Journal of Machine Learning Research 25 (2024) 1-49 Submitted 6/23; Revised 4/24; Published 5/24

# Statistical Inference for Fairness Auditing

**John J. Cherian** JCHERIAN@STANFORD.EDU
*Department of Statistics*
*Stanford University*
*Stanford, CA 94305, USA*

**Emmanuel J. Candès** CANDES@STANFORD.EDU
*Departments of Mathematics and Statistics*
*Stanford University*
*Stanford, CA 94305, USA*

**Editor:** Ilya Shpitser

## Abstract

Before deploying a black-box model in high-stakes problems, it is important to evaluate the model's performance on sensitive subpopulations. For example, in a recidivism prediction task, we may wish to identify demographic groups for which our prediction model has unacceptably high false positive rates or certify that no such groups exist. In this paper, we frame this task, often referred to as "fairness auditing," in terms of multiple hypothesis testing. We show how the bootstrap can be used to simultaneously bound performance disparities over a collection of groups with statistical guarantees. Our methods can be used to flag subpopulations affected by model underperformance, and certify subpopulations for which the model performs adequately. Crucially, our audit is model-agnostic and applicable to nearly any performance metric or group fairness criterion. Our methods also accommodate extremely rich—even infinite—collections of subpopulations. Further, we generalize beyond subpopulations by showing how to assess performance over certain distribution shifts. We test the proposed methods on benchmark datasets in predictive inference and algorithmic fairness and find that our audits can provide interpretable and trustworthy guarantees.

**Keywords:** algorithmic fairness, bootstrap, simultaneous inference, multiple testing, reproducing kernel Hilbert space

## 1. Introduction

While black-box models may demonstrate impressive accuracy on average, their performance can still vary substantially between subpopulations. For example, an algorithm deployed for recidivism prediction exhibits significantly higher false positive rates for African-American relative to Caucasian parolees (Angwin et al., 2016). Similar performance disparities have been documented in other high-stakes applications such as facial recognition and hiring (Buolamwini and Gebru, 2018; Dastin, 2018).

Motivated by this concern, numerous stakeholders have solicited methods, often referred to as "fairness audits," that can discover and quantify such disparities (Brundage et al., 2020; Schaake and Clark, 2022). Despite substantial prior work in this area (Morina et al., 2019; Xue et al., 2020; DiCiccio et al., 2020; Tramer et al., 2017; Taskesen et al., 2021; Si

©2024 John J. Cherian and Emmanuel J. Candès.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/23-0739.html.

et al., 2021; Yan and Zhang, 2022; von Zahn et al., 2023), the definition of fairness auditing remains fraught. Fairness auditing is often framed as a single statistical test that rejects in the case of *any* performance disparity over a limited set of sensitive subpopulations (DiCiccio et al., 2020; Tramer et al., 2017; Si et al., 2021; Morina et al., 2019; Taskesen et al., 2021; Xue et al., 2020; Roy and Mohapatra, 2023). While follow-up investigation to localize disparities is desired (and often performed), this task raises new challenges. For example, empirical parity across a limited collection of subgroups does not rule out substantial disparities among smaller subgroups (Kearns et al., 2018). Further, if we consider multiple performance metrics over a rich collection of subpopulations, discovering some disparity between two subgroups is hardly surprising. Unfortunately, existing methods for identifying localized (dis-)parities are accompanied by few statistical guarantees (von Zahn et al., 2023; Yan and Zhang, 2022; Schaake and Clark, 2022).

We develop a family of statistical methods that rigorously achieve two goals: (1) the "certification" of subpopulations for which the model performs adequately, and (2) the "flagging" of subpopulations that suffer harmful performance disparities. Formally, we approach these two tasks by allowing the auditor to define a "disparity" by comparing some measure of model performance on a subpopulation to a potentially data-dependent target. A certification audit then allows the auditor to identify subpopulations for which this disparity is acceptably low, while a flagging audit discovers subpopulations for which this disparity exceeds some prespecified threshold. Our proposed methods only require access to a so-called "audit trail," i.e., model predictions on a data set held out from training (Brundage et al., 2020), but not white-box access to the model itself. A Python package, `fairaudit`, implementing these methods is available to install from PyPI and can be downloaded at github.com/jjcherian/fairaudit.

### 1.1 Outline

The paper is organized as follows. We first describe the problem setting and associated notation in Section 2; we subsequently preview two applications of our methodology. Each section thereafter is devoted to an auditing method. In Section 3, we describe our procedures for certifying performance disparities over a collection of subpopulations. In Section 4, we show how to flag performance disparities. Lastly, in Section 5, we show how to extend our methodology from subpopulations to collections of distribution shifts.

## 2. Framework and preliminaries

### 2.1 Definitions and notation

We say that some prediction rule $f$ exhibits a performance disparity on a subpopulation $G$ if the mean of some metric $L(f(X), Y)$, conditional on $(X, Y)$ belonging to $G$, differs substantially from some target $\theta_P \in \mathbb{R}$.

Our statistical audits proceed by testing and/or constructing bounds on the group-wise performance disparity defined below.

**Definition 1** *We define a group-wise performance disparity by*

$$\underbrace{\epsilon(G)}_{\textit{disparity}} := \underbrace{\mathbb{E}_P[L(f(X), Y) \mid (X, Y) \in G]}_{\textit{group-specific}} - \underbrace{\theta_P}_{\textit{target}}. \tag{1}$$

Subgroup membership may be defined by a subset of the covariates (e.g., certain sensitive attributes) that the prediction rule does not directly use. For notational simplicity, however, we will refer to the same covariate vector $X$ when defining both the prediction rule and group membership.

Nearly every group fairness definition can be expressed in terms of $\epsilon(G)$. For example, when auditing for disparities in the positive classification rate of a binary predictor $f(x)$, we instantiate the flagging method by testing the null hypothesis $H_0(G) : \epsilon(G) \leq 0.05$ with

$$L(f(x), y) = \mathbf{1}\{f(x) = 1\} \quad \text{and} \quad \theta_P = \mathbb{P}(f(X) = 1).$$

See Appendix A for additional examples.

To evaluate these disparities, the audit we devise requires access only to a holdout data set $\mathcal{D} = \{X_i, Y_i\}_{i=1}^n \overset{\text{iid}}{\sim} P$; this is sometimes called an "audit trail" (Brundage et al., 2020).

If $\theta_P$ is not known a-priori, we will assume that it is possible to use this holdout set to construct a consistent estimator $\hat{\theta}$. We omit the argument specifying the data set used to construct $\hat{\theta}$ when it is clear from context. In the main text, we will make the following technical assumption regarding the estimator $\hat{\theta}$.

**Assumption 1** *The estimator of the target in Definition 1 is asymptotically linear, i.e.,*

$$\sqrt{n}(\hat{\theta}(\mathcal{D}) - \theta_P) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i, Y_i) + o_P(1).$$

**Remark 2** *Assumption 1 is satisfied by any estimator expressible as a differentiable function of averages, i.e., $\hat{\theta}(\mathcal{D}) := g(\sum_i h(x_i, y_i)/n)$ for some differentiable function $g$ and known features $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^k$. In this paper, we most often compare group-wise performance to the population average $\theta_P = \mathbb{E}_P[L(f(X), Y)]$, and consequently, $\hat{\theta}(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^n L(f(x_i), y_i)$ trivially satisfies this assumption.*

**Remark 3** *Asymptotic linearity is a canonical assumption for Gaussian approximation methods in statistics. It is also satisfied by any M or Z-estimator, e.g., if $\hat{\theta}$ is the empirical risk minimizer for a smooth convex loss (van der Vaart, 2000; Lehmann et al., 2005).*

We define the following terms for notational convenience. We replace $(X, Y) \in G$ with $G$ whenever the meaning is clear. For example, we let $\mathbb{P}_n(G) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{(x_i, y_i) \in G\}$ denote the empirical probability of $(X, Y)$ belonging to $G$ in the holdout set, and $\mathbb{P}(G) := \mathbb{P}((X, Y) \in G)$ denote the population probability of $(X, Y)$ belonging to $G$. To further simplify our notation, we will also replace $L(f(X), Y)$ by the abbreviation $L$. We also employ the shorthand $|G|$ to denote the cardinality of $G$ in the holdout set, i.e., $|G| := n \cdot \mathbb{P}_n(G)$. We thus define the plug-in estimator of the disparity, $\hat{\epsilon}(G) := |G|^{-1}\sum_{i \in G} L_i - \hat{\theta}$.

### 2.2 Preview of contributions

To motivate and summarize the main contributions of our paper, we preview two applications. We consider two concrete examples of performance metrics and subpopulations with which an auditor may test for either tolerable or excessive disparity compared to a target threshold. These examples illustrate the breadth of problems to which fairness auditing may be applicable. In our first example, we show how a certification audit provides practical guarantees on the performance of a popular predictive inference method. In our second example, we audit for false positive rate disparities in the canonical COMPAS data set.

#### 2.2.1 Certifying conditional coverage

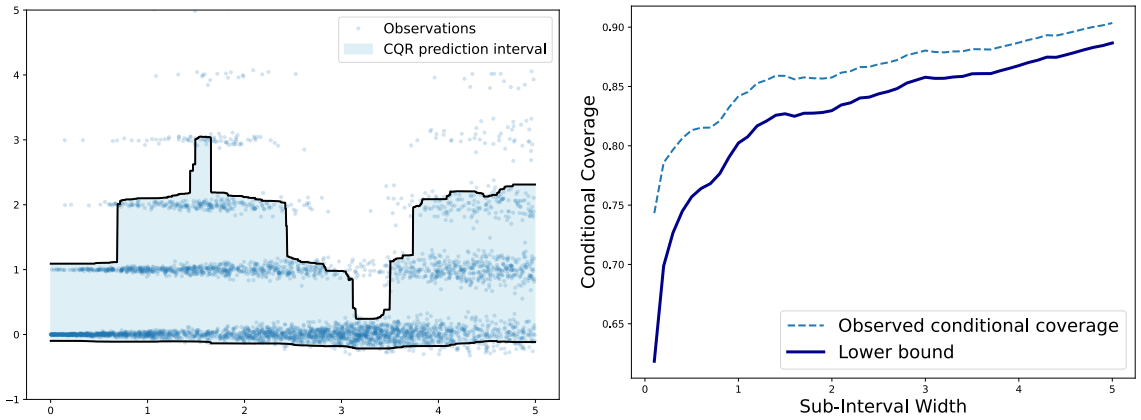
Consider a training set $\{(X_i, Y_i)\}_{i=1}^n$ and a test point $(X_{n+1}, Y_{n+1})$ sampled i.i.d. from some unknown distribution $P$. Using $\{X_i, Y_i\}_{i=1}^n \cup \{X_{n+1}\}$ as input, conformal prediction produces a set-valued function, denoted by $\hat{C}(\cdot)$, that satisfies the guarantee $\mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha$ *marginally* over the randomness in the training and test points. This marginal guarantee does not preclude loss of coverage, however, after we condition on $\hat{C}$ and $X \in G$. There may exist subsets $G \subseteq \mathcal{X}$ such that $\mathbb{P}(Y \in \hat{C}(X) \mid \hat{C}, X \in G) \ll 1 - \alpha$. In sensitive applications, for example, this implies that a conformal prediction set can cover the true label more than $(1 - \alpha)$-% of the time for some protected subgroups, while being untrustworthy for others.

In the language of fairness auditing, the "performance disparity" under consideration is the gap between the prediction set's unknown group-conditional coverage and known marginal coverage. More formally, we aim to certify conditional coverage over all sub-intervals with endpoints in $\{0, 0.1, 0.2, \ldots, 5\}$ by constructing a lower confidence bound on $\epsilon(G)$ with

$$L(\hat{C}(x), y) = \mathbf{1}\{y \in \hat{C}(x)\} \quad \text{and} \quad \theta_P = 0.9.$$

We re-visit the synthetic data experiment of Romano et al. (2019) and bound the size of this disparity over this large collection of groups. Figure 1a plots the data set used in this experiment and the prediction interval output by the conformalized quantile regression (CQR) method. Then, for each sub-interval, our audit issues simultaneously valid lower confidence bounds on the conditional coverage of CQR.

Since it is infeasible to display the 1,275 lower confidence bounds we output for all sub-intervals, we summarize our results by plotting a lower bound on the conditional coverage over all sub-intervals of a given width, i.e., $\min_{G:\text{width}(G)=w} \mathbb{P}(Y \in \hat{C}(X) \mid \hat{C}, X \in G)$. Figure 1b plots this bound (solid line) as well as the observed conditional coverage (dashed line), $\min_{G:\text{width}(G)=w} \hat{\mathbb{P}}_n(Y \in \hat{C}(X) \mid \hat{C}, X \in G)$. We define each point on the plotted bound using the smallest lower confidence bound on $\epsilon(G)$ over all subgroups of a particular width. Since the group-wise confidence bounds issued by our audit are simultaneously valid, the plotted curve is also simultaneously valid over all sub-interval widths with high probability. For example, we can say with 95% confidence that no sub-interval of length 1 has conditional coverage worse than 80.2%, and no sub-interval of length 2 has conditional coverage worse than 83.0%. Crucially, our guarantee is exact: in large samples, the probability that any lower bound is invalid converges to 5%.




(a) (b)

Figure 1: In Figure 1a, we plot a 90% conformal prediction set constructed using a quantile random forest on the synthetic dataset studied in (Romano et al., 2019). Figure 1b displays the prediction set's conditional coverage over all sub-intervals of a given width. The dashed line is the observed coverage, while the solid line plots the simultaneous lower bound obtained via Algorithm 2 ($\alpha = 0.05$, $p_* = 0.01$, $w_0 = \infty$).

In this example, the certification audit produces simultaneously valid lower bounds over a particular finite collection of sub-intervals. If the auditor has a sufficiently large computational budget, our method allows for richer collections of sub-intervals, up to and including the collection of *all* sub-intervals. See Section 3 for a complete description of our procedure for producing lower bounds.

#### 2.2.2 Flagging false positive rate disparities

Consider a district court evaluating whether the COMPAS recidivism prediction algorithm is biased. While the most notable previous work considers subpopulations defined by race (Angwin et al., 2016), the court is likely to be interested in discrimination against any groups formed by the intersections of legally protected attributes, e.g., age, gender, ethnicity. Over this larger collection of subpopulations, identifying the existence of some disparity is no longer of interest, but accurately localizing severe disparities is of great importance.

Following Angwin et al. (2016), we apply our auditing method to a data set obtained by ProPublica in 2016 that includes COMPAS risk scores ($f(X) \in \{\text{low-risk}, \text{high-risk}\}$), defendant demographics ($X_d$), and true recidivism outcomes after two years ($Y \in \{0, 1\}$) for $n = 6781$ individuals. In Figure 2, we shade in red the demographic groups flagged for having disparate false positive rates, i.e., those $G$ for which

$$\mathbb{P}(f(X) = \text{high-risk} \mid Y = 0, X_d \in G) - \mathbb{P}(f(X) = \text{high-risk} \mid Y = 0) > 0.05.$$

We can place this task in our auditing framework by defining $\epsilon(G)$ with

$$L(f(x), y) = \mathbf{1}\{f(x) = \text{high-risk}\} \quad \text{and} \quad \theta_P = \mathbb{P}(f(X) = \text{high-risk} \mid Y = 0)$$

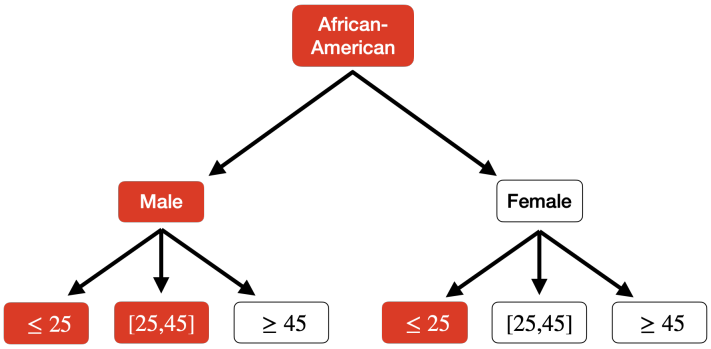



Figure 2: For the COMPAS algorithm, we audited 58 subpopulations formed by intersections of race, gender, and age. Here we plot subgroups of the African-American subpopulation. Boxes shaded in red denote groups flagged as having at least 5% higher-than-average false positive rates.

for all intersectional subgroups with $Y = 0$. Each flag then corresponds to a rejection of the null hypothesis $H_0(G) : \epsilon(G) \leq 0.05$.

To account for the inevitability of finding some false positive rate disparity when auditing over many groups, our method issues a formal guarantee on the rate at which the issued flags are invalid. Since falsely flagging a performance disparity is less consequential than false certification of fairness, we provide a less conservative guarantee than the simultaneous validity of the previous example. Instead, we provide an asymptotically valid upper bound on the "false discovery rate," i.e., the expected proportion of flags that are falsely issued. For the plotted example, we apply the flagging method described in Section 4 to control this proportion at 10%.

## 3. Certifying performance

### 3.1 Methods

#### 3.1.1 Bound certification

We first consider the problem of *certifying* subpopulations by providing a simultaneously valid confidence set for $\epsilon(G)$. To simplify our exposition, we construct a simultaneously valid *lower bound* on the group-wise disparity, i.e., we define $\epsilon_{\text{lb}}(G)$ such that

$$\lim_{n\to\infty} \mathbb{P}(\epsilon_{\text{lb}}(G) \leq \epsilon(G) \text{ for all } G \in \mathcal{G}) = 1 - \alpha. \tag{2}$$

Though it may seem counterintuitive to lower bound a performance disparity, recall the conditional coverage example previewed in Section 2.2.1. Upper confidence bounds and intervals for $\epsilon(G)$ are obtained via a trivial modification described in Appendix B.3.

Naively, we might define $\epsilon_{\text{lb}}(G)$ using an upper bound on the maximum deviation between our performance disparity estimator and the true disparity, i.e., the $(1-\alpha)$-quantile of $\sup_{G\in\mathcal{G}}\{\hat{\epsilon}(G) - \epsilon(G)\}$. Formally defining $\text{Quantile}(\alpha; X) := \inf\{x : \alpha \leq \mathbb{P}(X \leq x)\}$, we

could obtain a simultaneously valid lower bound on $\epsilon(G)$ via

$$\hat{\epsilon}(G) - \text{Quantile}\left(1-\alpha; \sup_{G\in\mathcal{G}}\{\hat{\epsilon}(G) - \epsilon(G)\}\right). \tag{3}$$

While it is straightforward to prove that (2) holds for the proposed bound, there are two crucial problems. First, the quantile in (3) cannot be estimated accurately: $\hat{\epsilon}(G)$ diverges for small groups, so when $\mathcal{G}$ is large, $\hat{\epsilon}(G)$ does not converge uniformly to $\epsilon(G)$. As a consequence, standard asymptotic methods (e.g., bootstrap) will fail. Second, even if we could estimate this quantile, we would obtain a constant correction to the naive estimator for all groups. Since any such correction must be large to achieve simultaneously validity over small groups, this approach would lead to impractically conservative lower bounds.

To circumvent the first of these obstacles, we show that it is possible to consistently estimate the distribution of

$$\sup_{G\in\mathcal{G}}\{\mathbb{P}(G)\cdot\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon(G))\}. \tag{4}$$

Intuitively, multiplying the naive process by $\mathbb{P}(G)\cdot\mathbb{P}_n(G)$ stabilizes its value for small groups. We then apply the bootstrap (Efron, 1979; Giné and Zinn, 1990) to estimate the $(1-\alpha)$-quantile of this process. Rigorously establishing bootstrap consistency requires a technical argument; see the proof of Theorem 4 for a detailed exposition.

