JAWS: Auditing Predictive Uncertainty Under Covariate Shift

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose **JAWS**, a series of wrapper methods for distribution-free uncertainty quantification tasks under covariate shift, centered on the core method **JAW**, the **JA**ckknife+ **W**eighted with data-dependent likelihood-ratio weights. JAWS also includes computationally efficient **A**pproximations of JAW using higher-order influence functions: **JAWA**. Theoretically, we show that JAW relaxes the jackknife+'s assumption of data exchangeability to achieve the same finite-sample coverage guarantee even under covariate shift. JAWA further approaches the JAW guarantee in the limit of either the sample size or the influence function order under mild assumptions. Moreover, we propose a general approach to repurposing any distribution-free uncertainty quantification method and its guarantees to the task of risk assessment: a task that estimates probability that the true label lies within a user-specified interval. We then propose **JAW-R** and **JAWA-R** as the repurposed versions of proposed methods for **R**isk assessment. Practically, JAWS outperform the state-of-the-art predictive inference baselines in a variety of biased real world data sets for both interval-generation and risk-assessment auditing tasks.

1 Introduction

Auditing the uncertainty under data shift Principled quantification of predictive uncertainty is crucial for enabling users to calibrate how much they should or should not trust a given prediction [Thiebes et al., 2021, Ghosh et al., 2021, Tomsett et al., 2020, Bhatt et al., 2021]. Uncertainty-based predictor auditing can be considered a type of uncertainty quantification performed *post-hoc*, for example by a regulator without detailed knowledge of a predictor's architecture and with limited resources [Schulam and Saria, 2019]. Data shift poses a major challenge to uncertainty quantification due to violation of the common assumption that the training and test data are independent and identically distributed (i.i.d.) [Ovadia et al., 2019, Ulmer et al., 2020, Zhou and Levine, 2021, Chan et al., 2020]. Therefore, it is essential to develop convenient tools for users or regulators to audit predictive uncertainty even when training data is biased.

Predictor auditing: interval generation In this work we distinguish between two types predictive uncertainty auditing. We describe a first type by the task of *interval generation*, which refers to a common goal in the distribution-free uncertainty quantification literature: to generate a predictive interval (or confidence interval) that covers the true label with at least some user-specified probability. For instance, an auditor may wish to generate confidence intervals around a prediction that contain the true label with at least, say 90% probability.

Predictor auditing: risk assessment While predictive interval generation has been a central focus of the distribution-free uncertainty quantification literature, in some applications the converse computation may be more actionable: estimating the probability that a user-specified interval around a prediction contains the true label. We describe this type of predictor auditing task as *risk assessment*

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

because a natural type of interval for a user to specify is one defined by an error threshold around a prediction, where the goal then becomes to estimate the probability that the true label is within the error threshold (or the "risk" that the label is outside the threshold). A concrete example of when such a risk-assessment audit may be more useful to a stakeholder than an interval-generation audit is the setting of chemical or radiation therapy dose prediction for cancer treatment, where administering the correct dosage within $5\% - 10\%$ is safety-critical (see Appendix A.1 for more details). As we will demonstrate, any interval generation method with a distribution-free coverage guarantee can be repurposed for the task of risk assessment with minor modifications (as we will see in Section 3.3).

Coverage We assume a standard regression setup with training data $\{(X_1, Y_1), ..., (X_n, Y_n)\}$ and a test point (X_{n+1}, Y_{n+1}) where $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ for all $i \in \{1, ..., n+1\}$. Also, we denote a predictor as $\widehat{\mu} = \mathcal{A}(\{(X_1, Y_1), ..., (X_n, Y_n)\})$, where $\mathcal{A}$ is the algorithm for fitting the model. For predictive or confidence interval (or region) $\widehat{C}^{\text{audit}}_{n,\alpha} : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}\}$, a *coverage guarantee* aims to give a lower bound to the probability that the interval covers a test point's true label:

$$\mathbb{P}\Big\{Y_{n+1} \in \widehat{C}^{\text{audit}}_{n,\alpha}(X_{n+1}) \geq 1 - \alpha\Big\} \tag{1}$$

The coverage guarantee provides the basis for both interval generation and risk assessment auditing, though it is important to note that in this work we focus on marginal rather than conditional coverage (see [Foygel Barber et al., 2021] for more details on this distinction). Standard conformal prediction methods [Vovk et al., 2005, Shafer and Vovk, 2008, Vovk, 2013] along with the jackknife+ and related methods [Barber et al., 2021], which we refer to together as "predictive inference" methods, provide a framework for generating predictive intervals with finite-sample guaranteed coverage.

Exchangeability Standard predictive inference methods including conformal prediction and the jackknife+ all rely on two crucial notions of *exchangeability*: data exchangeability, that is that the training and test data are all exchangeable (e.g., i.i.d.); and secondly that the model-fitting algorithm $\mathcal{A}$ treats the data symmetrically [Barber et al., 2022]. In common situations of dataset shift, however, the data exchangeability assumption is violated. Empirically, the coverage performance of standard conformal prediction methods can suffer under data shift[Tibshirani et al., 2019, Podkopaev and Ramdas, 2021]. In this work, we build on the jackknife+ method largely due to its beneficial compromise between other conformal prediction methods' statistical drawbacks (for split conformal) and computational limitations (for full conformal) [Barber et al., 2021]. However, jackknife+ coverage performance can still degrade in the presence of dataset shift, such as shown in Figure 1.

In this work, we develop JAWS, a series of wrapper methods for distribution-free uncertainty quantification under covariate shift. Table 1 provides a summary of the JAWs methods and a few of their key properties, which we will elaborate on in Section 3.

Our contributions can summarized as follows:

1. We develop JAW: a jackknife+ method with data-dependent likelihood-ratio weights for uncertainty quantification under covariate shift. We show that JAW achieves the same rigorous, finite-sample coverage guarantee as jackknife+ [Barber et al., 2021] while relaxing the data exchangeability assumption to allow for covariate shift.

2. We develop JAWA: a sequence of computationally efficient approximations to JAW that uses higher-order influence functions to avoid retraining. Under assumptions outlined in Giordano et al. [2019a] regarding the regularity of the data, Hessian of the objective (local strong convexity), and the existence and boundedness of higher order derivatives, we provide an asymptotic guarantee for the JAWA coverage in the limit of either the sample size or influence function order.

3. We propose a general approach to repurposing any distribution-free predictive inference method to the risk assessment task, with rigorous guarantees for the coverage probability estimation. Our approach applies to methods that assume exchangeable data and to methods like JAW and JAWA that allow for covariate shift—JAW-R and JAWA-R refer to the risk assessment versions.

