1}^{i} \exp(\phi_\theta(g_j))}{1 + \sum_{j=1}^{n} \exp(\phi_\theta(g_j))}$$

where g_i are random Gaussian seeds. This is of the same spirit as using random seeds in generative models. We find that a simple parameterized neural network model with 3 fully-connected hidden layers of dimension 64 is enough for good performance and robust to hyper-parameter settings. Take the upper bound optimization problem as an example; using the new parameterized model, we have

$$\min_{\{\theta\}_{i=1}^n} \sum_{i=1}^{n+1} \xi(X_{(i)}) \int_{L_{i-1}(\theta)}^{L_i(\theta)} \psi(p)dp \tag{1}$$

such that

$$n! \int_{L_n(\theta)}^{1} dx_n \int_{L_{n-1}(\theta)}^{x_n} dx_{n-1} \cdots \int_{L_1(\theta)}^{x_2} dx_1 \geq 1-\delta,$$

where $L_0 = 0$, $L_{n+1} = 1$, $X_{(n+1)} = \infty$ or a known upper bound for X. We can solve the above optimization problem using heuristic methods such as [9].

Post-processing for a rigorous guarantee for constraints. Notice that we may not ensure the constraint $n! \int_{L_n(\theta)}^{1} dx_n \int_{L_{n-1}(\theta)}^{x_n} dx_{n-1} \cdots \int_{L_1(\theta)}^{x_2} dx_1 \geq 1-\delta$ is satisfied in the above optimization because we may use surrogates like Langrange forms in our optimization processes. To make sure the constraint is strictly satisfied, we can do the following post-processing: let us denote the obtained L_i's by optimizing (1) as $L_i(\hat{\theta})$. Then, we look for $\gamma^* \in [0, L_n(\hat{\theta})]$ such that

$$\gamma^* = \inf\{\gamma : n! \upsilon(L_1(\hat{\theta}) - \gamma, \cdots, L_n(\hat{\theta}) - \gamma, 1) \geq 1-\delta, \gamma \geq 0\}.$$

Notice there is always a feasible solution as when $\gamma = L_n(\hat{\theta})$,

$$n! \upsilon(L_1(\hat{\theta}) - \gamma, \cdots, L_n(\hat{\theta}) - \gamma, 1) \geq \mathbb{P}\big(\forall i,\ U_{(i)} \geq 0\big) = 1$$

and $\upsilon(L_1(\hat{\theta}) - \gamma, \cdots, L_n(\hat{\theta}) - \gamma, 1)$ is a decreasing function of γ. We can use binary search to efficiently find (a good approximate of) γ^*.

B Other dispersion measures and calculation

B.1 Lorenz curve & the extended Gini family

Lorenz curve. In the main context, Lorenz curve has been mentioned in reference to Gini coefficient and Atkinson index. To be more complete, we further demonstrate the definition of Lorenz curve in its mathematical form.

Definition 5 (Lorenz curve)**.** *The definition of Lorenz curve is a function: for* $t \in [0, 1]$,

$$\mathcal{L}(t) = \frac{\int_0^t F^{-1}(p)\, dp}{\int_0^1 F^{-1}(p)\, dp}.$$

We can obtain a lower bound and an upper bound function for the Lorenz curve. Given a $(1-\delta)$-CBP $(\hat{F}^{\delta}_{n,L}, \hat{F}^{\delta}_{n,U})$ and $\hat{F}^{\delta}_{n,L} \succeq 0$, we can construct a lower bound function $\mathcal{L}^{\delta}_{L}(t)$:

$$\mathcal{L}^{\delta}_{L}(t) = \frac{\int_0^t \hat{F}^{\delta,-}_{n,U}(p)\, dp}{\int_0^1 \hat{F}^{\delta,-}_{n,L}(p)\, dp},$$

and an upper bound can be obtained by

$$\mathcal{L}^{\delta}_{U}(x) = \frac{\int_0^t \hat{F}^{\delta,-}_{n,L}(p)\, dp}{\int_0^1 \hat{F}^{\delta,-}_{n,U}(p)\, dp}.$$

With probability at least $1-\delta$, the true Lorenz curve sits between the upper bound function and the lower bound function for all $t \in [0, 1]$.