Mimicking our initial approach, we use an estimate of the $(1-\alpha)$-quantile of (4) to construct a simultaneously valid lower bound on the true disparity. For all $G\in\mathcal{G}$, we define $\epsilon_{\text{lb}}(G)$ such that $\mathbb{P}_n(G)^2\cdot(\hat{\epsilon}(G)-\epsilon_{\text{lb}}(G))$ equals the $(1-\alpha)$-quantile of (4). Letting $t^*$ denote the bootstrap estimate of this quantile, i.e., the output of Algorithm 1, we obtain a simplified definition of $\epsilon_{\text{lb}}(G)$:

$$\epsilon_{\text{lb}}(G) := \hat{\epsilon}(G) - \frac{t^*}{\mathbb{P}_n(G)^2}. \tag{5}$$

Given a valid estimate of $t^*$, the simultaneous validity of $\epsilon_{\text{lb}}(G)$ is a straightforward implication of our definition:

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}(\epsilon_{\text{lb}}(G)\le\epsilon(G)\,\text{for all}\,G\in\mathcal{G}) &= \lim_{n\to\infty}\mathbb{P}(\hat{\epsilon}(G)-t^*/\mathbb{P}_n(G)^2\le\epsilon(G)\,\text{for all}\,G\in\mathcal{G})\\ &= \lim_{n\to\infty}\mathbb{P}\left(\sup_{G\in\mathcal{G}}\mathbb{P}_n(G)^2\cdot(\hat{\epsilon}(G)-\epsilon(G))\le t^*\right).\end{aligned}$$

The first equality follows from our definition of $\epsilon_{\text{lb}}(G)$, and the second follows from rearrangement. Replacing $\mathbb{P}_n(G)$ with $\mathbb{P}(G)$ (by Slutsky's lemma) and applying the definition of $t^*$ then completes our argument for simultaneous validity:

$$\begin{aligned}&\lim_{n\to\infty}\mathbb{P}\left(\sup_{G\in\mathcal{G}}\mathbb{P}_n(G)^2\cdot(\hat{\epsilon}(G)-\epsilon(G))\le t^*\right)\\ &\qquad= \lim_{n\to\infty}\mathbb{P}\left(\sup_{G\in\mathcal{G}}\mathbb{P}(G)\cdot\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon(G))\le t^*\right)=1-\alpha.\end{aligned}$$

**Algorithm 1** Bootstrapping the lower confidence bound critical value

**Input:** Subpopulations $\mathcal{G}$, holdout set $\mathcal{D}$, level $\alpha$, number of bootstrap samples $B$
**for** $b = 1, \ldots, B$ **do**
  Let $\mathcal{D}_b^*$ be a sample with replacement of size $n$ from $\mathcal{D}$;
  Define $\mathbb{P}_b^*(G) := \frac{1}{n} \sum_{(x_i^*, y_i^*) \in \mathcal{D}_b^*} \mathbf{1}\left\{(x_i^*, y_i^*) \in G\right\}$;
  Define $\epsilon_b^*(G) := \frac{1}{\mathbb{P}_b^*(G) \cdot n} \sum_{(x_i^*, y_i^*) \in \mathcal{D}_b^*} L_i^* - \hat{\theta}(\mathcal{D}^*)$;
  $t^{(b)} = \max_{G \in \mathcal{G}} \left\{\mathbb{P}_n(G) \cdot \mathbb{P}_b^*(G) \cdot (\epsilon_b^*(G) - \hat{\epsilon}(G))\right\}$;
**end for**
**Return:** $t^* = \text{Quantile}(1 - \alpha; \{t^{(b)}\}_{b=1}^B)$

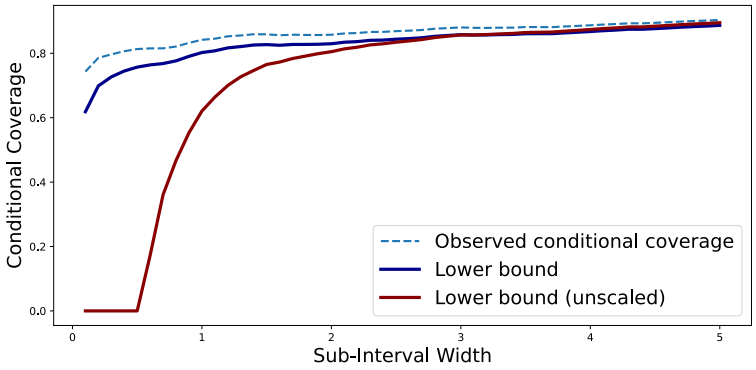



Figure 3: We reproduce the plot from Figure 1b. Here the red curve denotes the lower bound on conditional coverage obtained via (5). For smaller groups, it is substantially looser than the blue curve, i.e., the bound in Figure 1b. The blue curve is obtained by rescaling the bootstrap process using $p_* = 0.01$ and $w_0 = \infty$.

While this bound satisfies the validity condition given by (2), directly applying (5) leads to practically unusable lower bounds on $\epsilon(G)$ for small groups. This is caused by a suboptimal dependence on group size in (5). Asymptotically, our bound converges to

$$\epsilon_{\text{lb}}(G) = \hat{\epsilon}(G) - C_0 \cdot \left(\sqrt{\frac{1}{|G| \cdot \mathbb{P}_n(G)^3}}\right),$$

where $C_0$ is some group-independent constant. Even though the bound correction converges to 0 at the expected $1/\sqrt{|G|}$ rate, the group-dependent factor of $\mathbb{P}_n(G)^3$ catastrophically inflates our confidence set for even moderately sized groups. Figure 3 reproduces Figure 1b using (5) and shows that this definition produces vacuous lower bounds for all but the largest sub-intervals.

To motivate our revised approach, we observe that the classical Wald confidence bound for $\epsilon(G)$ has the following asymptotic behavior,

$$\epsilon_{\text{lb}}^{\text{ideal}}(G) = \hat{\epsilon}(G) - C_1 \cdot \frac{\sigma_G}{\sqrt{|G|}}; \tag{6}$$

$C_1$ is group-independent and $\sigma_G$ denotes the asymptotic standard deviation of $\sqrt{|G|}(\hat{\epsilon}(G) - \epsilon(G))$. The behavior of the Wald confidence bound is advantageous in two ways. First, the width of the confidence bound decays at the desired $1/\sqrt{|G|}$ rate with no additional size-dependent factors. And second, it is sensitive to heteroskedasticity: the $\sigma_G$ term accounts for whether the estimated performance disparity is more variable in some groups compared to others.

We can construct Wald-type bounds for all sufficiently large groups by rescaling the bootstrapped process. To understand how to define this scaling factor, we first study how rescaling the bootstrap process affects the resulting confidence bound. Let $s(G)$ and $\hat{s}(G)$ denote the population value and estimator of the scaling factor, respectively. Then, instead of bootstrapping the $(1-\alpha)$-quantile of $\sup_{G\in\mathcal{G}}\{\mathbb{P}(G)\cdot\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon(G))\}$, we estimate the $(1-\alpha)$-quantile of

$$\sup_{G\in\mathcal{G}}\left\{\frac{1}{s(G)}\cdot\mathbb{P}(G)\cdot\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon(G))\right\}. \tag{7}$$

Letting $t^*$ denote the $(1-\alpha)$-quantile of (7), we mimic the previous derivation and obtain a closed-form solution for $\epsilon_{\text{lb}}(G)$ in terms of $\hat{\epsilon}(G)$ and $\hat{s}(G)$,

$$\epsilon_{\text{lb}}(G) := \hat{\epsilon}(G) - t^*\cdot\frac{\hat{s}(G)}{\mathbb{P}_n(G)^2}. \tag{8}$$

Since $t^* = O_P(1/\sqrt{n})$, we can match the Wald-type bound in (6) by choosing $\hat{s}(G)$ to estimate $\mathbb{P}(G)^{3/2}\cdot\sigma_G$.

Naively rescaling the process reproduces the divergence for small groups problem we encountered in our first approach. In particular, bootstrap consistency requires $\{1/\hat{s}(G)\}_{G\in\mathcal{G}}$ to be a *uniformly consistent* estimator of $\{1/s(G)\}_{G\in\mathcal{G}}$. To this end, when $\mathbb{P}_n(G)$ is small, we forgo estimating $\mathbb{P}(G)^{3/2}\cdot\sigma_G$ exactly. Since we cannot consistently estimate the conditional variance of $L$ for arbitrarily small groups, we shrink our estimator towards a more naive but consistently estimable quantity: the population variance of our performance metric. Namely, as the group size goes to 0, we scale by the more naive estimator $p_*^{3/2}\cdot\sqrt{\widehat{\text{Var}}(L)}$ where $\widehat{\text{Var}}(L)$ denotes the sample variance.

To rigorously interpolate between these two group-size regimes, we set

$$\hat{s}(G) = \max\{\mathbb{P}_n(G), p_*\}^{3/2}\cdot\left(\frac{\mathbb{P}_n(G)}{\mathbb{P}_n(G)+w_0}\cdot\hat{\sigma}_G + \frac{w_0}{\mathbb{P}_n(G)+w_0}\cdot\widehat{\text{Var}}(L)^{1/2}\right), \tag{9}$$

where $w_0 > 0$ is a user-specified hyperparameter that controls the degree of shrinkage and $\hat{\sigma}_G$ is any point-wise, but not necessarily uniformly, consistent estimator of $\sigma_G$. We define such an estimator on Line 7 of Algorithm 2 by expanding the asymptotic variance of $\hat{\epsilon}(G)$ and replacing each term in the expansion by its plug-in estimator. For a more formal derivation, see the proof of Theorem 13 in the Appendix. By shrinking our estimate of the asymptotic variance of $\hat{\epsilon}(G)$ to a group-independent quantity, $\hat{s}(G)$ obtains the desired uniform consistency. In Algorithm 2, we show how to compute the $t^*$ used in (8).

In practice, we observe that estimating the asymptotic variance of $\hat{\epsilon}(G)$ can harm the finite-sample validity of $\epsilon_{\text{lb}}(G)$. We thus recommend setting $w_0 = \infty$ unless adaptivity to

**Algorithm 2** Bootstrapping the (rescaled) lower confidence bound critical value

**Input:** Subpopulations $\mathcal{G}$, holdout set $\mathcal{D}$, level $\alpha$, threshold $p_*$, weight $w_0$, number of bootstrap samples $B$
**for** $b = 1, \ldots, B$ **do**
  Let $\mathcal{D}_b^*$ be a sample with replacement of size $n$ from $\mathcal{D}$;
  Define $\mathbb{P}_b^*(G) := \frac{1}{n} \sum_{(x_i^*, y_i^*) \in \mathcal{D}_b^*} \mathbf{1}\{(x_i^*, y_i^*) \in G\}$;
  Define $\epsilon_b^*(G) := \frac{1}{\mathbb{P}_b^*(G) \cdot n} \sum_{(x_i^*, y_i^*) \in G} L_i^* - \hat{\theta}(\mathcal{D}_b^*)$;
**end for**
Define the asymptotic variance estimator by

$$\hat{\sigma}_G^2 := \widehat{\mathrm{Var}}(L \mid G) + \mathbb{P}_n(G) \left( \widehat{\mathrm{Var}}(\psi) - 2 \cdot \widehat{\mathrm{Cov}}(L, \psi \mid G) \right)$$

where $\widehat{\mathrm{Var}}(\cdot)$ and $\widehat{\mathrm{Cov}}(\cdot)$ correspond to the sample (conditional) variance and covariance;
Define $\hat{s}(G)$ by (9);
**for** $b = 1, \ldots, B$ **do**
  $t^{(b)} = \max_{G \in \mathcal{G}} \{ \frac{1}{\hat{s}(G)} \cdot \mathbb{P}_n(G) \cdot \mathbb{P}_b^*(G) \cdot (\epsilon_b^*(G) - \hat{\epsilon}(G)) \}$;
**end for**
**Return:** $t^* = \mathrm{Quantile}(1 - \alpha; \{t^{(b)}\}_{b=1}^B)$

the variance of $\hat{\epsilon}(G)$ is deemed critical. We also emphasize that our approach to rescaling the process is only one heuristic for improving the adaptivity of the constructed confidence bounds. Any uniformly consistent estimator $\hat{s}(G)$ that is bounded away from 0 would suffice.

Using the output of Algorithm 2 in (8), we obtain more practical confidence bounds. We produce the blue curve in Figure 3 using $p_* = 0.01$ and $w_0 = \infty$. There is no free lunch: observe that the red curve (unscaled) yields a tighter lower confidence bound for the largest sub-interval widths. Nevertheless, it is clear that the rescaled process produces a usable lower bound over a much wider range of group sizes.

Theorem 4 states sufficient conditions for bootstrap consistency and, therefore, simultaneous validity of the lower bounds defined in (5) or (8).

**Theorem 4 (Simultaneous validity)** *Assume that $L$ is bounded and that $L - \hat{\theta}$ is non-constant over at least one non-empty group. Further assume that $\mathcal{G}$ has finite Vapnik-Chernovenkis (VC) dimension. Then, for either definition presented, $\epsilon_{\mathrm{lb}}(G)$ is an asymptotic $(1-\alpha)$-lower confidence bound for $\epsilon(G)$ that is simultaneously valid for all $G \in \mathcal{G}$, i.e.,*

$$\lim_{n \to \infty} \mathbb{P}\left(\epsilon_{\mathrm{lb}}(G) \leq \epsilon(G) \text{ for all } G \in \mathcal{G}\right) = 1 - \alpha.$$

We make two remarks regarding our assumptions.

**Remark 5** *Our restriction that $\mathcal{G}$ has finite VC dimension is satisfied by most interpretable collections of groups: intervals, rectangles, halfspaces, etc.*

**Remark 6** *In typical fairness applications, $L$ is $\{0,1\}$-valued and satisfies the boundedness assumption; for unbounded metrics, the auditor might truncate $L$ or, if appropriate, assume*

*compactness of the domain. Furthermore, if $\mathcal{G}$ is a finite collection, we may relax this assumption to $Var(L) < \infty$.*

Even though Theorem 4 is an asymptotic result, a finite-sample approach, e.g., via empirical process concentration, can only satisfy a conservative coverage guarantee. By contrast, the simultaneous coverage of $\epsilon_{\text{lb}}(\cdot)$ converges to *exactly* $1-\alpha$ under any data-generating distribution. Even at small sample sizes, we show in Section 3.2 that the gap between nominal and realized coverage is minimal.

#### 3.1.2 Boolean certification

Next, we consider issuing a Boolean certificate for $G$ if $\epsilon(G)$ lies above some pre-specified tolerance $\epsilon$. The trivial extension of our methods to certifying $\epsilon(G) < \epsilon$ and $|\epsilon(G)| < \epsilon$ is described in Appendix B.3. While this approach returns strictly less information to the auditor compared to the confidence bound certificate, it offers computational benefits and can be more powerful in certain settings.

Formally, certifying $\epsilon(G) > \epsilon$ corresponds to testing the null hypothesis $\bar{H}_0(G) : \epsilon(G) \leq \epsilon$. We require that, in large samples, all issued certificates are simultaneously valid with probability $1-\alpha$. Equivalently,

$$\lim_{n\to\infty} \mathbb{P}\left(\text{there exists any falsely certified } G \in \mathcal{G}\right) \leq \alpha.$$

In the language of multiple testing, this desideratum is (asymptotic) strong family-wise error rate (FWER) control.

To construct such a test, we show that the bootstrap can conservatively estimate the $(1-\alpha)$-quantile of $\sup_{G\in\mathcal{G}}\{\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon)\}$. Letting $t^*$ denote the estimate output by Algorithm 3, we certify that $\epsilon(G) > \epsilon$ if $\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon) \geq t^*$. Simplifying, we reject the null when

$$\hat{\epsilon}(G) \geq \epsilon + \frac{t^*}{\mathbb{P}_n(G)}. \tag{10}$$

For example, if $L$ is the coverage indicator, we certify $G$ if the empirical conditional coverage on $G$ is sufficiently high.

Even though the scaling of the rejection threshold in (10) is sub-optimal for small groups, correcting this would eliminate the singular advantage of the Boolean certification procedure: efficient optimization over certain infinite group collections, e.g., intervals, slabs. Observe that

$$\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon) = \frac{1}{n}\sum_{i=1}^{n}(L_i - \hat{\theta} - \epsilon)\mathbf{1}\left\{(x_i, y_i) \in G\right\},$$

i.e., the process is linear in the group-indicator. Rescaling by $\hat{s}(G)$ does away with this linearity, and line 4 of Algorithm 3 then requires brute-force search over all $G \in \mathcal{G}$. For the sake of completeness, we describe this rescaled variant of the Boolean certification procedure in Algorithm 7.

Theorem 7 states that (10) produces valid certificates under the mild assumptions of Theorem 4.

**Algorithm 3** Bootstrapping the Boolean certificate critical value

**Input:** Subpopulations $\mathcal{G}$, disparity $\epsilon$, holdout set $\mathcal{D}$, level $\alpha$, number of bootstrap samples $B$
**for** $b = 1, \ldots, B$ **do**
    Let $\mathcal{D}_b^*$ be a sample with replacement of size $n$ from $\mathcal{D}$;
    $t^{(b)} = \max_{G \in \mathcal{G}} \{\mathbb{P}_b^*(G) \cdot (\epsilon_b^*(G) - \epsilon) - \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon)\}$;
**end for**
**Return:** $t^* = \text{Quantile}(1 - \alpha; \{t^{(b)}\}_{b=1}^B)$

**Theorem 7 (FWER control for certification)** *Retain the assumptions of Theorem 4. Then, the certificates issued by* (10) *satisfy*

$$\lim_{n \to \infty} \mathbb{P}\left(\text{there exists any falsely certified } G \in \mathcal{G}\right) \leq \alpha.$$

Our bootstrap procedure accounts for overlap between subpopulations and improves upon more naive FWER-controlling procedures such as the Bonferroni test. However, the asymptotic FWER is only exactly $\alpha$ when $\epsilon(G) = \epsilon$ for all $G$. This compares unfavorably to the simultaneous confidence set guarantee, which promises exact Type I error control under any data-generating distribution. We compare and contrast the finite-sample validity of the Boolean and confidence set-based certification methods in the sequel.

### 3.2 Empirical results

#### 3.2.1 Synthetic validation

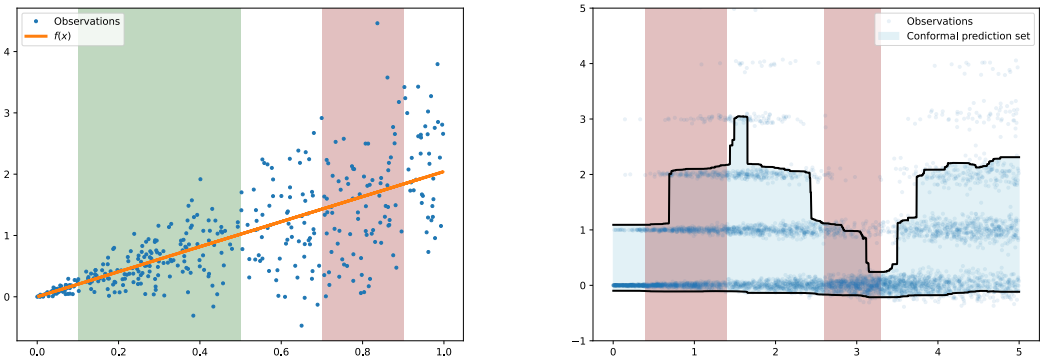
First, we verify that the (asymptotic) claims of Theorem 4 and Theorem 7 are accurate in finite samples. We consider three synthetic data experiments; the results from each are presented in Table 1.

In all of these experiments, we evaluate our method using 500 bootstrap samples; the computational complexity of the resulting procedure is non-trivial. While our procedure theoretically accommodates infinite collections of subpopulations, in practice, solving the optimization problems in Algorithm 3 and Algorithm 6 for each bootstrap sample is tantamount to solving a challenging 0-1 loss optimization problem hundreds of times. Even though a few infinite collections of subgroups, e.g., intervals and rectangles, admit efficient algorithms for Boolean certification, the confidence bound certification method realistically requires the auditor to limit their analysis to finite $\mathcal{G}$. Nevertheless, we show that it is possible to efficiently audit over large *finite* collections by first discretizing the space over which we define subpopulations. After this step, which still leads to $|\mathcal{G}| = 1275$ for the experiments considered in this subsection, running a certification audit at the largest sample size considered ($n = 1600$) takes under 7 seconds on a 2020 MacBook Pro.

We initially consider a homoskedastic linear model. We sample $(X_i, Y_i)$ from

$$X_i \overset{\text{iid}}{\sim} \text{Unif}(0, 1), \quad Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\beta_0 X_i, 1). \tag{11}$$

We then obtain $f(x) = \hat{\beta} x$ via ordinary least-squares on 1000 training points sampled from this distribution. The performance metric of interest is squared-error loss, i.e., $L(f(X), Y) =$




(a) (b)

Figure 4: Figure 4a displays a linear model over 400 holdout points generated from (12). Constructing a 90%-confidence bound for the expected MSE yields upper bounds of 0.5 (true MSE: 0.3) and 1.15 (true MSE: 0.80) for the green and red sub-intervals, respectively. Figure 4b displays prediction intervals from a conformalized quantile random forest on 400 holdout points from the synthetic data-generating process in Romano et al. (2019). Constructing a 90%-confidence bound for the conditional coverage yields lower bounds of 81.1% (true coverage: 92.4%) and 80.1% (true coverage: 89.5%) for the two red sub-intervals, respectively. These bounds are simultaneously valid over arbitrarily many such queries.

$(Y - f(X))^2$ and the target $\theta_P = 0$. Using held-out data sets of varying size, we issue Boolean certificates for sub-intervals $G \subseteq [0, 1]$ over which $\epsilon(G) < 1$.

In our second experiment, we validate our audit using a heteroskedastic linear model,

$$X_i \overset{\text{iid}}{\sim} \text{Unif}(0, 1), \quad Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\beta_0 X_i, X_i). \tag{12}$$

The model $f(X)$ and metric $L(f(X), Y)$ is obtained identically to the previous synthetic experiment. We then issue Boolean certificates for sub-intervals $G \subseteq [0, 1]$ over which $\epsilon(G) < \epsilon$ for $\epsilon \in \{0.4, 0.5\}$. To simplify the verification of issued certificates, we consider $G$ with endpoints belonging to $\{0, 0.1, \ldots, 1\}$. Figure 4a displays a trial experiment with an holdout set of sample size 400. For this setting, we observe that the nominal error rate overestimates the realized probability of false certification.

Last, we consider the synthetic dataset introduced by Romano et al. (2019) and displayed in Figure 4b. Here, the conformalized quantile regression (CQR) is guaranteed to have 90% marginal coverage, so we consider certifying sub-intervals for which the coverage exceeds 90%, e.g. $\theta_P = 0.9$ and $L = \mathbf{1}\{Y \in \hat{C}(X)\}$.

