JAW: Guaranteed Predictive Inference under Covariate Shift

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of uncertainty quantification with guaranteed coverage under distribution shift. In particular, we focus on covariate shift, where the mismatch between the training and testing distribution only occurs in the input marginal distribution, while the conditional label distribution stays the same. We propose **JAW**, the **JA**ckknife+ **W**eighted with likelihood ratio weights to estimate a valid, distribution-free predictive interval. Theoretically, we show that JAW achieves the same coverage guarantee as Jackknife+ even under covariate shift. Practically, we test our methods on a variety of biased real world data sets and under different magnitudes of covariate shift. JAW outperforms the state-of-the-art baselines in achieving a better trade-off between the coverage and confidence interval width.

1 Introduction

Auditing the uncertainty under data shift Principled quantification of predictive uncertainty is a key challenge in the development of machine learning for trustworthy AI, especially in high-stakes settings: appropriate uncertainty estimates enable 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]. Predictor auditing can be thought of as a type of uncertainty quantification that is performed *post-hoc*, for example by a regulator without detailed knowledge of a predictor's architecture and with limited resources, but potentially with access to the trained model, test data, and the data used for training [Schulam and Saria, 2019]. Data shift imposes significant challenges to the uncertainty auditing task due to violation of the common machine learning assumption of the training and test data being distributed identically and independently (IID) [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 estimate the model uncertainty even when training data is biased.

Coverage and exchangability Given a predictive model and a test data point, an uncertainty quantification-based auditing task might be to generate a predictive interval or confidence interval around the prediction that contains the true label with high probability. This is called the *coverage* guarantee. Standard conformal prediction methods[Vovk et al., 2005, Shafer and Vovk, 2008, Vovk, 2013] and Jackknife+ [Barber et al., 2021] methods all rely on the assumption of data exchangeability to achieve coverage guarantees, which is violated in oft-encountered situations of dataset shift. In particular, the framework of *conformal prediction* provides a frequentist approach to generating a predictive interval with guaranteed coverage when the data are known to be exchangeable (e.g., iid). The recently developed jackknife+ method provides an alternative to standard conformal prediction methods that does not require a labeled holdout set and has milder retraining requirements at the cost of a slightly weaker coverage guarantee [Barber et al., 2021]. Both of them inherit the exchangability requirement of the standard conformal prediction. Empirically, they also suffer from worse performance when dataset shift is present. For example, as shown in the Figure 1, Jackknife+ loss its coverage significantly under covariate shift.

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

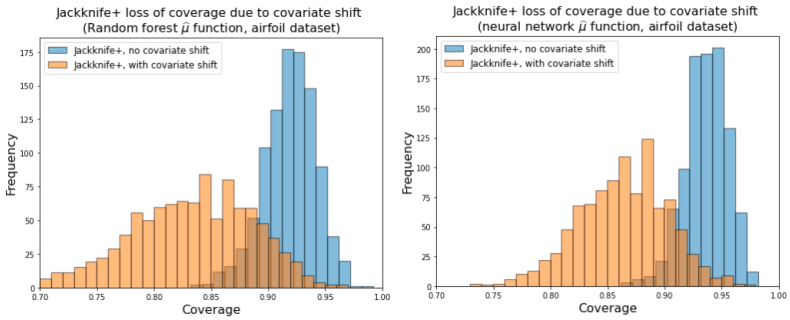


Figure 1: Jackknife+ coverage on the Airfoil dataset in the IID setting v.s. under covariate shift, using random forest and neural network predictors. **Left:** Jackknife+ loss of coverage for the random forest function; **Right:** Jackknife+ loss of coverage for the neural network function. The creation of the covariate shift is following the exponential tilting strategy. More details are in Sec. 4.

Covariate Shift *Covaraite shift* is a common assumption in dataset shift, which assumes the $Y|X$ distribution to be the same between training and test data but the marginal X distributions are allowed to change [Sugiyama et al., 2007, Shimodaira, 2000]. Recent work has extended the conformal prediction framework under covariate shift [Tibshirani et al., 2019], which is a weighted version of conformal prediction with coverage guarantees, with the weights given by likelihood ratios between the training and test distribution. However, it still suffers from the sample splitting of the split conformal prediction and the heavy computational burden of the full conformal prediction methods [Tibshirani et al., 2019, