# Robust Validation: Confident Predictions Even When Distributions Shift*

Maxime Cauchois[1], Suyash Gupta[1], Alnur Ali[2], and John C. Duchi[1, 2]

[1]Department of Statistics, Stanford University
[2]Department of Electrical Engineering, Stanford University
{maxcauch, suyash28, alnurali, jduchi}@stanford.edu

August 2020

**Abstract**

While the traditional viewpoint in machine learning and statistics assumes training and testing samples come from the same population, practice belies this fiction. One strategy—coming from robust statistics and optimization—is thus to build a model robust to distributional perturbations. In this paper, we take a different approach to describe procedures for robust predictive inference, where a model provides uncertainty estimates on its predictions rather than point predictions. We present a method that produces prediction sets (almost exactly) giving the right coverage level for any test distribution in an $f$-divergence ball around the training population. The method, based on conformal inference, achieves (nearly) valid coverage in finite samples, under only the condition that the training data be exchangeable. An essential component of our methodology is to estimate the amount of expected future data shift and build robustness to it; we develop estimators and prove their consistency for protection and validity of uncertainty estimates under shifts. By experimenting on several large-scale benchmark datasets, including Recht et al.'s CIFAR-v4 and ImageNet-V2 datasets, we provide complementary empirical results that highlight the importance of robust predictive validity.

## 1 Introduction

The central conceit of statistical machine learning is that data comes from a population, and that a model fit on a training set and validated on a held-out validation set will generalize to future data. Yet this conceit is at best debatable: indeed, Recht, Roelofs, Schmidt, and Shankar [29] create new test sets for the central image recognition CIFAR-10 and ImageNet benchmarks, and they observe that published accuracies drop by between 3–15% on CIFAR and more than 11% on ImageNet (increases in error rate of 50–100%), even though the authors follow the original dataset creation processes. Given this drop in accuracy—even in carefully reproduced experiments—shift in the data generating distribution is inevitable, and should be an essential focus, given the growing applications of machine learning.

*Research supported by the NSF under CAREER Award CCF-1553086 and HDR 1934578 (the Stanford Data Science Collaboratory), Office of Naval Research YIP Award N00014-19-2288, and the Stanford DAWN Consortium.

To address such distribution shifts and related challenges, a growing literature advocates fitting predictive models that adapt to changes in the data generating distribution. For example, researchers suggest reweighting data to match new test distributions when covariates shift [37, 15], while work on distributional robustness [4, 14, 11, 12, 6, 5, 32] considers fitting models that optimize losses under worst-case distribution changes. Yet the resulting models often are conservative, appear to sacrifice accuracy for robustness, and even more, they may not be robust to natural distribution shifts [38]. The models also come with few tools for validating their performance on new data.

Instead of seeking robust models, we instead advocate focusing on models that provide *validity* in their predictions: a model should be able to provide some calibrated notion of its confidence, even in the face of distribution shift. Consequently, in this paper we revisit cross validation, validity, and conformal inference [40] from the perspective of robustness, advocating for more robust approaches to cross validation and equipping predictors with valid confidence sets. We present a method for robust predictive inference under distributional shifts, borrowing tools both from conformal inference [40] and distributional robustness. Our method can allow valid inferences even when training and test distributions are distinct, and we provide a (in our view well-motivated, but still heuristic) methodology to estimate plausible amounts of shift to which we should be robust.

To formalize, consider a supervised learning problem of predicting labels $y \in \mathcal{Y}$ from data $x \in \mathcal{X}$, where we assume we have a putative predictive model that outputs scores $s(x, y)$ measuring error (so that $s(x, y) < s(x, y')$ means that the model assigns higher likelihood to $y$ than $y'$ given $x$). For example, for a probabilistic model $p(y \mid x)$, a typical choice is the negative log likelihood $s(x, y) = -\log(p(y \mid x))$. For a distribution $Q_0$ on $\mathcal{X} \times \mathcal{Y}$,[1] we observe $(X_i, Y_i)_{i=1}^n \stackrel{\text{iid}}{\sim} Q_0$. Future data may come from $Q_0$ or a distribution $Q$ near—in some appropriate sense, deriving from distribution shift—to $Q_0$, and we wish to output valid predictions for future instances $(X, Y) \sim Q$, where $Q$ is unknown. The goal of this paper is twofold: first, given a level $\alpha \in (0, 1)$ and an uncertainty set $\mathcal{Q}$ of plausible shifted distributions, we wish to construct *uniformly valid* confidence set mappings $\widehat{C} : \mathcal{X} \rightrightarrows \mathcal{Y}$ of the form $\widehat{C}(x) = \{y \in \mathcal{Y} \mid s(x, y) \le q\}$ for a threshold $q$, which provide $1 - \alpha$ coverage, satisfying

$$Q(Y \in \widehat{C}(X)) \ge 1 - \alpha \quad \text{for all } Q \in \mathcal{Q}. \tag{1}$$

Second, we propose a methodology for finding a collection $\mathcal{Q}$ of plausible shifts, providing convergence theory and a concomitant empirical validiation on real distribution shift problems.

### 1.1 Background: split conformal inference under exchangeability

To set the stage, we review conformal predictive inference [40, 22, 23, 24, 1]. The setting here is a supervised learning problem where we have exchangeable data $\{(X_i, Y_i)\}_{i=1}^{n+1} \subset \mathcal{X} \times \mathcal{Y}$, and for a given confidence level $\alpha \in (0, 1)$ we wish to provide a confidence set $\widehat{C}(X_{n+1})$ such that $\mathbb{P}(Y_{n+1} \in \widehat{C}(X_{n+1})) \ge 1 - \alpha$. Standard properties of quantiles make such a construction possible. Indeed, assume that $S_1, \ldots, S_{n+1} \in \mathbb{R}$ are exchangeable random variables; then, the rank $\mathrm{rank}(S_j)$ of any $S_j$ among $\{S_i\}_{i=1}^{n+1}$—its position if we sort the values of the $S_i$—is evidently uniform on $\{1, \ldots, n+1\}$, assuming ties are broken randomly. Thus, for probability distributions $P$ on $\mathbb{R}$, defining the familiar quantile

$$\mathrm{Quantile}(\beta; P) := \inf \left\{ s \in \mathbb{R} : P(S \le s) \ge \beta \right\}, \tag{2}$$

[1] We always write $Q$ for a probability on $\mathcal{X} \times \mathcal{Y}$ and $P$ for the induced distribution on $s(X, Y)$ for $(X, Y) \sim Q$.

and Quantile$(\beta; \{S_i\}_{i=1}^n)$ to be the corresponding empirical quantile on $\{S_i\}_{i=1}^n$, we have

$$\mathbb{P}\left(S_{n+1} \le \text{Quantile}\left(\left(1+n^{-1}\right)(1-\alpha), \{S_i\}_{i=1}^n\right)\right) \ge \mathbb{P}(\text{rank}(S_{n+1}) \le \lceil (n+1)\alpha \rceil) \ge 1-\alpha.$$

Using this idea to provide confidence sets is now straightforward [40, 24]. Let $\{(X_i, Y_i)\}_{i=1}^n$ be a validation set—we assume here and throughout that we have already fit a model on training data independent of the validation set $\{(X_i, Y_i)\}_{i=1}^n$—and assume we have a scoring function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where a large value of $s(x, y)$ indicates that the point $(x, y)$ is *non-conforming*. In typical supervised learning tasks, such a function is easy to construct. Indeed, assume we have a predictor function $\mu$ (fit on an independent training set); in the case of regression, $\mu : \mathcal{X} \to \mathbb{R}$ predicts $\mathbb{E}[Y \mid X]$, while for a multiclass classification problem $\mu : \mathcal{X} \to \mathbb{R}^k$, and $\mu_y(x)$ is large when the model predicts class $y$ to be likely given $x$. Then natural nonconformity scores are $s(x, y) = |\mu(x) - y|$ for regression and $s(x, y) = -\mu_y(x)$ for classification. As long as $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable, if we define $\widehat{\mathcal{Q}}_{n,1-\alpha} := \text{Quantile}\left((1+n^{-1})\alpha; \{s(X_i, Y_i)\}_{i=1}^n\right)$, the confidence set

$$\widehat{C}_n(x) := \left\{y \in \mathcal{Y} \mid s(x, y) \le \widehat{\mathcal{Q}}_{n,1-\alpha}\right\}, \tag{3}$$

immediately satisfies

$$\mathbb{P}(Y_{n+1} \in \widehat{C}_n(X_{n+1})) = \mathbb{P}\left(s(X_{n+1}, Y_{n+1}) \le \widehat{\mathcal{Q}}_{n,1-\alpha}\right) \ge 1-\alpha, \tag{4}$$

whatever the scoring function $s$ and distribution on $(X_i, Y_i)$ [40, 24]. The coverage statement (4) depends critically (as we shall see) on the exchangeability of the samples, failing if even the marginal distribution over $X$ changes, and it does not imply conditional coverage: we have no guarantee that $\mathbb{P}(Y \in \widehat{C}(X) \mid X) \ge 1-\alpha$.

### 1.2 Related work

The machine learning community has long identified distribution shift as a challenge, with domain adaptation strategies and covariate shift two major foci [37, 28, 2, 42, 26], though much of this work focuses on model estimation and selection strategies, and one often assumes access to data (or at least likelihood ratios) of data from the new distribution. We argue that a model should instead provide robust and valid estimates of its confidence rather than simply predictions that may or may not be robust. There is a growing body of work on distributionally robust optimization (DRO), which considers worst-case dataset shifts in neighborhoods of the training distribution; these have been important in finance and operations research, where one wishes to guard against catastrophic losses [30, 3, 4]. In DRO in statistical learning [33, 5, 13, 14, 35], the focus has also been on improving estimators rather than inferential predictive tasks. We extend this distributional robustness to apply in predictive inference.

Vovk et al. [40]'s conformal inference provides an important tool for valid predictions. The growing applications of machine learning and predictive analytics have renewed interest in predictive validity, and recent papers attempt to move beyond the standard exchangeability assumptions upon which conformalization reposes [39, 8], though this typically requires some additional assumptions for strict validity. Of particular relevance to our setting is Tibshirani et al.'s work [39], which considers conformal inference under covariate shift, where the marginal over $X$ changes while $P(Y \mid X)$ remains fixed. Validity in this setting requires knowing a likelihood ratio of the shift, which in high dimensions is challenging. In addition, as Jordan [19] argues, in typical practice covariate shifts are no more plausible than other (more general) shifts, especially in situations with unobserved confounders. For this reason, we take a more general approach and do not restrict to specific structured shifts.

### 1.3 A few motivating examples

Standard validation methodology randomly splits data into train/validation/test sets, artificially enforcing exchangeabilty). Thus, to motivate for the challenges in predictive validity even under simple covariate shifts—we only modify the distribution of $X$, returning later to more sophisticated real-world scenarios—we experiment on nine regression datasets from the UCI repository [10]. We repeat the following 50 times. We randomly partition each dataset into disjoint sets $D_{\text{train}}, D_{\text{val}}, D_{\text{test}}$, each consisting of 1/3 of the data. We fit a random forest predictor $\mu$ using $D_{\text{train}}$ and construct conformal intervals of the form (3) with $s(x, y) = |\mu(x) - y|$, so that $\widehat{C}_n(x) = \{y \mid |\mu(x) - y| \leq \hat{t}\}$ for a threshold $\hat{t}$ achieving coverage at nominal level $\alpha = .05$ on $D_{\text{val}}$, as is standard in split-conformal prediction [40, 24]. We evaluate coverage on tiltings of varying strength on $D_{\text{test}}$: letting $v$ be the top eigenvector of the test $X$-covariance $\Sigma_{\text{test}}$ and $\overline{x}_{\text{test}}$ be the mean of $X$ over $D_{\text{test}}$, we reweight $D_{\text{test}}$ by probabilities proportional to $w(x) = \exp(av^T(x - \overline{x}_{\text{test}}))$ for tilting parameters $a \in \pm\{0, .02, .04, .08, .16, .32, .64\}$. Essentially, this shift asks the following question: why would we *not* expect a shift along the principal directions of variation in $X$ on future data?

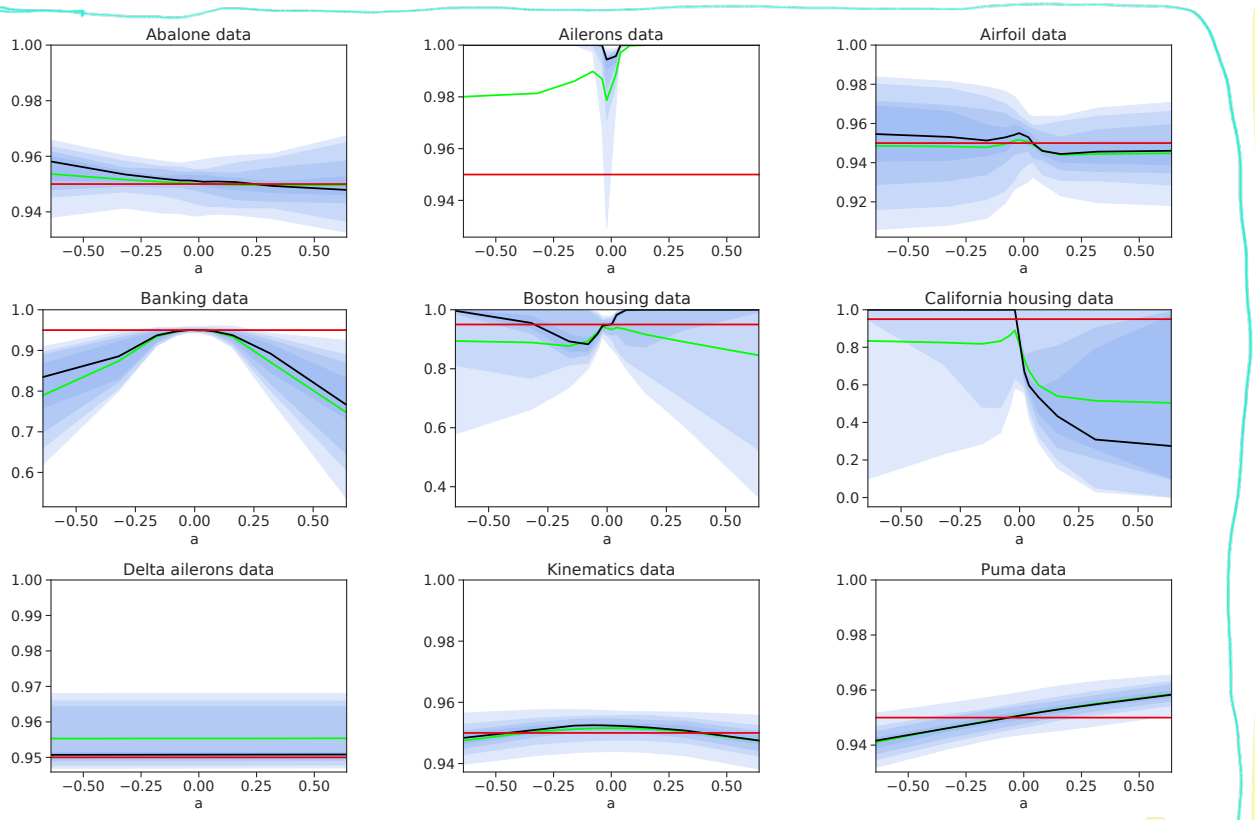


**Figure 1.** Empirical coverage for the prediction sets generated by the standard conformal methodology across nine regression data sets and 50 random splits of each data set, with an exponential tilting in $X$ space along the first principal component of $X$. The horizontal axis gives the value of the tilting parameter $a$; the vertical the coverage level. A green line marks the average coverage, a black line marks the median coverage, and the horizontal red line marks the nominal coverage .95. The blue bands show the coverage at deciles over 50 splits.

Figure 1 presents the results: even when the covariate shifts are small, which corresponds to tilting parameters $a$ with small magnitude, prediction intervals from the standard conformal methodology frequently fail to cover (sometimes grossly) the true response values. While this

is but a simple motivating (essentially) simulation, if we expect some shift in future data—say along the directions of principal variation in $X$, as the data itself is already variable along that axis—it seems that standard validation approaches [36, 16] provide too rosy of a picture of future validity [29], as they *enforce* exchangeability by randomly splitting data.

## 2 Robust predictive inference

Of course, standard cross validation and conformalization methodology makes no claims of validity without exchangeability [40, 1, 36], so their potential failure even under simple covariate shifts is not completely surprising. The coverage (4) relies on the exchangeability assumption between the training and test data, and evidently can quickly collapse when the test distribution violates that assumption, as Section 1.3 shows. These observations thus call for a notion of confidence more robust to potential future shifts.

Assume as usual that we have a score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and observe data $\{(X_i, Y_i)\}_{i=1}^n$ such that $\{S_i\}_{i=1}^n := \{s(X_i, Y_i)\}_{i=1}^n \stackrel{\text{iid}}{\sim} P_0$, so that $P_0$ is the push-forward of $(X, Y) \sim Q_0$ under $s(X, Y)$. For a set $\mathcal{P}(P_0)$ of potential future *score* distributions on $\mathbb{R}$, our goal is to achieve coverage (1) for all distributions $Q$ on pairs $(X, Y)$ that induce a distribution $P$ on $s(X, Y)$ such that $P \in \mathcal{P}(P_0)$, that is,

$$Q \in \mathcal{Q}(s, \mathcal{P}(P_0)) := \{Q \text{ s.t. for } (X, Y) \sim Q, \text{the score } s(X, Y) \sim P \in \mathcal{P}(P_0)\}.$$

Our focus is exclusively on validating our predictive model, not changing it, so we follow standard conformal practice [40, 1] and use confidence sets $\widehat{C}(x)$ to be of the form $\widehat{C}(x) = \{y \in \mathcal{Y} \mid S(x, y) \le t\}$ for a threshold $t \in \mathbb{R}$. For confidence sets of this form, the choice $t := \max_{P \in \mathcal{P}(P_0)} \text{Quantile}(1 - \alpha, P)$ is the smallest $q \in \mathbb{R}$ such that $P(S \le q) \ge 1 - \alpha$ for every distribution $P \in \mathcal{P}(P_0)$ of the scores. Our general problem to achieve coverage (1) with uncertainty $\mathcal{Q}(s, \mathcal{P}(P_0))$ thus reduces to the optimization problem

$$\text{maximize } \text{Quantile}(1 - \alpha; P) \quad \text{subject to } P \in \mathcal{P}(P_0). \tag{5}$$

In the next section, we characterize solutions to this problem, showing in Section 2.2 we show how to use the characterizations to achieve coverage on future data.