4. We demonstrate superior empirical performance of JAWS over other distribution-free predictive inference baselines on a variety of benchmark datasets under covariate shift.

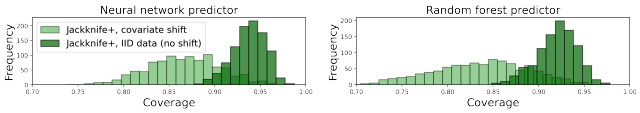


Figure 1: Jackknife+ loses coverage on the UCI airfoil dataset [Dua and Graff, 2017] due to covariate shift (creation of shift uses the exponential tilting strategy, more details are in Sec. 4).

Table 1: Summary of key properties for JAWS methods.

Method	Task	Guarantee (under covariate shift)		Avoids retraining
		Finite sample	Asymptotic	
JAW	Interval generation	✓	✓	✗
JAWA	Interval generation	✗	✓	✓
JAW-R	Risk assessment	✓	✓	✗
JAWA-R	Risk assessment	✗	✓	✓

2 Background and related Work

2.1 Standard conformal prediction

Conformal prediction initially arose in the 1990s and has since grown into a broad research field [Vovk et al., 2005, Shafer and Vovk, 2008, Vovk et al., 2009, Vovk, 2013, Burnaev and Vovk, 2014, Barber et al., 2022]. With the exchangeability assumptions in Section 1, standard conformal prediction methods apply a score function $\mathcal{S} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ (such as the residual score $\mathcal{S}(X_i, Y_i) = |Y_i - \widehat{\mu}(X_i)|$) to the data to generate a distribution of scores that determines the prediction interval. Conformal prediction uncertainty estimates are *distribution-free* [Lei and Wasserman, 2014, Lei et al., 2018] in that they apply to any data distribution with exchangeable datapoints.

Split conformal and full conformal are two main types of standard conformal prediction, each bears its own limitation [Vovk et al., 2005, Shafer and Vovk, 2008]. Split conformal generates scores on labeled holdout data and is computationally efficient due to not requiring retraining, but sample splitting to obtain the holdout set can reduce model accuracy Papadopoulos [2008], Lei et al. [2018], Vovk [2012]. On the other hand, full conformal prediction avoids the holdout set requirement, but at the expense of the heavy computational cost of retraining the model on every possible target value (or, in practice, on a fine grid of target values) [Ndiaye and Takeuchi, 2019, Zeni et al., 2020].

2.2 Covariate shift

Under the *covariate shift* assumption, the $Y|X$ distribution is assumed to be the same between training and test data but the marginal X distributions may change [Sugiyama et al., 2007, Shimodaira, 2000]:

$$(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P_X \times P_{Y|X}, i = 1, ..., n; \qquad (X_{n+1}, Y_{n+1}) \sim \tilde{P}_X \times P_{Y|X}, \text{ independently} \tag{2}$$

This is a strong but common assumption for many dataset shift problems. Rich literature exist in this domain—see Appendix A.2 for more details. Uncertainty quantification is relatively less explored under covariate shift, though recent work [Ovadia et al., 2019, Zhou and Levine, 2021, Chan et al., 2020] emphasizes its importance, especially in deep learning.

2.3 Conformal prediction under covariate shift and beyond exchangeability

Tibshirani et al. [2019] develop the idea of *weighted exchangeability* for adapting conformal prediction to the covariate shift setting. Random variables $V_1, ..., V_n$ are weighted exchangeable with weight functions $w_1, ..., w_n$ if their joint density f can be factorized as $f(v_1, ..., v_n) = \prod_{i=1}^n w_i(v_i) \cdot g(v_1, ..., v_n)$, where g is independent of ordering on its inputs. For covariate shift as in (2), the data $\{(X_i, Y_i)\}$ are weighted exchangeable with weights given by the likelihood ratio $w(X_i) = \mathrm{d}\tilde{P}_X(X_i)/\mathrm{d}P_X(X_i)$ [Tibshirani et al., 2019]

If $\{V_i\}$ represents a set of scores for standard conformal prediction, then we can represent the empirical distribution of $\{V_i\}$ as $\frac{1}{n+1}\sum_{i=1}^n \delta_{V_i} + \frac{1}{n+1}\delta_\infty$, where δ_{V_i} denotes a point mass at V_i.

By extension, weighted conformal prediction uses the *weighted* empirical distribution defined as $\sum_{i=1}^{n} p_i^w(x)\delta_{V_i} + p_{n+1}^w(x)\delta_\infty$, with weights given by

$$p_i^w(x) = \frac{w(X_i)}{\sum_{j=1}^{n} w(X_j) + w(x)}, i = 1, ..., n, \qquad \text{and} \qquad p_{n+1}^w(x) = \frac{w(x)}{\sum_{j=1}^{n} w(X_j) + w(x)} \tag{3}$$

where $p_i^w(x)$ can be thought of as a normalized likelihood ratio weight for each $i \in \{1, ..., n+1\}$. Corollary 1 in [Tibshirani et al., 2019] provides the coverage guarantee of weighted conformal prediction that takes the form of 1 but relaxes the exchangeable data assumption to allow for covariate shift. However, weighted split and weighted full conformal inherit the same statistical and computational limitations, respectively, from their standard (exchangeable) variants.

The recent work of [Barber et al., 2022] provides a novel extension of conformal prediction and the jackknife+ to unknown violations of the exchangeability assumption, including "nonexchangeable jackknife+" defined with fixed weights. The key difference between the nonexchangeable jackknife in Barber et al. [2022] and our proposed JAW method is that Barber et al. [2022] use fixed weights to compensate for unknown exchangeability violations (not limited to covariate shift) but at the expense of a bounded but generally nonzero "coverage gap" (drop in guaranteed coverage relative to if the data were exchangeable), whereas our JAW method with data-dependent weights assumes covariate shift but does not suffer from any similar coverage gap. See Appendix A.3 for more details.

2.4 Jackknife+

The jackknife+ [Barber et al., 2021] is an alternative to split and full conformal prediction that offers a compromise between those methods' respective statistical and computational drawbacks at the cost of a slightly weaker coverage guarantee. It leverages leave-one-out models like in cross validation [Butler and Rothman, 1980, Efron and Gong, 1983, Geisser, 1975, Stone, 1974] and is closely related to cross conformal prediction [Vovk et al., 2018]. The jackknife+ predictive interval can most easily be understood as a modification to a predictive interval based on the classic jackknife resampling method [Miller, 1974, Steinberger and Leeb, 2018, 2016] that it is inspired by.

Define the level α and $1-\alpha$ quantiles on the empirical distribution of some values $v_1, ..., v_n$ as $\widehat{q}^-_{n,\alpha}\{v_i\}$ and $\widehat{q}^+_{n,\alpha}\{v_i\}$ respectively. Then, letting $R_i^{LOO} = |Y_i - \widehat{\mu}_{-i}(X_i)|$ denote the leave-one-out residual for sample (X_i, Y_i), the prediction interval from the classic jackknife is defined as

$$\widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1}) = \left[\widehat{q}^-_{n,\alpha}\{\widehat{\mu}(X_{n+1}) - R_i^{LOO}\}, \widehat{q}^+_{n,\alpha}\{\widehat{\mu}(X_{n+1}) + R_i^{LOO}\}\right]. \tag{4}$$

In contrast, we obtain the jackknife+ predictive interval in Barber et al. [2021] by replacing the full model prediction $\widehat{\mu}(X_{n+1})$ in (4) with $\widehat{\mu}_{-i}(X_{n+1})$, where $\widehat{\mu}_{-i} = \mathcal{A}\big((X_1, Y_1), ..., (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), ..., (X_n, Y_n)\big)$, the model trained without the ith point:

$$\widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1}) = \left[\widehat{q}^-_{n,\alpha}\{\widehat{\mu}_{-i}(X_{n+1}) - R_i^{LOO}\}, \widehat{q}^+_{n,\alpha}\{\widehat{\mu}_{-i}(X_{n+1}) + R_i^{LOO}\}\right]. \tag{5}$$

[Barber et al., 2021] prove that the jackknife+ prediction interval satisfies

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})\} \geq 1 - 2\alpha \tag{6}$$

with the same exchangeability assumptions as in standard conformal prediction. Accordingly, the coverage guarantee (6) for the jackknife+ does not hold under dataset shift such as covariate shift.

2.5 Approximating leave-one-out models with higher-order influence functions

Influence functions (IFs) [Cook, 1977] have a long history in robust statistics for estimating how parameters would change due to reweighting the sample data, and in recent years IFs have become more widespread in machine learning for uses such as improving model interpretability Koh and Liang [2017]. Recent work has used IFs to approximate classic resampling uncertainty quantification methods including bootstrap [Schulam and Saria, 2019], jackknife, and leave-k-out cross validation [Giordano et al., 2019b] [Giordano et al., 2019a]. In each of these cases IFs enable approximation of the parameters that would be obtained by retraining the model on resampled data by instead estimating the effect of a corresponding reweighting. Alaa and Van Der Schaar [2020] proposed approximating the leave-one-out models required by the jackknife+ with higher-order IFs, but their work assumes exchangeable or i.i.d. train and test data.

Let $\hat{\theta}$ denote the fitted model parameters for predictor $\widehat{\mu}$ trained on the full training data. Given Assumptions 1-4 in Giordano et al. [2019a]—which require that $\hat{\theta}$ is a local minimum of the objective function, that the objective is $k+1$ times continuously differentiable with bounded norms, and that the objective is strongly convex in the neighborhood of $\hat{\theta}$—then the k-th order leave-one-out IF refers to the k-th order directional derivative of the model parameters $\hat{\theta}$ with respect to the data weights, in the direction of the leave-one-out change in weights (See A.4 for more details). With the kth order leave-one-out IF for each $k \in \{1, ..., K\}$, which we denote with condensed notation $\delta^k_{-i}\hat{\theta}$, we can construct a K-th order Taylor series approximation (or more specifically a Von Mises expansion, see Fernholz [2012]) to estimated the leave-one-out model parameters $\hat{\theta}_{-i}$:

$$\hat{\theta}^{\text{IF-}K}_{-i} := \hat{\theta} + \sum_{k=1}^{K} \frac{1}{k!} \delta^k_{-i}\hat{\theta} \tag{7}$$

In this work we implement the algorithm proposed by Giordano et al. [2019a] to compute higher-order IFs, a recursive procedure based on foreward-mode automatic differentiation [Maclaurin et al., 2015] for memory efficiency in computing higher-order directional derivatives. Our introduction of IFs is highly simplified—we refer to Appendix A.4 and to Giordano et al. [2019a] for more details.

3 Proposed approach and theoretical results

3.1 JAW: Jackknife+ weighted with data-dependent weights

We present **JAW**, the **JA**ckknife+ **W**eighted with data-dependent likelihood-ratio weights, defined by the following predictive interval:

$$\widehat{C}^{\text{JAW}}_{n,\alpha}(X_{n+1}) = \left[\widehat{q}^{-}_{n,\alpha}\{p^w_i(X_{n+1}) \cdot \delta_{\widehat{\mu}_{-i}(X_{n+1})-R^{LOO}_i}\},\ \widehat{q}^{+}_{n,\alpha}\{p^w_i(X_{n+1}) \cdot \delta_{\widehat{\mu}_{-i}(X_{n+1})+R^{LOO}_i}\}\right] \tag{8}$$

where $R^{LOO}_i = \left|\widehat{\mu}_{-i}(X_i) - Y_i\right|$, where $\widehat{q}^{+}_{n,\alpha}\{p^w_i(X_{n+1}) \cdot \delta_{\widehat{\mu}_{-i}(X_{n+1})+R^{LOO}_i}\}$ denotes the level $1-\alpha$ quantile of the empirical distribution $\sum_{i=1}^{n}\left[p^w_i(X_{n+1}) \cdot \delta_{\widehat{\mu}_{-i}(X_{n+1})+R^{LOO}_i}\right] + p^w_{n+1}(X_{n+1}) \cdot \delta_{\infty}$, and where $\widehat{q}^{-}_{n,\alpha}\{p^w_i(X_{n+1}) \cdot \delta_{\widehat{\mu}_{-i}(X_{n+1})-R^{LOO}_i}\}$ denotes the level α quantile of the empirical distribution $\sum_{i=1}^{n}\left[p^w_i(X_{n+1}) \cdot \delta_{\widehat{\mu}_{-i}(X_{n+1})-R^{LOO}_i}\right] + p^w_{n+1}(X_{n+1}) \cdot \delta_{-\infty}$. Here, $p^w_i(X)$ is the normalized likelihood ratio weights as defined in (3).