Table 1 displays the FWER of the certification audit over 200 trials. We set the nominal error rate to 0.1 and vary the sample size $n$ over $(100, 200, 400, 800, 1600)$. The results corroborate the predictions made by our theory. For large $n$, the nominal FWER matches or exceeds the realized level. When the null hypothesis does not hold at the boundary, i.e., $\epsilon(G)$ is not approximately equal to the threshold $\epsilon$, the nominal FWER can substantially

| | | Sample size ($n$) | | | | |
|---|---|---|---|---|---|---|
| | | 100 | 200 | 400 | 800 | 1600 |
| Model (11) ($\epsilon = 1$) | FWER | 0.145 | 0.145 | 0.17 | 0.125 | 0.095 |
| Model (12) ($\epsilon = 0.5$) | FWER | 0.105 | 0.105 | 0.1 | 0.075 | 0.065 |
| | Power | 0.145 | 0.268 | 0.411 | 0.562 | 0.709 |
| Model (12) ($\epsilon = 0.4$) | FWER | 0.045 | 0.04 | 0.015 | 0.025 | 0.02 |
| | Power | 0.036 | 0.079 | 0.21 | 0.375 | 0.57 |
| CQR ($\epsilon = 0.9$) | FWER | 0.081 | 0.034 | 0.019 | 0.012 | 0.003 |
| | Power | 0.019 | 0.008 | 0.009 | 0.02 | 0.062 |

Table 1: FWER and power of certificates issued by Algorithm 3 with $B = 500$ and $\alpha = 0.1$. All results are based on 200 trials. We see that, for large $n$, the simultaneous validity guarantee holds.

overestimate the true level. Table 1 also displays the power, i.e., the proportion of sub-intervals achieving the certification threshold that are actually certified by our method.

By contrast, recall that the confidence bound approach to certification guarantees (asymptotically) exact coverage under any data-generating distribution. To verify this claim, we construct simultaneous 90% confidence bounds for each of the aforementioned synthetic experiments.

Table 2 summarizes our results: we observe that Algorithm 1 and its rescaled variant Algorithm 2 ($p_* = 0.01, w_0 = \infty$) obtain the nominal coverage in large samples. For a fixed threshold $\epsilon$, we define the power of a confidence set as the proportion of sub-intervals with true error rates below $\epsilon$ for which the confidence set excludes $\epsilon$. Note that while the unscaled confidence bounds have low power, i.e., they are less likely to exclude the targeted error level, rescaling largely mitigates this issue. Indeed, even though these confidence bounds do not require a priori specification of a certification threshold, $\epsilon$, their power to certify error rates below some fixed $\epsilon$ is only slightly inferior to the Boolean certificates' in Table 1. This performance is attributable to the exact Type I error control of the bound certification method, as compared to the practically conservative error control of the Boolean certification approach.

### 3.2.2 Certifying COMPAS

Next, we reconsider previous analyses of the COMPAS recidivism prediction instrument (RPI) and show how our methods can be used to establish rigorous guarantees. The COMPAS algorithm assigns defendants risk scores ranging from 1 to 10 based on an estimated likelihood of re-offending. Prior work showed that African-American defendants are more likely to be mis-classified as high-risk when compared to Caucasian defendants (Angwin et al., 2016). In response to this finding, the creators of COMPAS, Northpointe Inc., ar-

| | | Sample size ($n$) | | | | |
|---|---|---|---|---|---|---|
| | | 100 | 200 | 400 | 800 | 1600 |
| Model (12) | Coverage | 0.84 | 0.83 | 0.86 | 0.895 | 0.905 |
| | Power ($\epsilon = 0.5$) | 0.043 | 0.081 | 0.15 | 0.256 | 0.378 |
| | Power ($\epsilon = 0.4$) | 0.003 | 0.005 | 0.012 | 0.058 | 0.187 |
| Model (12) (rescaled) | Coverage | 0.84 | 0.81 | 0.845 | 0.885 | 0.88 |
| | Power ($\epsilon = 0.5$) | 0.149 | 0.270 | 0.427 | 0.608 | 0.743 |
| | Power ($\epsilon = 0.4$) | 0.038 | 0.091 | 0.24 | 0.457 | 0.633 |
| CQR (rescaled) | Coverage | 0.815 | 0.87 | 0.845 | 0.865 | 0.91 |
| | Power ($\epsilon = 0.9$) | 0.025 | 0.011 | 0.007 | 0.017 | 0.055 |

Table 2: Simultaneous coverage of confidence bounds issued by Algorithm 1 and Algorithm 2 with $B = 500$, $\alpha = 0.1$, and $p_* = 0.01$. All results are based on 200 trials. We see that, for large $n$, the certification procedure satisfies the simultaneous validity guarantee.

gued that the algorithm is fair when evaluated by the predictive parity criterion (Dieterich et al., 2016; Flores et al., 2016). While they provide statistical evidence for the absence of a significant racial bias by this measure, our methods allow for the construction of an explicit bound on the true disparity.

The predictive parity criterion is satisfied for a single group, $G$, when the positive predictive value (PPV) of $f$ for $G$ matches the PPV for the complement of $G$, i.e.,

$$\mathbb{P}(Y = 1 \mid f(X) = 1, X \in G) = \mathbb{P}(Y = 1 \mid f(X) = 1, X \in G^c).$$

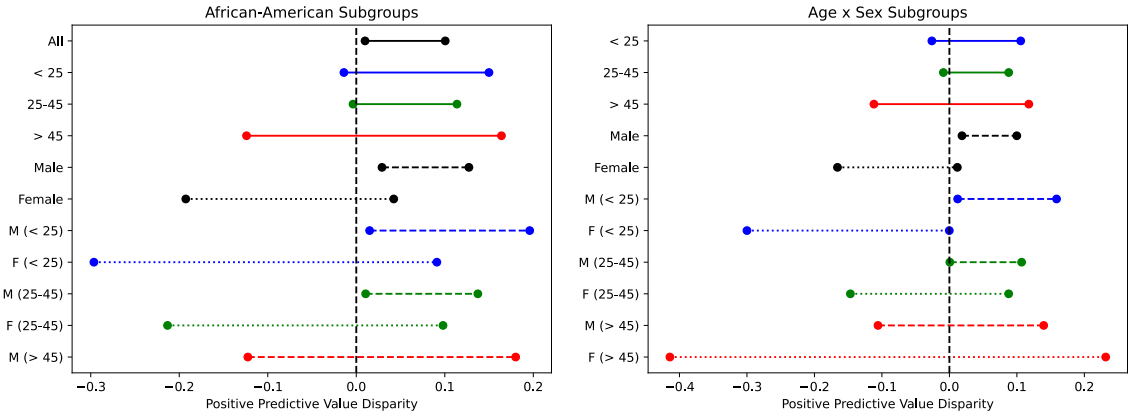
Intuitively, the PPV measures how informative a positive prediction is. For example, if COMPAS classifies a defendant as high-risk ($f(X) = 1$), the PPV corresponds to the probability that they actually recidivate ($Y = 1$).

Following prior work, we binarize the COMPAS scores by defining $f(X)$ to be 1 when the RPI score is 5 or higher. Then, to certify a lower bound on the gap between an African-American and Caucasian defendant's PPVs, we consider the subset ($n = 2525$) of the holdout set with $f(X) = 1$ and $X_{\text{race}} \in \{\text{African-American}, \text{Caucasian}\}$. We instantiate our audit with $L$ corresponding to the indicator that $Y$ matches $f(X)$, $\mathcal{G}$ containing just one group (African-American defendants), and $\theta_P$ denoting the PPV for White defendants:

$$L(f(X), Y) = Y, \quad \mathcal{G} = \{\{(X, Y) \mid X_{\text{race}} = \text{African-American}, f(X) = 1\}\},$$
$$\theta_P = \mathbb{E}[Y = 1 \mid X_{\text{race}} = \text{Caucasian}].$$

We estimate $\theta_P$ using the empirical conditional expectation, $\hat{\theta} = \hat{\mathbb{E}}_n[Y = 1 \mid X_{\text{race}} = \text{Caucasian}]$.

To verify the claim made by Dieterich et al. (2016) that there is no reduction in PPV for African-American defendants relative to Caucasian defendants, we construct a 90%-lower confidence bound for $\epsilon(G) := \mathbb{P}(Y = 1 \mid f(X) = 1, X \in G) - \mathbb{P}(Y = 1 \mid f(X) = 1, X \in G^c)$.




(a) (b)

Figure 5: In Figure 5a, we plot 90% confidence intervals for the difference in COMPAS PPV between each African-American subgroup and the entire Caucasian subgroup. In Figure 5b, we plot 90% confidence intervals for the difference in COMPAS PPV between each subgroup formed by intersections of age and sex and the entire Caucasian subgroup.

Using this approach, we can certify the previous claim: among defendants receiving high-risk predictions, an African-American defendant is at least 1.87% more likely to recidivate than a Caucasian defendant.

Our methodology is not essential to establishing this result since we only consider a single group. With these tools, however, we can establish PPV disparity bounds that hold simultaneously over every protected subgroup. Again using the COMPAS PPV on Caucasian defendants as our target, we construct simultaneously valid 90% confidence intervals on the PPV disparity for every subpopulation formed by the intersection of race, sex, and age ($|\mathcal{G}| = 58$). For this collection, the computational burden of this procedure is minimal: each audit takes under one second on a 2020 Macbook Pro. In Figure 5, we plot the intervals corresponding to subgroups of the African-American subpopulation and subgroups formed by intersections of sex and age alone. Our results validate Northpointe Inc.'s claims of PPV parity for several, albeit not all, African-American subpopulations. More generally, younger male subpopulations appear to have higher COMPAS PPV when compared to the Caucasian subpopulation.

## 4. Flagging performance disparities

### 4.1 Methods

In this section, we consider the other major subtask in fairness auditing: identifying subpopulations for which the predictive model exhibits substantial inaccuracy. Formally, we study the problem of *flagging* subpopulations for which the disparity exceeds some tolerance, i.e., when

$$H_0(G) : \epsilon(G) \leq \epsilon. \tag{13}$$

fails to hold. False flags are less problematic than false certificates of performance, so a weaker notion of error control than simultaneous validity suffices. A-priori, we expect most groups to satisfy the null hypothesis given by (13). This leads us to consider a notion of error rate that is defined relative to the number of flags issued instead of the number of subgroups tested; in particular, we control the asymptotic false discovery rate (FDR). In our setting, controlling the false discovery rate translates to upper bounding the expected proportion of falsely flagged subpopulations by $\alpha$. For instance, in Section 2.2.2, we controlled the FDR at 10%, implying that approximately 90% of the flagged subgroups have a false positive rate at least 5% higher than the population average.

Formalizing this criterion, we require

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{|\{\text{falsely flagged } G \in \mathcal{G}\}|}{|\{\text{flagged } G \in \mathcal{G}\}| \vee 1}\right] \leq \alpha.$$

Simultaneous error control (familywise-error rate control) implies control of the false discovery rate, making the latter a strictly weaker criterion.

To contextualize our work, we remark that there are many existing methods for computing performance disparities over various collections of subpopulations $\mathcal{G}$ (Saleiro et al., 2018; Zhang and Neill, 2016). However, it is unclear how we should interpret these naive estimates. If we search over many subpopulations, we should expect to observe disparities even if the underlying prediction rule treats each group equitably. Our method shows how one might prune the set of analyzed subpopulations to a shortlist where the proportion of falsely flagged groups is controlled.

To accomplish this task for a finite collection of subpopulations, we apply the well-known Benjamini-Hochberg procedure (denoted by $\mathsf{BH}(\alpha)$ in the sequel). This procedure takes as input a set of p-values that can be computed via the bootstrap for each tested hypothesis. Algorithm 4 describes how to compute bootstrap p-values for the flagging null hypothesis given by (13).

**Algorithm 4** Constructing p-values for $G \in \mathcal{G}$

**Input:** Subpopulations $\mathcal{G}$, holdout set $\mathcal{D}$, bootstrap samples $B$, tolerance $\epsilon$
**for** $b = 1, \ldots, B$ **do**
  Let $\mathcal{D}_b^*$ be a sample with replacement of size $n$ from $\mathcal{D}$;
  **for** $G \in \mathcal{G}$ **do**
    $t^{(b)}(G) = \epsilon_b^*(G) - \hat{\epsilon}(G)$;
  **end for**
**end for**
**for** $G \in \mathcal{G}$ **do**
  $s^*(G) = \frac{1}{\Phi(3/4)} \cdot \text{Quantile}(0.5; \{|t^{(b)}(G)|\}_{b=1}^B)$;
  $p(G) = 1 - \Phi\left((\hat{\epsilon}(G) - \epsilon)/s^*(G)\right)$;
**end for**
**Return:** $\{p(G)\}_{G\in\mathcal{G}}$.

The following proposition states that the $\mathsf{BH}(\alpha)$ procedure (applied to the output of Algorithm 4) controls the asymptotic false discovery rate under two cases of practical interest:

first, the case of mutually disjoint groups, and second, the case of binary outcomes with arbitrarily overlapping group structure. The validity of the procedure in the first case is not surprising since the p-values for disjoint groups are independent. The proof of the second case is more subtle: we show that the binary metric implies a certain positive dependency[1] among the p-values. Given this correlation structure, the BH procedure is known to be valid (Benjamini and Yekutieli, 2001).

**Proposition 8 (FDR control)** *Assume that* $\mathbb{P}(G)$ *and* $\mathrm{Var}\left(L \mid G\right)$ *are bounded away from* 0 *for all* $G \in \mathcal{G}$, $\theta_P$ *is a-priori known, and that at least one of the following conditions holds:*

*(i)* $\{G\}_{G\in\mathcal{G}}$ *are mutually disjoint;*
*(ii)* $L$ *takes values in* $\{0, 1\}$.

*If we flag the rejections of the* BH($\alpha$) *procedure on* $\{p(G)\}_{G\in\mathcal{G}}$, *then the false discovery rate is asymptotically controlled at level* $\alpha$.

We expect Theorem 8 to remain valid under violations of the stated assumptions. Even outside of the two cases stated in Theorem 8 (e.g., when we must estimate $\theta_P$), prior experiments with the BH algorithm and our own empirics suggest that this procedure will not violate FDR control (Fithian and Lei, 2022).

If flagging with FDR control over an infinite collection of subpopulations is desired, we suggest a generalization of the two-step procedure outlined in the independently-derived work of von Zahn et al. (2023). First, split the holdout set and use the first split to discover a finite sub-collection of interpretable subpopulations, e.g., von Zahn et al. (2023) fit a regression tree and let each leaf define a subpopulation of interest. Then, for the subpopulations identified with the first split, run the flagging procedure validated by Theorem 8 on p-values computed using the second split.

### 4.2 Empirical results

#### 4.2.1 Folktables

We evaluate the flagging methodology on an income prediction dataset derived from the 2018 Census American Community Survey Public Use Microdata and made available in the Folktables package (Ding et al., 2021; Flood, 2015). Using the California data set filtered to over-16 individuals who earned at least $100 in the past year, we aim to predict whether an individual's income exceeds $50,000. We include age, place of birth, education, race, marital status, occupation, sex, race, and hours worked in the fitted prediction rule.

To validate the (asymptotic) FDR control result in Theorem 8, we fit logistic and linear regression models to a training set of 1000 data points, and then sample holdout sets of varying size from the remaining data. We flag subpopulations formed by the intersection of age, race, and gender ($|\mathcal{G}| = 89$) for which: (1) the misclassification rate is higher than a fixed threshold of 0.5, (2) the misclassification rate is higher than the population average error rate, (3) the mean-squared error (MSE) of the predicted income is higher than the population mean-squared error. Each of these tasks sheds light on the relevance

1. The positive dependency we identify is formally termed positive regression dependence on a subset (PRDS) (Benjamini and Yekutieli, 2001).

| | | Sample size ($n$) | | | | | |
|---|---|---|---|---|---|---|---|
| | Task | 100 | 200 | 400 | 800 | 1600 | 3200 |
| Folktables | 1 | 0.045 | 0.038 | 0.032 | 0.025 | 0.021 | 0.014 |
| | 2 | 0.003 | 0.005 | 0.003 | 0.004 | 0.003 | 0.004 |
| | 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.002 |

Table 3: FDR of flags issued with $B = 500$ and $\alpha = 0.1$ for the tasks described in Section 4.2.1. All results are based on 1000 trials. We see that, for any $n$, the FDR guarantee is conservative.

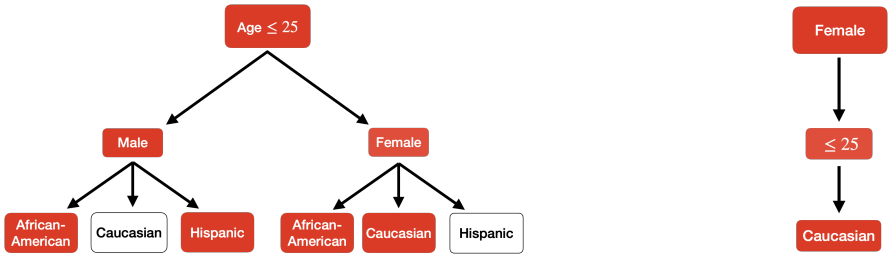
of Theorem 8. The first flagging task satisfies the assumptions of Theorem 8, while the other two violate the stated assumptions. Since we only consider a finite collection of subpopulations, these tasks are not computationally burdensome. For the largest sample size considered ($n = 3200$), the flagging audit considered takes under 1 second on a 2020 Macbook Pro.

Since the holdout sets are sampled with replacement, the data-generating distribution $P$ is the uniform distribution over the finite population of data held-out from model fitting. Therefore, a flag is falsely issued if the flagged subgroup's error rate on the entire held-out data set fails to exceed the stated threshold. Table 3 shows that over 1000 trials, the estimated FDR for each task is well below the nominal bound of 0.1 at every sample size tested. This is because the null p-values are, in practice, conservative, as the null hypothesis (13) rarely holds with equality.

### 4.2.2 COMPAS

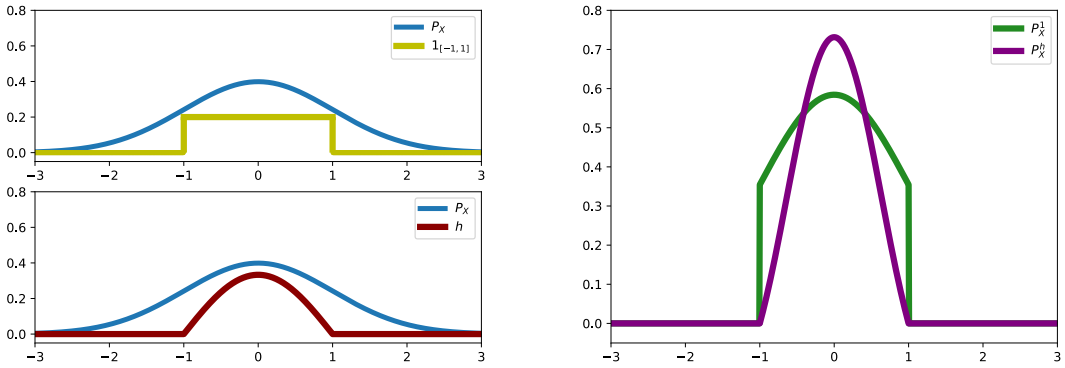
Prior analysis of the COMPAS RPI has shown that the false positive rate of the high-risk designation is substantially higher for African-American defendants compared to Caucasian defendants (Angwin et al., 2016). We revisit this often-studied example of fairness auditing to determine if we can identify any other demographic groups that suffer from false positive rates at least 5% higher than the average defendant. In particular, we audit over all intersections of race, sex, and age group ($n = 6781$, $|\mathcal{G}| = 48$). This flagging audit takes under 0.5 seconds on a 2020 Macbook Pro. Figure 2 plots the issued flags for subsets of the African-American subpopulation; we can further localize the false positive rate (FPR) disparity among African-American defendants to younger African-American defendants. As shown in Figure 6a, our method also flags nearly every under-25 subgroup, suggesting that this disparity affects young defendants more generally.

In the previous section, we *certified* that the positive predictive value (PPV) of the COMPAS RPI is higher for African-American defendants compared to Caucasian defendants. Since the COMPAS creators claim that PPV is a more appropriate measure of fairness (Dieterich et al., 2016), we investigate whether any other demographic groups suffer from harmful PPV disparities. As shown in Figure 6b, we are still able to flag certain subpopulations for having at least 5% lower PPV compared to the average.




(a) (b)

Figure 6: In Figure 6a, the red boxes correspond to groups flagged as having substantially higher-than-average false positive rates. Most under-25 subpopulations suffer from this disparity. In Figure 6b, the red boxes denote groups flagged as having substantially lower-than-average positive predictive values. Though we are able to flag only three nested subpopulations (Females, Females under 25, and Caucasian females under 25), these results suggest that certain subpopulations still face disparities according to this measure.




(a) (b)

Figure 7: In Figure 7a, we plot an indicator tilt that corresponds to conditioning on $X$ belonging to the interval $[-1, 1]$, and a non-negative tilt that resembles the indicator. Figure 7b plots the (similar) probability densities over $X$ produced by both tilts.

## 5. Beyond subpopulations

### 5.1 Methods

Auditing over subpopulations is equivalent to assessing performance over the class of distribution shifts indexed by the tilts, $\{\mathbf{1}\{(X, Y) \in G\}\}_{G \in \mathcal{G}}$. In particular,

$$\epsilon(G) = \mathbb{E}_P[L \mid G] - \theta_P = E_{P_G}[L] - \theta_P,$$

where $dP_G(x, y) \propto \mathbf{1}\{(x, y) \in G\}\, dP(x, y)$. In this section, we consider the natural generalization of these tilts from 0-1 valued functions to collections of non-negative functions. For

such a collection, $\mathcal{H}$, we might wish to bound the following discrepancy for each $h$,

$$\epsilon(h) = \mathbb{E}_{P_h}[L(f(X), Y)] - \theta_P,$$

where $dP_h(x, y) \propto h(x, y)dP(x, y)$. Figure 7 shows how non-negative functions that approximate an indicator can produce similar tilted distributions over $\mathcal{X}$.