The extended Gini family. The Gini coefficient can further give rise to the extended Gini family, which is a family of variability and inequality measures that depends on one parameter – the extended Gini parameter. The definition is as follows.

Definition 6 (The extended Gini family[37])**.** *The extended Gini coefficient is given by*

$$\begin{aligned}\mathcal{G}(\nu, X) :&= \frac{-\nu Cov(X, [1-F(X)]^{\nu-1})}{\mathbb{E}[X]} \\ &= 1 - \frac{\nu \int_0^1 (1-p)^{\nu-1} F^{-}(p) dp}{\int_0^1 F^{-}(p) dp},\end{aligned}$$

where $\nu > 0$ *is the extended Gini parameter and* $Cov(\cdot, \cdot)$ *is the covariance.*

For the extended Gini coefficient, choosing different ν's corresponds to different weighting schemes applied to the vertical distance between the egalitarian line and the Lorenz curve; and if $\nu = 2$, it is the standard Gini coefficient.

Given a $(1-\delta)$-CBP $(\hat{F}^{\delta}_{n,L}, \hat{F}^{\delta}_{n,U})$ and $\hat{F}^{\delta}_{n,L} \succeq 0$, we can construct upper bound for $\mathcal{G}$. Let

$$\mathcal{G}^{\delta}_{U}(\nu, X) := 1 - \frac{\nu \int_0^1 (1-p)^{\nu-1} \hat{F}^{\delta,-}_{n,U}(p)dp}{\int_0^1 \hat{F}^{\delta,-}_{n,L}(p)dp},$$

then $\mathcal{G}^{\delta}_{U}(\nu, X) \succeq \mathcal{G}(\nu, X)$ with probability at least $1-\delta$.

B.2 Generalized entropy index

The generalized entropy index [31] is another measure of inequality in a population. Specifically, the definition is: for real number α

$$GE(\alpha, X) := \begin{cases} \frac{1}{\alpha(\alpha-1)} \mathbb{E}\left[\left(\frac{X}{\mathbb{E}X}\right)^{\alpha} - 1\right], & \alpha \neq 0, 1 \\ \mathbb{E}\left[\frac{X}{\mathbb{E}X} \ln(\frac{X}{\mathbb{E}X})\right], & \text{if } \alpha = 1 \\ -\mathbb{E}\left[\ln(\frac{X}{\mathbb{E}X})\right], & \text{if } \alpha = 0. \end{cases}$$

It is not hard to further expand the expressions and write the generalized entropy index as:

$$GE(\alpha, X) := \begin{cases} \frac{1}{\alpha(\alpha-1)} \int_0^1 \left[\left(\frac{F^-(p)}{\int_0^1 F^-(p)dp}\right)^{\alpha} - 1\right] dp, & \alpha \neq 0, 1 \\ \int_0^1 \left[\frac{F^-(p)}{\int_0^1 F^-(p)dp} \ln(\frac{F^-(p)}{\int_0^1 F^-(p)dp})\right] dp, & \text{if } \alpha = 1 \\ -\int_0^1 \left[\ln(\frac{F^-(p)}{\int_0^1 F^-(p)dp})\right] dp, & \text{if } \alpha = 0. \end{cases}$$

Notice that $(\cdot)^{\alpha}$ is a monotonic function for the case $\alpha \neq 0, 1$, and $\ln(\cdot)$ is also a monotonic function, so the bound can be obtained similarly as in the case of Atkinson index. For instance, for $\alpha > 1$, given a $(1-\delta)$-CBP $(\hat{F}^{\delta}_{n,L}, \hat{F}^{\delta}_{n,U})$,

$$\frac{1}{\alpha(\alpha-1)} \int_0^1 \left[\left(\frac{F^-(p)}{\int_0^1 F^-(p)dp}\right)^{\alpha} - 1\right] dp \leq \frac{1}{\alpha(\alpha-1)} \int_0^1 \left[\left(\frac{\hat{F}^{\delta,-}_{n,L}(p)}{\int_0^1 \hat{F}^{\delta,-}_{n,U}(p)dp}\right)^{\alpha} - 1\right] dp.$$

Other cases can be tackled in a similar way, which we will not reiterate here.

B.3 Hoover index

The Hoover index [16] is equal to the percentage of the total population's income that would have to be redistributed to make all the incomes equal.

Definition 7 (Hoover index)**.** *For a non-negative random variable X, the Hoover index is defined as*

$$H(X) = \frac{\int_0^1 |F^-(p) - \int_0^1 F^-(q)dq|dp}{2\int_0^1 F^-(p)dp}$$

Hoover index involves forms like $|F^- - \mu|$ for $\mu = \int_0^1 F^-(p)dp$. This type of nonlinear structure can be dealt with the absolute function results mentioned in Appendix A.1.2.

For Hoover index and a $(1-\delta)$-CBP $(\hat{F}^{\delta}_{n,L}, \hat{F}^{\delta}_{n,U})$, let us define

$$H_U(X) = \frac{\int_0^1 \max\{|\hat{F}^{\delta,-}_{n,L}(p) - \int_0^1 \hat{F}^{\delta,-}_{n,U}(q)dq|, |\hat{F}^{\delta,-}_{n,U}(p) - \int_0^1 \hat{F}^{\delta,-}_{n,L}(q)dq|\}dp}{2\int_0^1 \hat{F}^{\delta,-}_{n,U}(p)dp}.$$

Then, with probability at least $1-\delta$, $H_U(,X)$ is an upper bound for $H(X)$.

B.4 Extreme observations & mean range

For example, a city may need to estimate the cost of damage to public amenities due to rain in a certain month. The loss for each day of a month is $X_1, \cdots, X_k$ i.i.d drawn from F, and the administration hopes to estimate and control the dispersion of the losses in a month so that they can accurately allocate resources. This involves quantities such as range ($\max_{i\in[k]} X_i - \min_{j\in[k]} X_j$) or quantiles of extreme observations ($\max_{i\in[k]} X_i$). The CDF of extreme observations such as $\max_{i\in[k]} X_i$ involves a nonlinear function of F, i.e. $(F(x))^k$.

Example 5 (Quantiles of extreme observations)**.** *The CDF of* $\max_{i\in[k]} X_i$ *is* F^k*. Thus, by the result of Appendix A.1.2, if given a* $(1-\delta)$*-CBP* $(\hat{F}^{\delta}_{n,L}, \hat{F}^{\delta}_{n,U})$ *and* $1 \succeq \hat{F}^{\delta,-}_{n,U} \succeq \hat{F}^{\delta,-}_{n,L} \succeq 0$*, with probability at least* $1-\delta$*,*

$$(\hat{F}^{\delta,-}_{n,L})^k \preceq F^k \preceq (\hat{F}^{\delta,-}_{n,U})^k.$$

We also have

$$(\hat{F}^{\delta,-}_{n,U})^k \preceq F^k \preceq (\hat{F}^{\delta,-}_{n,L})^k.$$

Similarly, for $\min_{i\in[k]} X_i$*, the CDF is* $1-(1-F)^k$*, thus, we have*

$$1-(1-\hat{F}^{\delta,-}_{n,U})^k \preceq F^k \preceq 1-(1-\hat{F}^{\delta,-}_{n,L})^k.$$

We also want to emphasize, even if, X *is* ***not*** *necessarily non-negative, we can apply the polynomial method in Appendix A.1.2 for* $\hat{F}^{\delta,-}_{n,U}$ *and* $\hat{F}^{\delta,-}_{n,L}$*.*

Example 6 (Mean range)**.** *By [13], if we further have prior knowledge that* X *is of continuous distribution, the mean of* $\max_{i\in[k]} X_i - \min_{j\in[k]} X_j$ *can be expressed as:*

$$k\int F^-(x)[F^{k-1}(x)-F^k(x)]dF(x) = k\int_0^1 F^-(F^-(p))[F^{k-1}(F^-(p))-F^k(F^-(p))]dp$$

Notice that both F *and* F^- *are increasing. Thus, if given a* $(1-\delta)$*-CBP* $(\hat{F}^{\delta}_{n,L}, \hat{F}^{\delta}_{n,U})$*,* $\hat{F}^{\delta}_{n,L} \succeq 0$*, then with probability at least* $1-\delta$*,*

$$\int_0^1 \hat{F}^{\delta,-}_{n,L}\big(\hat{F}^{\delta,-}_{n,L}(p)\big)\Big[(\hat{F}^{\delta}_{n,U})^k\big(\hat{F}^{\delta,-}_{n,L}(p)\big) - (\hat{F}^{\delta}_{n,L})^k\big(\hat{F}^{\delta,-}_{n,U}(p)\big)\Big]dp$$

is an upper bound of the mean range.

There are many other interesting societal dispersion measures that could be handled by our framework, such as those in [20]. For example, they study tail share that captures "the top 1% of people own X share of wealth", which could be easily handled with the tools provided here. We will leave those those examples to readers.

C Extension to multi-dimensional cases and applications

We briefly discuss extending our approach to multi-dimensional losses. Unfortunately, there is not a gold-standard definition of quantiles in the multi-dimensional case, and thus we only discuss functionals of CDFs and provide an example. For multi-dimensional samples $\{\boldsymbol{X}_i\}_{i=1}^n$, each of k dimensions, i.e. $\boldsymbol{X}_i = (X_1^i, \cdots, X_k^i)$, for any k-dimensional vector $\boldsymbol{x} = (x_1, \cdots, x_k)$, define empirical CDF

$$\hat{F}_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\boldsymbol{X}_i \preceq \boldsymbol{x}\}.$$

where we abuse the notation $\preceq$ to mean all of $\boldsymbol{X}_i$'s coordinates are smaller than $\boldsymbol{x}$'s.

By classic DKW inequality, we have with probability at least $1-\delta$,

$$|\hat{F}_n(\boldsymbol{x}) - F(\boldsymbol{x})| \le \sqrt{\frac{\ln(k(n+1)/\delta)}{2n}}.$$

Meanwhile, we can further adopt Frechet-Hoeffeding bound, which gives,

$$\max\{1-k+\sum_{i=1}^k F_i(x_i), 0\} \le F(\boldsymbol{x}) \le \min\{F_1(x_1), \cdots, F_k(x_k)\}$$

where F_i is the CDF of the i-th coordinate. Then, we can construct $(\hat{F}_{n,L}^{\delta/k,i}, \hat{F}_{n,U}^{\delta/k,i})$ such that $(\hat{F}_{n,L}^{\delta/k,i} \preceq F_i \preceq \hat{F}_{n,U}^{\delta/k,i})$, with probability at last $1-\delta/k$. Thus, by union bound,

$$\max\{1-k+\sum_{i=1}^{k}\hat{F}_{n,L}^{\delta/k,i}(x_i), 0\} \leq F(\boldsymbol{x}) \leq \min\{\hat{F}_{n,U}^{\delta/k,1}(x_1), \cdots, \hat{F}_{n,U}^{\delta/k,k}(x_k)\}$$

for all $\boldsymbol{x}$ with probability at last $1-\delta$.

We have

$$F(\boldsymbol{x}) \geq \max\{1-k+\sum_{i=1}^{k}\hat{F}_{n,L}^{\delta/k,i}(x_i), 0, \hat{F}_n(\boldsymbol{x}) - \sqrt{\frac{\ln(k(n+1)/\delta)}{2n}}\}$$

$$F(\boldsymbol{x}) \leq \min\{\hat{F}_{n,U}^{\delta/k,1}(x_1), \cdots, \hat{F}_{n,U}^{\delta/k,k}(x_k), \hat{F}_n(\boldsymbol{x}) + \sqrt{\frac{\ln(k(n+1)/\delta)}{2n}}\}$$

with probability at last $1-2\delta$.

Example 7 (Gini correlation coefficient [37])**.** *The Gini correlation coefficient for two non-negative random variable X and Y are defined as*

$$\Gamma_{X,Y} := \frac{Cov(X, F_Y(Y))}{Cov(X, F_X(X))} = \frac{\int\int \Big(F_{X,Y}(x,y) - F_X(x)F_Y(Y)\Big)dxdF_Y(y)}{Cov(X, F_X(X))},$$

where F_X, F_Y are marginal CDFs of X, Y and $F_{X,Y}$ is the joint CDF. One can use the multi-dimensional CDF bounds and our previous methods to provide bounds for the Gini correlation coeffiecient.

D Experiment details

This section contains additional details for the experiments in Section 5. We set $\delta = 0.05$ (before statistical corrections for multiple tests) in all experiments unless otherwise explicitly stated. Whenever we are bounding measures on multiple hypotheses, we perform a correction for the size of the hypothesis set. Additionally, when we bound measures on multiple distributions (e.g. demographic groups), we also perform a correction. Our code will be released publicly upon the publication of this article.

D.1 CivilComments (Section 5.1)

Our set of hypotheses are a toxicity model combined with a Platt scaler [24], where the model is fixed and we vary the scaling parameter in the range $[0.25, 2]$ while fixing the bias term to 0. We use a pre-trained toxicity model from the popular python library Detoxify [3] [10] and perform Platt Scaling using code from the python library released by [19] [4]. A Platt calibrator produces output according to:

$$h(v) = \frac{1}{1+\exp(wv+b)}$$

where w, b are learnable parameters and v is the log odds of the prediction. Thus we form our hypothesis set by varying the parameter w while fixing b to 0. Examples are drawn from the train split of CivilComments, which totals 269,038 data points.

The loss metric for our CivilComments experiments is the Brier Score. For n data points, Brier score is calculated as:

$$L = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2$$

where f_i is prediction confidence and o_i is the outcome (0 or 1).

[3] https://github.com/unitaryai/detoxify

[4] https://github.com/p-lambda/verified_calibration

D.1.1 Bounding complex objectives (Section 5.1.1)

We randomly sample 100,000 test points for calculating the empirical values in Table 1, and draw our validation points from the remaining data. We perform a Bonferroni correction on $\delta = 0.05$ for the size of the set of hypotheses as well as the number of distributions on which we bound our measures (in this case the number of groups, 4). We set $\lambda = 1.0$.

Numerical optimization details (including training strategy and hyperparameters) are the same as Section 5.1.2, explained below in Appendix D.1.2. For each group g we optimize the objective

$$\mathcal{O} = T_1(F_g) + T_2(F_g)$$

where F_g is the CDF bound for group, T_1 is expected loss, and T_2 is a smoothed version of a median with $a = 0.01$ (see Appendix D.1.2 and Figure 5).

For comparison, the DKW inequality is applied to get a CDF lower bound, which is then transformed to an upper bound via the reduction approach in Section 4.3.1. To get the lower bound $b^l_{1:n}$, we set:

$$b^l_i = \max(0, \frac{\#\text{ points} \leqslant \frac{i}{n}}{n} - \sqrt{\frac{\log(\frac{2}{\delta})}{2n}})$$

D.1.2 Numerical optimization examples (Section 5.1.2)

We parameterize the bounds with a fully connected network with 3 hidden layers of dimension 64. The n gaussian seeds are of size 32, which is also the input dimension for the network. Training is performed in two stages, where the network is first trained to approximate a Berk-Jones bound, and then optimized for some specified objective O. In both stages of training we aim to push the training error to zero or as close as possible (i.e. "overfit"), since we are optimizing a bound and do not seek generalization. The model is first trained for 100,000 epochs to output the Berk-Jones bound using a mean-squared error loss. Then optimization on O is performed for a maximum of 10,000 epochs, and validation is performed every 25 epochs, where we choose the best model according to the bound on O. Both stages of optimization use the Adam optimizer [17] with a learning rate 0.00005, and for the second stage the constraint weight is set to $\lambda = 0.00005$. We perform post-processing to ensure the constraint holds (see Section A.2.2). For some denominator m (in our case $m = 10^6$) we set $\gamma = \frac{1}{m}, \frac{2}{m}, \frac{3}{m}, \ldots$ and check the constraint until it is satisfied.

This approach is applied to both the experiments in Section 5.1.1 and Section 5.1.2. Details on the objective for Section 5.1.1 are above in Appendix D.1.1. In Section 5.1.2, we set $\delta = 0.01$ and our metrics for optimization are described below:

CVaR CVaR is a measure of the expected loss for the items at or above some quantile level β. We set $\beta = 0.75$, and thus we bound the expected loss for the worst-off 25% of the population.

VaR-Interval In the event that different stakeholders are interested in the VaR for different quantile levels β, we may want to select a bound based on some interval of the VaR $[\beta_{min}, \beta_{max}]$. We perform our experiment with $\beta_{min} = 0.5, \beta_{max} = 0.9$, which includes the median ($\beta = 0.5$) through the worst-case loss exluding a small batch of outliers ($\beta = 0.9$).

Quantile-Weighted We apply a weighting function to the quantile loss $\psi(p) = p$, such that the loss incurred by the worst-off members of a population are weighted more heavily.

Smoothed Median We study a more robust version of a median:

$$\psi(p; \beta) = \frac{1}{a\sqrt{\pi}} \exp(-\frac{(p-\beta)^2}{a^2})$$

with $\beta = 0.5$ and $a = 0.01$, similar to a normal distribution extremely concentrated around its mean. See Figure 5 for an illustration of such a weighting.

D.2 Bounds on standard measures (Section 5.2)

This section contains additional details for the experiments in Section 5.2.

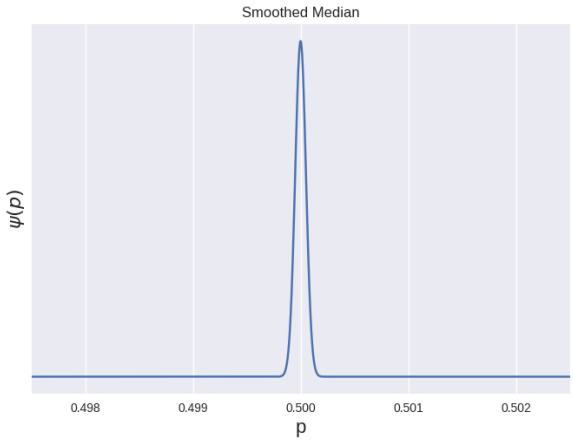

Figure 5: Plot of smoothed median function with $\beta = 0.5$ and $a = 0.01$

D.2.1 RxRx1 (Section 5.2.1)

We use the code released by [18][5] to pre-train a model on the train split of RxRx1 [33] and we evaluate our algorithm on the OOD val split with 9854 total samples. We randomly sample 2500 items for use in validation (bounding and model selection), and use the remainder of the data points for illustrating the empirical distribution induced by the different hypotheses. The thresholds which are combined with the pre-trained model to form our hypothesis set are evenly spaced in $[-8, 0]$ under the log transformation with base 10, thus leaving the thresholds in the range $[10^{-8}, 1]$.