We show that $\widehat{C}^{\text{JAW}}_{n,\alpha}(X_{n+1})$ satisfies the same coverage guarantee as the jackknife+ except relaxing the data exchangeability assumption to allow for covariate shift, which we state formally as follows:

Theorem 1. *With covariate shift between the training and test data, the JAW interval in (8) satisfies*

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}^{\text{JAW}}_{n,\alpha}(X_{n+1})\} \geq 1 - 2\alpha \tag{9}$$

Remark. The results from Tibshirani et al. [2019] do not directly imply Theorem 1. Applying the approach from Tibshirani et al. [2019] to the jackknife+ entails treating $\widehat{\mu}_{-i}(X_{n+1}) \pm R^{LOO}_i$ as implicit nonconformity scores. Tibshirani et al. [2019] assume that the nonconformity scores for the training data $\{V_1, ..., V_n\}$ are weighted exchangeable with the nonconformity score for the test point V_{n+1}. But, observe that for $i \in \{1, ..., n\}$, $\widehat{\mu}_{-i}$ is trained on $n-1$ datapoints, whereas $\widehat{\mu}_{-(n+1)} = \widehat{\mu}$ is trained on n datapoints. Thus, no reweighting can make $\widehat{\mu}_{-i}$ equivalent to $\widehat{\mu}$ in distribution, and therefore $\widehat{\mu}_{-i}(X_{n+1}) + R^{LOO}_i$ and $\widehat{\mu}(X_{n+1}) + R^{LOO}_{n+1}$ are not weighted exchangeable.

Proof sketch: Our proof technique for Theorem 1 follows the approach of Barber et al. [2021] and adapts weighted conformal prediction as in Tibshirani et al. [2019] to the jackknife+. It can also be seen as generalizing the jackknife+ coverage guarantee proof in Barber et al. [2021] to the covariate shift setting with likelihood ratio weights. The outline is as follows:

1. *Setup*: Following Barber et al. [2021], we define a set of leave-*two*-out models $\{\tilde{\mu}_{-(i,j)}\}$. We then generalize the notion of "strange" points described in Barber et al. [2021] to covariate shift.
2. *Bounding the total normalized weight of strange points*: We establish deterministically that the total normalized weight of strange points cannot exceed 2α.

3. *Weighted exchangeability using the leave-two-out models*: Using the leave-two-out model construction, we leverage weighted exchangeability to show that the probability that a test point $n+1$ is strange is thus bounded by 2α.
4. *Connection to JAW*: Lastly, we show that the JAW interval can only fail to cover the test label value Y_{n+1} if $n+1$ is a strange point.

We note that JAW assumes access to oracle likelihood ratio weights, but that in practice this information is often not available. In such cases, the likelihood ratios can be estimated through an approach such as probabilistic classification, moment matching, or minimization of ϕ-divergences (for a review of likelihood ratio estimation approaches see Sugiyama et al. [2012]). We refer to JAW with **E**stimated likelihood ratio weights as JAW-E to acknowledge that the method's coverage performance will depend on the quality of the likelihood ratio estimates.

3.2 JAWA predictive interval asymptotic coverage guarantee

For a computationally efficient approximation to JAW that avoids retraining n leave-one-out models, we propose the JAWA sequence, which approximates the leave-one-out models required by JAW using higher-order influence functions. For each training point $i \in \{1, ..., n\}$, define the K-th order influence function approximation to the leave-one-out refit parameters $\hat{\theta}_{-i}$, obtained from Algorithm 4 in Giordano et al. [2019a], as given by Equation (7), and let $\widehat{\mu}_{-i}^{\text{IF-}K}$ be the model with with these approximated parameters $\hat{\theta}_{-i}^{\text{IF-}K}$ for each $i \in \{1, ..., n\}$. Then, the K-th order JAWA prediction interval (i.e., for JAWA-K) is given by

$$\widehat{C}_{n,\alpha}^{\text{JAWA-}K}(X_{n+1}) = \Big[\widehat{q}_{n,\alpha}^{-}\{p_i^w(X_{n+1}) \cdot \delta_{\widehat{\mu}_{-i}^{\text{IF-}K}(X_{n+1}) - R_i^{\text{IF-}K,LOO}}\}, \\ \widehat{q}_{n,\alpha}^{+}\{p_i^w(X_{n+1}) \cdot \delta_{\widehat{\mu}_{-i}^{\text{IF-}K}(X_{n+1}) + R_i^{\text{IF-}K,LOO}}\}\Big] \tag{10}$$

where $R_i^{\text{IF-}K,LOO} = \left|\widehat{\mu}_{-i}^{\text{IF-}K}(X_i) - Y_i\right|$, $p_i^w(X)$ as in (3), and quantiles defined analogously to JAW.

We now provide an asymptotic coverage guarantee for $\widehat{C}_{n,\alpha}^{\text{JAWA-}K}(X_{n+1})$ that holds either in the limit of the sample size or in the limit of the influence function order, under regularity conditions formally described in Giordano et al. [2019a]. These assumptions concern the regularity and continuity of the training data, local convexity of the objective (or that the Hessian is strongly positive definite), and the existence and boundedness of the objective's 1st through $K+1$th order directional derivatives.

Theorem 2. *Let Assumptions 1 - 4 and either Condition 2 or Condition 4 from Giordano et al. [2019a] hold uniformly for all n. Then, with covariate shift between the training and test data, in either the limit of the training sample size $n \to \infty$ or in the limit of the influence function order $K \to \infty$, the JAWA-K interval in (10) satisfies*

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}_{n,\alpha}^{\text{JAWA-}K}(X_{n+1})\} \geq 1 - 2\alpha \tag{11}$$

We leave the proof to Appendix C.2, but the result follows by combining Propositions 2 and 3 in Giordano et al. [2019a] with our Theorem 1 stated above.

3.3 Risk assessment under covariate shift

We now propose a general approach to adapting any predictive inference method with a coverage guarantee under covariate shift to the task of risk assessment, that is, estimating the probability that a true test point label Y_{n+1} lies within a given arbitrary, finite interval $I = [a_L, a_U]$. We also provide a similar result for uniform-weighted methods that assume exchangeable data, but for brevity we defer the exchangeable data approach and result to Appendix B.1.