To motivate this auditing task, consider the problem of assessing an autonomous vehicle's performance over a diverse set of environments. While most work on distributional robustness focuses on assessing performance under the worst-case shift belonging to some set, such an appraisal of model robustness is inherently limited (Duchi and Namkoong, 2021). The worst-case over some tractable collection of tilts is unlikely to be a useful proxy for most real-world distribution shifts. By contrast, our methods allow the modeler to query any shift belonging to the audited collection. As new environments are encountered, the modeler can reliably assess which environments may have benign effects on performance and which may be especially problematic.

When $\mathcal{H}$ is finite, our methods for certifying and flagging performance are unchanged. For infinite $\mathcal{H}$, our previous results still apply: we can certify performance discrepancies over any VC class of tilts. While this is by no means an exhaustive list, such classes include the set of non-negative polynomials and any collection of bounded monotone tilts over a single covariate (van der Vaart, 2000).

Generalizing from binary to non-negative tilts expands auditing to certain function classes with *infinite* VC dimension. We describe sufficient conditions for bootstrap-based auditing over general function classes $\mathcal{F}$ in Appendix D, but here we highlight one collection in particular: the unit ball of a reproducing kernel Hilbert space (RKHS).

Let $h$ denote any non-negative function belonging to the unit ball of a RKHS; we denote the collection of such functions by $\mathcal{H}_1^+$. Define $\epsilon(h)$ as the disparity under this tilt, i.e.,

$$\epsilon(h) := \mathbb{E}_{P_h}[L] = \frac{\mathbb{E}_P\left[(L - \theta_P)h(X)\right]}{\mathbb{E}_P[h(X)]}.$$

Let $\hat{\mathbb{E}}_n[f(X)]$ denote the plug-in estimator for the expectation of $f$ under the empirical distribution. Then, we define $\hat{\epsilon}(h) := (\hat{\mathbb{E}}_n[h(X)])^{-1}\hat{\mathbb{E}}_n[(L - \hat{\theta})h(X)]$. We will assume hereafter that $\theta_P = 0$, but a generalization of our approach to estimated targets is given in Appendix D.

We highlight two important characteristics of the RKHS auditing task. First, for a suitably chosen RKHS, $h$ can approximate *any* smooth tilt of the covariate distribution, $P_X$, defined over a compact subset (Micchelli et al., 2006). Given sufficient data, this allows us to issue guarantees on model performance for essentially arbitrary groups and covariate shifts.

Auditing over the RKHS unit ball offers another advantage: we can construct a confidence bound for $\epsilon(h)$ without the onerous optimization present at each step of Algorithm 1. Recall that in Algorithm 1, we used the bootstrap to estimate the $(1 - \alpha)$-quantile of $\sup_{G \in \mathcal{G}}\{\mathbb{P}_n(G) \cdot \mathbb{P}(G) \cdot (\hat{\epsilon}(G) - \epsilon(G))\}$. Each iteration then required solving a challenging combinatorial optimization problem over $\mathcal{G}$. At first glance, the analogous task for RKHS-based auditing appears even more difficult. For each bootstrap sample, we must

**Algorithm 5** Bootstrapping the RKHS confidence set critical value

**Input:** Kernel $k$, holdout set $\mathcal{D}$, level $\alpha$, bootstrap samples $B$
Define $\mathbf{L} := \{L(f(x_i), y_i)\}_{i=1}^n$;
Define $\mathbf{K} := \{k(x_i, x_j)\}_{i,j=1}^n$;
**for** $b = 1, \ldots, B$ **do**
    Sample $\mathbf{w} \sim \text{Mult}\left(n; \frac{1}{n}, \ldots, \frac{1}{n}\right)$;
    $\mathbf{A} = \frac{1}{n^2}\left((\mathbf{w} \odot \mathbf{L})\,\mathbf{1}^\top - \mathbf{w}\mathbf{L}^\top\right)$;
    $t^{(b)} = \lambda_{\max}\left(\mathbf{K}^{1/2}\left(\frac{\mathbf{A}+\mathbf{A}^\top}{2}\right)\mathbf{K}^{1/2}\right)$;
**end for**
**Return:** $t^* = \text{Quantile}(1-\alpha; \{t^{(b)}\}_{b=1}^B)$

compute

$$\max_{h \in \mathcal{H}_1^+} \{\hat{\mathbb{E}}_n[h(X)] \cdot \hat{\mathbb{E}}_b^*[h(X)] \cdot (\epsilon_b^*(h) - \hat{\epsilon}(h))\}.$$

Naively, this optimization problem is intractable.

In Algorithm 5, however, we are able to reduce this task to computing the top eigenvalue of a low-rank matrix. Two observations are crucial to this reduction. First, the supremum of the process indexed by $h \in \mathcal{H}_1^+$ is upper bounded by the supremum of the same process indexed by the unrestricted unit ball $\mathcal{H}_1$, i.e.,

$$t^* \doteq \text{Quantile}\left(1-\alpha; \sup_{h \in \mathcal{H}_1} \{\mathbb{E}_P[h(X)] \cdot \hat{\mathbb{E}}_n[h(X)] \cdot (\hat{\epsilon}(h) - \epsilon(h))\}\right). \tag{14}$$

We can then simplify the maximization problem in each iteration of the bootstrap algorithm by exploiting the finite-dimensional representer theorem for RKHS functions. This theorem states that for any finite set $\{x_i\}_{i=1}^n$ and $h \in \mathcal{H}_1$, $\{h(x_i)\}_{i=1}^n = \mathbf{K}\mathbf{w}$ for $\{\mathbf{K}_{ij}\}_{i,j=1}^n = \{k(x_i, x_j)\}_{i,j=1}^n$ and some $\mathbf{w} \in \mathbb{R}^n$ (Steinwart and Christmann, 2008). The other steps of this reduction, which are technical but uninformative, are deferred to the proof of Theorem 9.

Using the output of Algorithm 5, we construct a lower confidence bound for $\epsilon(h)$ as

$$\epsilon_{\text{lb}}(h) := \hat{\epsilon}(h) - \frac{t^*}{\left(\frac{1}{n}\sum_{i=1}^n h(x_i)\right)^2}. \tag{15}$$

We remark that a rescaling similar to (7) can also be applied in this setting. The resulting bootstrap computation for the rescaled RKHS process, however, is prohibitively expensive. Nevertheless, we include a detailed description of the appropriate rescaling and bootstrap algorithm in Appendix D.

Theorem 9 states the assumptions under which $\epsilon_{\text{lb}}(h)$ is a simultaneously valid confidence bound.

**Theorem 9 (Simultaneous RKHS confidence bound validity)** *Assume that* $\text{Var}(L)$ *is bounded away from* 0, $\|L\|_\infty$ *and* $\|k(X,X)\|_\infty$ *are finite,* $k(\cdot, x)$ *is continuous for all* $x$*, and that* $k(\cdot,\cdot)$ *is a positive definite kernel. Then,*

$$\lim_{n\to\infty} \mathbb{P}\left(\epsilon_{\text{lb}}(h) \le \epsilon(h) \text{ for all } h \in \mathcal{H}_1^+\right) \ge 1 - \alpha.$$

| | | Sample size ($n$) | | | | |
|---|---|---|---|---|---|---|
| | $\sigma$ | 100 | 200 | 400 | 800 | 1600 |
| Model (16) | 1 | 0.92 | 0.945 | 0.94 | 0.89 | 0.91 |
| | 0.5 | 0.93 | 0.955 | 0.935 | 0.885 | 0.915 |
| | 0.1 | 0.95 | 0.95 | 0.935 | 0.93 | 0.95 |

Table 4: Realized percentile of $t^*$ output by Algorithm 5 with $B = 500$ and $\alpha = 0.1$. The nominal percentile is $1 - \alpha = 0.9$. All results are based on 200 trials. In small samples, the bootstrap approximation can be conservative.

Commonly used kernels, such as the Gaussian and Laplace kernels, satisfy the assumptions given in Theorem 9.

### 5.2 Empirical results

We first validate our coverage guarantee for RKHS-based confidence sets using a synthetic experiment in which the ground truth is known. Formally, we use the same data-generating process as in (12), but discretize $\mathcal{X}$ as follows,

$$X_i \overset{\text{iid}}{\sim} \text{Unif}(\{0, 0.01, 0.02, \ldots, 1\}), \quad Y_i = \mathcal{N}(\beta_0 X_i, X_i). \tag{16}$$

In each trial, we sample $\beta_0 \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and use 1000 training points, $(X_i, Y_i)$, to fit the OLS predictor, $f(x) = \hat{\beta}^\top x$. We then audit over covariate shifts corresponding to the non-negative functions belonging to the unit ball of a Gaussian RKHS with varying bandwidths $\sigma \in \{0.1, 0.5, 1\}$. Even though we now analyze performance over an infinite collection, reducing the optimization task to solving a low-rank eigenvalue problem in Algorithm 5 eases the computational burden substantially: each audit takes approximately 1.5 seconds on a 2020 MacBook Pro.

Since evaluating the coverage under all non-negative functions in the RKHS is infeasible, we instead check that the bootstrap approximation to (14) is valid. Recall that $t^*$ estimates the $(1 - \alpha)$-quantile of

$$\sup_{h \in \mathcal{H}_1} \{\mathbb{E}_P[h(X)] \cdot \hat{\mathbb{E}}_n[h(X)] \cdot (\hat{\epsilon}(h) - \epsilon(h))\}. \tag{17}$$

In Table 4, we compute the percentile of (17) realized by $t^*$ for the synthetic experiment, i.e., $\mathbb{P}(\sup_{h \in \mathcal{H}_1}\{\mathbb{E}_P[h(X)] \cdot \hat{\mathbb{E}}_n[h(X)] \cdot (\hat{\epsilon}(h) - \epsilon(h))\} \leq t^*)$. Our results show that in small samples, the RKHS audit may be conservative: the realized percentile is sometimes larger than the nominal level.

We also revisit the synthetic dataset of Romano et al. (2019) to practically demonstrate how (15) can provide coverage guarantees over complex subgroups. Figure 8 shows how an arbitrarily chosen union of 3 sub-intervals[2] can be approximated by a Gaussian RKHS function with bandwidth $\sigma = 0.5$. Given an holdout set of $n = 1000$ points, we issue

2. This subgroup is not an element of the sub-interval collection we previously considered.

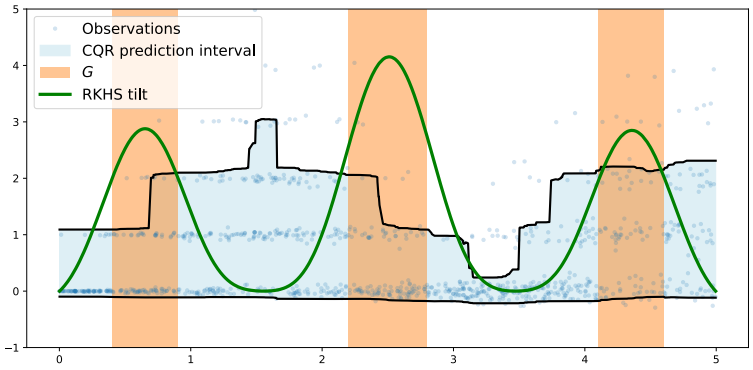



Figure 8: To provide an approximate coverage guarantee on the union of the three highlighted intervals, we tilt the covariate distribution by the depicted function belonging to the Gaussian RKHS with bandwidth $\sigma = 0.5$. For the holdout set displayed, we can lower bound the "tilted" coverage at 84.0%.

a conditional coverage lower bound of 84.0%; over infinitely many such tilts, the issued bounds are simultaneously valid with probability at least 90%.

## Acknowledgments

We thank Lihua Lei, Jonathan Taylor, Isaac Gibbs, Tim Morrison, and Anav Sood for helpful discussions. We especially thank Kevin Guo for sharing his notes on bootstrap and empirical process theory.

J.J.C. was supported by the John and Fannie Hertz Foundation. E.J.C. was supported by the Office of Naval Research grant N00014-20-1-2157, the National Science Foundation grant DMS-2032014, the Simons Foundation under award 814641, and the ARO grant 2003514594.

## References

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications, 2016.

M. A. Arcones and E. Gine. On the bootstrap of U and V statistics. *The Annals of Statistics*, pages 655–674, 1992.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, pages 1165–1188, 2001.

M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, et al. Toward trustworthy ai development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*, 2020.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018.

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

J. Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of Data and Analytics*, pages 296–299. Auerbach Publications, 2018.

C. DiCiccio, S. Vasudevan, K. Basu, K. Kenthapadi, and D. Agarwal. Evaluating fairness using permutation tests. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1467–1477, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984.

W. Dieterich, C. Mendoza, and T. Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(7.4):1, 2016.

F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

R. M. Dudley. A course on empirical processes. In *Ecole d'été de Probabilités de Saint-Flour XII-1982*, pages 1–142. Springer, 1984.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979.

W. Fithian and L. Lei. Conditional calibration for false discovery rate control under dependence. *The Annals of Statistics*, 50(6):3091–3118, 2022.

S. Flood. Integrated public use microdata series. *Current Population Survey: Version 4.0*, 2015.

A. W. Flores, K. Bechtel, and C. T. Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

P. Gaenssler, P. Molnár, and D. Rost. On continuity and strict increase of the cdf for the sup-functional of a gaussian process with applications to statistics. *Results in Mathematics*, 51:51–60, 2007.

E. Giné and J. Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, pages 929–989, 1984.

E. Giné and J. Zinn. Bootstrapping general empirical measures. *The Annals of Probability*, pages 851–869, 1990.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

U. Hébert-Johnson, M. Kim, O. Reingold, and G. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

M. Huskova and P. Janssen. Consistency of the generalized bootstrap for degenerate U-statistics. *The Annals of Statistics*, pages 1811–1823, 1993.

M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.

M. R. Kosorok. *Introduction to empirical processes and semiparametric inference.* Springer, 2008.

E. L. Lehmann, J. P. Romano, and G. Casella. *Testing statistical hypotheses*, volume 3. Springer, 2005.

D. J. Marcus. Relationships between donsker classes and sobolev spaces. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 69(3):323–330, 1985.

C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

G. Morina, V. Oliinyk, J. Waton, I. Marusic, and K. Georgatzis. Auditing and achieving intersectional fairness in classification problems. *arXiv preprint arXiv:1911.01468*, 2019.

Y. Romano, E. Patterson, and E. Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32:3543–3553, 2019.

A. Roy and P. Mohapatra. Fairness uncertainty quantification: How certain are you that the model is fair? *arXiv preprint arXiv:2110.01052*, 2023.

P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

M. Schaake and J. Clark. Stanford launches AI audit challenge, Jul 2022. URL `https://hai.stanford.edu/news/stanford-launches-ai-audit-challenge`.

N. Si, K. Murthy, J. Blanchet, and V. A. Nguyen. Testing group fairness via optimal transport projections. In *International Conference on Machine Learning*, pages 9649–9659. PMLR, 2021.

I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

B. Taskesen, J. Blanchet, D. Kuhn, and V. A. Nguyen. A statistical test for probabilistic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 648–665, 2021.

F. Tramer, V. Atlidakis, R. Geambasu, D. Hsu, J.-P. Hubaux, M. Humbert, A. Juels, and H. Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416. IEEE, 2017.

A. W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.

M. von Zahn, O. Hinz, and S. Feuerriegel. Locating disparities in machine learning. *arXiv preprint arXiv:2208.06680*, 2023.

B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017.

S. Xue, M. Yurochkin, and Y. Sun. Auditing ml models for individual bias and unfairness. In *International Conference on Artificial Intelligence and Statistics*, pages 4552–4562. PMLR, 2020.

T. Yan and C. Zhang. Active fairness auditing. In *International Conference on Machine Learning*, pages 24929–24962. PMLR, 2022.

Z. Zhang and D. B. Neill. Identifying significant predictive bias in classifiers. *arXiv preprint arXiv:1611.08292*, 2016.

## Appendix A. Connections to fairness

Recall that group fairness definitions ask for approximate parity of $L$ across all protected subpopulations (Dwork et al., 2012; Hardt et al., 2016; Corbett-Davies et al., 2017; Kearns et al., 2018). We show how to use our methods to audit two popular group fairness definitions below.

**Multicalibration** We say that a predictor $f(X)$ is calibrated if the conditional expectation of the binary label $Y$ given that $f(X) = v$ matches $v$, i.e., $\mathbb{E}[Y \mid f(X) = v] = v$. For the purpose of measuring classification quality across groups, we might require the binary classifier to be calibrated over many subgroups. Thus, for all subgroups $G \in \tilde{\mathcal{G}}$ and (potentially binned) predicted values $v$, the $\gamma$-multicalibration fairness criterion (Hébert-Johnson et al., 2018) requires that

$$|\mathbb{E}[Y \mid f(X) = v, X \in G] - v| \leq \gamma.$$

The formal multicalibration definition excludes groups smaller than some auditor-set threshold. Although the user may filter $\tilde{\mathcal{G}}$ as they see fit, a-priori, our methods do not exclude groups by their size. However, we can set a similar threshold in the rescaled certification method (Algorithm 2). This threshold does not prevent the auditor from querying any group for a bound on $\gamma$, but it ensures that confidence sets are narrower for groups above the threshold.

Choosing $L, \theta_P$, and $\mathcal{G}$ carefully, we can apply our methods to certify and flag multicalibration over $\tilde{\mathcal{G}}$. Letting $\mathcal{V}$ denote the set of unique values $f(X)$ can take, we set

$$L := Y - f(X) \qquad \theta_P := 0$$
$$\mathcal{G} := \{G \cap \{(X, Y) \mid f(X) = v\} \mid G \in \tilde{\mathcal{G}}, v \in \mathcal{V}\}.$$

Let $G_v$ denote each group in $\mathcal{G}$.

Certifying $\gamma$-multicalibration for $G$ is equivalent to establishing that $\max_v |\epsilon(G_v)| \leq \gamma$. Using our auditing methods, we can estimate $\gamma$ by constructing simultaneously valid confidence intervals for all $\{\epsilon(G_v)\}_{G_v \in \mathcal{G}}$. For a particular $G$, our bound on $\gamma$ then equals the maximum (absolute) value taken over all issued intervals for $\{\epsilon(G_v)\}_{v \in \mathcal{V}}$. For a fixed threshold $\gamma$, we could also run our Boolean certification method with $H_0(G_v) : |\epsilon(G_v)| \geq \gamma$ and certify $G$ if the null is rejected for all $\{G_v\}_{v \in \mathcal{V}}$.

Alternatively, we might flag $G$ for violation of $\gamma$-multicalibration by testing $H_0(G_v) : |\epsilon(G_v)| > \gamma + \epsilon$ for some disparity tolerance $\epsilon$. Using the method outlined in Section 4, we can construct p-values for each of these null hypotheses. Then, the BH procedure is run on all of the p-values computed. If any $H_0(G_v)$ is rejected, we flag $G$.

**Equalized odds** Given a collection of subsets of $\mathcal{X}$ denoted by $\mathcal{G}_X$, we say that a binary predictor $f$ satisfies the equalized odds criterion (Hardt et al., 2016; Woodworth et al., 2017) if for all $G \in \mathcal{G}_X$, both its true positive rates are equalized,

$$\mathbb{P}(f(X) = 1 \mid Y = 1, X \in G) = \mathbb{P}(f(X) = 1 \mid Y = 1),$$

and its false positive rates are equalized,

$$\mathbb{P}(f(X) = 1 \mid Y = 0, X \in G) = \mathbb{P}(f(X) = 1 \mid Y = 0).$$

A practitioner interested in this fairness criterion might then wish to audit the performance disparity $\epsilon(G) := \max(|\epsilon(G_0)|, |\epsilon(G_1)|)$ where

$$\epsilon(G_i) := \mathbb{P}(f(X) = 1 \mid Y = i, X \in G) - \mathbb{P}(f(X) = 1 \mid Y = i).$$

We can instantiate a certification audit for $\epsilon(G)$ by running our methods twice: once to bound $\epsilon(G_0)$ and a second time for $\epsilon(G_1)$. For the first audit, let

$$L := \mathbf{1}\{f(x) = 1\} \qquad \theta_P := \mathbb{P}(f(X) = 1 \mid Y = 0) \qquad \mathcal{G} = \{G \times \{0\} \mid G \in \mathcal{G}_X\}.$$

For the second audit, let

$$L := \mathbf{1}\{f(x) = 1\} \qquad \theta_P := \mathbb{P}(f(X) = 1 \mid Y = 1) \qquad \mathcal{G} = \{G \times \{1\} \mid G \in \mathcal{G}_X\}.$$

We construct confidence sets or Boolean certificates using a nominal Type I error threshold of $\alpha/2$ for each auditing task. We remark that this union bound is practically tight since the two data sets corresponding to these audits are disjoint. Our final certificate on $\epsilon(G)$ then consists of the "worse" of the two auditor outputs. For example, we might upper bound $\epsilon(G)$ by the maximum (absolute) value included in the two issued confidence intervals for $\epsilon(G_0)$ and $\epsilon(G_1)$.

For the flagging task, we test $H_0(G_i) : |\epsilon(G_i)| \leq \epsilon$. Using the method outlined in Section 4, we can construct p-values for each of these null hypotheses. Then, the BH procedure can be directly run on all of the p-values computed. We flag $G$ if either $H_0(G_0)$ or $H_0(G_1)$ is rejected.

**Individual fairness** Other fairness criteria fall into a category known as "individual fairness" measures. These quantify the intuition that similar inputs, $x$ and $x'$, should be treated by $f$ similarly (Dwork et al., 2012). While this definition of fairness cannot be tested using sampled model predictions, one might audit whether some notion of individual fairness holds with high probability among protected subgroups. Though we do not elaborate on such an extension here, we remark that prior work on the bootstrap of U-statistics and processes allows the natural extension of our auditing procedures to fairness measures defined over pairs of data points (Arcones and Gine, 1992; Huskova and Janssen, 1993).

## Appendix B. Certification audits

### B.1 Notation and review

We will begin by defining some relevant notation and reviewing certain basic results about the convergence of stochastic processes. Given $n$ i.i.d. samples, $\mathbb{P}_n := n^{-1} \sum_{i=1}^n \delta_{(X_i, Y_i)}$ denotes their empirical distribution. If $\{(X_i^*, Y_i^*)\}_{i=1}^n$ are i.i.d. samples from $\mathbb{P}_n$ conditional on $\{(X_i, Y_i)\}_{i=1}^n$, then $\mathbb{P}_n^* := n^{-1} \sum_{i=1}^n \delta_{(X_i^*, Y_i^*)}$ denotes their empirical distribution.

For a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^k$, $P[f]$ is shorthand for $\mathbb{E}_P[f(X, Y)]$, $\mathbb{P}_n[f]$ is shorthand for $n^{-1} \sum_{i=1}^n f(X_i, Y_i)$, and $\mathbb{P}_n^*[f]$ is shorthand for $n^{-1} \sum_{i=1}^n f(X_i^*, Y_i^*)$. We also write $(\mathbb{P}_n - P)[f]$ in place of $\mathbb{P}_n[f] - P[f]$. Given a class of functions $\mathcal{F}$, we think of $f \mapsto \sqrt{n}(\mathbb{P}_n - P)[f]$ as a mapping belonging to $\ell_\infty(\mathcal{F})$.

We will typically require and/or argue that $\mathcal{F}$ is a $P$-Donsker class. This means that the empirical process indexed by $f \in \mathcal{F}$ converges in distribution to a tight Gaussian limit in

$\ell_\infty(\mathcal{F})$. Formally, $\sqrt{n}(\mathbb{P}_n - P)[\cdot] \xrightarrow{d} \mathbb{G}[\cdot]$ where the limiting process $f \mapsto \mathbb{G}[f]$ is a Gaussian process that is also a tight Borel-measurable element of $\ell_\infty(\mathcal{F})$. If $\mathcal{F}$ is a $P$-Donsker class, then it is also $P$-Glivenko-Cantelli (van der Vaart, 2000), i.e., $\sup_{f\in\mathcal{F}} |(\mathbb{P}_n - P)[f]| \xrightarrow{p} 0$.

Next, we recall some results relating Donsker classes to VC dimension and bootstrap consistency. When we consider function classes of indicators indexed by subpopulations $\mathcal{G}$, i.e. $\mathcal{F} = \{\mathbf{1}\{(X,Y) \in G\} \mid G \in \mathcal{G}\}$, the Donsker property is *equivalent* to assuming that $\mathcal{G}$ is VC.

**Lemma 10 (Theorem 11.4.1 in Dudley (1984))** *Under suitable measurability assumptions, $\mathcal{F} := \{(x,y) \mapsto \mathbf{1}\{(x,y) \in G\} : G \in \mathcal{G}\}$ is Donsker if and only if $VC(\mathcal{G}) < \infty$.*

Last, we recall that the bootstrap approximation, $\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)[h]$, is valid if $\mathcal{H}$ is $P$-Donsker. Let $\mathbb{G}_n^*$ be shorthand for the bootstrap empirical process $\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)$ and $\mathbb{G}$ denote the limiting empirical process. Also note that the operator $\mathbb{E}^*$ refers to taking an expectation conditional on $X_1, \ldots, X_n$.

**Lemma 11 (Theorem 23.7 in van der Vaart (2000))** *For every Donsker class $\mathcal{H}$ of measurable functions with finite envelope function $F$, $\sup_{g\in BL_1(\ell^\infty(\mathcal{H}))} |\mathbb{E}^*[g(\mathbb{G}_n^*)] - \mathbb{E}[g(\mathbb{G})]| \xrightarrow{p} 0$ where $BL_1(\ell_\infty(\mathcal{H}))$ is the set of bounded 1-Lipschitz functions taking $\ell_\infty(\mathcal{H})$ into $\mathbb{R}$.*

### B.2 Proofs of certification theorems

**Estimating $\theta_P$.** We assume that the bootstrap and sampling distributions of $\hat{\theta}$ admit asymptotic linear expansions:

$$\sqrt{n}(\hat{\theta}(\mathcal{D}^*) - \hat{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i^*, Y_i^*) + o_{P_n}(1)$$
$$\sqrt{n}(\hat{\theta} - \theta_P) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i, Y_i) + o_P(1).$$

Here, $\psi$ is an influence function with mean zero and finite variance. It is easy to verify that this condition is satisfied by any $\hat{\theta}$ given in the main text. Our generalization enables auditing even if $\hat{\theta}$ is more complicated, e.g., it is the solution to some maximum likelihood estimation problem.

Next, we prove the main theorems of Section 3. As explained in the main text, the primary technical difficulty lies in proving that $t^*$ consistently estimates the $(1-\alpha)$-quantile of a particular stochastic process. There are two crucial preliminary results required to establish this result. First, we must show that bootstrap consistency implies consistency of the quantile estimate. Second, we must show that our proposed rescaling estimator $\hat{s}(G)$ is uniformly consistent for some estimand.

**Quantile consistency** First, to provide a unique definition of the quantile, we define the $\alpha$-quantile of a random variable $X$ as

$$\text{Quantile}(\alpha; X) := \inf_x \{x \mid \alpha \le \mathbb{P}(X \le x)\}.$$

The proof of the main theorem requires continuity and strict increase for the supremum of the limiting process at its $(1-\alpha)$-quantile. The following result establishes mild conditions under which this holds.

**Lemma 12** *Let $\mathbb{Z}(G)$ denote some Gaussian process that is the limit of an empirical process indexed by some countable class $\mathcal{G}$. Then, if $\mathrm{Var}(\mathbb{Z}(G)) > 0$ for some $G \in \mathcal{G}$, the distribution function of $\sup_{G\in\mathcal{G}} \mathbb{Z}(G)$ is continuous and strictly increasing on $\mathbb{R}_+$ and the distribution function of $\inf_{G\in\mathcal{G}} \mathbb{Z}(G)$ is continuous and strictly increasing on $\mathbb{R}_-$.*

**Proof** Because $\mathbb{Z}(G)$ is a centered process, we remark that $\sup_G \mathbb{Z}(G) \stackrel{d}{=} -\inf_G \mathbb{Z}(G)$. First, we prove that for all $x > 0$, the distribution function of $\sup_{G\in\mathcal{G}} \mathbb{Z}(G)$ is continuous at $x$. Under the stated assumptions, Gaenssler et al. (2007, Corollary 1.3) prove that $\sup_G |\mathbb{Z}(G)|$ has a continuous distribution function at $x$, i.e., $\mathbb{P}(\sup_G |\mathbb{Z}(G)| = x) = 0$. Then,

$$0 = \mathbb{P}(\sup_G |\mathbb{Z}(G)| = x) = \mathbb{P}(\{\sup_G \mathbb{Z}(G) = x\} \cup \{\inf_G \mathbb{Z}(G) = -x\}),$$

which implies $\mathbb{P}(\sup_G \mathbb{Z}(G) = x) = \mathbb{P}(\inf_G \mathbb{Z}(G) = -x) = 0$. This is equivalent to the desired continuity statement.

Next, to prove strict increase at $x$, we argue by contradiction. If $\sup_G \mathbb{Z}(G)$ did *not* have a strictly increasing distribution function on $\mathbb{R}_+$, then there exists $0 < x_1 < x_2$ such that

$$\mathbb{P}(\sup_G \mathbb{Z}(G) \le x_1) = \mathbb{P}(\sup_G \mathbb{Z}(G) \le x_2) \iff \mathbb{P}(\inf_G \mathbb{Z}(G) \ge -x_1) = \mathbb{P}(\inf_G \mathbb{Z}(G) \ge -x_2).$$

Gaenssler et al. (2007, Corollary 1.3) also prove that $\sup_G |\mathbb{Z}(G)|$ satisfies $\mathbb{P}(\sup_G |\mathbb{Z}(G)| \le x_1) < \mathbb{P}(\sup_G |\mathbb{Z}(G)| \le x_2)$. Since

$$\begin{aligned}\mathbb{P}(\sup_G |\mathbb{Z}(G)| \le x) = \mathbb{P}(\sup_G \mathbb{Z}(G) \le x) + \mathbb{P}(\inf_G \mathbb{Z}(G) \ge -x) \\ - \mathbb{P}(\{\sup_G \mathbb{Z}(G) \le x\} \cap \{\inf_G \mathbb{Z}(G) \ge -x\}),\end{aligned}$$

we conclude that

$$\mathbb{P}(\{\sup_G \mathbb{Z}(G) \le x_1\} \cap \{\inf_G \mathbb{Z}(G) \ge -x_1\}) > \mathbb{P}(\{\sup_G \mathbb{Z}(G) \le x_2\} \cap \{\inf_G \mathbb{Z}(G) \ge -x_2\}).$$

This is a contradiction since the event on the LHS is a subset of the event on the RHS. We conclude that both $\sup_G \mathbb{Z}(G)$ and $\inf_G \mathbb{Z}(G)$ have strictly increasing distribution functions on the positive and negative reals, respectively. ■

We make two comments on Theorem 12. First, the cited result of Gaenssler et al. (2007) is stated for empirical processes indexed by a VC class, but their argument applies to any limiting Gaussian process indexed by a $P$-Donsker class. Second, the countable assumption is not crucial, and is only included to avoid verifying certain technical measurability conditions. In this article, we expect the "pointwise measurability" (approximately equivalent to well-approximation by a dense countable subset) condition given in van der Vaart and Wellner (1996) to hold, but this condition must be checked for each function class. As a consequence, for simplicity, we will also assume throughout that the function classes we work with are countable.

**Uniform consistency of $\hat{s}$** Recall that we define

$$\hat{s}(G) := \max(\mathbb{P}_n(G), p_*)^{3/2} \cdot \hat{\sigma}(G, w_0)$$

for

$$\hat{\sigma}(G, w_0) := \left(\frac{\mathbb{P}_n(G)}{\mathbb{P}_n(G) + w_0}\right) \cdot \hat{\sigma}_G + \left(\frac{w_0}{\mathbb{P}_n(G) + w_0}\right) \cdot \sqrt{\widehat{\mathrm{Var}}(L)}$$

and

$$\hat{\sigma}_G^2 := \widehat{\mathrm{Var}}(L \mid G) + \mathbb{P}_n(G)\left(\widehat{\mathrm{Var}}(\psi) - 2 \cdot \widehat{\mathrm{Cov}}(L, \psi \mid G)\right).$$

**Lemma 13** *Assume that* $0 < \mathrm{Var}(L) < \infty$ *and* $VC(\mathcal{G}) < \infty$. *If* $p_*, w_0 > 0$, *then* $\sup_{G \in \mathcal{G}} |1/\hat{s}(G) - 1/s(G)| \xrightarrow{p} 0$, *where*

$$s(G) = \max(\mathbb{P}(G), p_*)^{3/2} \cdot \left[\left(\frac{\mathbb{P}(G)}{\mathbb{P}(G) + w_0}\right) \cdot \sigma_G + \left(\frac{w_0}{\mathbb{P}(G) + w_0}\right) \cdot \sqrt{\mathrm{Var}(L)}\right].$$

**Proof** To simplify notation, we will replace $\mathbf{1}\{(X, Y) \in G\}$ with $\mathbf{1}_G$ throughout.

We first compute the asymptotic variance of $\sqrt{|G|}(\hat{\epsilon}(G) - \epsilon(G))$ by linearizing the random quantity:

$$\begin{aligned}
\sqrt{|G|}\left(\hat{\epsilon}(G) - \epsilon(G)\right) &= \sqrt{|G|}\left[\left(\frac{1}{|G|}\sum_{(X_i, Y_i) \in G} L_i - \hat{\theta}\right) - \left(\mathbb{E}[L \mid G] - \theta_P\right)\right] \\
&= \sqrt{|G|}\left[\frac{1}{|G|}\sum_{i=1}^{n}(L_i - \mathbb{E}[L \mid G]) \cdot \mathbf{1}_G - \frac{1}{n}\sum_{i=1}^{n}\psi_i\right] \\
&= \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{L_i - \mathbb{E}[L \mid G]}{\sqrt{\mathbb{P}(G)}} \cdot \mathbf{1}_G - \sqrt{\mathbb{P}(G)} \cdot \psi_i\right] + o_P(1).
\end{aligned}$$

So, we conclude that

$$\begin{aligned}
\sigma_G^2 &:= \mathrm{Var}\left(\frac{L - \mathbb{E}[L \mid G]}{\sqrt{\mathbb{P}(G)}} \cdot \mathbf{1}_G - \sqrt{\mathbb{P}(G)} \cdot \psi\right) \\
&= \mathrm{Var}\left(\frac{L - \mathbb{E}[L \mid G]}{\sqrt{\mathbb{P}(G)}} \cdot \mathbf{1}_G\right) + \mathrm{Var}\left(\sqrt{\mathbb{P}(G)} \cdot \psi\right) - 2 \cdot \mathrm{Cov}\left(\frac{L - \mathbb{E}[L \mid G]}{\sqrt{\mathbb{P}(G)}} \cdot \mathbf{1}_G, \sqrt{\mathbb{P}(G)} \cdot \psi\right) \\
&= \mathrm{Var}(L \mid G) + \mathbb{P}(G) \cdot \left(\mathrm{Var}(\psi) - 2 \cdot \mathrm{Cov}(L, \psi \mid G)\right).
\end{aligned}$$

Next, after some rearrangement, observe that $1/\hat{s}^2(G) - 1/s^2(G)$ equals

$$\frac{\mathbb{P}_n(G) + w_0}{\max(\mathbb{P}_n(G), p_*)^{3/2} \cdot \left(\mathbb{P}_n(G) \cdot \hat{\sigma}_G + w_0 \cdot \sqrt{\widehat{\mathrm{Var}}(L)}\right)} - \frac{\mathbb{P}(G) + w_0}{\max(\mathbb{P}(G), p_*)^{3/2} \cdot \left(\mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)}\right)}.$$

Combining the two fractions yields a numerator of

$$(\mathbb{P}_n(G) + w_0)\left(\max(\mathbb{P}(G), p_*)^{3/2} \cdot \left(\mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)}\right)\right) - (\mathbb{P}(G) + w_0)\left(\max(\mathbb{P}_n(G), p_*)^{3/2} \cdot \left(\mathbb{P}_n(G) \cdot \hat{\sigma}_G + w_0 \cdot \sqrt{\widehat{\mathrm{Var}}(L)}\right)\right)$$

and a denominator of

$$\max(\mathbb{P}_n(G), p_*)^{3/2} \cdot \max(\mathbb{P}(G), p_*)^{3/2} \cdot \left(\mathbb{P}_n(G) \cdot \hat{\sigma}_G + w_0 \cdot \sqrt{\widehat{\mathrm{Var}}(L)}\right) \cdot \left(\mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)}\right).$$

To prove uniform consistency, it is sufficient to prove that

$$\lim_{n\to\infty} \inf_{G\in\mathcal{G}} \max(\mathbb{P}_n(G), p_*)^{3/2} \cdot \max(\mathbb{P}(G), p_*)^{3/2} \cdot \left(\mathbb{P}_n(G) \cdot \hat{\sigma}_G + w_0 \cdot \sqrt{\widehat{\mathrm{Var}}(L)}\right) \cdot \left(\mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)}\right) > 0$$

and

$$\lim_{n\to\infty} \sup_{G\in\mathcal{G}} \Big|(\mathbb{P}_n(G) + w_0)\left(\max(\mathbb{P}(G), p_*)^{3/2} \cdot \left(\mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)}\right)\right) - (\mathbb{P}(G) + w_0)\left(\max(\mathbb{P}_n(G), p_*)^{3/2} \cdot \left(\mathbb{P}_n(G) \cdot \hat{\sigma}_G + w_0 \cdot \sqrt{\widehat{\mathrm{Var}}(L)}\right)\right)\Big| = 0.$$

For the first of these tasks, observe that the denominator is lower bounded by

$$p_*^3 \cdot \left(w_0^2 \cdot \sqrt{\widehat{\mathrm{Var}}(L)} \cdot \sqrt{\mathrm{Var}(L)}\right) \to p_*^3 \cdot w_0^2 \cdot \mathrm{Var}(L) > 0.$$

The numerator requires a more careful analysis. We distribute the $\mathbb{P}_n(G)+w_0$ and $\mathbb{P}(G)+w_0$ prefactors and analyze the first term in the numerator obtained by matching the terms beginning with $\mathbb{P}_n(G)$ and $\mathbb{P}(G)$:

$$\left|\mathbb{P}_n(G) \cdot \max(\mathbb{P}(G), p_*)^{3/2} \cdot \left(\mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)}\right) - \mathbb{P}(G) \cdot \max(\mathbb{P}_n(G), p_*)^{3/2} \cdot \left(\mathbb{P}_n(G) \cdot \hat{\sigma}_G + w_0 \cdot \sqrt{\widehat{\mathrm{Var}}(L)}\right)\right|.$$

Observe that if this term goes to 0 uniformly over $G \in \mathcal{G}$, the second term obtained by replacing both $\mathbb{P}_n(G)$ and $\mathbb{P}(G)$ with $w_0$ will also uniformly converge to 0. Thus, it suffices to analyze this expression.

Adding and subtracting $\mathbb{P}(G)\cdot\max(\mathbb{P}(G), p_*)^{3/2}\cdot(\mathbb{P}(G)\cdot\sigma_G+w_0\cdot\sqrt{\mathrm{Var}(L)})$ and applying a triangle inequality, we obtain:

$$\begin{aligned}
&|\mathbb{P}_n(G) - \mathbb{P}(G)| \cdot \max(\mathbb{P}(G), p_*)^{3/2} \cdot (\mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)}) \\
&\quad + \left|\max(\mathbb{P}(G), p_*)^{3/2} - \max(\mathbb{P}_n(G), p_*)^{3/2}\right| \cdot \mathbb{P}(G) \cdot (\mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)}) \\
&\quad + \mathbb{P}(G) \cdot \max(\mathbb{P}_n(G), p_*)^{3/2} \cdot |\mathbb{P}(G) \cdot \sigma_G - \mathbb{P}_n(G) \cdot \hat{\sigma}_G| \\
&\quad + \mathbb{P}(G) \cdot \max(\mathbb{P}_n(G), p_*)^{3/2} \left|w_0 \cdot (\sqrt{\mathrm{Var}(L)} - \sqrt{\widehat{\mathrm{Var}}(L)})\right|.
\end{aligned}$$

Since the sample variance of $L$ does not depend on $G$ and is consistent under the stated assumptions, the uniform convergence of the fourth term to 0 is easy to see. To prove uniform convergence of the first and second terms, we must show that the $n$-independent term is bounded, i.e.,

$$\begin{aligned}(\mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)}) &\leq \mathbb{P}(G)^{1/2} \sqrt{\mathbb{E}[(L - \mathbb{E}[L \mid G])^2 \cdot \mathbf{1}_G]} + w_0 \cdot \sqrt{\mathrm{Var}(L)}) \\ &\leq \mathbb{P}(G)^{1/2} \sqrt{\mathbb{E}[(L - \mathbb{E}[L])^2 \cdot \mathbf{1}_G]} + w_0 \cdot \sqrt{\mathrm{Var}(L)}) \\ &\leq \mathbb{P}(G)^{1/2} \sqrt{\mathbb{E}[(L - \mathbb{E}[L])^2]} + w_0 \cdot \sqrt{\mathrm{Var}(L)}) \\ &\leq (1 + w_0) \cdot \sqrt{\mathrm{Var}(L)} = C < \infty.\end{aligned}$$

Thus, since $\mathrm{VC}(\mathcal{G}) < \infty$, $(\mathbb{P}_n - P)[\mathbf{1}_G]$ is $P$-Glivenko-Cantelli, and

$$\begin{aligned}&\sup_{G \in \mathcal{G}} \left| (\mathbb{P}_n(G) - \mathbb{P}(G)) \cdot \max(\mathbb{P}(G), p_*)^{3/2} \cdot (\mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)}) \right| \\ &\qquad \leq C \cdot \sup_{G \in \mathcal{G}} |\mathbb{P}_n(G) - \mathbb{P}(G)| \xrightarrow{p} 0.\end{aligned}$$

Since a $P$-Glivenko Cantelli class is preserved under truncation (Example 2.10.11 in van der Vaart and Wellner (1996)), we then also conclude that

$$\begin{aligned}&\sup_{G \in \mathcal{G}} |\max(\mathbb{P}(G), p_*)^{3/2} - \max(\mathbb{P}_n(G), p_*)^{3/2}| \cdot \mathbb{P}(G) \cdot \left( \mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)} \right) \\ &\qquad \leq C \cdot \sup_{G \in \mathcal{G}} |\max(\mathbb{P}(G), p_*)^{3/2} - \max(\mathbb{P}_n(G), p_*)^{3/2}| \xrightarrow{p} 0.\end{aligned}$$

Last, we show the third term converges to 0. The $n$-independent part is bounded by 1, so we can ignore that. So, we need to show that $\sup_{G \in \mathcal{G}} |\mathbb{P}(G) \cdot \sigma_G - \mathbb{P}_n(G) \cdot \hat{\sigma}_G| \xrightarrow{p} 0$. To avoid writing the square root, we will square the two terms for the time being. Then,

$$\begin{aligned}\mathbb{P}(G)^2 \cdot \sigma_G^2 &= \mathbb{P}(G)^2 \left[ \mathrm{Var}(L \mid G) + \mathbb{P}(G) \cdot (\mathrm{Var}(\psi) - 2 \cdot \mathrm{Cov}(L, \psi \mid G)) \right] \\ &= \mathbb{P}(G)^2 (\mathbb{E}[L^2 \mid G] - \mathbb{E}[L \mid G]^2) \\ &\quad + \mathbb{P}(G)^3 \cdot (\mathrm{Var}(\psi) - 2 \cdot (\mathbb{E}[L \cdot \psi \mid G] - \mathbb{E}[L \mid G] \mathbb{E}[\psi \mid G])) \\ &= \mathbb{P}(G) \mathbb{E}[L^2 \cdot \mathbf{1}_G] - \mathbb{E}[L \cdot \mathbf{1}_G]^2 + \mathbb{P}(G)^3 \cdot \mathrm{Var}(\psi) \\ &\quad - 2 \cdot \mathbb{P}(G)^2 \cdot \mathbb{E}[(L \cdot \psi) \mathbf{1}_G] \\ &\quad - 2 \cdot \mathbb{P}(G) \cdot \mathbb{E}[L \cdot \mathbf{1}_G] \mathbb{E}[\psi \cdot \mathbf{1}_G]\end{aligned}$$

We can obtain an analogous expansion of $\mathbb{P}_n(G)^2 \hat{\sigma}_G^2$. Then, we need to show that each matching term converges uniformly. We will only explicitly work out the argument for one pair of terms, but the argument for the remainder should be clear. First, observe that $\mathcal{F}_g = \{g \cdot \mathbf{1}_G : G \in \mathcal{G}\}$ is a $P$-Glivenko-Cantelli class so long as $P[|g|] < \infty$ (Corollary 3 in Giné and Zinn (1984)). Then,

$$\begin{aligned}&\sup_{G \in \mathcal{G}} \left| \mathbb{P}_n(G)^2 \cdot \mathbb{P}_n[(L \cdot \psi) \cdot \mathbf{1}_G] - \mathbb{P}(G)^2 \cdot P[(L \cdot \psi) \cdot \mathbf{1} \{X, Y) \in G\}] \right| \\ &\qquad \leq \sup_{G \in \mathcal{G}} \left| \mathbb{P}_n(G)^2 \cdot (\mathbb{P}_n[(L \cdot \psi) \cdot \mathbf{1}_G] - P[(L \cdot \psi) \cdot \mathbf{1}_G]) \right| \\ &\qquad\qquad + \sup_{G \in \mathcal{G}} \left| (\mathbb{P}_n(G)^2 - \mathbb{P}(G)^2) \cdot P[(L \cdot \psi) \cdot \mathbf{1}_G] \right|\end{aligned}$$

We can further upper bound the RHS by

$$\underbrace{\sup_{G\in\mathcal{G}} |(\mathbb{P}_n - P)[(L\cdot\psi)\cdot\mathbf{1}_G]|}_{o_P(1)} + 2\cdot \underbrace{P[|L\cdot\psi|]}_{\leq\sqrt{P[L^2]P[\psi^2]}<\infty} \cdot \underbrace{\sup_{G\in\mathcal{G}} |\mathbb{P}_n(G) - \mathbb{P}(G)|}_{o_P(1)}.$$

A similar argument can be made for each of the other terms in this expression.

Thus, we conclude that the numerator converges uniformly over all $G$ to 0. This yields the claimed uniform consistency of $1/\hat{s}(G)$. ■

**Bound certification** First, we prove that our lower confidence bound construction is valid. Algorithm 6 restates our method for defining the critical value $t^*$. Then, using the output of Algorithm 6, we construct an asymptotically valid lower confidence bound for arbitrary $G \in \mathcal{G}$ by setting

$$\epsilon_{\text{lb}}(G) = \hat{\epsilon}(G) - t^* \cdot \frac{\hat{s}(G)}{\mathbb{P}_n(G)^2}.$$

Theorem 14 restates Theorem 4.

**Algorithm 6** Bootstrapping the (rescaled) lower confidence bound critical value

**Input:** Subpopulations $\mathcal{G}$, holdout set $\mathcal{D}$, level $\alpha$, threshold $p_*$, weight $w_0$, number of bootstrap samples $B$
**for** $b = 1, \ldots, B$ **do**
  Let $\mathcal{D}_b^*$ be a sample with replacement of size $n$ from $\mathcal{D}$;
  Define $\mathbb{P}_b^*(G) := \frac{1}{n}\sum_{(x_i^*,y_i^*)\in\mathcal{D}_b^*} \mathbf{1}\{(x_i^*, y_i^*) \in G\}$;
  Define $\epsilon_b^*(G) := \frac{1}{\mathbb{P}_b^*(G)\cdot n}\sum_{(x_i^*,y_i^*)\in G} L_i^* - \hat{\theta}(\mathcal{D}_b^*)$;
**end for**
Define the asymptotic variance estimator by

$$\hat{\sigma}_G^2 := \widehat{\text{Var}}(L \mid G) + \mathbb{P}_n(G)\left(\widehat{\text{Var}}(\psi) - 2\cdot\widehat{\text{Cov}}(L, \psi \mid G)\right)$$

where $\widehat{\text{Var}}(\cdot)$ and $\widehat{\text{Cov}}(\cdot)$ correspond to the sample (conditional) variance and covariance;
Define $\hat{s}(G)$ by (9);
**for** $b = 1, \ldots, B$ **do**
  $t^{(b)} = \max_{G\in\mathcal{G}}\{\frac{1}{\hat{s}(G)}\cdot\mathbb{P}_n(G)\cdot\mathbb{P}_b^*(G)\cdot(\epsilon_b^*(G) - \hat{\epsilon}(G))\}$;
**end for**
**Return:** $t^* = \text{Quantile}(1-\alpha; \{t^{(b)}\}_{b=1}^B)$

**Theorem 14** *Assume that $L$ is bounded and that $L - \hat{\theta}$ is non-constant over at least one non-empty group. Further assume that $\mathcal{G}$ has finite Vapnik-Chernovenkis (VC) dimension. Then, $\epsilon_{\text{lb}}(G)$ is an asymptotic $(1-\alpha)$-lower confidence bound for $\epsilon(G)$ that is simultaneously valid for all $G \in \mathcal{G}$.*

**Proof** To simplify notation, we will replace $\mathbf{1}\{(X,Y) \in G\}$ with $\mathbf{1}_G$ throughout. We first restate the result we aim to prove.

$$\lim_{n\to\infty} \mathbb{P}\left(\exists\, G \in \mathcal{G} \text{ s.t. } \epsilon_{\text{lb}}(G) > \epsilon(G)\right) = \lim_{n\to\infty} \mathbb{P}\left(\exists\, G \in \mathcal{G} \text{ s.t. } \hat{\epsilon}(G) - t^* \cdot \frac{s^*(G)}{\mathbb{P}_n(G)^2} > \epsilon(G)\right)$$

Rearranging the latter event, we obtain

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\left(\sup_{G\in\mathcal{G}} \left\{\left(\frac{\mathbb{P}_n(G)^2}{\hat{s}(G)}\right) \cdot (\hat{\epsilon}(G) - \epsilon(G))\right\} > t^*\right) \\ = \lim_{n\to\infty} \mathbb{P}\left(\sup_{G\in\mathcal{G}} \left\{\left(\frac{\mathbb{P}_n(G)}{\hat{s}(G)}\right) \cdot \mathbb{P}_n[(L - \hat{\theta} - \epsilon(G)) \cdot \mathbf{1}_G]\right\} > t^*\right).\end{aligned}$$

Then, we claim that $\sup_{G\in\mathcal{G}} |\mathbb{P}_n(G)/\hat{s}(G) - \mathbb{P}(G)/s(G)| \xrightarrow{p} 0$. This follows by observing that

$$\begin{aligned}\sup_{G\in\mathcal{G}} \left|\frac{\mathbb{P}_n(G)}{\hat{s}(G)} - \frac{\mathbb{P}(G)}{s(G)}\right| &\le \sup_{G\in\mathcal{G}} \left|\frac{\mathbb{P}_n(G)}{\hat{s}(G)} - \frac{\mathbb{P}_n(G)}{s(G)}\right| + \sup_{G\in\mathcal{G}} \left|\frac{\mathbb{P}_n(G)}{s(G)} - \frac{\mathbb{P}(G)}{s(G)}\right| \\ &\le \sup_{G\in\mathcal{G}} |\mathbb{P}_n(G)| \cdot \sup_{G\in\mathcal{G}} \left|\frac{1}{\hat{s}(G)} - \frac{1}{s(G)}\right| + \left|\frac{1}{\inf_{G\in\mathcal{G}} s(G)}\right| \cdot \sup_{G\in\mathcal{G}} |\mathbb{P}_n(G) - \mathbb{P}(G)|.\end{aligned}$$

Observe that $\sup_{G\in\mathcal{G}} |\mathbb{P}_n(G)| \le 1$ and $\inf_{G\in\mathcal{G}} s(G) \ge p_*^{3/2} \cdot (w_0/(1+w_0)) \cdot \sqrt{\text{Var}(L)} > 0$. Then, by Theorem 13 and the assumption that $\mathcal{G}$ is $P$-Donsker, and thus also $P$-Glivenko-Cantelli, we conclude the desired uniform consistency result.

To apply Slutsky's lemma to the uniformly consistent estimator, we must prove that $\sqrt{n} \cdot \mathbb{P}_n[(L - \hat{\theta} - \epsilon(G)) \cdot \mathbf{1}_G] = O_P(1)$. Note that

$$\begin{aligned}&\left|\sup_{G\in\mathcal{G}} \sqrt{n} \cdot \mathbb{P}_n[(L - \hat{\theta} - \epsilon(G)) \cdot \mathbf{1}_G]\right| \\ &\quad= \left|\sup_{G\in\mathcal{G}} \left\{\sqrt{n} \cdot (\mathbb{P}_n - P)[(L - \theta_P - \epsilon(G)) \cdot \mathbf{1}_G] - \sqrt{n}(\hat{\theta} - \theta_P) \cdot \mathbb{P}_n[\mathbf{1}_G]\right\}\right| \\ &\quad\le \left|\sup_{G\in\mathcal{G}} \sqrt{n}(\mathbb{P}_n - P)[(L - \theta_P - \epsilon(G)) \cdot \mathbf{1}_G]\right| + \left|\sqrt{n}(\hat{\theta} - \theta_P)\right| \cdot \sup_{G\in\mathcal{G}} \mathbb{P}_n(G).\end{aligned}$$

By standard Donsker preservation results (van der Vaart and Wellner, 1996, Section 2.10), the function class $\{(L-\theta_P-\epsilon(G))\cdot\mathbf{1}_G \mid G \in \mathcal{G}\}$ is $P$-Donsker and hence $\sup_{G\in\mathcal{G}} \sqrt{n}(\mathbb{P}_n - P)[(L - \theta_P - \epsilon(G))\mathbf{1}_G] = O_P(1)$. Meanwhile, $|\sqrt{n}(\hat{\theta} - \theta_P)| \cdot \sup_{G\in\mathcal{G}} \mathbb{P}_n(G) \le |\sqrt{n}(\hat{\theta} - \theta_P)| = O_P(1)$ by assumption. Thus, this upper bound is $O_P(1)$.

Then, applying Slutsky's lemma, we obtain

$$\begin{aligned}&\left(\frac{\mathbb{P}_n[\mathbf{1}_G]}{\hat{s}(G)}\right) \cdot \sqrt{n} \cdot \mathbb{P}_n\left[\left(L - \hat{\theta} - \epsilon(G)\right)\mathbf{1}_G\right] \\ &\quad= \frac{1}{s(G)} \cdot \sqrt{n} \cdot \mathbb{P}_n\left[\left((L - \hat{\theta})P[\mathbf{1}_G] - P[(L - \theta_P)\mathbf{1}_G]\right)\mathbf{1}_G\right] + o_P(1) \\ &\quad= \frac{1}{s(G)} \cdot \sqrt{n}\left(\mathbb{P}_n[(L - \hat{\theta})\mathbf{1}_G] \cdot P[\mathbf{1}_G] - P[(L - \theta_P)\mathbf{1}_G] \cdot \mathbb{P}_n[\mathbf{1}_G]\right) + o_P(1).\end{aligned}$$

Our goal is to now prove that the bootstrap analogue to this process is consistent, i.e., we must show that this process is indexed by a $P$-Donsker class. To this end, observe that we can rewrite the process as

$$\frac{P[\mathbf{1}_G]}{s(G)} \cdot \sqrt{n} \left( \mathbb{P}_n[(L - \hat{\theta}) \cdot \mathbf{1}_G] - P[(L - \theta_P) \cdot \mathbf{1}_G] \right) - \frac{P[(L - \theta_P)\mathbf{1}_G]}{s(G)} \cdot \sqrt{n}(\mathbb{P}_n - P)[\mathbf{1}_G] + o_P(1)$$

We then linearize the first term by observing that

$$\begin{aligned} \sqrt{n} \left( \mathbb{P}_n[(L - \hat{\theta}) \cdot \mathbf{1}_G] - P[(L - \theta_P) \cdot \mathbf{1}_G] \right) &= \sqrt{n}(\mathbb{P}_n - P)[L \cdot \mathbf{1}_G] - \theta_P \cdot \sqrt{n}(\mathbb{P}_n - P)[\mathbf{1}_G] \\ &\quad - P[\mathbf{1}_G] \cdot \sqrt{n}(\hat{\theta} - \theta_P) + \underbrace{\frac{1}{\sqrt{n}} \cdot \sqrt{n}(\mathbb{P}_n - P)[\mathbf{1}_G] \cdot \sqrt{n}(\hat{\theta} - \theta_P)}_{o_P(1)}. \end{aligned}$$

Replacing this in the previous expansion, we simplify and obtain

$$\begin{aligned} &\sqrt{n}(\mathbb{P}_n - P) \left[ \frac{P[\mathbf{1}_G]}{s(G)} \cdot L \cdot \mathbf{1}_G \right] - \sqrt{n}(\mathbb{P}_n - P) \left[ \frac{P[L \cdot \mathbf{1}_G]}{s(G)} \cdot \mathbf{1}_G \right] - \sqrt{n}(\mathbb{P}_n - P) \left[ \frac{P[\mathbf{1}_G]^2}{s(G)} \cdot \psi \right] \\ &\qquad + o_P(1). \end{aligned}$$

We claim that the three empirical processes in this display are all indexed by Donsker classes. Using the first process as an example, we observe that $\{P[\mathbf{1}_G]/s(G) \mid G \in \mathcal{G}\}$ is a universal, uniformly bounded Donsker class, while $L \cdot \mathbf{1}_G$ is a $P$-Donsker class because the product of a bounded measurable function and a Donsker class is Donsker. The pairwise product of two uniformly bounded Donsker classes is Donsker. Clearly, $\{P[\mathbf{1}_G]/s(G) \cdot L \cdot \mathbf{1}_G \mid G \in \mathcal{G}\}$ is the subset of such a product, so by Theorem 2.10.1 in van der Vaart and Wellner (1996), the first process is indexed by a Donsker class. Near-identical arguments prove the last two processes are also $P$-Donsker. Recall that the Donsker property is preserved under addition. So, we conclude that

$$\sqrt{n}(\mathbb{P}_n - P) \left[ \frac{P[\mathbf{1}_G]}{s(G)} \cdot L \cdot \mathbf{1}_G - \frac{P[L \cdot \mathbf{1}_G]}{s(G)} \cdot \mathbf{1}_G - \frac{P[\mathbf{1}_G]^2}{s(G)} \cdot \psi \right] + o_P(1) \tag{18}$$

is indexed by a subset of a Donsker class, i.e., a Donsker class.

Via an identical derivation, we also show that the bootstrap process in Algorithm 6, $\frac{1}{\hat{s}(G)} \cdot \mathbb{P}_n(G) \cdot \mathbb{P}_n^*(G) \cdot (\epsilon^*(G) - \hat{\epsilon}(G))$ is equal to

$$\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n) \left[ \frac{P[\mathbf{1}_G]}{s(G)} \cdot L \cdot \mathbf{1}_G - \frac{P[L \cdot \mathbf{1}_G]}{s(G)} \cdot \mathbf{1}_G - \frac{P[\mathbf{1}_G]^2}{s(G)} \cdot \psi \right] + o_P(1).$$

Applying Theorem 11, we conclude that the bootstrap is consistent. Due to the continuous mapping theorem, we may take a sup over $G \in \mathcal{G}$ and conclude that the process sampled in Algorithm 6 is asymptotically equivalent to the process stated in the probability of interest,

$$\lim_{n \to \infty} \mathbb{P} \left( \sup_{G \in \mathcal{G}} \left\{ \sqrt{n} \cdot \left( \frac{\mathbb{P}_n(G)}{\hat{s}(G)} \right) \cdot \mathbb{P}_n[(L - \hat{\theta} - \epsilon(G)) \cdot \mathbf{1}_G] \right\} \geq t^* \cdot \sqrt{n} \right).$$

Last, we must show that $t^* \cdot \sqrt{n}$ consistently estimates the $(1-\alpha)$-quantile of the supremum of (18). Recall that Lehmann et al. (2005, Lemma 11.2.1(ii)) establishes quantile consistency whenever the distribution function is continuous and strictly increasing at the point of interest. Our assumption that $L - \hat{\theta}$ is non-constant for some non-empty $G$ implies that the asymptotic variance of $\sqrt{n} \cdot \frac{1}{s(G)} \cdot \mathbb{P}_n(G) \cdot \mathbb{P}(G) \cdot (\hat{\epsilon}(G) - \epsilon(G))$ is non-zero for some $G$. Then, the $(1-\alpha)$-quantile of the supremized process is strictly greater than 0 and Theorem 12 implies the desired result.

Completing the proof,

$$\lim_{n\to\infty} \mathbb{P}\left(\exists\, G \in \mathcal{G} \text{ s.t. } \epsilon_{\text{lb}}(G) > \epsilon(G)\right) = \lim_{n\to\infty} \left( \sup_{G\in\mathcal{G}} \left\{ \frac{1}{\hat{s}(G)} \cdot \mathbb{P}_n(G) \cdot \mathbb{P}_n^*(G) \cdot (\epsilon^*(G) - \hat{\epsilon}(G)) \right\} > t^* \right) = \alpha.$$

■

**Boolean certification** We next consider Boolean certification, in which we test the null

$$H_0(G) : \epsilon(G) \leq \epsilon.$$

In Algorithm 7, we define a rescaled variant of Algorithm 3. Recall that we issue a certificate for $G$ when

$$\hat{\epsilon}(G) \geq \epsilon + \frac{t^*}{\mathbb{P}_n(G)}.$$

Theorem 15 then proves the issued certificates are simultaneously valid with probability $1-\alpha$.