### 2.1 Characterizing and computing quantiles over $f$-divergence balls

It remains to specify a set of distributions $\mathcal{P}(P_0)$ that makes problem (5) computationally tractable and statistically meaningful. We thus consider various restrictions on the likelihood ratio $dP/dP_0$ for $P \in \mathcal{P}(P_0)$. Following the distributionally robust optimization literature (DRO) [3, 33, 5, 11, 14, 35], we consider $f$-divergence balls. Given a closed convex function $f : \mathbb{R} \to \mathbb{R}$ satisfying $f(1) = 0$ and $f(t) = +\infty$ for $t < 0$, the $f$-divergence [9] between probability distributions $P$ and $Q$ on a set $\mathcal{Z}$ is

$$D_f(P \| Q) := \int_{z \in \mathcal{Z}} f\left(\frac{dP(z)}{dQ(z)}\right) dQ(z).$$

Jensen's inequality guarantees that $D_f(P \| Q) \ge 0$ always, and familiar examples include $f(z) = z \log z$, which induces the KL-divergence, and $f(t) = \frac{1}{2}(t-1)^2$, which gives $\chi^2$-divergence. We consider a slight restriction of the collection of all $f$-divergences and consider

those for which $f$ is 1-coercive, meaning that $\lim_{z\to\infty} f(z)/z = \infty$. We study problem (5) in the case where $\mathcal{P}(P_0)$ is an $f$-divergence ball, where $f$ is 1-coercive, of radius $\rho$ around $P_0$:

$$\mathcal{P}_{f,\rho}(P_0) := \{P \text{ s.t. } D_f(P\|P_0) \le \rho\}. \tag{6}$$

Unlike most work in the DRO literature, instead of trying to build a model minimizing a DRO-type loss, we assume we already have a model and wish to robustly validate it: no matter the model's form, we wish to provide predictive confidence sets that are valid and robust to distribution shifts. By the data processing inequality, all distributions $Q$ on $(X, Y)$ satisfying $D_f(Q\|Q_0) \le \rho$ induce a distribution $P$ on $s(X, Y)$ satisfying $D_f(P\|P_0) \le \rho$, so solving problem (5) with $\mathcal{P}_{f,\rho}(P_0)$ provides coverage for all sufficiently small shifts on $(X, Y) \sim Q_0$.

We show how to solve problem (5) for fixed $f$ and $\rho$ defining the constraint (6) by characterizing worst-case quantiles, essentially reducing the problem to a one-parameter (Bernoulli) problem. The choice of $f$ and $\rho$ determine plausible amounts of shift—appropriate chioces are a longstanding problem [13]—and we defer approaches for selecting them to the sequel. For $\alpha \in (0, 1)$ and any distribution $P$ on the real line, we define the $(\alpha, \rho, f)$-worst-case quantile

$$\text{Quantile}^{\text{WC}}_{f,\rho}(\alpha; P) := \sup_{D_f(P_1\|P)\le\rho} \text{Quantile}(\alpha; P_1).$$

Key to our results both here and on valid coverage in Section 2.2 is that this worst-case quantile is a standard quantile of $P$ at a level that depends only on $f, \rho$, and $\alpha$, but not on $P$.

**Proposition 1.** *Define the function $g_{f,\rho} : [0, 1] \to [0, 1]$ by*

$$g_{f,\rho}(\beta) := \inf\left\{z \in [0, 1] : \beta f\left(\frac{z}{\beta}\right) + (1-\beta) f\left(\frac{1-z}{1-\beta}\right) \le \rho\right\}.$$

*Then the inverse*

$$g^{-1}_{f,\rho}(\tau) = \sup\{\beta \in [0, 1] : g_{f,\rho}(\beta) \le \tau\}$$

*guarantees that for all distributions $P$ on $\mathbb{R}$ and $\alpha \in (0, 1)$,*

$$\text{Quantile}^{\text{WC}}_{f,\rho}(\alpha; P) = \text{Quantile}(g^{-1}_{f,\rho}(\alpha); P).$$

See Appendix A.1 for a proof of the proposition.

Proposition 1 shows that it is easy to compute $g_{f,\rho}$ and $g^{-1}_{f,\rho}$, as they are both solutions to one-dimensional convex optimization problems and therefore admit efficient binary search procedures. In some cases, we have closed forms; for example for $f(t) = (t-1)^2$, we have $g_{f,\rho}(\beta) = [\beta - \sqrt{2\rho\beta(1-\beta)}]_+$. Letting $g = g_{f,\rho}$ for shorthand, we sketch how to compute $g^{-1}$ efficiently. Computing the inverse $g^{-1}(\tau)$ is equivalent to solving the optimization problem

$$\underset{0\le\beta,z\le 1}{\text{maximize}} \ \beta \quad \text{subject to } z \le \tau, \ \ \beta f\left(\frac{z}{\beta}\right) + (1-\beta) f\left(\frac{1-z}{1-\beta}\right) \le \rho.$$

We seek the largest $\beta \ge \tau$ feasible for this problem (as $\beta = \tau$ is feasible); because $h(\beta, z) = \beta f(z/\beta) + (1-\beta) f((1-z)/(1-\beta))$ is convex and minimized at any $z = \beta$ with $h(z, z) = 0$, for $\beta \ge \tau$ it is evident that $\inf_{0\le z\le\tau} h(\beta, z) = h(\beta, \tau)$. Thus may equivalently write

$$g^{-1}_{f,\rho}(\tau) = \sup\left\{\beta \in [\tau, 1] \mid \beta f\left(\frac{\tau}{\beta}\right) + (1-\beta) f\left(\frac{1-\tau}{1-\beta}\right) \le \rho\right\},$$

which a quick binary search over feasible $\beta \in [\tau, 1]$ solves to accuracy $\epsilon$ in time $\log \frac{1-\tau}{\epsilon}$.

## 2.2 Achieving coverage with empirical estimates

With the characterization of $\text{Quantile}^{\text{WC}}$, we can define the corresponding prediction set

$$C_{f,\rho}(x; P) := \{y \in \mathcal{Y} \mid s(x, y) \le \text{Quantile}^{\text{WC}}_{f,\rho}(1 - \alpha; P)\}. \tag{7}$$

As we observe only a sample $\{S_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_0$, we use the empirical plug-in to develop confidence sets (7) (and therefore in problem (5)), considering $\widehat{C}_{n,f,\rho}(x) := C_{f,\rho}(x; \hat{P}_n)$. By doing this, Proposition 1 allows us to derive guarantees for the prediction set (7) from standard quantile statistics. In particular, the next proposition, whose proof we give in Appendix A.2, lower bounds future coverage conditionally on the validation set $\{(X_i, Y_i)\}_{i=1}^n$ and relates future test coverage to the amount of shift.

**Proposition 2.** *Let $S_{n+1} = s(X_{n+1}, Y_{n+1}) \sim P_{\text{test}}$ is independent of $\{S_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_0$, and let $\rho^\star = D_f(P_{\text{test}} \| P_0) \in [0, \infty)$. Let $F_0$ be the c.d.f. of $P_0$. Then the confidence set $\widehat{C}_{n,f,\rho}(x) := C_{f,\rho}(x; \hat{P}_n)$ satisfies*

$$\begin{aligned}\mathbb{P}\left(Y_{n+1} \in \widehat{C}_{n,f,\rho}(X_{n+1}) \mid \{(X_i, Y_i)\}_{i=1}^n\right) &\ge g_{f,\rho^\star}\left(F_0\left(\text{Quantile}^{\text{WC}}_{f,\rho}(1 - \alpha; \hat{P}_n)\right)\right) \\ &= g_{f,\rho^\star}\left(F_0\left(\text{Quantile}(g^{-1}_{f,\rho}(1 - \alpha); \hat{P}_n)\right)\right).\end{aligned}$$

With the two preceding propositions, we turn to the main coverage theorem and a few corollaries, which provide the validity of coverage as long as the true shift between $P_0$ and $P_{\text{test}}$ is no more than our guess. We provide the proof of the theorem in Appendix A.3.

**Theorem 1.** *Assume that $S_{n+1} = s(X_{n+1}, Y_{n+1}) \sim P_{test}$ is independent of $\{S_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_0$, and let $\rho^\star = D_f(P_{test} \| P_0) < \infty$. Then*

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_{n,f,\rho}(X_{n+1})\right) \ge g_{f,\rho^\star}\left(\frac{\left\lceil n g^{-1}_{f,\rho}(1 - \alpha)\right\rceil}{n + 1}\right).$$

The theorem as stated is a bit unwieldy, so we develop a few relatively straightforward corollaries, whose proofs we provide in Appendix A.4. In each, we assume that the $\rho$ we use to construct the confidence sets (7) satisfies $\rho \ge \rho^\star = D_f(P_{\text{test}} \| P_0)$, which guarantees validity.

**Corollary 2.1.** *Let the conditions of Theorem 1 hold, but additionally assume that $\rho^\star = D_f(P_{test} \| P_0) \le \rho$. Then for $c_{\alpha,\rho,f} := g^{-1}_{f,\rho}(1 - \alpha) g'_{f,\rho}(g^{-1}_{f,\rho}(1 - \alpha)) < \infty$, we have*

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}_{n,f,\rho}(X_{n+1})\right) \ge 1 - \alpha - \frac{c_{\alpha,\rho,f}}{n + 1}.$$

If instead we replace $\alpha$ in the definition (7) of the confidence set $C_{f,\rho}(x; P)$ with

$$\alpha_n := 1 - g_{f,\rho}\left((1 + 1/n) g^{-1}_{f,\rho}(1 - \alpha)\right) = \alpha - O(1/n),$$

we can construct the corrected empirical confidence set

$$\widehat{C}^{\text{corr}}_{n,f,\rho}(x) := \left\{y \in \mathcal{Y} \mid s(x, y) \le \text{Quantile}^{\text{WC}}_{f,\rho}(1 - \alpha_n; \hat{P}_n)\right\}.$$

We then have the correct level $\alpha$ coverage:

**Corollary 2.2.** *Let the conditions of Corollary 2.1 hold. Then*

$$\mathbb{P}\left(Y_{n+1} \in \widehat{C}^{\text{corr}}_{n,f,\rho}(X_{n+1})\right) \geq 1 - \alpha.$$

Summarizing, the empirical prediction sets $\widehat{C}_{n,f,\rho}$ and $\widehat{C}^{\text{corr}}_{n,f,\rho}$ achieve nearly or better than $1-\alpha$ coverage if the $f$-divergence between the new distribution $P_{\text{test}}$ and the current distribution $P_0$ remains below $\rho$. When this fails, Theorem 1 shows graceful degradation in converage as long as the divergence between $P_{\text{test}}$ and the validation population $P_0$ is not too large.

## 3 Adaptive procedures for estimating future distribution shift

While the results in the previous section apply for a fixed shift amount $\rho$, a fundamental challenge is—given a validation data set—to determine the amount of shift against which to protect. We suggest a methodology to identify shifts motivated by two (somewhat oppositional) perspectives: first, the variability in predictions in current data is suggestive of the amount of variability we might expect in the future; second, from the perspective of protection against future shifts, that there is no reason future data would *not* shift as much as we can observe in a given validation set. As a motivating thought experiment, consider the case that the data is a mixture of distinct sub-populations. Should we provide valid coverage for each of these sub-populations, we expect our coverage to remain valid if the future (test) distribution remains any mixture of the same sub-populations. In empirical risk minimization (ERM)-based models, we expect rarer sub-populations to have higher non-conformity scores than average, and building on this intuition, our procedures look for regions in validation data with high non-conformity scores, choosing $\rho$ to give valid coverage in these regions.

We adopt a two-step procedure to describe the set of shifts we consider. Abstractly, let $\mathcal{V}$ be a (potentially infinite) set indexing "directions" of possible shifts, and to each $v \in \mathcal{V}$ associate a collection $\mathcal{R}_v$ of subsets of $\mathcal{X}$. (Typically, we take $\mathcal{V} \subset \mathbb{R}^d$ when $\mathcal{X} \subset \mathbb{R}^d$.) Then for each $R \in \mathcal{R} := \bigcup_{v\in\mathcal{V}} \mathcal{R}_v \subset \mathcal{P}(\mathcal{X})$, we consider the shifted distribution

$$dQ_R(x, y) = \frac{1\{x \in R\}}{Q_0(X \in R)} dQ_0(x, y) = dQ_0(x, y \mid x \in R), \tag{8}$$

which restricts $X$ to a smaller subset $R$ of the feature space without changing the conditional distribution of $Y \mid X$. The intuition behind the approach is twofold: first, conditionally valid preditors remain valid under covariate shifts of only $X$ (so that we hope to identify failures of validity under such shifts), and second, there may exist privileged directions of shift in the $\mathcal{X}$-space (e.g. time in temporal data or protected attributes in data with sensitive features) for which we wish to provide appropriate $1 - \alpha$ coverage.

**Example 1** (Slabs and Euclidean balls)**:** Our prototypical example is slabs (hyperplanes) and Euclidean balls, where we take $\mathcal{V} \subset \mathbb{R}^d$, both of which have VC-dimension $O(d)$. In the slab case, for $v \in \mathbb{R}^d$ we define the collection of slabs orthogonal to $v$,

$$\mathcal{R}_v = \left\{\{x \in \mathbb{R}^d \mid a \leq v^T x \leq b\} \text{ s.t. } a < b\right\}.$$

In the Euclidean ball case, we consider $\mathcal{R}_v = \{\{x \in \mathbb{R}^d \mid \|x - v\|_2 \leq r\} \text{ s.t. } r > 0\}$, the collection of $\ell_2$-balls centered at $v \in \mathcal{V} = \mathbb{R}^d$. $\diamond$

Given $\delta \in (0, 1)$, we define the *worst coverage* for a confidence set mapping $C : \mathcal{X} \rightrightarrows \mathcal{Y}$ over $\mathcal{R}$-sets of size $\delta$ by

$$\text{WC}(C, \mathcal{R}, \delta; Q) := \inf_{R\in\mathcal{R}} \{Q\left(Y \in C(X) \mid X \in R\right) \text{ s.t. } Q(X \in R) \geq \delta\} \tag{9}$$

Our goal is to find a (tight) confidence set $\widehat{C}$ such that $\mathrm{WC}(\widehat{C}, \mathcal{R}, \delta; Q_0) \geq 1 - \alpha$, which, in the setting of Section 2, corresponds to choosing $\rho > 0$ such that

$$\mathrm{WC}(\widehat{C}_{n,f,\rho}, \mathcal{R}, \delta;\, Q_0) \geq 1 - \alpha.$$

That is, we seek $1 - \alpha$ coverage over all large enough subsets of $X$-space.

Unfortunately, the computation of the worst coverage (9) is usually challenging when the dimension $d$ of the problem grows (e.g. in Example 1), since it typically involves minimizing a non-convex function over a $d$-dimensional domain. This makes the estimation of quantity (9) intractable for large values of $d$, and also hints that requiring such coverage to hold uniformly over all directions $v \in \mathcal{V}$ may be too stringent for practical purposes. However, for a fixed $v \in \mathbb{R}^d$, both sets $\mathcal{R}_v$ in Example 1 admit $O(d{\cdot}n)$-time algorithms for computing $\mathrm{WC}(C, \mathcal{R}_v, \delta; \widehat{Q}_n)$ for any empirical distribution $\widehat{Q}_n$ with support on $n$ points, which in the slab case is the maximum density segment problem [27]. Thus instead of the full worst coverage (9), we typically resort to a slightly weaker notion of robust coverage, where we require coverage to hold for "most" distributions of the form (8). In the next two sections, we therefore consider two approaches: one that samples directions $v \in \mathcal{V}$, seeking good coverage with high probability, and the other that proposes surrogate convex optimization problems to find the worst direction $v$, which we can show under (strong) distributional assumptions is optimal.

### 3.1 High-probability coverage over specific classes of shifts

Our first approach is to let $\mathbb{P}_{\mathsf{v}}$ be a distribution on $v \in \mathcal{V}$ that models plausible future shifts. A natural desiderata here is to provide coverage with high probability, that is, conditional on $\widehat{C}$, to guarantee that for a hyperparameter $0 < \alpha_{\mathsf{v}} < 1$ and for $v \sim \mathbb{P}_{\mathsf{v}}$,

$$\mathbb{P}_{\mathsf{v}}\left[\mathrm{WC}(\widehat{C}, \mathcal{R}_v, \delta;\, Q_0) \geq 1 - \alpha\right] \geq 1 - \alpha_{\mathsf{v}}. \tag{10}$$

That is, we require that with $\mathbb{P}_{\mathsf{v}}$-probability $1 - \alpha_{\mathsf{v}}$ over the direction $v$ of shift, the confidence set $\widehat{C}(X)$ provides $1-\alpha$ coverage over all $R \in \mathcal{R}_v$ satisfying $Q_0(X \in R) \geq \delta$. The coverage (10) becomes more conservative as $\alpha_{\mathsf{v}}$ decreases to 0, reducing to condition (9) at $\alpha_{\mathsf{v}} = 0$.

Before presenting the procedure, we index the confidence sets by the threshold $q$ for the score function $s$, providing a complementary condition via the robust prediction set (7).

**Definition 3.1.** *For $q \in \mathbb{R}$, the* prediction set at level $q$ *is*

$$C^{(q)}(x) := \{y \in \mathcal{Y} \mid s(x, y) \leq q\}.$$

*For a distribution $P$ on $\mathbb{R}$, the value $\rho$ provides* sufficient divergence for threshold $q$ *if*

$$C_{f,\rho}(x; P) \supset C^{(q)}(x) \text{ for all } x \in \mathcal{X}.$$

By the definition (7) of $C_{f,\rho}$ and Proposition 1, we see that $\rho$ gives sufficient divergence for threshold $q$ if and only if

$$\mathrm{Quantile}^{\mathrm{WC}}_{f,\rho}(1 - \alpha; P) = \mathrm{Quantile}(g^{-1}_{f,\rho}(1 - \alpha; P)) \geq q.$$

To output a confidence set $\widehat{C}$ satisfying the high probability worst-coverage (10), we wish to find $q \in \mathbb{R}$ such that $\mathbb{P}_{\mathsf{v}}[\mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta; Q_0) \geq 1 - \alpha] \geq 1 - \alpha_{\mathsf{v}}$. Notably, any choice of $\rho$ satisfying $\mathrm{Quantile}^{\mathrm{WC}}_{f,\rho}(1 - \alpha; P_0) \geq q$ yields a prediction set $C_{f,\rho}(\cdot; P_0)$ that both provides

**Algorithm 1** Worst-subset validation procedure

**Input:** sample $\{(X_i, Y_i)\}_{i=1}^n$ with empirical distribution $\hat{Q}_n$; score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ with empirical distribution $\hat{P}_n$ on $\{s(X_i, Y_i)\}_{i=1}^n$; levels $\alpha, \alpha_{\mathsf{v}} \in (0, 1)$; divergence function $f : \mathbb{R}_+ \to \mathbb{R}$; smallest subset size $\delta \in (0, 1)$; number of sampled directions $k \geq 1$.