The key idea is to leverage a predictive inference method's predictive *distribution* (whose quantiles determine the method's predictive interval) to find the largest predictive interval contained within I, and then to use the target coverage of this largest predictive sub-interval to estimate the coverage of I. Then, for a predictive inference method with data-dependent weights for generating intervals $\widehat{C}_{n,\alpha}^{\text{audit}}(X_{n+1}) = \left[q_{n,\alpha}^{-}\{p_i^w(x)\delta_{V_i^L}\},\ q_{n,\alpha}^{+}\{p_i^w(x)\delta_{V_i^U}\}\right]$ under covariate shift, define $\alpha_I^{\text{w-audit}}$ as

$$\alpha_I^{\text{w-audit}} = \min\left(\left\{\alpha' \ :\ a_L \leq\ q_{n,\alpha'}^{-}\{p_i^w(x)\delta_{V_i^L}\}\ ,\ q_{n,\alpha'}^{+}\{p_i^w(x)\delta_{V_i^U}\} \leq a_U\right\}\right). \tag{12}$$

Then, we estimate the coverage probability for interval I as the predictive inference method's target coverage for the interval corresponding to $\alpha_I^{\text{w-audit}}$:

$$\widehat{p}\{Y_{n+1} \in I\} \geq \begin{cases} 1 - \alpha_I^{\text{w-audit}} & \text{if } \alpha_I^{\text{w-audit}} \text{ exists} \\ 0 & \text{otherwise} \end{cases}. \tag{13}$$

Figure 2 illustrates the terms involved in computing $\alpha_I^{\text{w-audit}}$.

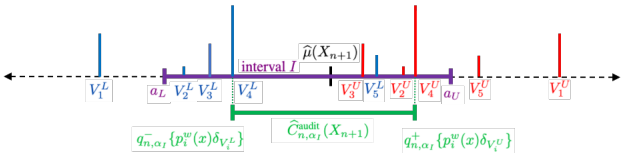


Figure 2: Illustration of terms involved in computing $\alpha_I^{\text{w-audit}}$. The interval $I = [a_L, a_U]$ is shown in violet, the values $\{V_i^L\}$ in blue, the values $\{V_i^U\}$ in red, and the interval $\widehat{C}^{\text{audit}}_{n,\alpha}(X_{n+1})$ in green. Each vertical line at a location V_i on the real line represents a point mass δ_{V_i} with height corresponding to the normalized likelihood ratio weight $p_i^w(x)$.

The following theorem gives the worst-case risk assessment guarantee under covariate shift (See Appendix B.1 for the analogous result for exchangeable data.)

Theorem 3. *Given a predictive inference method for generating intervals* $\widehat{C}^{audit}_{n,\alpha}(X_{n+1}) = \left[q^-_{n,\alpha}\{p_i^w(x)\delta_{V_i^L}\},\ q^+_{n,\alpha}\{p_i^w(x)\delta_{V_i^U}\}\right]$ *with coverage guarantee* $\mathbb{P}\{Y_{n+1} \in \widehat{C}^{audit}_{n,\alpha}(X_{n+1})\} \geq 1 - c_1\alpha - c_2$ *under covariate shift where* $c_1, c_2 > 0$ *are constants, suppose we are given some arbitrary interval* $I = [a_L, a_U] \subseteq \mathbb{R}$ *(assuming* $a_L \leq a_U$ *WLOG) whose coverage we are interested in estimating. Define* $\alpha_I^{w\text{-}audit}$ *as follows:*

$$\alpha_I^{w\text{-}audit} = \min\left(\left\{\alpha' \ :\ a_L \leq\ q^-_{n,\alpha'}\{p_i^w(x)\delta_{V_i^L}\}\ ,\ q^+_{n,\alpha'}\{p_i^w(x)\delta_{V_i^U}\} \leq a_U\right\}\right). \tag{14}$$

Then, I has coverage given by

$$\mathbb{P}\{Y_{n+1} \in I\} \geq \begin{cases} 1 - c_1\alpha_I^{w\text{-}audit} - c_2 & \text{if } \alpha_I^{w\text{-}audit} \text{ exists and } \alpha_I^{w\text{-}audit} < \frac{1-c_2}{c_1} \\ 0 & \text{otherwise} \end{cases}. \tag{15}$$

The proof for Theorem 3 is given in Appendix C.3.

4 Experiments[1]

4.1 Datasets and creation of covariate shift

We conduct experiments on five UCI datasets Dua and Graff [2017] with various numbers of features: airfoil self-noise (airfoil), red wine quality prediction (wine) [Cortez et al., 2009], wave energy converters (wave), superconductivity (superconduct) [Hamidieh, 2018], and communities and crime (communities) [Redmond and Baveja, 2002]. Table 2 shows their statistics.

We use exponential tilting to induce covariate shift on the test data, similar to as used in Tibshirani et al. [2019]. We first randomly split the data into two portions. The first portion was used for training and consisted of 200 samples in all experiments. The test distribution is then sampled from the remaining datapoints that are not used for training with probabilities proportional to exponential tilting weights, thus producing a biased test set with total number of test points equal to half the number of points not used for training. For the relatively lower dimensional airfoil and wine datasets, the weights took the form $w(x) = \exp(x^T\beta)$, while for the relatively higher dimensional datasets the weights took the form $w(x) = \exp(x_{\text{PCA}}^T\beta)$ where x_{PCA} is some PCA-based representation of the covariates data x. See Appendix D.1 for additional details on induced covariate shift.

[1] Additional experimental analysis and a link to code for all experiments and included in Appendix D and E

Table 2: Statistics for the UCI datasets. Only the first 2000 samples were used for the wave and superconductivity datasets (for wave, the first 2000 samples of Adelaide data).

Dataset	# of samples	# of features	label range
Airfoil	1503	5	[103.38, 140.987]
Red wine quality	1599	11	[3, 8]
Wave energy converters	2000	48	[1226969, 1449349]
Superconductivity	2000	81	[0.2, 136.0]
Communities and crime	1994	99	[0, 1]

4.2 Baselines

Baselines for comparison to JAW We compared JAW to the following baselines:

1. **Naive** estimates are based on training data residuals $|Y_i - \widehat{\mu}(X_i)|$, which suffers from overfitting.
2. **Jackknife** uses the classic Jackknife resampling as in (4).
3. **Jackknife+** follows (5), which replaces the prediction $\widehat{\mu}(X_{n+1})$ in jackknife with $\widehat{\mu}_{-i}(X_{n+1})$.
4. **Jackknife-mm** is a more conservative alternative to the jackknife+ method that guarantees coverage at the $1-\alpha$ level with exchangeable data, but usually with overly-wide intervals.

$$\widehat{C}^{\text{jackknife-mm}}_{n,\alpha}(X_{n+1}) = \Big[\min_{i=1,\dots,n} \widehat{\mu}_{-i}(X_{n+1}) - \widehat{q}^{+}_{n,\alpha}\{R_i^{LOO}\}, \max_{i=1,\dots,n} \widehat{\mu}_{-i}(X_{n+1}) + \widehat{q}^{+}_{n,\alpha}\{R_i^{LOO}\}\Big] \tag{16}$$

5. **CV+** (cross validation) similar to jackknife+ except splits data into k portions and replaces the $\widehat{\mu}_{-i}(X_{n+1})$ with $\widehat{\mu}_{-k}(X_{n+1})$, the model trained with the kth subset removed.
6. **Split** method follows the split conformal prediction and use half of the data for training and the other half of the data for generating the nonconformal scores.
7. **Weighted split** version of split conformal with likelihood ratio weights for covariate shift, as in Tibshirani et al. [2019].

Baselines for comparison to JAWA For influence function orders $K \in \{1, 2, 3\}$, we compared the proposed JAWA-K method with K-th order influence function approximations of the jackknife-based baselines that we used as comparisons to JAW—we thus refer to these approximations as IF-K jackknife, IF-K jackknife+, and IF-K jackknife-mm. Each baseline compared to JAWA-K is thus also approximated with the same K-th order leave-one-out influence function models.

4.3 Experimental results

Here we report experimental results on the interval generation task for both JAW and JAWA and on the risk assessment task for JAW, compared to baselines. Additional experimental details and supplementary experiments can be found in Appendix D, including experiments with estimated likelihood ratio weights in D.5, ablation study with shift magnitudes in D.6).

4.3.1 Interval generation results for JAW: coverage and interval width

Figure 4 compares JAW and its baselines, firstly regarding mean coverage and secondarily regarding median interval width, on all five UCI datasets for both neural network and random forest predictors, averaged over 200 experimental replicates. See Appendix D.2 for predictor function details. Meeting the target coverage level of $1-\alpha$ is the primary goal of the interval-generation audit task, so the first indication of a high quality method is whether the mean coverage reaches this target coverage level ($1-\alpha = 0.9$ noted by the dashed line in Figure 4). Then, for a method that meets or nearly meets the target coverage level, smaller interval widths are more informative than larger interval widths. In other words, an interval that achieves target coverage and small width is ideal, an interval with low coverage and small width is misleading, and any interval with overly large width can be uninformative in practice (e.g., $\mathbb{R}$ trivially has coverage of 1.0 but conveys no information). Figure 3 also demonstrates an example histogram comparison between JAW and jackknife+.

As seen in Figure 4, the JAW predictive interval coverage is above the target level of 0.9 across all datasets, for both random forest and neural network $\widehat{\mu}$ functions. In each of these conditions, JAW has higher coverage than the jackknife+ and all other baselines aside from jackknife-mm and weighted split. JAW's interval widths are generally smaller and thus more informative than those of

jackknife-mm (which are often overly large, as noted in Barber et al. [2021]). Weighted split and JAW perform similarly on mean coverage and median interval width (both methods have coverage guarantees under covariate shift), but JAW avoids sample splitting and as a result has lower coverage variance than weighted split for all dataset and predictor conditions (see Appendix D.3), which suggests that JAW's predictive intervals are more reliable.

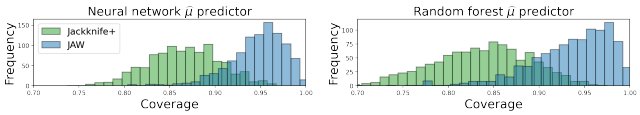
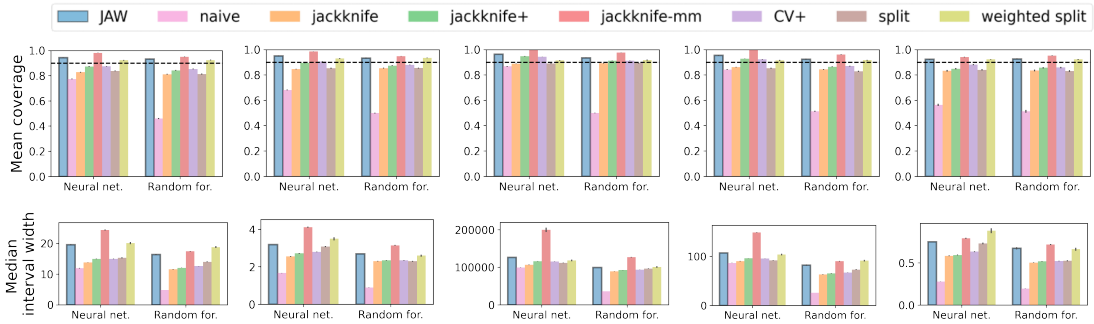


Figure 3: Jackknife+ versus JAW coverage under covariate shift for the airfoil dataset, when $\beta = (-1, 0, 0, 0, 1)$, for 1000 replicates. JAW maintains the high coverage under covariate shift.



(a) Airfoil (b) Wine (c) Wave (d) Superconduct (e) Communities

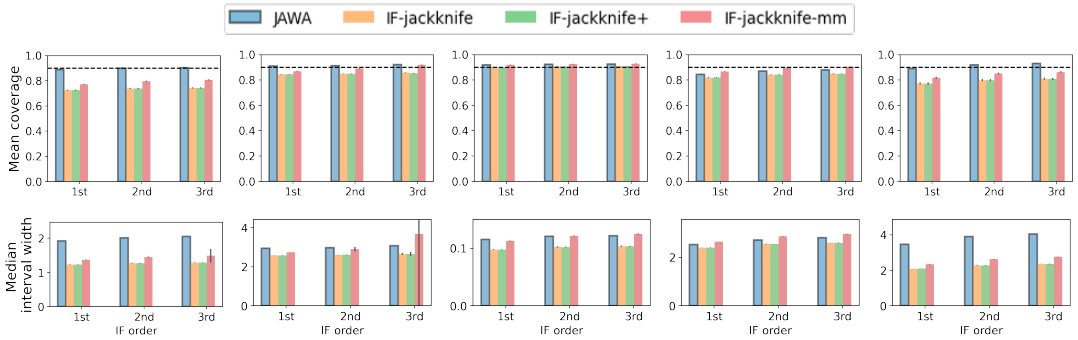
Figure 4: Mean coverage (first row) and median interval width (second row) for neural network and random forest predictors on UCI datasets. Dashed line is the target coverage level ($1 - \alpha$ = 0.9). Error bars show the standard error of 200 repeated experiments.

4.3.2 Interval generation results for JAWA: coverage and interval width

Figure 5 evaluates the coverage and interval width compared to its baselines for IF orders $K \in \{1, 2, 3\}$ with a neural network predictor (see Appendix D.2 for predictor details). As with the main JAW coverage and interval width experiments, coverage at the target level of $1 - \alpha = 0.9$ is the primary goal, while secondarily, smaller intervals are more informative for methods that meet or nearly meet the target coverage. For three of the five datasets (airfoil, wine, and communities) JAWA is the only method that consistently reaches or nearly reaches the target coverage level. JAWA and all the baselines perform well on the wave datasets, and in the superconduct dataset JAWA still outperforms approximations of jackknife and jackknife+ for all IF orders. Appendix D.7 provides an an example empirical comparison of JAWA and JAW runtimes, which demonstrates that JAWA can be orders of magnitude faster to compute than JAW.

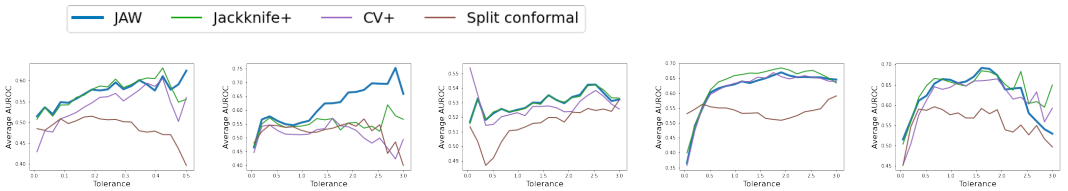
4.3.3 Risk assessment results for JAW: AUC

We now turn to the risk-assessment audit task where the goal is to evaluate a method's ability to estimate the coverage probability of a given interval, or to estimate the risk that the true label lies outside some error threshold τ around the original prediction $\widehat{\mu}(X_{n+1})$. Given prediction $\widehat{\mu}(X_{n+1})$ and error threshold τ, define the interval of interest as $I(\widehat{\mu}(X_{n+1}), \tau) = [\widehat{\mu}(X_{n+1}) - \tau,\ \widehat{\mu}(X_{n+1}) + \tau]$. Then, the goal is to estimate the probability that $\widehat{\mu}(X_{n+1})$ is correct, i.e., $|Y_{n+1} - \widehat{\mu}(X_{n+1})| \leq \tau \iff Y_{n+1} \in I(\widehat{\mu}(X_{n+1}), \tau)$; or incorrect, i.e., $|Y_{n+1} - \widehat{\mu}(X_{n+1})| > \tau \iff Y_{n+1} \notin I(\widehat{\mu}(X_{n+1}), \tau)$. We conduct these risk assessment experiments experiments on all five UCI datasets and for four predictive interval-generation methods that we repurpose to the risk assessment task (where "R" stands for the new task): JAW-R, jackknife+R, cross validation+R, and split conformal-R. Figure 6 reports the area under the receiver operating characteristic curve (AUROC) for values of τ selected with uniform spacing in the interval $[q^{-}_{n,0.05}\{R^{LOO}_i\},\ q^{+}_{n,0.05}\{R^{LOO}_i\}]$ (between the 0.05 and 0.95 quantiles of the leave-one-out residual prediction errors) for five repeated experiments. Better performing methods have higher overall AUROC values for all values τ.



(a) Airfoil (b) Wine (c) Wave (d) Superconduct (e) Communities

Figure 5: Mean coverage (first row) and median interval width (second row) for JAWA and baselines for influence function orders $K \in \{1, 2, 3\}$. Dashed line is the target coverage level ($1 - \alpha = 0.9$). JAWA achieves the most consistent coverage compared to the baselines across all influence function orders and across most datasets. Error bar shows the standard error of 200 repeated experiments.



(a) Airfoil (b) Wine (c) Wave (d) Superconduct (e) Communities

Figure 6: AUROC values for different tolerance levels τ across the three datasets for the random forest predictor, averaged across 5 experiment replicates, with random forest $\widehat{\mu}$. See Appendix D.4 for additional experiments with neural network predictor.

JAW achieves AUROC values comparable to jackknife+ and CV+ for most tolerance levels and datasets. In addition, JAW achieves higher AUROC values than all other methods in the wine dataset with the random forest predictor. These results seem to suggest that JAW may not offer a consistent benefit over the jackknife+ method in risk assessment, but that there may be a marginal benefit of JAW compared to jackknife+ for some datasets and conditions. We speculate that the seeming lack of substantial benefit of JAW over jackknife+ for the risk assessment may be due to the increased variance of the JAW predictive intervals under covariant shift (see Appendix D.6).

5 Conclusion and Future Work

In this paper, we develop JAWS, a series of wrapper methods for distribution-free predictive uncertainty auditing tasks when the data exchangeability assumption is violated due to covariate shift. We also propose a general approach to repurposing any distribution-free predictive inference method to the risk assessment task. We provide rigorous finite-sample guarantees for JAW and JAW-R on the interval generation and risk assessment tasks respectively, and analogous asymptotic guarantees for the computationally efficient JAWA and JAWA-R. We moreover demonstrate superior performance of the JAWS series on a variety of datasets. In practice, we do observe that JAW appears to have high variance in its coverage under covariate shift relative to the unweighted jackknife+ applied to exchangeable data. Additionally, we note that in extreme cases of covariate shift where the normalized likelihood ratio of the test point approaches or exceeds α, JAW and JAWA may produce overly large predictive intervals, though this is also a limitation of weighted conformal prediction [Tibshirani et al., 2019]. In the future, we are interesting in addressing the problems of reducing the variance in coverage and improving the sharpness of the predictive intervals.

References

Scott Thiebes, Sebastian Lins, and Ali Sunyaev. Trustworthy artificial intelligence. *Electronic Markets*, 31(2):447–464, 2021.

Soumya Ghosh, Q Vera Liao, Karthikeyan Natesan Ramamurthy, Jiri Navratil, Prasanna Sattigeri, Kush R Varshney, and Yunfeng Zhang. Uncertainty quantification 360: A holistic toolkit for quantifying and communicating the uncertainty of ai. *arXiv preprint arXiv:2106.01410*, 2021.

Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. Rapid trust calibration through interpretable and uncertainty-aware ai. *Patterns*, 1(4):100049, 2020.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, 2021.