**Theorem 15** *Assume that $L$ has finite variance and that $L - \hat{\theta}$ is non-constant over at least one non-empty group. Further assume that $\mathcal{G}$ has finite VC dimension. Then,*

$$\lim_{n\to\infty} \mathbb{P}(\text{there exists any falsely certified } G \in \mathcal{G}) \leq \alpha.$$

**Proof** To simplify notation, we will replace $\mathbf{1}\{(X,Y) \in G\}$ with $\mathbf{1}_G$ throughout.

First, we restate the result we need to prove.

$$\lim_{n\to\infty} \mathbb{P}\left( \sup_{G \in \{G : H_0(G) \text{ holds}\}} \hat{\epsilon}(G) \geq \epsilon + t^* \cdot \frac{\hat{s}(G)}{\mathbb{P}_n(G)} \right) \leq \alpha.$$

Rearranging, we obtain

$$\lim_{n\to\infty} \mathbb{P}\left( \sup_{G \in \{G : H_0(G) \text{ holds}\}} \hat{\epsilon}(G) \geq \epsilon + t^* \cdot \frac{\hat{s}(G)}{\mathbb{P}_n(G)} \right) = \lim_{n\to\infty} \mathbb{P}\left( \sup_{G \in \{G : H_0(G) \text{ holds}\}} \frac{1}{\hat{s}(G)} \cdot \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) \geq t^* \right).$$

**Algorithm 7** Bootstrapping the Boolean certificate critical value (rescaled)

**Input:** Subpopulations $\mathcal{G}$, disparity $\epsilon$, holdout set $\mathcal{D}$, level $\alpha$, threshold $p_*$, weight $w_0$, number of bootstrap samples $B$
**for** $b = 1, \ldots, B$ **do**
  Let $\mathcal{D}_b^*$ be a sample with replacement of size $n$ from $\mathcal{D}$;
  Define $\mathbb{P}_b^*(G) := \frac{1}{n} \sum_{(x_i^*, y_i^*) \in \mathcal{D}_b^*} \mathbf{1}\{(x_i^*, y_i^*) \in G\}$;
  Define $\epsilon_b^*(G) := \frac{1}{\mathbb{P}_b^*(G) \cdot n} \sum_{(x_i^*, y_i^*) \in G} L_i^* - \hat{\theta}(\mathcal{D}_b^*)$;
**end for**
Define the asymptotic variance estimator by

$$\hat{\sigma}_G^2 := \widehat{\mathrm{Var}}(L \mid G) + \mathbb{P}_n(G) \left( \widehat{\mathrm{Var}}(\psi) - 2 \cdot \widehat{\mathrm{Cov}}(L, \psi \mid G) \right)$$

where $\widehat{\mathrm{Var}}(\cdot)$ and $\widehat{\mathrm{Cov}}(\cdot)$ correspond to the sample (conditional) variance and covariance;
Define $\hat{s}(G)$ by (9);
**for** $b = 1, \ldots, B$ **do**
  $t^{(b)} = \max_{G \in \mathcal{G}} \left\{ \frac{1}{\hat{s}(G)} \cdot (\mathbb{P}_b^*(G) \cdot (\epsilon_b^*(G) - \epsilon) - \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon)) \right\}$;
**end for**
**Return:** $t^* = \mathrm{Quantile}(1 - \alpha; \{t^{(b)}\}_{b=1}^B)$

Under $H_0(G)$, by assumption, $\mathbb{P}(G) \cdot (\epsilon(G) - \epsilon) \leq 0$, so it follows that

$$\begin{aligned}
&\lim_{n \to \infty} \mathbb{P}\left( \sup_{G \in \{G : H_0(G) \text{ holds}\}} \left\{ \frac{1}{\hat{s}(G)} \cdot \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) \right\} \geq t^* \right) \\
&\quad \leq \lim_{n \to \infty} \mathbb{P}\left( \sup_{G \in \{\mathcal{G} : H_0(G) \text{ holds}\}} \left\{ \frac{1}{\hat{s}(G)} \cdot (\mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) - \mathbb{P}(G) \cdot (\epsilon(G) - \epsilon)) \right\} \geq t^* \right) \\
&\quad \leq \mathbb{P}\left( \sup_{G \in \mathcal{G}} \left\{ \frac{1}{\hat{s}(G)} \cdot (\mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) - \mathbb{P}(G) \cdot (\epsilon(G) - \epsilon)) \right\} \geq t^* \right)
\end{aligned}$$

For $B = \infty$, recall that $t^*$ is the $(1 - \alpha)$-quantile of $\sup_{G \in \mathcal{G}} \{(1/\hat{s}(G)) \cdot (\mathbb{P}_n^*(G) \cdot (\epsilon^*(G) - \epsilon) - \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon))\}$. We must show that $t^*$ is consistent for the $(1 - \alpha)$-quantile of $\sup_{G \in \mathcal{G}} \{(1/\hat{s}(G)) \cdot (\mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) - \mathbb{P}(G) \cdot (\epsilon(G) - \epsilon))\}$.

Since we established that $1/\hat{s}(G)$ is uniformly consistent for $1/s(G)$ in the proof of Theorem 14, we can apply Slutsky's lemma and replace $1/\hat{s}(G)$ with $\frac{1}{s(G)}$ so long as $\mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) - \mathbb{P}(G) \cdot (\epsilon(G) - \epsilon)$ is $O_P(1)$. We establish this by showing that it is asymptotically equivalent to an empirical process indexed by some $P$-Donsker function class. To this end, we rewrite $\sqrt{n}[\mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) - \mathbb{P}(G) \cdot (\epsilon(G) - \epsilon)]$ as

$$\sqrt{n}(\mathbb{P}_n - P)[(L - \epsilon) \cdot \mathbf{1}_G] - \sqrt{n} \left( \hat{\theta} \cdot \mathbb{P}_n[\mathbf{1}_G] - \theta_P \cdot P[\mathbf{1}_G] \right).$$

We can further expand $\sqrt{n}\left(\hat{\theta}\cdot\mathbb{P}_n[\mathbf{1}_G]-\theta_P\cdot P[\mathbf{1}_G]\right)$ to obtain

$$\theta_P\cdot\left(\sqrt{n}(\mathbb{P}_n-P)[\mathbf{1}_G]\right)+\mathbb{P}(G)\cdot\left(\sqrt{n}(\hat{\theta}-\theta_P)\right)+\underbrace{\frac{1}{\sqrt{n}}\left(\sqrt{n}(\mathbb{P}_n-P)[\mathbf{1}_G]\right)\left(\sqrt{n}(\hat{\theta}-\theta_P)\right)}_{o_P(1)}.$$

Combining these results and applying the asymptotic linearity of $\sqrt{n}(\hat{\theta}-\theta_P)$, we can now linearize $\sqrt{n}\left[\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon)-\mathbb{P}(G)\cdot(\epsilon(G)-\epsilon)\right]$ to yield the (asymptotically) equivalent process,

$$\sqrt{n}(\mathbb{P}_n-P)[(L-\epsilon)\cdot\mathbf{1}_G]-\sqrt{n}(\mathbb{P}_n-P)[\theta_P\cdot\mathbf{1}_G]-\sqrt{n}(\mathbb{P}_n-P)[\mathbb{P}(G)\cdot\psi]+o_P(1).$$

Proceeding identically, we can also linearize the bootstrap analogue to this process, i.e.,

$$\sqrt{n}(\mathbb{P}_n^*-\mathbb{P}_n)[(L-\epsilon)\cdot\mathbf{1}_G]-\sqrt{n}(\mathbb{P}_n^*-\mathbb{P}_n)[\theta_P\cdot\mathbf{1}_G]-\sqrt{n}(\mathbb{P}_n^*-\mathbb{P}_n)[\mathbb{P}(G)\cdot\psi]+o_{P_n}(1).$$

The three terms are empirical or bootstrap empirical processes, respectively, indexed by the following function classes:

$$\mathcal{F}_1:=\{(L-\epsilon)\cdot\mathbf{1}_G\mid G\in\mathcal{G}\}\qquad\mathcal{F}_2:=\{\theta_P\cdot\mathbf{1}_G\mid G\in\mathcal{G}\}\qquad\mathcal{F}_3:=\{\mathbb{P}(G)\cdot\psi\mid G\in\mathcal{G}\}.$$

Applying Lemma 9.9(vi) of Kosorok (2008) with $g(x,y)=L-\epsilon$ and $\mathcal{F}=\{\mathbf{1}\{x\in G\}:G\in\mathcal{G}\}$ implies that $\mathcal{F}_1$ is VC. Then, our assumption that $\mathbb{E}_P[L^2]<\infty$ implies that $\mathcal{H}_1$ is $P$-Donsker (Theorem 2.10.20 in van der Vaart and Wellner (1996)). $\mathcal{F}_2$ is $P$-Donsker by Theorem 10. Last, observe that the uniform entropy integral of $\mathcal{F}=\{\mathbb{P}(G)\mid G\in\mathcal{G}\}$ is finite. By the same argument as $\mathcal{F}_1$, we conclude from Theorem 2.10.20 in van der Vaart and Wellner (1996) that $\mathcal{F}_3$ is $P$-Donsker.

Since the Donsker property is preserved under pointwise addition (Example 2.10.7 in van der Vaart and Wellner (1996)) and when taking subsets (Theorem 2.10.1 in van der Vaart and Wellner (1996)), we conclude that $\mathcal{F}_1+\mathcal{F}_2+\mathcal{F}_3$ is also $P$-Donsker. Thus, applying Slutsky's lemma and Theorem 11, we have shown that that the bootstrap distribution $\sqrt{n}\left[(1/\hat{s}(G))\cdot(\mathbb{P}_n^*(G)\cdot(\epsilon^*(G)-\epsilon)-\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon))\right]$ consistently estimates the limiting distribution of $\sqrt{n}\left[(1/\hat{s}(G))\cdot(\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon)-\mathbb{P}(G)\cdot(\epsilon(G)-\epsilon))\right]$. Since the sup is continuous, we can immediately claim by the continuous mapping theorem that

$$\sup_{G\in\mathcal{G}}\sqrt{n}\left[(1/\hat{s}(G))\cdot(\mathbb{P}_n^*(G)\cdot(\epsilon^*(G)-\epsilon)-\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon))\right]$$

consistently estimates its limiting distribution.

To conclude that $t^*$ is a valid critical value, we need to prove that the distribution function of the limiting distribution of $\sup_{G\in\mathcal{G}}\{\sqrt{n}\left[(1/\hat{s}(G))\cdot(\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon)-\mathbb{P}(G)\cdot(\epsilon(G)-\epsilon))\right]\}$ is strictly increasing and continuous at its $(1-\alpha)$-quantile.

The existence of some non-empty $G$ such that $L-\hat{\theta}$ is non-constant implies this fact. In particular, this assumption implies that the asymptotic variance of the process evaluated at $G$ is greater than 0. Thus, the $(1-\alpha)$-quantile of the limiting process is strictly greater than 0. Then, Theorem 12 yields the desired result.

Summarizing our argument, we have proven that the asymptotic probability of false certification can be upper bounded in the following manner,

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left(\sup_{G\in\{G:H_0(G) \text{ holds}\}} \hat{\epsilon}(G) \geq \epsilon + t^* \cdot \frac{\hat{s}(G)}{\mathbb{P}_n(G)}\right) \\
&\quad \leq \lim_{n\to\infty} \mathbb{P}\left(\sup_{G\in\mathcal{G}} \left\{\frac{1}{\hat{s}(G)} \cdot (\mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) - \mathbb{P}(G) \cdot (\epsilon(G) - \epsilon))\right\} \geq t^*\right) \\
&\quad = \lim_{n\to\infty} \mathbb{P}\left(\sup_{G\in\mathcal{G}} \left\{\frac{1}{\hat{s}(G)} \cdot (\mathbb{P}_n^*(G) \cdot (\epsilon^*(G) - \epsilon) - \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon))\right\} \geq t^*\right) = \alpha.
\end{aligned}$$

■

Theorem 7 follows immediately from Theorem 15 when $\hat{s}(G) := 1$. We remark also that any looseness in our upper bound of the Type I error disappears if $\epsilon(G) = \epsilon$ for all $G$. As alluded to in the main text, this corresponds to the condition under which $t^*$ achieves exact Type I error control.

### B.3 Alternative certification goals

Here, we describe how to extend the certification procedures described in the previous section and the main text to alternative notions of performance disparities.

When constructing *upper* confidence bounds or certifying $\epsilon(G) < \epsilon$, we simply multiply $t^*$ by $-1$. So, now $\epsilon_{\text{ub}}(G) = \hat{\epsilon}(G) + t^*\hat{s}(G)/\mathbb{P}_n(G)^2$, and we certify when $\hat{\epsilon}(G) \leq \epsilon - t^*/\mathbb{P}_n(G)$.

Observe that the argument for validity of the upper confidence bound / certificate goes through if we replace the $(1-\alpha)$-quantile of the sup-process with the $\alpha$-quantile of the inf-process. The latter, however, is just $-1$ times the former. Recall that the infimum of a centered Gaussian process is equal in distribution to $-1$ times the supremum of that process.

Next, we consider the problem of constructing confidence intervals. We propose to bootstrap the absolute process,

$$\sup_{G\in\mathcal{G}} \left|\frac{1}{\hat{s}(G)} \cdot \mathbb{P}(G) \cdot \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon(G))\right|.$$

Then, if $t^*$ denotes the bootstrap estimate of the $(1-\alpha)$-quantile of this process, the confidence set is constructed via

$$\left[\hat{\epsilon}(G) - t^* \cdot \frac{\hat{s}(G)}{\mathbb{P}_n(G)^2}, \hat{\epsilon}(G) + t^* \cdot \frac{\hat{s}(G)}{\mathbb{P}_n(G)^2}\right].$$

Last, for the interval certification task, we test the null hypothesis $H_0(G) : |\epsilon(G)| \geq \epsilon$. Equivalently, we test:

$$\bar{H}_0(G) : \{\epsilon(G) \geq \epsilon\} \bigcup \{\epsilon(G) \leq -\epsilon\}.$$

This is also known as a "bioequivalence" null and it can be tested by running the one-sided Boolean certification procedures for $H_0(G) : \epsilon(G) \geq \epsilon$ and $H_0(G) : \epsilon(G) \leq -\epsilon$ and

certifying when both are rejected. Formally, observe that we would test the first null at level $\alpha$ by rejecting when $\hat{\epsilon}(G) \leq \epsilon - t_1^*/\mathbb{P}_n(G)$ and the second null at level $\alpha$ by rejecting when $\hat{\epsilon}(G) \geq -\epsilon + t_2^*/\mathbb{P}_n(G)$. Then, we certify $G$ as satisfying $|\epsilon(G)| < \epsilon$ whenever both of these inequalities simultaneously hold.

## Appendix C. Flagging audits

To prove that our flagging procedure is asymptotically valid, we first establish that the proposed test statistic is asymptotically normal. Recall that $\hat{\epsilon}(G) := |G|^{-1} \sum_{i \in G} L_i - \hat{\theta}$. We denote the vector of statistics $\{\hat{\epsilon}(G)\}_{G \in \mathcal{G}}$ as $\hat{\epsilon}$, and the vector of corresponding true disparities as $\epsilon$.

**Proposition 16** *Assume that* $\mathbb{P}(G)$ *and* $\mathrm{Var}\left(L \mid G\right)$ *are bounded away from* 0 *for all* $G \in \mathcal{G}$. *If* $m = |\mathcal{G}| < \infty$, *then* $\sqrt{n}\left(\hat{\epsilon} - \epsilon\right) \rightsquigarrow \mathcal{N}_m\left(\mathbf{0}_m, \Sigma\right)$ *for some symmetric positive semi-definite* $\Sigma$.

**Proof** Apply the CLT to the asymptotic linear expansion derived in Theorem 13. ■

For any group $G_j \in \mathcal{G}$, Theorem 16 implies that the p-values

$$p_1(G_j; \epsilon) := \Phi\left(\frac{\sqrt{n}\left(\hat{\epsilon}(G_j) - \epsilon\right)}{\sqrt{\Sigma_{jj}}}\right), \quad p_2(G_j; \epsilon) := 1 - \Phi\left(\frac{\sqrt{n}\left(\hat{\epsilon}(G_j) + \epsilon\right)}{\sqrt{\Sigma_{jj}}}\right),$$
$$p_3(G_j) := p_1(G_j; -\epsilon) \wedge p_2(G_j; -\epsilon)$$

are marginally asymptotically valid for the null hypotheses $H_1(G_j) : \epsilon(G_j) \geq \epsilon$, $H_2(G_j) : \epsilon(G_j) \leq -\epsilon$, $H_3(G_j) : |\epsilon(G_j)| \leq \epsilon$, respectively, i.e., $\limsup_{n \to \infty} \mathbb{P}_{H_i}(p_i(G_j) \leq u) \leq u$.

$\Sigma_{jj}$ is not known, but by Slutsky's lemma, we can replace $\Sigma_{jj}$ with any consistent estimator. We can compute such an estimator analytically, but below we rely on the bootstrap to construct such an estimator. In particular, let

$$s_j^* := \mathrm{Quantile}\left(0.5, \sqrt{n}\left|\epsilon^*(G) - \hat{\epsilon}(G)\right|\right).$$

Then, Theorem 17 establishes conditions under which $s_j^*$ is a consistent estimator of $\sqrt{\Sigma_{jj}}$.

**Lemma 17** *Retain the assumptions of Theorem 16. Then,* $s_j^*/\Phi^{-1}(3/4) \xrightarrow{p} \sqrt{\Sigma_{jj}}$.

**Proof** Applying the bootstrap delta method and continuous mapping theorem we can conclude that the bootstrap distribution, $\sqrt{n}\left|\epsilon^*(G_j) - \hat{\epsilon}(G_j)\right|$ consistently estimates the distribution of $\sqrt{n}\left|\hat{\epsilon}(G_j) - \epsilon(G_j)\right|$. Since the limiting distribution of the latter is continuous and strictly increasing everywhere (Theorem 16), we can apply van der Vaart (2000, Lemma 5.10) to conclude that the bootstrap estimate of the median absolute deviation is consistent for the asymptotic median absolute deviation. Then, the result follows by recalling the well-known fact that the median absolute deviation of a Gaussian distribution with variance $\sigma^2$ is equal to $\Phi^{-1}(3/4) \cdot \sigma$. ■

Theorem 17 thus implies that we can replace $\sqrt{\Sigma_{jj}}$ in the p-value definitions with the Monte Carlo estimate given by $s_j^*$.

With these results, we can prove Theorem 8 after recalling some definitions and well-known conditions regarding the validity of the BH procedure.

**Definition 18** *We say that $X$ has positive regression dependency on a subset $I_0$ (PRDS on $I_0$) if for any increasing set $D$ and for each $i \in I_0$, $\mathbb{P}(X \in D \mid X_i = x)$ is increasing in $x$.*

**Example 1 (Case 3.1 in Benjamini and Yekutieli (2001))** *The one-sided Gaussian p-values obtained by testing $H_0 : \mu \leq \mu^*$ or $H_0 : \mu \geq \mu^*$, using $T \sim \mathcal{N}(\mu, \Sigma)$ are PRDS on $I_0$ if $\Sigma_{ij} \geq 0$ for all $i \in I_0$ and $j \in [m]$.*

**Theorem 19 (Theorem 1.2 in Benjamini and Yekutieli (2001))** *If the joint distribution of the test statistics is PRDS on the subset of test statistics corresponding to the $m_0$ true null hypotheses, the BH($\alpha$) procedure controls the FDR at level less than or equal to $\frac{m_0}{m}\alpha$.*

For the reader's convenience, we restate Theorem 8 before completing its proof.

**Proposition 20** *Assume that $\mathbb{P}((X,Y) \in G)$ and $\mathrm{Var}\left(L(X,Y) \mid (X,Y) \in G\right)$ is bounded away from 0 for all $G \in \mathcal{G}$, $\theta_P$ is a-priori known, and that one of the following conditions holds:*

*(i) $\{G\}_{G\in\mathcal{G}}$ are mutually disjoint;*
*(ii) $L$ takes values in $\{0,1\}$.*

*If we flag the rejections of the BH($\alpha$) procedure on $\{p(G)\}_{G\in\mathcal{G}}$, then the false discovery rate is asymptotically controlled at level $\alpha$.*

**Proof** For clearer indexing, we let $\mathcal{G} = \{G_j\}_{j=1}^m$. We assume w.l.o.g. that $\theta_P = 0$. Then, Theorem 16 shows that $\sqrt{n}\left\{\hat{\epsilon}(G_j) - \epsilon(G_j)\right\}_{j=1}^m \rightsquigarrow \mathcal{N}\left(\mathbf{0}_m, \Sigma\right)$. The off-diagonal entries of $\Sigma$ equal

$$\frac{\mathbb{E}\left[(L - \epsilon(G_j))(L - \epsilon(G_k))\mathbf{1}\left\{G_j \cap G_k\right\}\right]}{\mathbb{P}(G_j)\mathbb{P}(G_k)}.$$

Under condition (i), the off-diagonal entries are all 0 since the indicator in the numerator always equals 0.

Consider condition (ii). In this setting, we can rewrite the covariance expression as

$$\left(\epsilon(G_j \cap G_k)\left(1 - \epsilon(G_j) - \epsilon(G_k)\right) + \epsilon(G_j)\epsilon(G_k)\right)\frac{\mathbb{P}(G_j \cap G_k)}{\mathbb{P}(G_j)\mathbb{P}(G_k)}.$$

Then, assume for the sake of contradiction that this expression is negative. Then,

$$\begin{aligned} 0 &> \epsilon(G_j \cap G_k)\left(1 - \epsilon(G_j) - \epsilon(G_k)\right) + \epsilon(G_j)\epsilon(G_k) \\ &\geq 1 - \epsilon(G_j) - \epsilon(G_k) + \epsilon(G_j)\epsilon(G_k) \\ &= (1 - \epsilon(G_j))(1 - \epsilon(G_k)). \end{aligned}$$

But since the last expression is non-negative under condition (ii), we obtain a contradiction and must conclude that the asymptotic covariance is non-negative.

In both cases, we showed that the asymptotic distribution of these test statistics is a multivariate Gaussian with non-negative covariance. As a consequence, the one-sided p-values output by Algorithm 4 are asymptotically PRDS on the set of nulls. Then, applying the bounded convergence theorem and Theorem 19 yields the desired result. ■

While we do not have a proof of FDR control when testing the two-sided null, it is widely speculated that the BH procedure enjoys FDR control in this setting Fithian and Lei (2022). Moreover, even when the PRDS property does not hold (i.e., outside of the conditions given in Theorem 8), it is, in practice, quite challenging to obtain substantial violations of FDR control when applying the BH procedure to asymptotic Gaussian p-values.

## Appendix D. Auditing distribution shifts

Before considering the problem of auditing over shifts belonging to the unit ball of an RKHS, we observe that neither the proof of Theorem 14 nor the proof of Theorem 15 rely on $\mathcal{G}$ being a VC class. Since $\mathcal{G}$ is VC, $\mathcal{F} = \{\mathbf{1}\{(X,Y) \in G\} \mid G \in \mathcal{G}\}$ is a $P$-Donsker class. This implies that the proofs employed above are valid even if we replace $\mathcal{G}$ with some generic Donsker function class.

Extending our results to the RKHS setting, however, require some additional work. To prove that our method for constructing confidence sets on $\epsilon(h)$ are valid, we must prove some preliminary results regarding the unit ball of an RKHS.

**Lemma 21** *Assume that $\|k(X,X)\|_\infty$ is finite, and that $k(\cdot, x)$ is continuous. Then, the unit ball of the RKHS induced by $k$, which we denote by $\mathcal{H}_1$, is a $P$-Donsker class.*

**Proof** Lemma 4.28 and Lemma 4.33 of Steinwart and Christmann (2008) show that the assumptions are sufficient to guarantee that $\mathcal{H}_1$ is a separable Hilbert space and a subset of the space of bounded and continuous functions. The conclusion then follows from Theorem 1.1 of Marcus (1985) with $T$ chosen to be the identity. The identity mapping is trivially linear, and also meets the assumption of continuity in the sup-norm because the former is dominated by the RKHS norm. ■

**Lemma 22** *Retain the assumptions of Theorem 21. Then, for uniformly bounded functions $L$ and $M$, $\tilde{\mathcal{H}}_1 := \{P[L \cdot h] \cdot h \mid h \in \mathcal{H}_1\}$ is a $P$-Donsker class.*

**Proof** First, observe that $P[L \cdot h]$ is uniformly bounded for all $h \in \mathcal{H}_1$:

$$P[L \cdot h] \leq \|L\|_\infty \|k(X,X)\|_\infty \|h\|_{\mathcal{H}} \leq \|L\|_\infty \|k(X,X)\|_\infty =: C.$$

Thus, $\tilde{\mathcal{H}}_1$ is a subset of $\mathcal{H}_C$, i.e., it is a dilation of $\mathcal{H}_1$. Since $C \cdot \mathcal{H}_1$ is $P$-Donsker (Theorem 2.10.6 in van der Vaart and Wellner (1996)) and any subset of a $P$-Donsker class is $P$-Donsker (Theorem 2.10.1 in van der Vaart and Wellner (1996)), we conclude that $\tilde{\mathcal{H}}_1$ is $P$-Donsker. ■

Here we generalize the main theorem to include disparities that are defined relative to an estimated threshold $\hat{\theta}$; this threshold is assumed to satisfy the asymptotic linearity and bootstrap consistency assumptions stated in Appendix B.2. Algorithm 8 modifies Algorithm 5 so that the bootstrap accounts for estimation error in $\hat{\theta}$.

**Algorithm 8** Bootstrapping the RKHS confidence set critical value with estimated threshold

**Input:** Kernel $k$, holdout set $\mathcal{D}$, level $\alpha$, bootstrap samples $B$
Define $\mathbf{L} := \{L(f(x_i), y_i)\}_{i=1}^n$;
Define $\mathbf{K} := \{k(x_i, x_j)\}_{i,j=1}^n$;
**for** $b = 1, \ldots, B$ **do**
    Sample $\mathbf{w} \sim \text{Mult}\left(n; \frac{1}{n}, \ldots, \frac{1}{n}\right)$;
    Estimate bootstrap threshold deviation $t = \frac{1}{n}\sum_{i=1}^n (\mathbf{w}_i - 1) \cdot \psi_i$;
    $\mathbf{A} = \frac{1}{n^2}\left((\mathbf{w} \odot \mathbf{L})\,\mathbf{1}^\top - \mathbf{w}\mathbf{L}^\top - t \cdot \mathbb{I}_n\right)$;
    $t^{(b)} = \lambda_{\max}\left(\mathbf{K}^{1/2}\left(\frac{\mathbf{A}+\mathbf{A}^\top}{2}\right)\mathbf{K}^{1/2}\right)$;
**end for**
**Return:** $t^* = \text{Quantile}(1-\alpha; \{t^{(b)}\}_{b=1}^B)$

**Lemma 23** *For $B = \infty$, the $t^*$ output by Algorithm 8 equals the $(1-\alpha)$-quantile of*

$$\sup_{h \in \mathcal{H}_1} (\mathbb{P}_n^* - \mathbb{P}_n)[\mathbb{P}_n[h] \cdot L \cdot h - \mathbb{P}_n[L \cdot h] \cdot h - \mathbb{P}_n[h]^2 \cdot \psi].$$

**Proof** First, note that we can rewrite the process of interest using a multinomial variable,

$$\sup_{h \in \mathcal{H}_1} \mathbb{P}_n[W \cdot (L \cdot \mathbb{P}_n[h] - \mathbb{P}_n[L \cdot h])h] - \mathbb{P}_n[h]^2 \cdot \mathbb{P}_n[(W-1) \cdot \psi],$$

for $W \sim \text{Mult}(n, 1/n)$.

We can rewrite the process in terms of inner products between the unknown function $h \in \mathcal{H}$ and the kernel function,

$$\sup_{h \in \mathcal{H}_1} (\langle \mathbb{P}_n[W \cdot L \cdot k(X, \cdot)], h\rangle \langle \mathbb{P}_n[k(X, \cdot)], h\rangle - \langle \mathbb{P}_n[L \cdot k(X, \cdot)], h\rangle \langle \mathbb{P}_n[W \cdot k(X, \cdot)], h\rangle) - \mathbb{P}_n[(W-1) \cdot \psi] \langle \mathbb{P}_n[k(X, \cdot)], h\rangle^2.$$

Since we know that the optimal $h^*$ must be of the form $\sum_{i=1}^n \alpha_k k(\cdot, x_i)$ (recall the direct sum decomposition of any RKHS), we can rewrite the above supremum as

$$\sup_{\alpha : \alpha^\top \mathbf{K} \alpha \leq 1} \frac{1}{n^2}\left(\alpha^\top \mathbf{K}(\mathbf{w} \odot \mathbf{L})\mathbf{1}^\top \mathbf{K}\alpha - \alpha^\top \mathbf{K}\mathbf{w}\mathbf{L}^\top \mathbf{K}\alpha - t \cdot \alpha^T \mathbf{K}\mathbf{K}\alpha\right),$$

where $\mathbf{K} = \{k(x_i, x_j)\}_{i,j=1}^n$, $\mathbf{w} = (W_1, \ldots, W_n)^\top$, $\mathbf{L} = (L_1, \ldots, L_n)^\top$, and $t = \mathbb{P}_n[(W-1) \cdot \psi]$.

Letting $\mathbf{A} = \frac{1}{n^2}[(\mathbf{w} \odot \mathbf{L})\mathbf{1}^\top - \mathbf{w}\mathbf{L}^\top - t\mathbb{I}]$, we obtain the equivalent objectives,

$$\begin{aligned}\sup_{\alpha:\alpha^\top \mathbf{K}\alpha \leq 1} \frac{1}{2}\left(\alpha^\top \mathbf{K}\left(\mathbf{A}+\mathbf{A}^\top\right)\mathbf{K}\alpha\right) &= \sup_{\beta:\|\beta\|_2 \leq 1} \frac{1}{2}\left(\beta^\top \mathbf{K}^{1/2}\left(\mathbf{A}+\mathbf{A}^\top\right)\mathbf{K}^{1/2}\beta\right) \\ &= \frac{1}{2}\lambda_{\max}\left(\mathbf{K}^{1/2}\left(\mathbf{A}+\mathbf{A}^\top\right)\mathbf{K}^{1/2}\right).\end{aligned}$$

■

We remark that if $\theta_P$ is not estimated, $t = 0$ and the identity matrix in the definition of $A$ can be ignored. This greatly simplifies the computation required for the maximum eigenvalue, since $A + A^T$ is now low rank. As a consequence, we generally do not recommend using RKHS-based confidence sets for performance metrics that are defined relative to an unknown threshold.

With these preliminary results in hand, we can now prove our main theorem regarding RKHS confidence set validity. Given $t^*$ output by Algorithm 8, recall that we obtain a lower confidence bound by setting

$$\epsilon_{\text{lb}}(h) := \hat{\epsilon}(h) - \frac{t^*}{\left(\frac{1}{n}\sum_{i=1}^n h(x_i)\right)^2}.$$

For notational convenience, we let $\mathcal{H}_1^+$ denote the set of all non-negative functions belonging to $\mathcal{H}_1$.

**Theorem 24** *Assume that* $\text{Var}(L) > 0$, $\|L\|_\infty$ *and* $\|k(X,X)\|_\infty$ *are finite,* $k(\cdot, x)$ *is continuous, and that* $k(\cdot,\cdot)$ *is a positive definite kernel. Then,*

$$\lim_{n\to\infty} \mathbb{P}\left(\epsilon_{\text{lb}}(h) \leq \epsilon(h) \text{ for all } h \in \mathcal{H}_1^+\right) \geq 1 - \alpha.$$

**Proof** The desired result is equivalent to

$$\lim_{n\to\infty} \mathbb{P}\left(\exists\, h \in \mathcal{H}_1^+ \text{ s.t. } (\mathbb{P}_n[h] \cdot \mathbb{P}_n[h] \cdot (\hat{\epsilon}(h) - \epsilon(h)) > t^*\right) \leq \alpha.$$

Multiplying through by $\mathbb{P}_n[h]$, the process on the LHS can be rewritten as

$$\mathbb{P}_n[h] \cdot \mathbb{P}_n[(L - \hat{\theta} - \epsilon(h)) \cdot h].$$

We claim that $\mathbb{P}_n[h]$ is uniformly consistent for $\mathbb{P}[h]$. To see this, first recall that Theorem 21 implies that $\mathcal{H}_1$ is $P$-Donsker. The Donsker property is preserved for any subset, so if $\mathcal{H}_1$ is $P$-Donsker, then so is $\mathcal{H}_1^+$. Uniform consistency follows from $P$-Donsker $\implies$ $P$-Glivenko-Cantelli.

To apply Slutsky's lemma and replace $\mathbb{P}_n[h]$ with $\mathbb{P}[h]$, we must also show that $\sqrt{n} \cdot \mathbb{P}_n[(L - \hat{\theta} - \epsilon(h)) \cdot h]$ is $O_P(1)$. To this end, observe that

$$\begin{aligned}\left|\sup_{h\in\mathcal{H}_1^+} \sqrt{n} \cdot \mathbb{P}_n[(L - \hat{\theta} - \epsilon(h)) \cdot h]\right| \leq &\left|\sup_{h\in\mathcal{H}_1^+} \sqrt{n}(\mathbb{P}_n - P)[(L - \theta_P - \epsilon(h)) \cdot h]\right| \\ &+ \left|\sup_{h\in\mathcal{H}_1^+} \sqrt{n}(\mathbb{P}_n - P)[\psi] \cdot \mathbb{P}_n[h]\right|\end{aligned}$$

We can show that both of the terms on the RHS are $O_P(1)$. First, we claim that $\tilde{\mathcal{H}}_1 = \{(L - \epsilon(h))h \mid h \in \mathcal{H}_1^+\}$ is $P$-Donsker. First, observe that

$$\tilde{\mathcal{H}}_1 \subseteq \{L \cdot h \mid h \in \mathcal{H}_1^+\} - \{\epsilon(h) \cdot h \mid h \in \mathcal{H}_1^+\}.$$

Recalling that any subset of a $P$-Donsker class is also $P$-Donsker (Theorem 2.10.1 in van der Vaart and Wellner (1996)), we need to show that the RHS of this display is $P$-Donsker. Each function class on the RHS is $P$-Donsker. The first is because $\|L\|_\infty < \infty$ (Example 2.10.10 in van der Vaart and Wellner (1996)). Then, the second is $P$-Donsker because it is a subset of the elementwise product of two uniformly bounded $P$-Donsker classes (Example 2.10.8 in van der Vaart and Wellner (1996)). Last, elementwise addition of two $P$-Donsker classes yields a $P$-Donsker class (Example 2.10.7 in van der Vaart and Wellner (1996)). Thus, we conclude that $\sqrt{n}(\mathbb{P}_n - P)[(L - \epsilon(h))h] = O_P(1)$. We can upper bound the second term by $|\sqrt{n}(\mathbb{P}_n - P)[\psi]|$, so this term is also $O_P(1)$.

We now apply Slutsky's lemma and obtain:

$$\begin{aligned} P[h] \cdot \sqrt{n} \cdot \mathbb{P}_n[(L - \hat{\theta} - \epsilon(h))h] &= \sqrt{n} \cdot (\mathbb{P}_n[\{(L - \hat{\theta}) \cdot P[h] - P[(L - \theta_P) \cdot h]\}h]) \\ &= P[h] \cdot \sqrt{n}(\mathbb{P}_n[(L - \hat{\theta}) \cdot h] - P[(L - \theta_P) \cdot h]) \\ &\quad - P[(L - \theta_P) \cdot h] \cdot \sqrt{n}(\mathbb{P}_n - P)[h] \end{aligned}$$

Next, we show that this process is $P$-Donsker and converges to a tight Gaussian limit. To do so, we linearize the first term of the process:

$$\begin{aligned} P[h] \cdot \sqrt{n}(\mathbb{P}_n[(L - \hat{\theta}) \cdot h] - P[(L - \theta_P) \cdot h]) &= P[h] \cdot \sqrt{n}(\mathbb{P}_n - P)[L \cdot h] \\ &\quad - P[h] \cdot \sqrt{n}(\hat{\theta} \cdot \mathbb{P}_n[h] - \theta_P \cdot P[h])) \\ &= P[h] \cdot \sqrt{n}(\mathbb{P}_n - P)[L \cdot h] \\ &\quad - P[h] \cdot \theta_P \sqrt{n}(\mathbb{P}_n - P)[h] \\ &\quad - P[h]^2 \cdot \sqrt{n}(\mathbb{P}_n - P)[\psi] \\ &\quad + \underbrace{\frac{P[h]}{\sqrt{n}} \left( \sqrt{n}(\hat{\theta} - \theta_P) \cdot \sqrt{n}(\mathbb{P}_n - P)[h] \right)}_{o_P(1)}. \end{aligned}$$

Combining both terms, we conclude that the process is equivalent to

$$\begin{aligned} &\sqrt{n}(\mathbb{P}_n - P)[P[h] \cdot L \cdot h] - \sqrt{n}(\mathbb{P}_n - P)[P[L \cdot h] \cdot h] - \sqrt{n}(\mathbb{P}_n - P)[P[h]^2 \cdot \psi] \\ &\qquad = \sqrt{n}(\mathbb{P}_n - P)[P[h] \cdot L \cdot h - P[L \cdot h] \cdot h - P[h]^2 \cdot \psi] \end{aligned}$$

Observe that the function class indexing this process can be written as a subset of an elementwise sum of three classes,

$$\begin{aligned} &\{P[h] \cdot L \cdot h - P[L \cdot h] \cdot h - P[h]^2 \cdot \psi \mid h \in \mathcal{H}_1^+\} \\ &\qquad \subseteq \{P[h] \cdot L \cdot h \mid h \in \mathcal{H}_1^+\} - \{P[L \cdot h] \cdot h \mid h \in \mathcal{H}_1^+\} - \{P[h]^2 \cdot \psi \mid h \in \mathcal{H}_1^+\}, \end{aligned}$$

each of which is Donsker.

We check the first class is Donsker as an example; the arguments for the other two proofs follow identically. Note that $\{L \cdot h \mid h \in \mathcal{H}_1^+\}$ is a $P$-Donsker class because $L$ is uniformly bounded and $\{h \mid h \in \mathcal{H}_1\}$ is a uniformly bounded Donsker class (Example 2.10.10 in van der Vaart and Wellner (1996)). Then, $\{P[h] \cdot L \cdot h \mid h \in \mathcal{H}_1^+\}$ is a Donsker class because the subset of an elementwise product of two uniformly bounded Donsker classes is a Donsker class (Example 2.10.8 in van der Vaart and Wellner (1996)).

Thus, by the definition of a Donsker class and the continuous mapping theorem,

$$\sup_{h \in \mathcal{H}_1^+} \sqrt{n}(\mathbb{P}_n - P)[P[h] \cdot L \cdot h - P[L \cdot h] \cdot h - P[h]^2 \cdot \psi]$$

converges to a tight limit.

Then, since $\mathcal{H}_1^+ \subseteq \mathcal{H}_1$, we can upper bound

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\left(\sup_{h \in \mathcal{H}_1^+} \sqrt{n}(\mathbb{P}_n - P)[P[h] \cdot L \cdot h - P[L \cdot h] \cdot h - P[h]^2 \cdot \psi] > t^* \cdot \sqrt{n}\right) \\ \leq \lim_{n\to\infty} \mathbb{P}\left(\sup_{h \in \mathcal{H}_1} \sqrt{n}(\mathbb{P}_n - P)[P[h] \cdot L \cdot h - P[L \cdot h] \cdot h - P[h]^2 \cdot \psi] > t^* \cdot \sqrt{n}\right)\end{aligned}$$

Because this function class is $P$-Donsker, we also know that the analogous bootstrap process (replacing $\mathbb{P}_n$ with $\mathbb{P}_n^*$ and $P$ with $\mathbb{P}_n$) is consistent for

$$\sup_{h \in \mathcal{H}_1} \sqrt{n}(\mathbb{P}_n - P)[P[h] \cdot L \cdot h - P[L \cdot h] \cdot h - P[h]^2 \cdot \psi].$$

Thus, we conclude by Theorem 11 and the continuous mapping theorem that the supremum of the bootstrap process sampled in Algorithm 8 is consistent for the distribution of the supremum of the limit process.

Last, we need to show that the bootstrap quantile $t^*$ is a consistent estimator of the true limiting quantile. We establish consistency of the bootstrap quantile by verifying that the limiting distribution has a continuous and strictly increasing CDF at its $(1-\alpha)$-quantile (Lehmann et al., 2005, Lemma 11.2.1(ii)). The variance assumption on $L$ guarantees that for at least some $h \in \mathcal{H}$, the limiting distribution of

$$\sqrt{n}(\mathbb{P}_n - P)[P[h] \cdot L \cdot h - P[L \cdot h] \cdot h - P[h]^2 \cdot \psi]$$

is a non-trivial Gaussian, which then implies the desired CDF property. Theorem 12 implies consistency of $t^*$ and, thus, our desired claim regarding simultaneous coverage. ■

We might adapt the bootstrap process for the RKHS so that the confidence bound width scales more naturally with the "complexity" of the shift chosen. Here we do not consider defining Wald-style confidence bounds, but rather simply aim to adjust the process so that the confidence bound width scales more naturally with the "complexity" of the queried shift. For example, the current bound scales as

$$\epsilon_{\text{lb}}(h) = \hat{\epsilon}(h) - \frac{C}{\sqrt{n} \cdot \mathbb{P}_n[h]^2},$$

while we might wish the bound to scale as

$$\epsilon_{\text{lb}}(h) = \hat{\epsilon}(h) - \frac{C}{\sqrt{n_{\text{eff}}}}$$

where $n_{\text{eff}}$ quantifies the "effective sample size" of the reweighted metric.

We motivate our choice of $n_{\text{eff}} = (\sum_i h(x_i))^2 / \sum_i h(x_i)^2$ by observing that the variance of $(\sum_{i=1}^n z_i \cdot h(x_i))/(\sum_{i=1}^n h(x_i))$ for $z_i \overset{\text{iid}}{\sim} \text{Bern}(0.5)$ scales as $1/\sqrt{n_{\text{eff}}}$.

To motivate our choice of $\hat{s}(h)$, observe that rescaling the process by $1/\hat{s}(h)$ yields a bound of the form

$$\epsilon_{\text{lb}}(h) = \hat{\epsilon}(h) - C\frac{\hat{s}(h)}{\mathbb{P}_n[h]^2 \cdot \sqrt{n}}.$$

So, if we solve for $\hat{s}(\cdot)$ that yields the desired $\sqrt{n_{\text{eff}}}$ denominator, we obtain $|\mathbb{P}_n[h]| \cdot \sqrt{\mathbb{P}_n[h^2]}$. Truncating to ensure uniform consistency, we define

$$\hat{s}(h) := \max\left(|\mathbb{P}_n[h]| \cdot \sqrt{\mathbb{P}_n[h^2]},\, h_*\right),$$

for some threshold $h_*$. Unlike $p_*$ in (9), note that $h_*$ is not interpretable.

Besides the lack of interpretability, the rescaled RKHS process can no longer be efficiently bootstrapped. Even if we assume that $\theta_P$ is known, we must now compute in line 8 of Algorithm 8,

$$t^{(b)} = \sup_{h \in \mathcal{H}_1} \frac{\mathbb{P}_n^b[L \cdot h] \cdot \mathbb{P}_n[h] - \mathbb{P}_n[L \cdot h] \cdot \mathbb{P}_n^b[h]}{\max\left(|\mathbb{P}_n[h]| \cdot \sqrt{\mathbb{P}_n[h^2]},\, h_*\right)}.$$

While one might hope to mimic our previous approach and reduce this computation to some eigenvalue problem, applying the finite-dimensional representation of the RKHS function only yields

$$\sup_{\beta : \|\beta\|_2 \le 1} \frac{\left(\beta^\top \mathbf{K}^{1/2} \left(\mathbf{A} + \mathbf{A}^\top\right) \mathbf{K}^{1/2} \beta\right)}{2 \cdot \max\left(|\mathbf{1}^\top K^{1/2} \beta| \cdot \sqrt{\beta^\top K \beta},\, h_*\right)}.$$

The $\sqrt{\beta^\top K \beta}$ term in the denominator makes optimizing $\beta$ extremely challenging. Since $K$ is full-rank for a positive definite kernel, we cannot rely on any low-rank structure in $A$ to simplify this problem. At best, a rank-$m$ approximation to $K$ yields an intractable (and inaccurate) optimization problem over the surface of a $(m+4)$-dimensional hypersphere. If we simply drop the $\sqrt{\beta^\top K \beta}$ term from $\hat{s}(G)$, the bootstrap step can be reduced to an optimization problem over the surface of a 4-dimensional hypersphere. We can solve that problem via a brute-force search, but, in practice, we find that the resulting confidence bounds are not improved. As a consequence, we recommend against rescaling the RKHS process.