**Do:** Sample $\{v_j\}_{j=1}^k \overset{\text{iid}}{\sim} \mathbb{P}_{\mathsf{v}}$, and let $\hat{\mathbb{P}}_{\mathsf{v},k}$ be their empirical distribution and set

$$\widehat{q}_\delta := \inf\left\{ q \in \mathbb{R} : \hat{\mathbb{P}}_{\mathsf{v},k}\Big(\mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta; \widehat{Q}_n) \geq 1 - \alpha\Big) \geq 1 - \alpha_{\mathsf{v}} \right\}. \tag{11}$$

Set $\hat{\rho}_\delta$ to be any sufficient divergence level for threshold $\widehat{q}_\delta$.

**Return:** confidence set mapping $\widehat{C} : \mathcal{X} \rightrightarrows \mathcal{Y}$ with $\widehat{C}(x) := C^{(\widehat{q}_\delta)}(x)$ or $\widehat{C}(x) := C_{f,\hat{\rho}_\delta}(x; \hat{P}_n)$.

coverage for covariate shifts $Q_R$ of the form (8) across most directions $v \in \mathcal{V}$, in agreement with (10), and enjoys the protection against distribution shift we establish in Section 2 for the given value $\rho$ (including against more than covariate shifts). Algorithm 1 performs this using plug-in empirical estimators for $P_0$, $Q_0$ and $\mathbb{P}_{\mathsf{v}}$.

We can show that procedure 1 approaches uniform $1-\alpha$ coverage if the subsets in $\mathcal{R}$ have finite VC-dimension. To achieve almost exact coverage, we will sometimes require

**Assumption A1** (Score continuity)**.** *The distribution of the scores under $P_0$ is continuous.*

**Theorem 2.** *Let $\widehat{C}$ be the prediction set Alg. 1 returns. Assume that $\mathcal{R} = \bigcup_{v \in \mathcal{V}} \mathcal{R}_v$ has VC-dimension $\mathrm{VC}(\mathcal{R}) < \infty$. Then there exists a universal constant $c < \infty$ such that the following holds. For all $t > 0$, defining*

$$\alpha_{t,n}^{\pm} := \alpha \pm c\sqrt{\frac{\mathrm{VC}(\mathcal{R})\log n + t}{\delta n}}, \quad \text{and} \quad \delta_{t,n}^{\pm} = \delta \pm c\sqrt{\frac{\mathrm{VC}(\mathcal{R})\log n + t}{n}},$$

*then with probability at least $1 - e^{-t}$ over $\{X_i, Y_i\}_{i=1}^n \overset{\text{iid}}{\sim} Q_0$ and $\{v_i\}_{i=1}^k \overset{\text{iid}}{\sim} \mathbb{P}_{\mathsf{v}}$,*

$$\mathbb{P}_{\mathsf{v}}\Big(\mathrm{WC}(\widehat{C}, \mathcal{R}_v, \delta_{t,n}^{+}; Q_0) \geq 1 - \alpha_{t,n}^{+}\Big) \geq 1 - \alpha_{\mathsf{v}} - c\sqrt{\frac{1+t}{k}}.$$

*If additionally Assumption A1 holds, then*

$$\mathbb{P}_{\mathsf{v}}\Big(\mathrm{WC}(\widehat{C}, \mathcal{R}_v, \delta_{t,n}^{-}; Q_0) \leq 1 - \alpha_{t,n}^{-}\Big) \geq \alpha_{\mathsf{v}} - \frac{1}{k} - c\sqrt{\frac{1+t}{k}}.$$

See Appendix B for a proof of the theorem.

Theorem 2 shows that procedure 1 approaches uniform $1-\alpha$ coverage if the subsets in $\mathcal{R}$ have finite VC-dimension. More precisely, the estimate $\hat{\rho}_\delta$ almost achieves the randomized worst-case coverage (10): with probability nearly $1-\alpha_{\mathsf{v}}$ over the direction $v \sim \mathbb{P}_{\mathsf{v}}$, $\widehat{C}$ provides coverage at level $1 - \alpha - O(1/\sqrt{n})$ for all shifts $Q_R$ (as in Eq. (8)) satisfying $R \in \mathcal{R}_v$ and $Q_0(X \in R) \geq \delta$. The second statement in the theorem is an insurance against drastic overcoverage: while we cannot guarantee that the worst coverage is always no more than $1-\alpha$, we can guarantee that—if the scores are continuous—then the empirical set $\widehat{C}$ has worst coverage *no more* than $1 - \alpha + O(1/\sqrt{n})$ for at least a fraction $\alpha_{\mathsf{v}}$ of directions $v \sim \mathbb{P}_{\mathsf{v}}$. In a sense, this is unimprovable: if the worst coverage $W = \mathrm{WC}(C, \mathcal{R}_v, \delta; Q_0)$ is continuous in $v$, the best we could expect is that $\mathbb{P}_{\mathsf{v}}(W \geq 1-\alpha) = 1 - \alpha_{\mathsf{v}}$ while $\mathbb{P}_{\mathsf{v}}(W < 1-\alpha) = \alpha_{\mathsf{v}}$.

As a last remark, we note that when the scores are distinct, there is a complete equivalence between thresholds $q$ and divergence levels $\rho$ in Algorithm 1; see Appendix B.1 for a proof.

**Lemma 3.1.** *Assume that the scores $s(X_i, Y_i)$ are all distinct. Define $\rho_{f,\alpha}(q; P) := \sup\{\rho \geq 0 \mid \text{Quantile}^{\text{WC}}_{f,\rho}(1-\alpha; P) \leq q\}$ and let $\widehat{\rho}_\delta = \rho_{f,\alpha}(\widehat{q}_\delta, \hat{P}_n)$. Then $C^{(\widehat{q}_\delta)} = C_{f,\widehat{\rho}_\delta}(\cdot; \hat{P}_n)$.*

### 3.2 Finding directions of maximal shift

In this section, we revisit worst potential shifts, designing a (heuristic) methodology to estimate the worst direction and protect against it, additionally providing sufficient conditions for consistency. For a confidence set mapping $C : \mathcal{X} \rightrightarrows Y$, we define the worst shift direction

$$v_\star(C) := \operatorname*{argmin}_{v \in \mathcal{V}} \text{WC}(C, \mathcal{R}_v, \delta; Q_0), \tag{12}$$

which evidently satisfies

$$\text{WC}(C, \mathcal{R}_{v_\star(C)}, \delta; Q_0) = \text{WC}(C, \mathcal{R}, \delta; Q_0) := \inf_{v \in \mathcal{V}} \text{WC}(C, \mathcal{R}_v, \delta; Q_0).$$

If we could identify such a worst direction and it is consistent across thresholds $q$ in our typical definition $C(x) = \{y \in \mathcal{Y} \mid s(x, y) \leq q\}$ (a strong condition), then the procedures in the preceding sections allow us to choose thresholds to guarantee coverage. The intuition here is that there may exist a preferred direction with higher variance in predictions, for example, time in a temporal system. A more explicit example comes from heteroskedastic regression:

**Example 2** (Heteroskedastic regression)**:** Let the data $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ follow the model

$$Y = \mu^\star(X) + h(v_{\text{var}}^T X)\varepsilon$$

where $h : \mathbb{R} \to \mathbb{R}_+$ is non-decreasing, $\varepsilon \sim \mathsf{N}(0, 1)$ independent of $X$, which generalizes the standard regression model to have heteroskedastic noise, with the noise increasing in the direction $v_{\text{var}}$. Evidently the oracle (smallest length) conditional confidence set for $Y \mid X = x$ is the interval $[\pm z_{1-\alpha/2}\sqrt{h(v_{\text{var}}^T x)}]$ where $z_{1-\alpha}$ is the standard normal quantile. The standard split conformal methodology (Section 1.1) will undercover for those $x$ such that $v_{\text{var}}^T x$ is large—shifts of $X$ in the direction $v_\star = v_{\text{var}}$ may decrease coverage significantly. $\diamond$

With this example as motivation, we propose identifying challenging directions for dataset shift by separating those datapoints $(X_i, Y_i)$ with large nonconformity scores $s(X_i, Y_i)$ from those with lower scores. One could use any M-estimator to find such a discriminator; for simplicity in Algorithm 2 we make this concrete with a linear regression estimator. To enable our coming analysis, we elaborate slightly and modify notation to reflect that the scoring function $s_n$ may change with sample size $n$. We also refine Definition 3.1 of the confidence sets to explicitly depend on both the threshold $q$ and score $s$.

**Definition 3.2.** *For $q \in \mathbb{R}$ and a score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the* $s$-prediction set at level $q$ *is*

$$C^{(q,s)}(x) := \{y \in \mathcal{Y} \mid s(x, y) \leq q\}. \tag{13}$$

The intuition behind Algorithm 2 is simple: we seek a direction $v$ in which shifts in $X$ make the given nonconformity score $s_n$ large, then guarantee coverage for shifts in that direction and, via the distributionally robust confidence set $C_{f,\hat{\rho}}$ the procedure returns, any future distributional shift for which the distribution $P_{\text{new}}$ of scores $s(X, Y)$ satisfies $D_f\left(P_{\text{new}} \| P_0\right) \leq \hat{\rho}$. Because we need only solve a single M-estimation problem—rather than a large sample of direction $v$ as in Alg. 1—the estimation methodology is more computationally efficient.

**Algorithm 2** Worst-direction validation given a score function

**Input:** sample $\{(X_i, Y_i)\}_{i=1}^n$; score function $s_n : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ independent of the sample; coverage rate $1-\alpha \in (0,1)$; divergence function $f : \mathbb{R}_+ \to \mathbb{R}$; smallest subset size $\delta \in (0,1)$.
**Initialize:** Split sample $\{(X_i, Y_i)\}_{i=1}^n$ into $\{(X_i, Y_i)\}_{i=1}^{n_1}$, $\{(X_i, Y_i)\}_{i=n_1+1}^{n_1+n_2}$ with empirical distributions $\hat{Q}_{n_1}$, and $\hat{Q}_{n_2}$ (resp. $\hat{P}_{n_1}$ and $\hat{P}_{n_2}$ for the scores).
**Do:** Regress $s_n(X, Y)$ against $X$ on the first sample distribution $\hat{Q}_{n_1}$:

$$\widehat{v} = \operatorname*{argmin}_{v \in \mathbb{R}^d} \left\{ \frac{1}{n_1} \sum_{i=1}^{n_1} \left( s_n(X_i, Y_i) - v^T X_i \right)^2 \right\}.$$

Use the second subsample to set the threshold $\widehat{q}_\delta$ to

$$\widehat{q}_\delta := \inf \left\{ q \in \mathbb{R} : \mathrm{WC}(C^{(q,s_n)}, \mathcal{R}_{\hat{v}}, \delta; \hat{Q}_{n_2}) \geq 1 - \alpha \right\}. \tag{14}$$

Set $\hat{\rho}_\delta := \rho_{f,\alpha}(\widehat{q}_\delta; \hat{P}_{n_2}) = \sup\{\rho \geq 0 \mid \mathrm{Quantile}^{\mathrm{WC}}_{f,\rho}(1-\alpha; \hat{P}_{n_2}) \leq q\}$ as in Lemma 3.1.
**Return:** the confidence set mapping $\widehat{C}_n(x) = C^{(\widehat{q}_\delta, s_n)}(x) = C_{f,\hat{\rho}_\delta}(x; \hat{P}_{n_2})$.

### 3.2.1 Population-level consistency of the worst direction

The consistency of Procedure 2 or a similar M-estimation procedure seeking worst directions requires strong assumptions, somewhat oppositional to the distribution-free coverage we seek (though again we still have the distributionally robust protections). Yet it is still of interest to understand conditions under which the method 2 is consistent; as we show here, in examples such as the heteroskedastic regression (Ex. 2), this holds. We turn to our assumptions.

We restrict ourselves to linear shifts, taking $\mathcal{V} = \mathbb{S}^{d-1} = \{v \in \mathbb{R}^d \mid \|v\|_2 = 1\}$, and defining

$$\mathcal{R}_v := \left\{ \{x \in \mathbb{R}^d \mid v^T x \geq a\} \right\}_{a \in \mathbb{R}}.$$

A challenge is that the worst direction $v_\star(C^{(q,s)})$ may vary substantially in $q$. One condition sufficient to ameliorate this reposes on stochastic orders [34], where for random variables $U$ and $V$ on $\mathbb{R}$, we say $U$ stochastically dominates $V$, written $U \succeq V$, if $\mathbb{P}(U \geq t) \geq \mathbb{P}(V \geq t)$ for all $t \in \mathbb{R}$. Letting $\mathcal{L}$ denote the law of a random variable, we write $\mathcal{L}(U) \succeq \mathcal{L}(V)$ if $U \succeq V$.

**Assumption A2.** *There is a direction $v^\star \in \mathcal{V}$ such that*

$$\mathcal{L}(s(X,Y) \mid X^T v^\star \geq \tau)$$

*is increasing in the stochastic dominance order as $\tau$ increases. Moreover, for all $u \in \mathcal{V}, t \in \mathbb{R}$, with $\mathbb{P}(X^T u \geq t) \geq \mathbb{P}(X^T v^\star \geq \tau)$, we have*

$$\mathcal{L}(s(X,Y) \mid X^T v^\star \geq \tau) \succeq \mathcal{L}(s(X,Y) \mid X^T u \geq t).$$

The intuition is that covariate shifts in direction $v^\star$ increase nonconformity and that $v^\star$ is the worst such direction. Under Assumption A2, confidence sets share the same worst shift $v^\star$:

**Lemma 3.2.** *Let Assumption A2 hold and the distribution of $X^T v^\star$ be continuous. Then for all $q \in \mathbb{R}$, the worst shift (12) for the confidence set (13) satisfies $v_\star(C^{(q,s)}) = v^\star$.*

We present the (nearly immediate) proof of Lemma 3.2 in Appendix C.1. While Assumption A2 is admittedly strong, the next lemma (whose proof we provide in Appendix C.2) shows that it holds in the heteroskedastic regression case of Example 2.

**Lemma 3.3.** *Assume the regression model of Example 2, $Y = \mu^\star(X) + h(X^T v_{\text{var}})\varepsilon$, with noncomformity score*

$$s(x, y) = (y - \mu^\star(x))^2 \quad or \quad s(x, y) = |y - \mu^\star(x)|.$$

*If $X \sim \mathsf{N}(0, I)$, then $v^\star = v_{\text{var}}$ satisfies Assumption A2.*

We can also show that various M-estimators can identify the direction $v^\star$ when Assumption A2 holds. We present one such plausible result here, providing the proof in Appendix C.2.1.

**Proposition 3.** *Let Assumption A2 hold and assume that for some $\Sigma \succ 0$, $\Sigma^{-1/2}X$ is rotationally invariant with finite second moments. Additionally, suppose that $s(X, Y) \geq 0$, $\mathbb{E}\left[s(X, Y)X\right] \neq 0$, and $\mathbb{E}[s(X, Y)^2] < \infty$. Then $v^\star$ is proportional to the least-squares solution*

$$v^\star \propto \operatorname*{argmin}_{v \in \mathbb{R}^d} \mathbb{E}\left[\left(s(X, Y) - v^T X\right)^2\right]. \tag{15}$$

Example 2 with $X \sim \mathsf{N}(0, \Sigma)$ and typical nonconformity scores satisfies the conditions of Proposition 3. While in more general models least squares estimation need not find the worst shift direction, Proposition 3 suggests it may be a reasonable heuristic.

### 3.2.2 Asymptotic estimation of the worst direction

With the population level recovery guarantees we establish in Proposition 3, it is now of interest to understand when we may recover the optimal worst direction and corresponding confidence set $C$ using Algorithm 2, which has access only to samples from the population $Q_0$. An immediate corollary of Theorem 2 ensures that, under the same conditions, Algorithm 2 returns a confidence set mapping $\widehat{C}_n$ that satisfies, conditionally on $s_n$ and $\hat{v}$ and with probability $1 - e^{-t}$ over the second half of the validation data,

$$\mathrm{WC}(\widehat{C}_n, \mathcal{R}_{\hat{v}}, \delta^+_{t,n_2}; Q_0) \geq 1 - \alpha^+_{t,n_2} \text{ and } \mathrm{WC}(\widehat{C}_n, \mathcal{R}_{\hat{v}}, \delta^-_{t,n_2}; Q_0) \leq 1 - \alpha^-_{t,n_2}. \tag{16}$$

Recalling the definition (9), it remains to understand how close we can expect the uniform quantity $\mathrm{WC}(\widehat{C}_n, \mathcal{R}, \delta; Q_0)$ to be to $1 - \alpha$. By the bounds (16), if the worst coverage is continuous in $v \in \mathcal{V}$ and $s_n$ and $\hat{v}$ are appropriately consistent, we should expect a uniform $1 - \alpha$ coverage guarantee in the limit as $n \to \infty$.

To present such a consistency result, we require a few additional assumptions.

**Assumption A3** (Consistency of scores and directions)**.** *As $n \to \infty$, we have*

$$\|s_n - s\|^2_{L^2(Q_0)} := \int_{\mathcal{X} \times \mathcal{Y}} (s_n(x, y) - s(x, y))^2 \, dQ_0(x, y) = o_P(1) \quad and \quad \|\widehat{v} - v^*\| = o_P(1).$$

**Assumption A4** (Continuous distributions)**.** *For $(X, Y) \sim Q_0$, the random variables $s(X, Y)$ and $v^T X$ have continuous distributions for all $v \in \mathcal{V}$.*

**Assumption A5** (Distinct scores)**.** *The scores are asymptotically distinct in probability,*

$$Q_0^n \left[\text{there exist } i, j \in [n], i \neq j \text{ s.t. } s_n(X_i, Y_i) = s_n(X_j, Y_j)\right] \xrightarrow{p} 0.$$

Assumption A5 is a technical assumption that will typically hold whenever Assumption A4 holds, for example, if $s_n$ belongs to a parametric family.

Under these assumptions, Theorem 3 proves that we asymptotically provide uniform coverage at level $1-\alpha$ over all shifts $Q_R$, $R \in \mathcal{R}$. See Appendix C.3 for a proof.

**Theorem 3.** *Let Assumptions A2, A3, and A4 hold. Then Algorithm 2 returns a confidence set mapping $\widehat{C}_n$ that satisfies*

$$\mathrm{WC}(\widehat{C}_n, \mathcal{R}, \delta; Q_0) = 1 - \alpha + u_n + \varepsilon_n$$

*where $u_n \geq 0$ and $\varepsilon_n \xrightarrow{p} 0$ as $n \to \infty$. If additionally Assumption A5 holds, then $u_n \xrightarrow{p} 0$.*

To conclude, we see that the M-estimation-based procedure 2 to find the worst shift direction can be consistent. Yet even without the (strong) assumptions Theorem 3 requires, we believe the methodology in Algorithm 2 (and Alg. 1) is valuable: it is important to look for variation in coverage within a dataset and to protect against possible future dataset shifts.

# 4 Empirical analysis

Given the challenges arising in the practice of machine learning and statistics, this paper argues that methodology equipping models with a notion of validity in their predictions—e.g., conformalization procedures as in this paper—is essential to any modern prediction pipeline. Section 1.3 illustrates the need for these sorts of procedures, showing that the standard conformal methodology is sensitive to even small shifts in the data, through (semi-synthetic) experiments on data from the UCI repository. In Section 2, we propose methods for robust predictive inference, giving methodology that estimates the amount of shift to which we should be robust. Fuller justification requires a more careful empirical study that highlights both the failures of non-robust prediction sets on real data as well as the potential to handle such shifts using the methodology here. To that end, we turn to experimental work. We begin our empirical study with fully synthetic experiments (Sec. 4.1)—where we know the precise model—to demonstrate a few successes and failures. Resuming the evaluation on the semi-synthetic data from Section 1.3, we assess our own methodology on the UCI datasets in Section 4.2, before turning to an evaluation centered around the new MNIST, CIFAR-10, and ImageNet test sets in Section 4.3. Importantly, these datasets exhibit real-world distributional shifts, as we discuss. In the first two sets of experiments, we use slabs (recall Example 1) for the plausible shifts, while in the last, we use halfspaces.

## 4.1 Synthetic data

We begin by evaluating our procedures on synthetic data to illustrate the issues at play. As any inferential procedure must trade between a confidence set's coverage and size, we wish to understand how our procedures for robust predictive inference trade these criteria under distributional shifts. We pay special attention to the scoring function $s(x, y)$'s influence on these criteria; its (mis)specification is critical to the formation of our prediction sets.

We consider the following setting. The validation data follow a standard regression model,

$$y_{\mathrm{val},i} = x_{\mathrm{val},i}^T \theta_0 + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the predictors $x_{\mathrm{val},i} \in \mathbb{R}^{10}$ and the noise $\varepsilon_i$ are independent draws from a standard normal distribution. To isolate and probe the impact of the (mis)specification of the scoring

function, we consider the following family of squared error scoring functions. Let $\theta_0, \theta_1 \in \mathbb{R}^{10}$ be a pair of orthonormal vectors, and let $\theta_t = \sqrt{1-t^2}\theta_0 + t\theta_1$, $t \in [0,1]$, interpolate between $\theta_0, \theta_1$. For $t \in [0,1]$, we consider the scoring function

$$s_t(x,y) = (y - x^T\theta_t)^2,$$

which moves from well-specified ($t = 0$) to completely misspecified ($t = 1$).

We consider the standard conformal methodology (Sec. 1.1) and our robust variants. Letting $\widehat{\mathcal{Q}}_{n,1-\alpha,t}$ denote the $(1+n^{-1})\alpha$ quantile of the scores for a fixed value of $t$, the standard split conformal confidence set (3) at a test point $x_{\text{test}}$ is

$$\hat{C}_{n,t}(x_{\text{test}}) = \left\{ y \in \mathbb{R} \mid s_t(x_{\text{test}}, y) \le \widehat{\mathcal{Q}}_{n,1-\alpha,t} \right\} = x_{\text{test}}^T\theta_t \pm \widehat{\mathcal{Q}}^{1/2}_{n,1-\alpha,t}. \tag{17}$$

When the model is correct ($t = 0$), the oracle confidence set is $x_{\text{test}}^T\theta_0 \pm z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard Gaussian, to which the interval (17) converges.

At each test point and misspecification parameter $t$, we also form our robust prediction sets $C_{f,\rho,t}$, which replace the $(1-\alpha)$ quantile in (17) with the worst-case quantile (see Section 2.1). To finish the specification of our prediction sets, we set $\rho$ by sampling $k = 5000$ directions of shift from the uniform distribution on the sphere (see Alg. 1 in Section 3.1), choosing the smallest value of $\rho$ with worst-case coverage at the level $\alpha_{\mathsf{v}} = .05$ across all subsets of size at least $\delta = 1/3$. We form the prediction sets with the chi-square divergence (i.e. $f(t) = (t-1)^2$).

To simulate distribution shift, we generate the responses according to the model

$$y_{\text{test}} = x_{\text{test}}^T\theta_0 + \varepsilon,$$

but, importantly, we now have that $x_{\text{test}} \sim \mathsf{N}(2e_1, I)$. We evaluate the coverage and size of the bands (17) on this shifted version of the validation set, averaging the results over 20 runs. The validation and test sets each contain 5000 samples (there is no training set in this setup).

Figure 2 presents the results. When $t = 0$, the model is correct, and so in this case we expect the prediction sets coming from the (oracle) standard conformal methodology as well as our robust prediction sets to attain coverage at the nominal level .95. Because our methodology seeks protection from worst-case shifts, even when the scoring function is well-specified, random fluctuations in the score across directions $v$ necessitate somewhat wider robust prediction sets relative to the standard conformal methodology. The figure bears out both of these intuitions—and, in particular, we see that the robust sets are actually only slightly larger than those from standard conformal.

On the other hand, when $t > 0$, the model is incorrect. We expect the prediction set sizes to grow as the level of misspecification increases, and indeed the figure reflects this behavior. Now, if the validation and test points were in fact exchangeable, then we would expect standard conformal to nonetheless give proper coverage (just with potentially larger prediction sets). In a distribution shift setup, however, the standard conformal guarantees break down, whereas our procedures should maintain validity as long as the divergence between the validation and test populations is not too large. The results broadly match our expectations. In fact, from the figure, we can see a serious degradation in coverage for standard conformal as the level of misspecification $t$ grows even in this simple covariate shift setup; moreover, even for a given uncertainty set size (right figure), the standard split-conformal methodology provides lower coverage than our distributionally robust confidence sets. In contrast, our procedures give proper coverage across a reasonably wide range of misspecification values $t$, providing some protection against covariate shift even with misspecification.

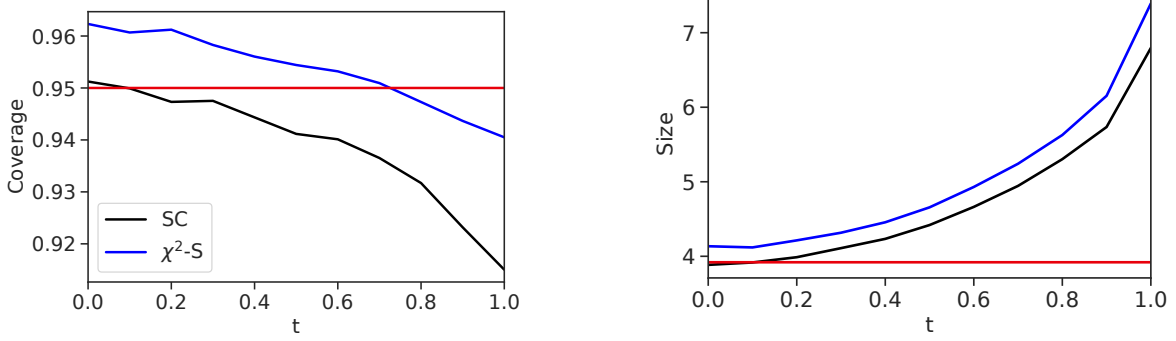


**Figure 2.** Empirical coverage and average size for the prediction sets generated by the standard conformal methodology ("SC") and the chi-squared divergence, across 20 random splits of the synthetic data. We set $\rho$ according to the sampling-based strategy ("$\chi^2$-S") for estimating the amount of shift that we describe in Section 3.1, Alg. 1. The horizontal red lines mark the marginal coverage .95, and length of the oracle prediction set (i.e., $2z_{1-\alpha/2}$).

### 4.2 UCI datasets

We now revisit the experiments in Section 1.3, focusing on evaluating our methodology for robust predictive inference. Throughout these experiments, to illustrate the various issues at play, we fix the desired robustness level $\rho = .01$, corresponding (approximately) to the median chi-squared divergence between the natural and tilted empirical distributions across the nine data sets and values of the tilting parameter $a$. We therefore expect Algorithm 1, which emphasizes robustness to worst-case shifts, to restore the coverage level for the tiltings from Section 1.3 that possess (roughly) this level of shift.

Figure 3 presents the results for the chi-squared divergence (the results for the Kullback-Leibler divergence are similar). Although not perfect, we see that the methodology often restores validity for the shifts from Section 1.3. In particular, we see clearly improved performance over the standard conformal methodology on the abalone, delta ailerons, kinematics, puma, and airfoil datasets (cf. Figure 1); indeed, on all of these datasets, our methodology consistently yields average coverage above the nominal level, whereas standard conformal fails to do so on all of these datasets. In line with our expectations, the chi-squared divergence between the natural and shifted distributions is roughly in the expected range for each of these datasets (the divergence values are .03, .02, .04, .05, and 3.65, respectively), while the level of divergence for the remaining datasets (ailerons—which still covers—banking, and Boston and California housing) is higher. We note in passing that in other experiments we omit for brevity, the trends above hold for other types of shifts.

### 4.3 CIFAR-10, MNIST, and ImageNet datasets

We finally consider an evaluation of our procedures on the CIFAR-10, ImageNet, and MNIST datasets [20, 31, 21], which continue to play a central role in the evaluation of machine learning methodology. Yet concerns about overfitting to these benchmarks motivate Recht, Roelofs, Schmidt, and Shankar [29] to create new test sets for both CIFAR-10 and ImageNet by carefully following the original dataset creation protocol. Though these new test sets strongly resemble the original datasets, as Recht et al. observe, the natural variation arising in the creation of the new test sets yield evidently significant differences, giving organic dataset shifts on which to evaluate our procedures.

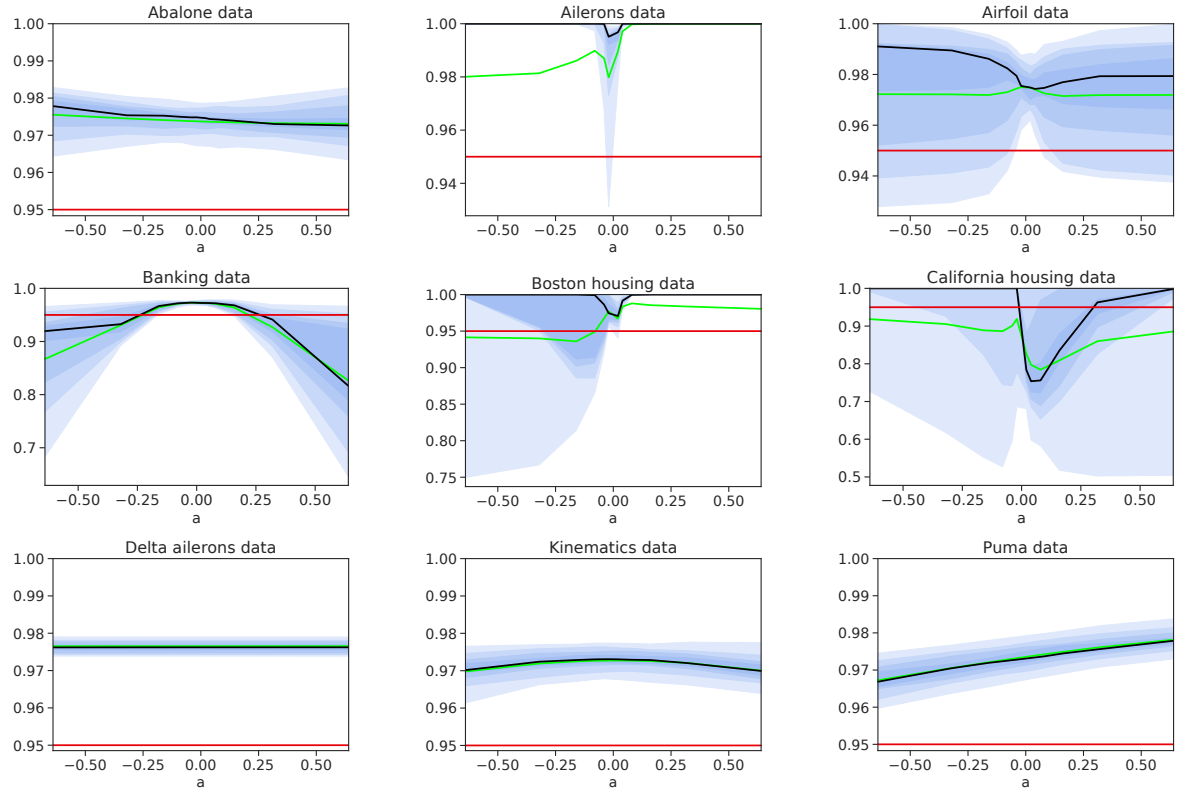


**Figure 3.** Empirical coverage for the prediction sets generated by the chi-squared divergence, following the same experimental setup from Section 1.3. The horizontal axis gives the value of the tilting parameter $a$; the vertical the coverage level. A green line marks the average coverage, a black line marks the median coverage, and the horizontal red line marks the nominal coverage .95. The blue bands show the coverage at various deciles.

We evaluate on the three datasets as follows. We use 70% of the original CIFAR-10, MNIST, and ImageNet datasets for training, and treat the remaining 30% as a validation set. We fit a standard ResNet-50 [17] to the training data, and use the negative log-likelihood $s(x, y) = -\log p_\theta(y \mid x)$, where $p_\theta(y \mid x)$ is the output of the (top) sigmoid layer of the network, as the scoring function on the validation data for our conformalization procedures. We compare our procedures to the split conformal methodology on three new datasets nominally generated identically to the initial datasets: the CIFAR-10.1 v4 dataset [29], which consists of 2,000 32×32 images from 10 different classes; the QMNIST50K data, which extends MNIST to consist of 50,000 28×28 images from 10 classes [43]; and the ImageNetV2 Threshold0.7 data [29], consisting of 10,000 images from 200 classes. In each test of robust predictive inference, we set the level of robustness according to the data-driven strategies we detail in Section 3: sampling directions of shift from the uniform distribution on the unit sphere (Alg. 1), estimating the shift direction via regression (Alg. 2) or via classification, which replaces the regression step in Alg. 2 with a support vector machine (SVM) to separate the largest 50% of scores $s(X_i, Y_i)$ from the smallest. Finally, for the CIFAR-10 and MNIST datasets, we set the level $\alpha = .05$; for the ImageNet dataset, we set $\alpha = .1$.

Figures 4, 5, and 6 present the results for each setup over 20 random splits of the data. As is apparent from the figures, we see that the standard conformal methodology fails to correctly cover. As both the new CIFAR-10 and ImageNet test sets exhibit larger degradations in classifier performance (increased error) than does the MNIST test set [29], we expect the

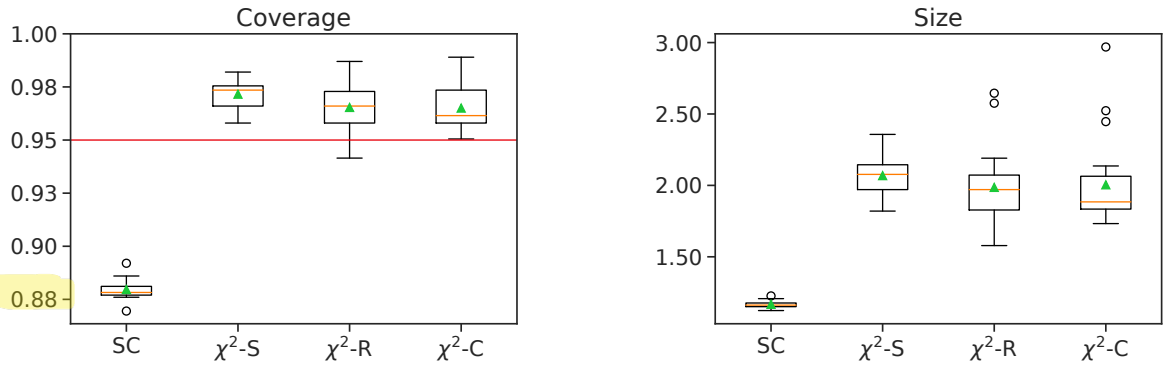


**Figure 4.** Empirical coverage and average size for the prediction sets generated by the standard conformal methodology ("SC") and the chi-squared divergence, across 20 random splits of the CIFAR-10 data. We set $\rho$ according to the sampling ("$\chi^2$-S"), regression ("$\chi^2$-R"), and classification-based ("$\chi^2$-C") strategies for estimating the amount of shift that we describe in Section in 3. The horizontal red line marks the marginal coverage .95.

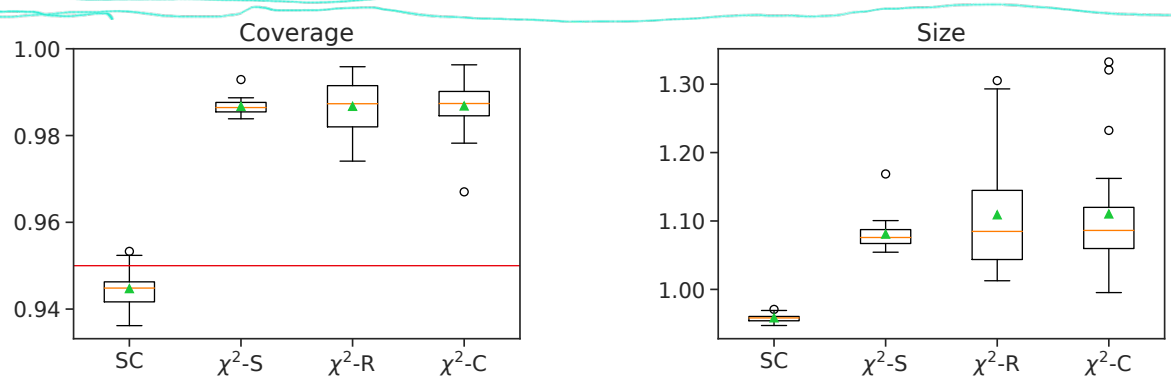


**Figure 5.** Empirical coverage and average size for the prediction sets generated by the standard conformal methodology ("SC") and the chi-squared divergence, across 20 random splits of the MNIST data. We set $\rho$ according to the sampling ("$\chi^2$-S"), regression ("$\chi^2$-R"), and classification-based ("$\chi^2$-C") strategies for estimating the amount of shift that we describe in Section in 3. The horizontal red line marks the marginal coverage .95.

failure of standard conformal to be pronounced on these two datasets. Indeed, the split conformal method (Sec. 1.1) provides especially poor coverage on these datasets, where it yields average coverage .88 (instead of the nominal .95) and .8 (instead of the nominal .9) on the new CIFAR-10 and ImageNet test sets, respectively. On the other hand, our inferential methodology consistently gives more coverage regardless of the strategy used to estimate the amount of divergence $\rho$. The uniformity in coverage across the three strategies is notable, as our procedures for estimating the amount of shift assume some structure for the underlying shift, which is unlikely to be consistent with the provenance of the new test sets.

In our experiments, estimating the direction of shift using either regression or classification (Alg. 2) is faster than sampling directions (Alg. 1); the former takes time $O(nd \min\{n, d\})$ and the latter $O(knd)$, where $k$ is the number of sampled directions $v$, using a linear-time implementation for computing the worst coverage (maximum density segment) along a direction $v$ [27]. The difference of course depends on the desired sampling frequency $k$.

Finally, the aforementioned validity does not (apparently) come with a significant loss in statistical efficiency: Figures 4, 5, and 6 show that our confidence sets are not substantially larger than those coming from standard conformal inference—which may be somewhat

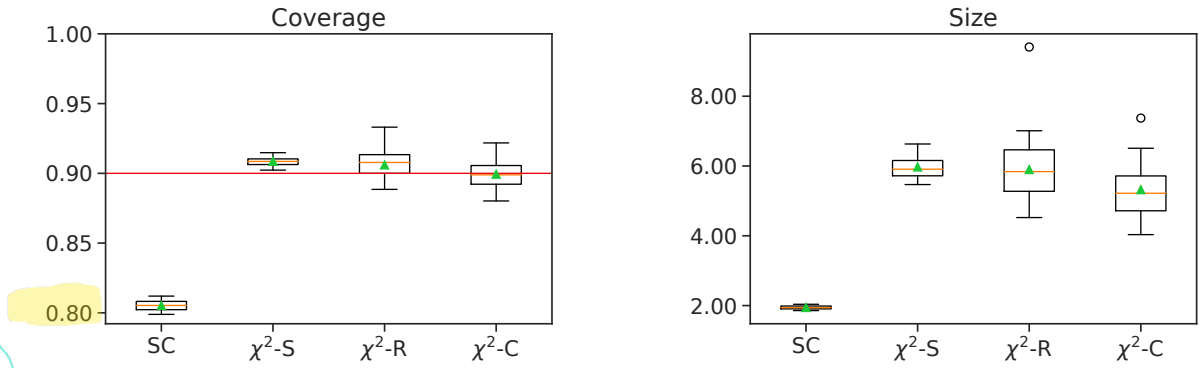


**Figure 6.** Empirical coverage and average size for the prediction sets generated by the standard conformal methodology ("SC") and the chi-squared divergence, across 20 random splits of the ImageNet data. We set $\rho$ according to the sampling ("$\chi^2$-S"), regression ("$\chi^2$-R"), and classification-based ("$\chi^2$-C") strategies for estimating the amount of shift that we describe in Section in 3. The horizontal red line marks the marginal coverage .9.

surprising, given the relatively large number of classes present in the ImageNet dataset.

## 5 Discussion and conclusions

We have presented methods and motivation for robust predicitve inference, seeking protection against distribution shift. Our arguments and perspective are somewhat different from the typical approach in distributional robustness [4, 14, 13, 6, 32], as we wish to maintain validity in prediction. A number of future directions and questions remain unanswered. Perhaps the most glaring is to fully understand the "right" level of robustness. While this is a longstanding problem [13], we present approaches to leverage the available validation data. Alternatives might be compare new covariates and test data $X$ to the available validation data. Tibshirani et al. [39] suggest an importance-sampling approach for this, reweighting data based on likelihood ratios, which may sometimes be feasible but is likely impossible in high-dimensional scenarios. It would be interesting, for example, to use projections of the data to match $X$-statistics on new test data, using this to generate appropriate distributional robustness sets. We hope that the perspective here inspires renewed consideration of predictive validity.

## References


[1] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference*, to appear, 2019.

[2] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[3] A. Ben-Tal, D. den Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[4] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming, Series A*, 167(2):235–292, 2018.

[5] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[6] J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[7] M. Cauchois, S. Gupta, and J. Duchi. Knowing what you know: valid confidence sets in multiclass and multilabel prediction. *arXiv:2004.10181 [stat.ML]*, 2020.

[8] V. Chernozhukov, K. Wuthrich, and Y. Zhu. Exact and robust conformal inference methods for predictive machine learning with dependent data. *arXiv:1802.06300 [stat.ML]*, 2018.

[9] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientifica Mathematica Hungary*, 2:299–318, 1967.

[10] D. Dua and C. Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[11] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv:1810.08750 [stat.ML]*, 2018.

[12] J. C. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

[13] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, to appear, 2020. URL https://arXiv.org/abs/1810.08750.

[14] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming, Series A*, 171(1–2):115–166, 2018.

[15] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. In J. Q. nonero Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, editors, *Dataset Shift in Machine Learning*, chapter 8, pages 131–160. MIT Press, 2009.

[16] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[18] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer, New York, 1993.

[19] M. I. Jordan. Artificial intelligence—the revolution hasnt happened yet. *Harvard Data Science Review*, 1(1), 2019.

[20] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[21] Y. LeCun, C. Cortes, and C. Burges. MNIST handwritten digit database, 1998. URL http://yann.lecun.com/exdb/mnist. ATT Labs [Online].

[22] J. Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 2014.

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

[24] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[25] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

[26] Z. C. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[27] K. min Chung and H.-I. Lu. An optimal algorithm for the maximum-density segment problem. In *Proceedings of the 11th Annual European Symposium on Algorithms*, 2003.

[28] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[29] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[30] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge.

*International Journal of Computer Vision*, 115(3):211–252, 2015.

[32] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *Proceedings of the Eighth International Conference on Learning Representations*, 2020.

[33] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 28*, pages 1576–1584, 2015.

[34] M. Shaked and J. G. Shanthikumar. *Stochastic Orders*. Springer Series in Statistics. Springer, 2007.

[35] A. Sinha, H. Namkoong, and J. Duchi. Certifying some distributional robustness with principled adversarial training. In *Proceedings of the Sixth International Conference on Learning Representations*, 2018.

[36] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36(2):111–147, 1974.

[37] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

[38] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. When robustness doesn't promote robustness: Synthetic vs. natural distribution shifts on ImageNet. under review, 2020. URL https://openreview.net/forum?id=HyxPIyrFvH.

[39] R. J. Tibshirani, R. F. Barber, E. J. Candès, and A. Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems 32*, 2019.

[40] V. Vovk, A. Grammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

[41] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[42] J. Wen, C.-N. Yu, and R. Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *Proceedings of the 31st International Conference on Machine Learning*, pages 631–639, 2014.

[43] C. Yadav and L. Bottou. Cold case: The lost MNIST digits. In *Advances in Neural Information Processing Systems 32*, 2019.

# A Proofs of results on robust inference

## A.1 Proof of Proposition 1

We provide several properties of $g_{f,\rho}(\beta) = \inf\{z \in [0,1] : \beta f(\frac{z}{\beta}) + (1-\beta) f(\frac{1-z}{1-\beta}) \le \rho\}$, deferring their proof to Sec. A.1.1.

**Lemma A.1** (Properties of $g_{f,\rho}$)**.** *Let $f$ be continuous near 1 and 1-coercive. Then the function $g_{f,\rho}$ satisfies the following.*

*(a)* *$(\beta, \rho) \mapsto g_{f,\rho}(\beta)$ is a convex function, continuous on $\beta \in [0,1]$ and in $\rho > 0$.*

*(b)* *For $\beta \in [0,1]$ and $\rho > 0$, $g_{f,\rho}(\beta) \le \beta$. For $\rho > 0$, equality holds for $\beta \in \{0,1\}$ and strict inequality holds for $\beta \in (0,1)$ and $\rho > 0$.*

*(c)* *$g_{f,\rho}$ is non-increasing in $\rho$ and non-decreasing in $\beta$. Moreover, for all $\rho > 0$, there exists $\beta_0(\rho) := \sup\{\beta \in (0,1) \mid g_{f,\rho}(\beta) = 0\}$, and $g_{f,\rho}$ is strictly increasing for $\beta > \beta_0(\rho)$.*

*(d)* *Let $g_{f,\rho}^{-1}(t) = \sup\{\beta : g_{f,\rho}(\beta) \le t\}$ as in the statement of Proposition 1. Then for $\beta \in (0,1)$, $g_{f,\rho}(\tau) \ge \beta$ if and only if $g_{f,\rho}^{-1}(\beta) \le \tau$.*

We now define the worst-case cumulative distribution function, which generalizes the c.d.f. of a distribution in the same way the worst-case quantile generalizes standard quantiles.

**Definition A.1** ($f$-worst-case c.d.f.)**.** *Let $\rho > 0$ and consider any distribution $P$ on the real line. The* $(f, \rho)$-worst-case cumulative distribution function *is*

$$F_{f,\rho}^{\mathrm{WC}}(t;\, P) := \inf \left\{ P_1(S \le t) \mid S \sim P,\ D_f(P_1 \| P) \le \rho \right\}. \tag{18}$$

Proposition 1 will then follow from the coming lemma.

**Lemma A.2.** *Let $P$ be a distribution on $\mathbb{R}$ with c.d.f. $F$. Then*

$$F_{f,\rho}^{\mathrm{WC}}(t;\, P) = g_{f,\rho}(F(t)). \tag{19}$$

Deferring the proof of this lemma as well (see Sec. A.1.2), let us see how it implies Proposition 1. Observe that for all $\beta \in (0, 1)$, and any real distribution $P$ with c.d.f. $F$, we have

$$\begin{aligned} \mathrm{Quantile}_{f,\rho}^{\mathrm{WC}}(\beta;\, P) &= \inf \left\{ q \in \mathbb{R} \mid F_{f,\rho}^{\mathrm{WC}}(q, P) \ge \beta \right\} \\ &\overset{(i)}{=} \inf \left\{ q \in \mathbb{R} \mid g_{f,\rho}(F(q)) \ge \beta \right\} \\ &\overset{(ii)}{=} \inf \left\{ q \in \mathbb{R} \mid F(q) \ge g_{f,\rho}^{-1}(\beta) \right\} = \mathrm{Quantile}(g_{f,\rho}^{-1}(\beta);\, P), \end{aligned}$$

where equality $(i)$ uses Lemma A.2 and $(ii)$ follows because by Lemma A.1, as $g_{f,\rho}(\tau) \ge \beta$ if and only if $g_{f,\rho}^{-1}(\beta) \le \tau$.

### A.1.1 Proof of Lemma A.1

It is no loss of generality to assume that $f'(1) = 0$ and $f \ge 0$, as replacing $f$ by $f_0(t) := f(t) + f'(1)(t-1)$ generates the same $f$-divergence and evidently $\inf_t f_0(t) = f_0(1) = 0$.

(a) Let $f_{\mathrm{per}}(t, \beta) = \beta f(t/\beta)$ be the perspective of $f$, which is convex. Then $g_{f,\rho}(\beta)$ is the partial minimization of the convex function $(\rho, \beta, z) \mapsto z + \mathbf{I}(f_{\mathrm{per}}(z, \beta) + f_{\mathrm{per}}(z, 1-\beta) \le \rho)$ and hence convex, where $\mathbf{I}(\cdot)$ is the convex indicator function, $+\infty$ if its argument is false and 0 otherwise. Any convex function is continuous on the interior of its domain. (See [18, Ch. IV] for proofs of each of these claims.)

To see that $g_{f,\rho}$ is continuous from the left at $\beta = 1$ and that $g_{f,\rho}(1) = 1$, we use the 1-coercivity of $f$. For $\beta$ near 1, we use our w.l.o.g. assumption that $f \ge 0$ to note that

$$\lim_{\beta \uparrow 1} \left\{ \beta f\left(\frac{z}{\beta}\right) + (1-\beta) f\left(\frac{1-z}{1-\beta}\right) \right\} \ge \lim_{\beta \uparrow 1} (1-\beta) f\left(\frac{1-z}{1-\beta}\right) = \lim_{t \to \infty} \frac{1}{t} f(t(1-z)),$$

and the last quantity is infinite unless $z = 1$. Therefore, for each $\epsilon > 0$ there exists $\delta > 0$ such that $\beta \ge 1 - \delta$ implies that $(1-\beta) f((1-z)/(1-\beta)) > \rho$ for all $z < 1 - \delta$, and $g$ is continuous at $\beta = 1$.

That $g_{f,\rho}$ is right continuous $\beta = 0$ is immediate because $g$ is non-decreasing and convex.

(b) The non-strict inequality is immediate by considering $z = \beta$ and using that $f(1) = 0$. The strict inequality is immediate because $f$ is continuous near 1, while the equalities are as in the previous part of the proof.

(c) That $\rho \mapsto g_{f,\rho}(\beta)$ is non-increasing is evident. As $g$ is nonnegative, convex, and $g_{f,\rho}(0) = 0$ (where we use the standard convention that $0 f(0/0) = 0$), it must therefore be non-decreasing. That $g_{f,\rho}(\beta) > 0$ is strictly increasing in $\beta > \beta_0(\rho)$ is again immediate by convexity as $g_{f,\rho}(0) = 0$.

(d) Let $g = g_{f,\rho}$ for shorthand. Suppose that $g(\tau) \geq \beta > 0$. Then as $g$ is strictly increasing when it is positive, we have $g(t) > g(\tau) \geq \beta$ for all $t > \tau$, so that $g^{-1}(\beta) \leq t$ for any $t > \tau$, or $g^{-1}(\beta) \leq \tau$.

Now, assume the converse, that is, that $g^{-1}(\beta) \leq \tau$, and assume for the sake of contradiction that $g(\tau) < \beta$. By part (b), we must therefore have $\tau < 1$. As $g$ is continuous by part (a), we have $g(\tau + \epsilon) \leq \beta$ for all sufficiently small $\epsilon > 0$, contradicting that $g^{-1}(\beta) \leq \tau$. Thus we must have $g(\tau) \geq \beta$.

### A.1.2 Proof of Lemma A.2

Recall that $P$ is a real distribution with c.d.f. $F$. The result is immediate if $F(t) \in \{0, 1\}$, as the 1-coercivity of $f$ implies that any $P_1$ satisfying $D_f(P_1 \,||\, P) < \infty$ is absolutely continuous with respect to $P$, guaranteeing that $P_1(S \leq t) = F(t) \in \{0, 1\}$, so that $F^{\mathrm{WC}}_{f,\rho}(t; P) = F(t) = g_{f,\rho}(F(t))$. From this point onward, fix $t \in \mathbb{R}$ so that $0 < F(t) = P(S \leq t) < 1$.

The inequality $F^{\mathrm{WC}}_{f,\rho}(t;\, P) \leq g_{f,\rho}(F(t), \rho)$ is immediate:

$$\begin{aligned}
&\inf \left\{ P_1(S \leq t) \mid D_f\left(P_1 \| P\right) \leq \rho \right\} \\
&\leq \inf \left\{ P_1(S \leq t) \mid D_f\left(P_1 \| P\right) \leq \rho,\ \frac{dP_1}{dP} \text{ is constant on } \{S \leq t\} \text{ and } \{S > t\} \right\}.
\end{aligned}$$

The reverse inequality is a consequence of the data processing inequality [25]. Fix $t \in \mathbb{R}$. Let $P_1$ be a distribution satisfying $D_f\left(P_1 \| P\right) \leq \rho$. We show how to construct $\tilde{P}$ with $D_f(\tilde{P} \| P) \leq D_f(P_1 \| P)$ and $\tilde{P}(S \leq t) = P_1(S \leq t)$. Indeed, define the Markov kernel $K$ by

$$K\left(ds' \mid s\right) \propto \begin{cases} dP(s') 1\{s' \leq t\}, & \text{if } s \leq t \\ dP(s') 1\{s' > t\}, & \text{if } s > t. \end{cases}$$

Then $P = K \cdot P$, while $\tilde{P} \coloneqq K \cdot P_1$ satisfies

$$D_f(\tilde{P} \| P) = D_f(K \cdot P_1 \| K \cdot P) \leq D_f(P_1 \| P) \leq \rho$$

by the data processing inequality. Now we observe that

$$d\tilde{P}(s) = \left( \frac{P_1(S \leq t)}{P(S \leq t)} 1\{S \leq t\} + \frac{P_1(S > t)}{P(S > t)} 1\{S > t\} \right) dP(s).$$

By construction, $\tilde{P}(S \leq t) = P(S \leq t)$, and it is immediate that

$$D_f(\tilde{P} \| P) = P(S \leq t) f\left( \frac{P'(S \leq t)}{P(S \leq t)} \right) + P(S > t) f\left( \frac{P'(S > t)}{P(S > t)} \right).$$

Matching the expression of $D_f(\tilde{P} \| P)$ to the definition of $g_{f,\rho}$ gives $g_{f,\rho}(F(t)) \leq P_1(S \leq t)$. Taking the infimum over all possible distributions $P_1$ concludes the proof.

## A.2 Proof of Proposition 2

Since $\rho^\star = D_f(P_{\mathrm{test}} \| P_0) < \infty$, the definition of $F^{\mathrm{WC}}_{f,\rho}$ and Lemma A.2 imply that for all $q \in \mathbb{R}$,

$$F_{\mathrm{test}}(q) \geq F^{\mathrm{WC}}_{f,\rho^\star}(q, P_0) = g_{f,\rho^\star}(F_0(q)).$$

Applying this inequality with $q := \text{Quantile}^{\text{WC}}_{f,\rho}(1-\alpha; \hat{P}_n) = \text{Quantile}(g^{-1}_{f,\rho}(1-\alpha); \hat{P}_n)$, we obtain

$$\begin{aligned}\mathbb{P}\Big(Y_{n+1} \in \hat{C}_{n,f,\rho}(X_{n+1}) \mid \{(X_i,Y_i)\}_{i=1}^n\Big) &\overset{(i)}{=} F_{\text{test}}(\text{Quantile}^{\text{WC}}_{f,\rho}(1-\alpha; \hat{P}_n)) \\ &\geq g_{f,\rho^\star}(F_0(\text{Quantile}^{\text{WC}}_{f,\rho}(1-\alpha; \hat{P}_n))) \\ &\overset{(ii)}{=} g_{f,\rho^\star}(F_0(\text{Quantile}(g^{-1}_{f,\rho}(1-\alpha); \hat{P}_n))),\end{aligned}$$

where equality $(i)$ uses that $s(X_{n+1}, Y_{n+1}) \sim P_{\text{test}}$ is independent of $\{(X_i,Y_i)\}_{i=1}^n$ and $(ii)$ is Proposition 1.

### A.3 Proof of Theorem 1

We require the following lemma to prove the theorem.

**Lemma A.3** (Quantile coverage [40, 24, 1])**.** *Assume that $\{S_i\}_{i=1}^n \overset{\text{iid}}{\sim} P_0$ with c.d.f. $F_0$, and let $\hat{P}_n$ be their empirical distribution. Then for all $\beta \in (0,1)$,*

$$\mathbb{E}\left[F_0\left(\text{Quantile}(\beta; \hat{P}_n)\right)\right] \geq \frac{\lceil n\beta \rceil}{n+1}.$$

We include the brief proof of Lemma A.3 below for completeness, giving the proof of Theorem 1 here. By Proposition 2, for $\rho^\star = D_f(P_{\text{test}} \| P_0) < \infty$, we have

$$\mathbb{P}\Big(Y_{n+1} \in \hat{C}_{n,f,\rho}(X_{n+1}) \mid \{(X_i,Y_i)\}_{i=1}^n\Big) \geq g_{f,\rho^\star}(F_0(\text{Quantile}(g^{-1}_{f,\rho}(1-\alpha); \hat{P}_n))).$$

Marginalizing over $(X_i, Y_i)$, this implies that

$$\begin{aligned}\mathbb{P}\Big(Y_{n+1} \in \hat{C}_{n,f,\rho}(X_{n+1})\Big) &\geq \mathbb{E}\left[g_{f,\rho^\star}(F_0(\text{Quantile}(g^{-1}_{f,\rho}(1-\alpha); \hat{P}_n))\right] \\ &\overset{(i)}{\geq} g_{f,\rho^\star}\left(\mathbb{E}\left[F_0(\text{Quantile}(g^{-1}_{f,\rho}(1-\alpha); \hat{P}_n)\right]\right) \\ &\overset{(ii)}{\geq} g_{f,\rho^\star}\left(\frac{\left\lceil n g^{-1}_{f,\rho}(1-\alpha)\right\rceil}{n+1}\right),\end{aligned}$$

where inequality $(i)$ is a consequence of Jensen's inequality applied to $g_{f,\rho^\star}$ (recall Lemma A.1(a)), while inequality $(ii)$ uses Lemma A.3 and that $\beta \mapsto g_{f,\rho}(\beta)$ is non-decreasing.

**Proof of Lemma A.3** Let $S_{n+1} \sim P_0$ independent of $\{S_i\}_{i=1}^n$. Then

$$\begin{aligned}\mathbb{E}\left[F_0\left(\text{Quantile}(\beta; P_n)\right)\right] &= \mathbb{P}\left(S_{n+1} \leq \text{Quantile}(\beta; P_n)\right) \\ &\geq \mathbb{P}(\text{Rank of } S_{n+1} \text{ in } \{S_i\}_{i=1}^{n+1} \leq \lceil n\beta \rceil) = \frac{\lceil n\beta \rceil}{n+1},\end{aligned}$$

where we break ties uniformly at random to define the rank of $S_{n+1}$ in $\{S_i\}_{i=1}^{n+1}$, ensuring by exchangeability that it is uniform on $\{1, \ldots, n+1\}$. □

### A.4 Proof of Corollaries 2.1 and 2.2

When $\rho^\star = D_f(P_{\text{test}} \| P_0) \le \rho$, Lemma A.1 guarantees that $g_{f,\rho} \ge g_{f,\rho^\star}$, so Theorem 1 gives

$$\mathbb{P}(Y_{n+1} \in \hat{C}_{n,f,\rho}) \ge g_{f,\rho}\left(\frac{\left\lceil n g_{f,\rho}^{-1}(1-\alpha)\right\rceil}{n+1}\right). \tag{20}$$

To prove Corollary 2.1, note that as $g_{f,\rho}$ in convex, it has (at least) a left derivative $g'_{f,\rho}$, which satisfies

$$g_{f,\rho}\left(\frac{\left\lceil n g_{f,\rho}^{-1}(1-\alpha)\right\rceil}{n+1}\right) \ge g_{f,\rho}\left(\frac{n g_{f,\rho}^{-1}(1-\alpha)}{n+1}\right) \ge 1-\alpha - \frac{g_{f,\rho}^{-1}(1-\alpha) g'_{f,\rho}(g_{f,\rho}^{-1}(1-\alpha))}{n+1}.$$

This gives the first corollary.

For the second corollary, replacing $\hat{C}$ in Eq. (20) with $\hat{C}^{\text{corr}}$ gives

$$\begin{aligned}\mathbb{P}(Y_{n+1} \in \hat{C}^{\text{corr}}_{n,f,\rho}) &\ge g_{f,\rho}\left(\frac{\left\lceil n g_{f,\rho}^{-1}\left(g_{f,\rho}\left((1+1/n) g_{f,\rho}^{-1}(1-\alpha)\right)\right)\right\rceil}{n+1}\right) \\ &= g_{f,\rho}\left(\frac{\left\lceil n(1+1/n) g_{f,\rho}^{-1}(1-\alpha)\right\rceil}{n+1}\right) \ge g_{f,\rho}\left(g_{f,\rho}^{-1}(1-\alpha)\right) \ge 1-\alpha.\end{aligned}$$

## B Proof of Theorem 2

Throughout the proof, we will typically not assume that the scores $s(X_i, Y_i)$ are distinct, and thus will not make Assumption A1. Some inequalities will require the assumption, which implies the distinctness of the scores, and we will highlight those.

Recall that $\{(X_i, Y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} Q_0$ and $\{s(X_i, Y_i)\}_{i=1}^n \overset{\text{iid}}{\sim} P_0$, that for all $q \in \mathbb{R}$

$$C^{(q)}(x) = \{y \in \mathcal{Y} \mid s(x,y) \le q\},$$

and that we use $P_0(\cdot \mid X \in R)$ as shorthand for the law of $s(X,Y)$ for $(X,Y) \sim Q_0(\cdot \mid X \in R)$. We also use $\hat{Q}_n$ and $\hat{P}_n$ for the empirical distributions f $Q$ and $P$, respectively. Observe that for all $q \in \mathbb{R}$ and $0 < \delta < 1$, $\text{WC}(C^{(q)}, \mathcal{R}_v, \delta; Q_0) \ge 1-\alpha$ if and only if

$$\sup_{R \in \mathcal{R}_v : Q_0(R) \ge \delta} \text{Quantile}(1-\alpha; P_0(\cdot \mid X \in R)) \le q.$$

By a VC-covering argument (cf. [7, Sec. A.4] or [1, Thm. 5]), there exists a universal constant $C_\varepsilon < \infty$ such that the following holds. For $t > 0$, define $\epsilon_n(t) \coloneqq C_\varepsilon \sqrt{\frac{\text{VC}(\mathcal{R}) \log(n) + t}{n}}$. Then with probability at least $1 - \frac{1}{2}e^{-t}$ over $\{X_i, Y_i\}_{i=1}^n$, the following equations hold simultaneously for all $v \in \mathcal{V}$:

$$\sup_{s \in \mathbb{R}} \left| \inf_{\substack{R \in \mathcal{R}_v \\ \hat{Q}_n(R) \ge \delta}} \hat{P}_n\left(s(X,Y) \le s \mid X \in R\right) - \inf_{\substack{R \in \mathcal{R}_v \\ \hat{Q}_n(R) \ge \delta}} P_0\left(s(X,Y) \le s \mid X \in R\right) \right| \le \frac{\varepsilon_n(t)}{\sqrt{\delta}} \tag{21}$$

and

$$\sup_{R\in\mathcal{R}_v}\left|\hat{Q}_n(X\in R)-Q_0(X\in R)\right|\le\varepsilon_n(t). \tag{22}$$

We assume for the remainder of the proof that inequalities (21) and (22) hold.

Define the empirical quantile

$$\widehat{q}_n(v,\delta)\coloneqq\sup_{R\in\mathcal{R}_v}\left\{\text{Quantile}(1-\alpha;\hat{P}_n(\cdot\mid X\in R))\text{ s.t. }\hat{Q}_n(X\in R)\ge\delta\right\}.$$

We first give a lemma on its coverage.

**Lemma B.1.** *Let the bounds (21) and (22) hold. Then*

$$\begin{aligned}\text{WC}(C^{(\widehat{q}_n(v,\delta))},\mathcal{R}_v,\delta_n^+(t);\,Q_0)&\;\ge\;1-\alpha_n^+(t)\\ \text{WC}(C^{(\widehat{q}_n(v,\delta))},\mathcal{R}_v,\delta_n^-(t);\,Q_0)&\overset{\text{(A1)}}{\le}1-\alpha_n^-(t)\end{aligned} \tag{23}$$

*simultaneously for all $v\in\mathcal{V}$, where the second inequality requires Assumption A1.*

**Proof** Applying the bounds (21), we can bound the worst-case quantiles via

$$\begin{aligned}&\sup_{R\in\mathcal{R}_v}\left\{\text{Quantile}(1-\alpha_n^+(t);P_0(\cdot\mid X\in R))\text{ s.t. }\hat{Q}_n(X\in R)\ge\delta\right\}\\ &\quad\le\widehat{q}_n(v,\delta)\le\sup_{R\in\mathcal{R}_v}\left\{\text{Quantile}(1-\alpha_n^-(t);P_0(\cdot\mid X\in R))\text{ s.t. }\hat{Q}_n(X\in R)\ge\delta\right\}.\end{aligned} \tag{24}$$

The inclusions

$$\begin{aligned}\{R\in\mathcal{R}\mid Q_0(X\in R)\ge\delta-\varepsilon_n(t)\}&\subset\{R\in\mathcal{R}\mid\hat{Q}_n(X\in R)\ge\delta\}\\ &\subset\{R\in\mathcal{R}\mid Q_0(X\in R)\ge\delta+\varepsilon_n(t)\}\end{aligned}$$

are an immediate consequence of inequality (22), and, in turn, imply that for all $\alpha\in(0,1)$,

$$\begin{aligned}\sup_{\substack{R\in\mathcal{R}_v\\ Q_0(X\in R)\ge\delta_n^+(t)}}\text{Quantile}(1-\alpha;P_0(\cdot\mid X\in R))&\le\sup_{\substack{R\in\mathcal{R}_v\\ \hat{Q}_n(X\in R)\ge\delta}}\text{Quantile}(1-\alpha;P_0(\cdot\mid X\in R))\\ &\le\sup_{\substack{R\in\mathcal{R}_v\\ Q_0(X\in R)\ge\delta_n^-(t)}}\text{Quantile}(1-\alpha;P_0(\cdot\mid X\in R)).\end{aligned}$$

Combining these inclusions with the inequalities (24), we thus obtain

$$\begin{aligned}&q_n^{\inf}(v)\coloneqq\sup_{R\in\mathcal{R}_v}\left\{\text{Quantile}(1-\alpha_n^+(t);P_0(\cdot\mid X\in R))\text{ s.t. }Q_0(X\in R)\ge\delta_n^+(t)\right\}\\ &\le\widehat{q}_n(v,\delta)\\ &\le\sup_{R\in\mathcal{R}_v}\left\{\text{Quantile}(1-\alpha_n^-(t);P_0(\cdot\mid X\in R))\text{ s.t. }Q_0(X\in R)\ge\delta_n^-(t)\right\}=:q_n^{\sup}(v).\end{aligned} \tag{25}$$

The infimum and supremum quantiles satisfy

$$\begin{aligned}\text{WC}(C^{(q_n^{\inf}(v))},\mathcal{R}_v,\delta_n^+(t);\,Q_0)&\;\ge\;1-\alpha_n^+(t)\\ \text{WC}(C^{(q_n^{\sup}(v))},\mathcal{R}_v,\delta_n^-(t);\,Q_0)&\overset{\text{(A1)}}{=}1-\alpha_n^-(t),\end{aligned}$$

where the inequality always holds and the equality requires Assumption A1.

We now observe that for any fixed $(v, \delta) \in \mathcal{V} \times (0,1)$, the function $q \mapsto \mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta; Q_0)$ is non-decreasing, since the confidence sets $C^{(q)}(x)$ increase as $q$ increases. Recalling inequalities (25), we conclude that

$$\mathrm{WC}(C^{(\widehat{q}_n(v,\delta)}, \mathcal{R}_v, \delta_n^+(t); Q_0) \geq \mathrm{WC}(C^{(q_n^{\inf}(v))}, \mathcal{R}_v, \delta_n^+(t); Q_0) \geq 1 - \alpha_n^+(t)$$

and

$$\mathrm{WC}(C^{(\widehat{q}_n(v,\delta))}, \mathcal{R}_v, \delta_n^-(t); Q_0) \leq \mathrm{WC}(C^{(q_n^{\sup}(v))}, \mathcal{R}_v, \delta_n^-(t); Q_0) = 1 - \alpha_n^-(t),$$

simultaneously for all $v \in \mathcal{V}$, with the second inequality requiring Assumption A1. □

Recall that $\widehat{q}_\delta$ in Algorithm 1 is the $(1-\alpha_{\mathsf{v}})$-empirical quantile of $\{\widehat{q}_n(v_i, \delta)\}_{i=1}^k$. Then inequalities (23) in Lemmma B.1 and that $\mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta; Q_0)$ is non-decreasing in $q$ imply

$$\hat{\mathbb{P}}_{\mathsf{v},k}\left[\mathrm{WC}(C^{(\widehat{q}_\delta)}, \mathcal{R}_v, \delta_n^+(t); Q_0) \geq 1 - \alpha_n^+(t)\right] \geq \hat{\mathbb{P}}_{\mathsf{v},k}\left[\widehat{q}_\delta \geq \widehat{q}_n(v_i, \delta)\right] \geq 1 - \alpha_{\mathsf{v}},$$

while under Assumption A1, we have the converse lower bound

$$\hat{\mathbb{P}}_{\mathsf{v},k}\left[\mathrm{WC}(C^{(\widehat{q}_\delta)}, \mathcal{R}_v, \delta_n^-(t); Q_0) \leq 1 - \alpha_n^-(t)\right] \geq \hat{\mathbb{P}}_{\mathsf{v},k}\left[\widehat{q}_\delta \leq \widehat{q}_n(v, \delta)\right] \geq \alpha_{\mathsf{v}} - \frac{1}{k},$$

using the second inequality of Lemma B.1.

For $q \in \mathbb{R}$, define the functions $f_q^+(v) \coloneqq 1\{\mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta_n^+(t); Q_0) \geq 1 - \alpha_n^+(t)\} \in \{0,1\}$ for all $q \in \mathbb{R}$. The set of functions $\{f_q^+\}_{q\in\mathbb{R}}$ is uniformly bounded (by 1) and each is non-decreasing in $q \in \mathbb{R}$ so that its VC-dimension cannot exceed 1. Thus, there exists a universal constant $C < \infty$ such that, with probability $1 - 4^{-1}e^{-t}$ [e.g. 41, Thm. 4.10, Ex. 5.24],

$$\sup_{q\in\mathbb{R}} \left|\hat{\mathbb{P}}_{\mathsf{v},k} f_q^+ - \mathbb{P}_{\mathsf{v}} f_q^+\right| \leq C\sqrt{\frac{1+t}{k}}.$$

Similarly, if we define $f_q^-(v) \coloneqq 1\{\mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta_n^-(t); Q_0) \leq 1 - \alpha_n^-(t)\} \in \{0,1\}$, then with probability at least $1 - 4^{-1}e^{-t}$, we have $\sup_{q\in\mathbb{R}} |\hat{\mathbb{P}}_{\mathsf{v},k} f_q^- - \mathbb{P}_{\mathsf{v}} f_q^-| \leq C\sqrt{\frac{1+t}{k}}$. Combining the statements, we see that with probability $1 - e^{-t}$ over the draw $(X_i, Y_i)_{i=1}^n \overset{\mathrm{iid}}{\sim} Q_0$ and $\{v_i\}_{i=1}^k \overset{\mathrm{iid}}{\sim} \mathbb{P}_{\mathsf{v}}$, we have

$$\mathbb{P}_{\mathsf{v}} f_{\widehat{q}_\delta}^+ = \mathbb{P}_{\mathsf{v}}\left[\mathrm{WC}(C^{(\widehat{q}_\delta)}, \mathcal{R}_v, \delta_n^+(t); Q_0) \geq 1 - \alpha_n^+(t)\right] \geq 1 - \alpha_{\mathsf{v}} - C\sqrt{\frac{1+t}{k}},$$

and under Assumption A1,

$$\mathbb{P}_{\mathsf{v}} f_{\widehat{q}_\delta}^- = \mathbb{P}_{\mathsf{v}}\left[\mathrm{WC}(C^{(\widehat{q}_\delta)}, \mathcal{R}_v, \delta_n^-(t); Q_0) \leq 1 - \alpha_n^-(t)\right] \overset{\text{(A1)}}{\geq} \alpha_{\mathsf{v}} - \frac{1}{k} - C\sqrt{\frac{1+t}{k}}.$$

## B.1 Proof of Lemma 3.1

Let $S_i = s(X_i, Y_i)$ for shorthand, and assume w.l.o.g. that $S_1 \leq \cdots \leq S_n$. We will show that if $\widehat{q} \in \{S_i\}_{i \geq \lceil n(1-\alpha)\rceil}$, then if

$$\widehat{\rho} = \rho_{f,\alpha}(\widehat{q}; \hat{P}_n) \quad \text{then} \quad \widehat{q} = \mathrm{Quantile}^{\mathrm{WC}}_{f,\widehat{\rho}}(1-\alpha; \hat{P}_n). \tag{26}$$

Evidently this implies that $C^{(\widehat{q})}(x) = C_{f,\widehat{\rho}}(x; \hat{P}_n)$ for all $x \in \mathcal{X}$, giving the lemma, so for the remainder, we show the equivalence (26).

Recall the definition $g_{f,\rho}^{-1}(\tau) = \sup\{\beta \in [\tau, 1] \mid \beta f(\frac{\tau}{\beta}) + (1-\beta) f(\frac{1-\tau}{1-\beta}) \le \rho\}$ in the discussion following Proposition 1. Suppose that $\widehat{q} = S_j$, where $j \in [n]$, which immediately implies that, for all $(j-1)/n < \beta \le j/n$, $\widehat{q} = \text{Quantile}(\beta; \hat{P}_n)$. By Proposition 1, we therefore see that if $\rho \ge 0$ satisfies $(j-1)/n < g_{f,\rho}^{-1}(1-\alpha) \le j/n$, then

$$\text{Quantile}_{f,\rho}^{\text{WC}}(1-\alpha; \hat{P}_n) = \widehat{q}.$$

In addition, as the scores $S_i$ are all distinct, $\text{Quantile}_{f,\rho}^{\text{WC}}(1-\alpha; \hat{P}_n) > \widehat{q}$ if $g_{f,\rho}^{-1}(1-\alpha) > j/n$, making $\rho_{f,\alpha}$ in this case equal to

$$\rho_{f,\alpha}(\widehat{q}; \hat{P}_n) = \sup\{\rho \ge 0 \mid g_{f,\rho}^{-1}(1-\alpha) \le j/n\}.$$

The mapping $\rho \mapsto g_{f,\rho}^{-1}(\tau)$ is concave and nonnegative. As $f$ is 1-coercive by assumption, we also have that it is defined on $\mathbb{R}_+$, and it is continuous strictly increasing on $\mathbb{R}_{++}$. Its inverse (as a function of $\rho$) is therefore continuous, which implies in particular that $g^{-1}_{f,\rho_{f,\alpha}(\widehat{q};\hat{P}_n)}(1-\alpha) = j/n$, and hence equality (26) holds.

# C Proofs related to finding worst shift directions

## C.1 Proof of Lemma 3.2

Fix $q \in \mathbb{R}$ and $\delta \in (0,1)$. For $(u,t) \in \mathbb{S}^{d-1} \times \mathbb{R}$ such that $\mathbb{P}(X^T u \ge t) \ge \delta$, there exists $\tau$ such that $\mathbb{P}(X^T v^\star \ge \tau) = \mathbb{P}(X^T u \ge t)$ by continuity of the distribution of $X^T v^\star$. Then

$$\mathbb{P}(s(X,Y) > q \mid X^T v^\star \ge \tau) \ge \mathbb{P}(s(X,Y) > q \mid X^T u \ge t)$$

by Assumption A2, which implies that

$$\text{WC}(C^{(q)}, \mathcal{R}_{v^\star}, \delta; Q_0) \le \mathbb{P}(s(X,Y) \le q \mid X^T u \ge t).$$

The result follows by taking the infimum over all $(u,t) \in \mathbb{S}^{d-1} \times \mathbb{R}$ such that $\mathbb{P}(X^T u \ge t) \ge \delta$.

## C.2 Proof of Lemma 3.3

The first condition of the assumption is immediate, as for any $\mathbb{R}$-valued random variable $Z$, we have $\mathcal{L}(Z \mid Z \ge \tau) \succeq \mathcal{L}(Z \mid Z \ge t)$ for $\tau \ge t$ (cf. [34, Thm. 1.A.15]). I now claim that if $(Z_1, Z_2) \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$, then for any $\beta \in [-1,1]$ and $\tau \in \mathbb{R}$,

$$\mathcal{L}(Z_1 \mid Z_1 \ge \tau) \succeq \mathcal{L}(\beta Z_1 + \sqrt{1-\beta^2} Z_2 \mid Z_1 \ge \tau). \tag{27}$$

The result is clear when $\beta = 1$, so let $\beta < 1$. Let $\gamma$ denote the 2-dimensional Gaussian measure, let $t > \tau$, and define the halfspaces $A_1 = \{z_1 \ge t\}$, $A_2 = \{\beta z_1 + \sqrt{1-\beta^2} z_2 \ge t\}$, and $B = \{z_1 \ge \tau\}$. Then $\gamma(A_1) = \gamma(A_2)$ by the rotational symmetry of Gaussians, and $A_1 \cap B = A_1$ while $A_2 \cap B \subsetneq A_2$. Thus

$$\mathbb{P}(Z_1 \ge t \mid Z_1 \ge \tau) = \frac{\gamma(A_1 \cap B)}{\gamma(B)} > \frac{\gamma(A_2 \cap B)}{\gamma(B)} = \mathbb{P}(\beta Z_1 + \sqrt{1-\beta^2} Z_2 \ge t \mid Z_1 \ge \tau).$$

If $t < \tau$, then $\mathbb{P}(Z_1 \geq t \mid Z_1 \geq \tau) = 1 \geq \mathbb{P}(\beta Z_1 + \sqrt{1-\beta^2} Z_2 \geq t \mid Z_1 \geq \tau)$, proving the dominance (27).

We can now show the second condition of Assumption A2. Without loss of generality, let $u \in \mathbb{S}^{d-1}$. Then writing $X = (I - uu^T)X + uu^T X$, we have

$$v_{\text{var}}^T X = v_{\text{var}}^T (I - uu^T) X + v_{\text{var}}^T u u^T X \stackrel{\text{dist}}{=} \sqrt{1-\beta^2} Z_2 + \beta Z_1$$

for $\beta = v_{\text{var}}^T u$ and $Z_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0,1)$. Notably, if $\mathbb{P}(u^T X \geq t) \geq \mathbb{P}(v_{\text{var}}^T X \geq \tau)$, then $t \leq \tau$, and therefore

$$\mathcal{L}(v_{\text{var}}^T X \mid v_{\text{var}}^T X \geq \tau) \succeq \mathcal{L}(v_{\text{var}}^T X \mid v_{\text{var}}^T X \geq t) \stackrel{(\star)}{\succeq} \mathcal{L}(v_{\text{var}}^T X \mid u^T X \geq t)$$

where inequality $(\star)$ is the dominance (27). As $h$ is increasing, we therefore obtain that

$$\begin{aligned}\mathcal{L}(s(X,Y) \mid v_{\text{var}}^T X \geq \tau) &= \mathcal{L}(h(v^T X)^2 \varepsilon^2 \mid v_{\text{var}}^T X \geq \tau) \\ &\succeq \mathcal{L}(h(v_{\text{var}}^T X)^2 \varepsilon^2 \mid u^T X \geq t) = \mathcal{L}(s(X,Y) \mid u^T X \geq t)\end{aligned}$$

as desired.

### C.2.1 Proof of Proposition 3

We begin with a few preliminaries on stochastic orders on random vectors. For random vectors $U$ and $V \in \mathbb{R}^d$, we say that $U$ stochastically dominates $V$ in the *upper orthant order* if for each $t \in \mathbb{R}^d$ we have $\mathbb{P}(U \geq t) \geq \mathbb{P}(V \geq t)$, written $U \succeq_{\mathsf{uo}} V$ (see [34, Ch. 6], where this is called the *usual stochastic order*). We also have the following.

**Lemma C.1.** *Let $U, V \in \mathbb{R}^2$. Then $U \succeq_{\mathsf{uo}} V$ if and only if for all non-negative and non-decreasing functions $f, g$,*

$$\mathbb{E}[f(V_1)g(V_2)] \leq \mathbb{E}[f(U_1)g(U_2)]. \tag{28}$$

*If additionally $U_1 \stackrel{\text{dist}}{=} V_1$ and $\mathbb{E}[|f(V_1)g(V_2)|]$ and $\mathbb{E}[|f(U_1)g(U_2)|] < \infty$, then $\mathbb{E}[f(V_1)g(V_2)] \leq \mathbb{E}[f(U_1)g(U_2)]$ for all non-negative and non-decreasing $f$ and non-decreasing $g$.*

**Proof** The equivalence of inequality (28) and $U \succeq_{\mathsf{uo}} V$ is [34, Eq. (6.B.4)]. For the second result, consider the sequence $g_m(x) := [g(x) + m]_+ - m$ for $m = 1, 2, \ldots$. Then $g_m \downarrow g$, while

$$\mathbb{E}[f(U_1)g_m(U_2)] \geq \mathbb{E}[f(V_1)\left[g(V_2) + m\right]_+] - m\mathbb{E}[g(V_1)] = \mathbb{E}[f(V_1)g_m(V_2)].$$

Dominated convergence gives the result. □

We now show that

$$(s(X,Y), X^T v^\star) \succeq_{\mathsf{uo}} (s(X,Y), X^T u) \tag{29}$$

for any vector $u$ satisfying $\|\Sigma^{1/2} v^\star\|_2 = \|\Sigma^{1/2} u\|_2$. Without loss of generality, we assume $\|\Sigma^{1/2} v^\star\|_2 = 1$. Then for all $q \in \mathbb{R}$ and $t \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}(s(X,Y) \geq q, X^T v^\star \geq t) &= \mathbb{P}(s(X,Y) \geq q \mid X^T v^\star \geq t)\mathbb{P}(X^T v^\star \geq t) \\ &\stackrel{(\star)}{\geq} \mathbb{P}(s(X,Y) \geq q \mid X^T u \geq t)\mathbb{P}(X^T u \geq t) \\ &= \mathbb{P}(s(X,Y) \geq q, X^T u \geq t),\end{aligned}$$

where inequality ($\star$) uses Assumption A2 and that $\tilde{X} \coloneqq \Sigma^{-1/2}X$ has an isotropic disitribution, so that $\mathbb{P}(X^Tu \geq t) = \mathbb{P}(\tilde{X}^T\Sigma^{1/2}u \geq t) = \mathbb{P}(\tilde{X}^T\Sigma^{1/2}v^\star \geq t) = \mathbb{P}(X^Tv^\star \geq t)$ and $X^Tu \stackrel{\text{dist}}{=} X^Tv^\star$. In particular, Lemma C.1 yields

$$\mathbb{E}\left[s(X,Y)X^Tu\right] \leq \mathbb{E}\left[s(X,Y)X^Tv^\star\right]$$

for all $u \in \mathbb{R}^d$ such that $\|\Sigma^{1/2}u\|_2 = \|\Sigma^{1/2}v^\star\|_2$, because $\mathbb{E}[|s(X,Y)X^Tu|] < \infty$ by Cauchy-Schwarz. As a result, using the assumption in the proposition that $\mathbb{E}[s(X,Y)X] \neq 0$, we have the fixed point

$$v^\star = \operatorname*{argmin}_{u\in\mathbb{R}^d}\left\{\mathbb{E}\left[s(X,Y)X^Tu\right] \mid \|\Sigma^{1/2}u\|_2 = \|\Sigma^{1/2}v^\star\|_2\right\}.$$

By a direct change of variables via $\tilde{X} = \Sigma^{-1/2}X$, this is equivalent to

$$\Sigma^{1/2}v^\star = \operatorname*{argmin}_{\tilde{u}\in\mathbb{R}^d}\left\{\tilde{u}^T\mathbb{E}\left[s(X,Y)\tilde{X}\right] \mid \|\tilde{u}\|_2 = \left\|\Sigma^{1/2}v^\star\right\|_2\right\}.$$

Rewriting, we obtain

$$v^\star \propto \Sigma^{-1}\mathbb{E}\left[Xs(X,Y)\right] = \mathbb{E}\left[XX^T\right]^{-1}\mathbb{E}\left[Xs(X,Y)\right] = \operatorname*{argmin}_u \mathbb{E}[(s(X,Y) - X^Tu)^2].$$

## C.3 Proof of Theorem 3

The proof of the theorem is technical, so we state and prove several lemmas on worst coverage regularity and convergence (Section C.3.1), combining all the pieces in Section C.3.2.

### C.3.1 Lemmas on worst coverage estimation

**Lemma C.2.** *Let Assumption A4 hold. Then the function $(q,v,\delta) \mapsto \mathrm{WC}(C^{(q,s)}, \mathcal{R}_v, \delta; Q_0)$ is continuous on $\mathbb{R}\times\mathcal{V}\times(0,1)$ and uniformly continuous on $\mathbb{R}\times\mathcal{V}\times[\delta,1)$ for any $\delta > 0$.*

**Proof** We use $C^{(q)}$ as shorthand for $C^{(q,s)}$, and we consider a sequence $\{(q_n,v_n,\delta_n)\}_{n\geq1} \to (q,v,\delta) \in \mathbb{R}\times\mathcal{V}\times(0,1)$. We will show that $\{\mathrm{WC}(C^{(q_n)}, \mathcal{R}_{v_n}, \delta_n; Q_0)\}_{n\geq1}$ converges by proving that the sequence has a unique accumulation point. We therefore assume without loss of generality that

$$\mathrm{WC}(C^{(q_n)}, \mathcal{R}_{v_n}, \delta_n; Q_0) \underset{n\to\infty}{\longrightarrow} \ell \in [0,1], \tag{30}$$

and we successively prove that $\ell \leq \mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta; Q_0)$ and $\mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta; Q_0) \leq \ell$. Combining claims C.1 and C.2 immediately gives the continuity claim in Lemma C.2.

**Claim C.1.** *The limit $\ell$ in Eq. (30) satisfies $\ell \leq \mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta; Q_0)$.*

**Proof** Let $\varepsilon > 0$, and consider $t \in \mathbb{R}$ such that $\mathbb{P}(X^Tv \geq t) \in (\delta, 1)$ and

$$Q_0\left(s(X,Y) \leq q \mid X^Tv \geq t\right) \leq \mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta; Q_0) + \varepsilon.$$

Next, consider $t_n \in \mathbb{R}$ such that $Q_0(X^Tv_n \geq t_n) = \max\{\delta_n, Q_0(X^Tv \geq t)\}$. As we may consider a subsequence, we assume without loss of generality that $\{t_n\}_{n\geq1}$ converges to $\tilde{t} \in [-\infty,\infty]$. Then we have by Slutsky's lemma that $X^Tv_n - t_n \xrightarrow{d} X^Tv - \tilde{t}$, and thus

$$Q_0(X^Tv \geq \tilde{t}) = \lim_{n\to\infty} Q_0(X^Tv_n \geq t_n) = Q_0(X^Tv \geq t) \geq \delta,$$

as $X^T v$ has a continuous distribution (the above relation also proves that $\tilde{t} \in \mathbb{R}$, since $0 < Q_0(X^T v \geq t) < 1$). Since we either have $\{X^T v \geq \tilde{t}\} \subset \{X^T v \geq t\}$ or $\{X^T v \geq t\} \subset \{X^T v \geq \tilde{t}\}$, the above relation also shows that

$$Q_0(s(X,Y) \leq q \mid X^T v \geq \tilde{t}) = Q_0(s(X,Y) \leq q \mid X^T v \geq t) \leq \mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta;\, Q_0) + \varepsilon.$$

Finally, if we define $\Delta_{n,v} \coloneqq X^T v_n - X^T v - t_n + \tilde{t} \xrightarrow{p} 0$, we have

$$\begin{aligned}
&\left|Q_0(s(X,Y) \leq q_n, X^T v_n \geq t_n) - Q_0(s(X,Y) \leq q, X^T v \geq \tilde{t})\right| \\
&\leq Q_0\left(|s(X,Y) - q| \leq |q_n - q|\right) + Q_0(\tilde{t} - \Delta_{n,v} \leq X^T v < \tilde{t}) + Q_0(\tilde{t} \leq X^T v < \tilde{t} - \Delta_{n,v}) \quad (31) \\
&\underset{n\to\infty}{\longrightarrow} 0,
\end{aligned}$$

where the first (resp. second and third) term converges to 0 as the distribution of $s(X,Y)$ (resp. $X^T v$) is continuous under $Q_0$. This proves that

$$Q_0(s(X,Y) \leq q_n \mid X^T v_n \geq t_n) \underset{n\to\infty}{\longrightarrow} Q_0(s(X,Y) \leq q \mid X^T v \geq \tilde{t}) \leq \mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta;\, Q_0) + \varepsilon,$$

and thus $\ell \leq \mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta;\, Q_0) + \varepsilon$. As $\varepsilon > 0$ was arbitrary, we have the claim. □

**Claim C.2.** *The limit $\ell$ in Eq. (30) satisfies $\ell \leq \mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta; Q_0)$.*

**Proof** By definition of the worst-coverage, we can find $\{t_n\}_{n\geq 1}$ such that $Q_0(X^T v_n \geq t_n) \geq \delta_n$ for all $n \geq 1$, and

$$Q_0(s(X,Y) \leq q_n \mid X^T v_n \geq t_n) \underset{n\to\infty}{\longrightarrow} \ell.$$

As we may always consider a subsequence, we again assume that $t_n \to t \in [-\infty, \infty]$. Next, observe that, by continuous mapping and Slutsky's lemma, $X^T v_n - t_n \xrightarrow{d} X^T v - t$ (where the limit distribution is continuous but potentially infinite if $t \in \{-\infty, \infty\}$), so

$$Q_0(X^T v \geq t) = \lim_n Q_0(X^T v_n \geq t_n) \geq \delta \quad (32)$$

by the Portmanteau theorem, which also proves that $t < \infty$.

If $t = -\infty$, then $Q_0(X^T v_n \geq t_n) \to 1$. As the distribution of $s(X,Y)$ is continuous under $Q_0$, this ensures that

$$Q_0(s(X,Y) \leq q_n \mid X^T v_n \geq t_n) \to Q_0(s(X,Y) \leq q) \geq \mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta;\, Q_0),$$

and proves that $\ell \geq \mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta;\, Q_0)$. If $t \in \mathbb{R}$, then with derivation *mutatis mutandis* identical to that to develop the convergence (31), we obtain that

$$Q_0(s(X,Y) \leq q_n, X^T v_n \geq t_n) - Q_0(s(X,Y) \leq q, X^T v \geq t) \underset{n\to\infty}{\to} 0.$$

With equation (32), this directly shows that

$$\begin{aligned}
\mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta;\, Q_0) &\leq Q_0(s(X,Y) \leq q \mid X^T v \geq t) \\
&= \lim_{n\to\infty} Q_0(s(X,Y) \leq q \mid X^T v_n \geq t_n) = \ell
\end{aligned}$$

as desired. □

For the proof of the uniform continuity claim, let $\bar{\mathbb{R}} = [-\infty, \infty]$ be the usual compactification of $\mathbb{R}$. We extend the worst-coverage function to $\bar{\mathbb{R}} \times \mathcal{V} \times (0,1]$ by setting $\mathrm{WC}(C^{(-\infty)}, \mathcal{R}_v, \delta; Q_0) = 0$, and $\mathrm{WC}(C^{(+\infty)}, \mathcal{R}_v, \delta; Q_0) = 1$ for all $(v, \delta) \in \mathcal{V} \times (0,1]$, and $\mathrm{WC}(C^{(q)}, \mathcal{R}_v, 1; Q_0) = P_0(S \le q)$ for all $(q, v) \in \mathbb{R} \times \mathcal{V}$. The bound

$$1 - \delta^{-1} P_0(S > q) \le \mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta; Q_0) \le \delta^{-1} P_0(S \le q),$$

valid for all $(q, v, \delta) \in \mathbb{R} \times \mathcal{V} \times (0,1)$, ensures that the extension itself is continuous on $\bar{\mathbb{R}} \times \mathcal{V} \times (0,1]$, as if $\{(q_n, v_n, \delta_n)\}_{n \ge 1} \to (q, v, \delta)$, with $q \in \{-\infty, \infty\}$ or $\delta = 1$, we still have

$$\mathrm{WC}(C^{(q_n)}, \mathcal{R}_{v_n}, \delta_n; Q_0) \underset{n \to \infty}{\longrightarrow} \mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta; Q_0) = \begin{cases} P_0(S \le q) & \text{if } \delta = 1 \text{ and } q \in \mathbb{R} \\ 0 & \text{if } q = -\infty \\ 1 & \text{if } q = \infty. \end{cases}$$

The set $\bar{\mathbb{R}} \times \mathcal{V} \times [\delta, 1]$ is compact, whence Heine's theorem ensures that the function $(q, v, \delta) \mapsto \mathrm{WC}(C^{(q)}, \mathcal{R}_v, \delta; Q_0)$ is uniformly continuous on $\bar{\mathbb{R}} \times \mathcal{V} \times [\delta, 1]$. The restriction of the function to $\mathbb{R} \times \mathcal{V} \times [\delta, 1)$ is then also uniformly continuous. □

**Lemma C.3.** *As $n \to \infty$ $(n_1, n_2 \to \infty)$, the confidence set mapping $\widehat{C}_n$ from Alg. 2 satisfies*

$$1 - \alpha \le \mathrm{WC}(\widehat{C}_n, \mathcal{R}_{\hat{v}}, \delta; \widehat{Q}_{n_2}) \le 1 - \alpha + u_n,$$

*where $u_n \in [0, \alpha]$, and $u_n \xrightarrow{p} 0$ if Assumption A5 holds.*

**Proof** The lower bound is immediate by definition of $\widehat{C}_n$. For the upper bound, we have

$$\mathrm{WC}(\widehat{C}_n, \mathcal{R}_{\hat{v}}, \delta; \widehat{Q}_{n_2}) \le 1 - \alpha + \frac{1}{n_2 \delta}$$

whenever the scores $\{s_n(X_i, Y_i)\}_{i=n_1+1}^{n}$ are all distinct, which occurs eventually with high probability under Assumption A5. □

**Lemma C.4.** *Let Assumption A4 hold. Then as $n \to \infty$, the worst coverages under $\widehat{Q}_{n_2}$ and $Q_0$ satisfy the Glivenko-Cantelli result*

$$\sup_{q \in \mathbb{R}} \left| \mathrm{WC}(C^{(q, s_n)}, \mathcal{R}_{\hat{v}}, \delta; \widehat{Q}_{n_2}) - \mathrm{WC}(C^{(q, s_n)}, \mathcal{R}_{\hat{v}}, \delta; Q_0) \right| \xrightarrow{a.s.} 0.$$

**Proof** Let $\varepsilon > 0$ be arbitrary. Recalling equations (21) and (22) in the proof of Theorem 2, there exists a universal constant $c < \infty$ such that conditionally on $s_n$ and the first half of the validation set (hence $\hat{v}$), we have with probability at least $1 - \varepsilon$ over $\{(X_i, Y_i)\}_{i=n_1+1}^{n}$ that

$$\sup_{q \in \mathbb{R}} \left| \inf_{\substack{R \in \mathcal{R}_{\hat{v}} \\ Q_{n_2}(X \in R) \ge \delta}} Q_{n_2}\left(s_n(X,Y) \le q \mid X \in R\right) - \inf_{\substack{R \in \mathcal{R}_{\hat{v}} \\ Q_{n_2}(X \in R) \ge \delta}} Q_0\left(s_n(X,Y) \le q \mid X \in R\right) \right| \le c \sqrt{\frac{\log(n_2/\varepsilon)}{n_2 \delta}}$$

and

$$\sup_{q\in\mathbb{R},R\in\mathcal{R}_{\hat{v}}} |Q_{n_2}(X \in R) - Q_0(X \in R)| \le c\sqrt{\frac{\log(n_2/\varepsilon)}{n_2}}.$$

Setting $\delta_n^{\pm} := \delta \pm c\sqrt{\frac{\log(n_2/\varepsilon)}{n_2}}$, these two statements ensure that with probability $1-\varepsilon$ over $\{(X_i, Y_i)\}_{i=n_1+1}^n$, simultaneously for all $q \in \mathbb{R}$,

$$\begin{aligned}\mathrm{WC}(C^{(q,s_n)}, \mathcal{R}_{\hat{v}}, \delta_n^-; Q_0) - c\sqrt{\frac{\log(n_2/\varepsilon)}{n_2\delta}} &\le \mathrm{WC}(C^{(q,s_n)}, \mathcal{R}_{\hat{v}}, \delta; \widehat{Q}_{n_2}) \\ &\le \mathrm{WC}(C^{(q,s_n)}, \mathcal{R}_{\hat{v}}, \delta_n^+; Q_0) + c\sqrt{\frac{\log(n_2/\varepsilon)}{n_2\delta}}.\end{aligned}$$

To conclude, we claim that for all $q \in \mathbb{R}$, $v \in \mathcal{V}$, scores $s$, and $0 < \delta_0 < \delta_1 < 1$, we have

$$\mathrm{WC}(C^{(q,s)}, \mathcal{R}_v, \delta_1; Q_0) - \frac{\delta_1 - \delta_0}{\delta_1} \le \mathrm{WC}(C^{(q,s)}, \mathcal{R}_v, \delta_0; Q_0). \tag{33}$$

Temporarily deferring the proof of inequality (33), this shows in particular that for all $q \in \mathbb{R}$,

$$\mathrm{WC}(C^{(q,s_n)}, \mathcal{R}_{\hat{v}}, \delta_n^-; Q_0) \ge \mathrm{WC}(C^{(q,s_n)}, \mathcal{R}_{\hat{v}}, \delta; Q_0) - \frac{\delta_n^- - \delta}{\delta},$$

and that

$$\mathrm{WC}(C^{(q,s_n)}, \mathcal{R}_{\hat{v}}, \delta_n^+; Q_0) \le \mathrm{WC}(C^{(q,s_n)}, \mathcal{R}_{\hat{v}}, \delta; Q_0) + \frac{\delta_n^+ - \delta}{\delta}.$$

We thus have, conditionally on $s_n$ and $\hat{v}$, which are independent of the sample $\{(X_i, Y_i)\}_{i=n_1+1}^n$, that with probability at least $1-\varepsilon$

$$\sup_{q\in\mathbb{R}} |\mathrm{WC}(C^{(q,s_n)}, \mathcal{R}_{\hat{v}}, \delta; \widehat{Q}_{n_2}) - \mathrm{WC}(C^{(q,s_n)}, \mathcal{R}_{\hat{v}}, \delta; Q_0)| \le c\sqrt{\frac{\log(n_2/\varepsilon)}{n_2}} \left(\delta^{-1/2} + \delta^{-1}\right).$$

The Borel-Cantelli lemma then gives the almost sure convergence.

We return to demonstrate the claim (33). We have by definition that

$$\mathrm{WC}(C^{(q,s)}, \mathcal{R}_v, \delta_0; Q_0) = \min \left\{ \begin{array}{l} \inf_{R\in\mathcal{R}_v}\{Q_0(s(X,Y) \le q \mid X \in R) : \delta_1 \le Q_0(X \in R)\}, \\ \inf_{R\in\mathcal{R}_v}\{Q_0(s(X,Y) \le q \mid X \in R) : \delta_0 \le Q_0(X \in R) < \delta_1\} \end{array} \right\}.$$

If the topmost term achieves the minimum, the claim (33) is immediate, so we may instead assume that the bottom term achieves it. Assumption A4 ensures the existence of $a_1 \in \mathbb{R}$ such that $Q_0(X^T v \ge a_1) = \delta_1$ satisfying

$$\mathrm{WC}(C^{(q,s)}, \mathcal{R}_v, \delta_1; Q_0) \le Q_0(s(X,Y) \le q \mid X^T v \ge a_1)$$

as WC is an infimum over all such shifts. Then for any $a_0 \ge a_1$ such that $Q_0(X^T v \ge a_0) \ge \delta_0$, we in turn have

$$\begin{aligned}Q_0(s(X,Y) \le q \mid X^T v \ge a_1) &= \delta_1^{-1} Q_0(s(X,Y) \le q, X^T v \ge a_1) \\ &\le \delta_1^{-1} \left(Q_0(s(X,Y) \le q, X^T v \ge a_0) + Q_0(a_1 \le X^T v < a_0)\right) \\ &\le Q_0(s(X,Y) \le q \mid X^T v \ge a_0) + \frac{\delta_1 - \delta_0}{\delta_1},\end{aligned}$$

where we have used that $Q_0(X^T v \geq a_0) \leq \delta_1$. Taking an infimum over all such $a_0$ gives the statement (33) above. □

**Lemma C.5.** *Let Assumptions A3 and A4 hold. Then the score functions $s_n$ and $s$ offer uniformly close worst coverage in the sense that*

$$\sup_{q,v} \left\{ \left| \mathrm{WC}(C^{(q,s_n)}, \mathcal{R}_v, \delta; Q_0) - \mathrm{WC}(C^{(q,s)}, \mathcal{R}_v, \delta; Q_0) \right| \mid q \in \mathbb{R}, v \in \mathcal{V} \right\} = o_P(1).$$

**Proof** We need to show

$$\sup_{q,v} \left| \inf_{a: Q_0(X \in R_{v,a}) \geq \delta} P_0(s_n \leq q \mid X \in R_{v,a}) - \inf_{a: Q_0(X \in R_{v,a}) \geq \delta} P_0(S \leq q \mid X \in R_{v,a}) \right| = o_P(1),$$

for which it is sufficient to prove that

$$\sup_{a} \left\{ |Q_0(s_n(X,Y) \leq q, X \in R_{v,a}) - Q_0(s(X,Y) \leq q, X \in R_{v,a})| \mid Q_0\left(X^T v \geq a\right) \geq \delta \right\} \xrightarrow{p} 0.$$

Fix $\varepsilon > 0$. Under Assumption A4, the distribution of $S$ is continuous, so that $q \mapsto P_0(S \leq q)$ is continuous, monotone, and has finite limits in $\pm\infty$, so that it is uniformly continuous. Thus, there exists $\eta = \eta(\varepsilon) > 0$ such that

$$\sup_{q \in \mathbb{R}} P_0(q < S \leq q + \eta) \leq \varepsilon.$$

Now, define

$$B_n \coloneqq \{(x,y) \in \mathcal{X} \times \mathcal{Y} \mid |s_n(x,y) - s(x,y)| \geq \eta\},$$

and observe that for all $q \in \mathbb{R}$, $v \in \mathcal{V}$ and $a \in \mathbb{R}$, we have

$$\begin{aligned} Q_0(s_n(X,Y) \leq q, X^T v \geq a) &\leq Q_0(B_n) + Q_0(s(X,Y) \leq q + \eta, X^T v \geq a) \\ &\leq Q_0(B_n) + Q_0(s(X,Y) \leq q + \eta, X^T v \geq a) \\ &\leq Q_0(B_n) + Q_0(s(X,Y) \leq q, X^T v \geq a) + \varepsilon, \end{aligned}$$

and similarly

$$Q_0(s(X,Y) \leq q, X^T v \geq a) \leq Q_0(B_n) + Q_0(s_n(X,Y) \leq q, X^T v \geq a) + \varepsilon.$$

These imply that

$$\begin{aligned} \sup_{a} \left\{ \left| Q_0(s_n(X,Y) \leq q, X^T v \geq a) - Q_0(s(X,Y) \leq q, X^T v \geq a) \right| \mid Q_0\left(X^T v \geq a\right) \geq \delta \right\} \\ \leq \varepsilon + Q_0(B_n), \end{aligned}$$

and we conclude using Markov's inequality and Assumption A3 that

$$Q_0(B_n) \leq \frac{\|s_n - s\|^2_{L^2(Q_0)}}{\eta^2} \xrightarrow{p} 0,$$

which gives the result. □

**Lemma C.6.** *Let Assumptions A3 and A4 hold. Then as $n_1 \to \infty$,*

$$\sup_q |\mathrm{WC}(C^{(q,s)}, \mathcal{R}_{\hat{v}}, \delta;\, Q_0) - \mathrm{WC}(C^{(q,s)}, \mathcal{R}_{v^\star}, \delta;\, Q_0)| = o_p(1).$$

**Proof** Fix $\varepsilon > 0$. By Lemma C.2, $(q, v) \mapsto \mathrm{WC}(C^{(q,s)}, \mathcal{R}_v, \delta;\, Q_0)$ is uniformly continuous, so there exists $\eta > 0$ such that for all $v, v' \in \mathcal{V}$ with $\|v - v'\|_2 \le \eta$,

$$\sup_q |\mathrm{WC}(C^{(q,s)}, \mathcal{R}_{v'}, \delta;\, Q_0) - \mathrm{WC}(C^{(q,s)}, \mathcal{R}_{v}, \delta;\, Q_0)| \le \varepsilon,$$

and thus

$$\mathbb{P}\left[\sup_q |\mathrm{WC}(C^{(q,s)}, \mathcal{R}_{\hat{v}}, \delta;\, Q_0) - \mathrm{WC}(C^{(q,s)}, \mathcal{R}_{v^\star}, \delta;\, Q_0)| > \varepsilon\right] \le \mathbb{P}\left[\|\hat{v} - v^\star\|_2 > \eta\right].$$

Assumption A3 that $\hat{v} \xrightarrow{p} v^\star$ then gives the lemma. □

### C.3.2 Finalizing the proof of Theorem 3

Lemma C.5 shows that $\widehat{C}_n = C^{(\hat{q}_\delta, s_n)}$ satisfies

$$\sup_{v \in \mathcal{V}} |\mathrm{WC}(\widehat{C}_n, \mathcal{R}_v, \delta;\, Q_0) - \mathrm{WC}(C^{(\hat{q}_\delta, s)}, \mathcal{R}_v, \delta;\, Q_0)| = o_p(1),$$

which implies

$$|\mathrm{WC}(\widehat{C}_n, \mathcal{R}, \delta;\, Q_0) - \mathrm{WC}(C^{(\hat{q}_\delta, s)}, \mathcal{R}, \delta;\, Q_0)| = o_p(1). \tag{34}$$

Combining Lemmas C.4, C.5 and C.6, we additionally see that

$$\begin{aligned}
\mathrm{WC}(C^{(\hat{q}_\delta, s)}, \mathcal{R}_{v^\star}, \delta;\, Q_0) &\overset{C.6}{=} \mathrm{WC}(C^{(\hat{q}_\delta, s)}, \mathcal{R}_{\hat{v}}, \delta;\, Q_0) + o_P(1) \\
&\overset{C.5}{=} \mathrm{WC}(C^{(\hat{q}_\delta, s_n)}, \mathcal{R}_{\hat{v}}, \delta;\, Q_0) + o_P(1) \\
&\overset{C.4}{=} \mathrm{WC}(C^{(\hat{q}_\delta, s_n)}, \mathcal{R}_{\hat{v}}, \delta;\, \widehat{Q}_{n_2}) + o_P(1).
\end{aligned}$$

As $\mathrm{WC}(C^{(\hat{q}_\delta, s_n)}, \mathcal{R}_{\hat{v}}, \delta;\, \widehat{Q}_{n_2}) = 1 - \alpha + u_n$ for some $u_n \ge 0$ by Lemma C.3, where $u_n \xrightarrow{p} 0$ under Assumption A5, we have

$$\mathrm{WC}(C^{(\hat{q}_\delta, s)}, \mathcal{R}_{v^\star}, \delta;\, Q_0) = 1 - \alpha + u_n + o_P(1). \tag{35}$$

With Lemma 3.2, Assumption A2 ensures that $\mathrm{WC}(C^{(\hat{q}_\delta, s)}, \mathcal{R}_{v^\star}, \delta;\, Q_0) = \mathrm{WC}(C^{(\hat{q}_\delta, s)}, \mathcal{R}, \delta;\, Q_0)$, so we can conclude that

$$\begin{aligned}
\mathrm{WC}(\widehat{C}_n, \mathcal{R}, \delta;\, Q_0) &\overset{(34)}{=} \mathrm{WC}(C^{(\hat{q}_\delta, s)}, \mathcal{R}, \delta;\, Q_0) + o_p(1) \\
&\overset{\text{Lem. } 3.2}{=} \mathrm{WC}(C^{(\hat{q}_\delta, s)}, \mathcal{R}_{v^\star}, \delta;\, Q_0) + o_p(1) \overset{(35)}{=} 1 - \alpha + u_n + o_p(1).
\end{aligned}$$