Peter Schulam and Suchi Saria. Can you trust this prediction? auditing pointwise reliability after learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1022–1031. PMLR, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Dennis Ulmer, Lotta Meijerink, and Giovanni Cinà. Trust issues: Uncertainty estimation does not enable reliable ood detection on medical tabular data. In *Machine Learning for Health*, pages 341–354. PMLR, 2020.

Aurick Zhou and Sergey Levine. Amortized conditional normalized maximum likelihood: Reliable out of distribution uncertainty estimation. In *International Conference on Machine Learning*, pages 12803–12812. PMLR, 2021.

Alex Chan, Ahmed Alaa, Zhaozhi Qian, and Mihaela Van Der Schaar. Unlabelled data improves bayesian uncertainty calibration under covariate shift. In *International Conference on Machine Learning*, pages 1392–1402. PMLR, 2020.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

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

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Vladimir Vovk. Transductive conformal predictors. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 348–360. Springer, 2013.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2022.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pages 844–853. PMLR, 2021.

Ryan Giordano, Michael I Jordan, and Tamara Broderick. A higher-order swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*, 2019a.

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

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. On-line predictive linear regression. *The Annals of Statistics*, pages 1566–1590, 2009.

Evgeny Burnaev and Vladimir Vovk. Efficiency of conformalized ridge regression. In *Conference on Learning Theory*, pages 605–622. PMLR, 2014.

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

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Harris Papadopoulos. *Inductive conformal prediction: Theory and application to neural networks*. INTECH Open Access Publisher Rijeka, 2008.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

Eugene Ndiaye and Ichiro Takeuchi. Computing full conformal prediction set with approximate homotopy. *Advances in Neural Information Processing Systems*, 32, 2019.

Gianluca Zeni, Matteo Fontana, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *arXiv preprint arXiv:2005.07972*, 2020.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Ronald Butler and Edward D Rothman. Predictive intervals based on reuse of the sample. *Journal of the American Statistical Association*, 75(372):881–889, 1980.

Bradley Efron and Gail Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328, 1975.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alexander Gammerman. Cross-conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*, pages 37–51. PMLR, 2018.

Rupert G Miller. The jackknife-a review. *Biometrika*, 61(1):1–15, 1974.

Lukas Steinberger and Hannes Leeb. Conditional predictive inference for high-dimensional stable algorithms. *arXiv preprint arXiv:1809.01412*, 2018.

Lukas Steinberger and Hannes Leeb. Leave-one-out prediction intervals in linear regression models with many variables. *arXiv preprint arXiv:1602.05801*, 2016.

R Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

Ryan Giordano, William Stephenson, Runjing Liu, Michael Jordan, and Tamara Broderick. A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR, 2019b.

Ahmed Alaa and Mihaela Van Der Schaar. Discriminative jackknife: Quantifying uncertainty in deep learning via higher-order influence functions. In *International Conference on Machine Learning*, pages 165–174. PMLR, 2020.

Luisa Turrin Fernholz. *Von Mises calculus for statistical functionals*, volume 19. Springer Science & Business Media, 2012.

Dougal Maclaurin, David Duvenaud, and Ryan P Adams. Autograd: Effortless gradients in numpy. In *ICML 2015 AutoML workshop*, volume 238, 2015.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4): 547–553, 2009.

Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018.

Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

Mary Feng, Gilmer Valdes, Nayha Dixit, and Timothy D Solberg. Machine learning in radiation oncology: opportunities, requirements, and needs. *Frontiers in oncology*, 8:110, 2018.

Elizabeth Huynh, Ahmed Hosny, Christian Guthier, Danielle S Bitterman, Steven F Petit, Daphne A Haas-Kogan, Benjamin Kann, Hugo JWL Aerts, and Raymond H Mak. Artificial intelligence in radiation oncology. *Nature Reviews Clinical Oncology*, 17(12):771–781, 2020.

Saul N Weingart, Lulu Zhang, Megan Sweeney, and Michael Hassett. Chemotherapy medication errors. *The Lancet Oncology*, 19(4):e191–e199, 2018.

Marcel Van Herk. Errors and margins in radiotherapy. In *Seminars in radiation oncology*, volume 14, pages 52–64. Elsevier, 2004.

Howard Gurney. How to calculate the dose of chemotherapy. *British journal of cancer*, 86(8): 1297–1302, 2002.

Michael R Cohen, Roger W Anderson, Richard M Attilio, Laurence Green, Raymond J Muller, and Jane M Pruemer. Preventing medication errors in cancer chemotherapy. *American Journal of Health-System Pharmacy*, 53(7):737–746, 1996.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

Yaoliang Yu and Csaba Szepesvári. Analysis of kernel mean matching under covariate shift. *arXiv preprint arXiv:1206.4650*, 2012.

Kai Zhang, Vincent Zheng, Qiaojun Wang, James Kwok, Qiang Yang, and Ivan Marsic. Covariate shift in hilbert space: A solution via sorrogate kernels. In *International Conference on Machine Learning*, pages 388–395. PMLR, 2013.

Yin Zhao, Longjun Cai, et al. Reducing the covariate shift by mirror samples in cross domain alignment. *Advances in Neural Information Processing Systems*, 34, 2021.

Anqi Liu and Brian Ziebart. Robust classification under sample selection bias. *Advances in neural information processing systems*, 27, 2014.

Xiangli Chen, Mathew Monfort, Anqi Liu, and Brian D Ziebart. Robust covariate shift regression. In *Artificial Intelligence and Statistics*, pages 1270–1279. PMLR, 2016.

John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts. *Under review*, 2, 2019.

Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D Ziebart. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9419–9427, 2021.

Sashank Reddi, Barnabas Poczos, and Alex Smola. Doubly robust covariate shift correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. *arXiv preprint arXiv:2006.14651*, 2020.

Fengpei Li, Henry Lam, and Siddharth Prusty. Robust importance weighting for covariate shift. In *International Conference on Artificial Intelligence and Statistics*, pages 352–362. PMLR, 2020.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Appendix E.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We discuss the limitations of our method in the experiments (Section 4) and the conclusion (Section 5). We aim to resolve them in our future work.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work are motivated from the predictor auditing perspective, which aims to facilitate more convenient quality estimation of the predictive models for the users and regulators.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] We include a proof sketch of Theorem 1 in Section 3.1, and have the full proof of all our theoretical results in Appendix C.
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include an anonymous link to our code repository in the supplemental material, Appendix E.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4 and Appendix D.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 4 and Figure 5. We repeat our experiments 200 times and report the standard error as error bars.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We include the computational details in the supplemental materials, Appenidx E

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include an anonymous link to our code repository in the supplemental material.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]