Balanced accuracy is calculated as:

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

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

where Y is the set of ground truth labels (which in this experiment will always be one label), $\hat{Y}$ is a set of predictions, and k is the number of classes.

D.2.2 MovieLens-1M (Section 5.2.2)

MovieLens-1M [12] is a publicly available dataset. We filter all ratings below 5 stars, a typical pre-processing step, and filter any users with less than 15 5-star ratings, leaving us with 4050 users. For each user, the 5 most recently watched items are added to the test set, while the remaining (earlier) items are added to the train set. We train a user/item embedding model using the popular python recommender library LightFM [6] with a WARP ranking loss for 30 epochs and an embedding dimension of 16.

For recommendation set $\hat{I}$ we compute a loss combining recall and precision against a user test set I of size k:

$$L = \alpha l_r(\hat{I}, I)^2 + (1 - \alpha) l_p(\hat{I}, I)^2, \text{ where}$$

$$l_r(\hat{I}, I) = 1 - \frac{1}{k} \sum_{i \in I} \mathbb{1}\{i \in \hat{I}\} \text{ and } l_p(\hat{I}, I) = 1 - \frac{1}{|\hat{I}|} \sum_{i \in \hat{I}} \mathbb{1}\{i \in I\}$$

where $\alpha = 0.5$. We randomly sample 1500 users for validation, and use the remaining users to plot the empirical distributions. The 100 hypotheses tested are evenly spaced between the minimum and maximum scores of any user/item pair in the score matrix.

[5] https://github.com/p-lambda/wilds
[6] https://github.com/lyst/lightfm

E Additional results for numerical optimization (Section 5.1.2)

Figure 6 compares the learned bounds G_{opt} to the Berk-Jones (G_{BJ}) and Truncated Berk-Jones (G_{BJ-t}) bounds, as well as the empirical CDF of the real loss distribution.

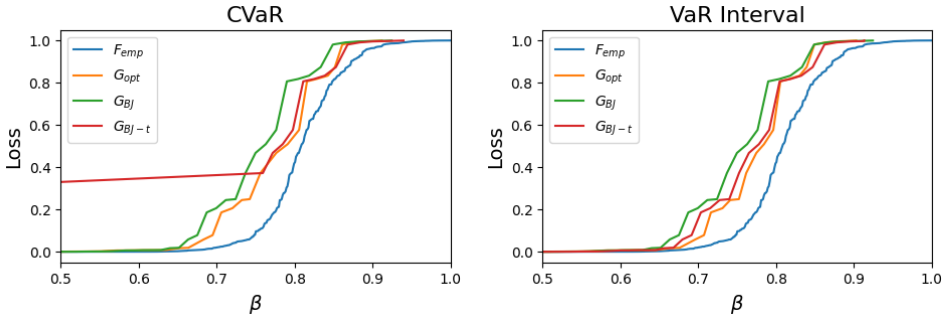

Figure 6: Learning tighter bounds on functionals of interest for protected groups. On the left, a bound is optimized for CVaR with $\beta = 0.75$, and on the right a bound is optimized for the VaR Interval $[0.5, 0.9]$. In both cases the optimized bounds are tightest on both the target metric as well as the mean, illustrating the power of adaptation both to particular quantile ranges as well as real loss distributions.

References

[1] 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*, November 2021.

[2] Anthony B Atkinson et al. On the measurement of inequality. *Journal of economic theory*, 2(3):244–263, 1970.

[3] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-Free, Risk-Controlling Prediction Sets. *arXiv:2101.02703*, August 2021.

[4] Mohammed Berkhouch, Ghizlane Lakhnati, and Marcelo Brutti Righi. Spectral risk measures and uncertainty. *arXiv preprint arXiv:1905.07716*, 2019.

[5] Neil Bhutta, Andrew Chang, Lisa Dettling, and Joanne Hsu. Disparities in wealth by race and ethnicity in the 2019 survey of consumer finances. *FEDS Notes*, 2020, 09 2020.

[6] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification, 2019.

[7] Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. Two-sided fairness in rankings via lorenz dominance, 2021.

[8] Kevin Dowd and David Blake. After VaR: The Theory, Estimation, and Insurance Applications of Quantile-Based Risk Measures. *Journal of Risk & Insurance*, 73(2):193–229, June 2006.

[9] Chengyue Gong and Xingchao Liu. Bi-objective trade-off with dynamic barrier gradient descent. *NeurIPS 2021*, 2021.

[10] Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

[11] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[12] F Maxwell Harper and Joseph A Konstan. The MovieLens datasets: History and context, December 2015.

[13] HO Hartley and HA David. Universal bounds for mean range and extreme observation. *The Annals of Mathematical Statistics*, pages 85–99, 1954.

[14] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-Identifiable) masses. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR, 2018.

[15] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 2018.

[16] Bruce P Kennedy, Ichiro Kawachi, and Deborah Prothrow-Stith. Income distribution and mortality: cross sectional ecological study of the robin hood index in the united states. *Bmj*, 312(7037):1004–1007, 1996.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. 2015. arXiv: 1412.6980.

[18] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts, 2021.

[19] Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration, 2020.

[20] Tomo Lazovich, Luca Belli, Aaron Gonzales, Amanda Bower, Uthaipon Tantipongpipat, Kristian Lum, Ferenc Huszar, and Rumman Chowdhury. Measuring disparate outcomes of content recommendation algorithms with distributional inequality metrics. *Patterns*, 3(8):100568, 2022.

[21] Jean-Yves Le Boudec. Rate adaptation, congestion control and fairness: A tutorial. *Web page, November*, 4, 2005.

[22] Amit Moscovich, Boaz Nadler, and Clifford Spiegelman. On the exact Berk-Jones statistics and their p-value calculation. *Electronic Journal of Statistics*, 10(2):2329–2354, 2016.

[23] Amit Moscovich, Boaz Nadler, and Clifford Spiegelman. On the exact Berk-Jones statistics and their p-value calculation. *Electronic Journal of Statistics*, 10(2), January 2016.

[24] John C Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, 1999. tex.organization: Citeseer.

[25] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[26] 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.

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

[28] Halsey Lawrence Royden and Patrick Fitzpatrick. *Real analysis*, volume 32. Macmillan New York, 1988.

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

[30] Galen R. Shorack and Jon A. Wellner. *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics, January 2009.

[31] Anthony F Shorrocks. The class of additively decomposable inequality measures. *Econometrica: Journal of the Econometric Society*, pages 613–625, 1980.

[32] Jake C Snell, Thomas P Zollo, Zhun Deng, Toniann Pitassi, and Richard Zemel. Quantile risk control: A flexible framework for bounding the probability of high-loss predictions. *arXiv preprint arXiv:2212.13629*, 2022.

[33] James Taylor, Berton Earnshaw, Ben Mabey, Mason Victors, and Jason Yosinski. Rxrx1: An image set for cellular morphological variation across many experimental batches. In *International Conference on Learning Representations (ICLR)*, volume 22, page 23, 2019.

[34] Vladimir Vovk, Ilia Nouretdinov, Akimichi Takemura, and Glenn Shafer. Defensive forecasting for linear protocols. *arXiv:cs/0506007*, September 2005. arXiv: cs/0506007.

[35] Robert Williamson and Aditya Menon. Fairness risk measures. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6786–6797. PMLR, June 2019.

[36] Shlomo Yitzhaki. Relative deprivation and the gini coefficient. *The quarterly journal of economics*, 93(2):321–324, 1979.

[37] Shlomo Yitzhaki and Edna Schechtman. *The Gini methodology: a primer on a statistical methodology*. Springer, 2013.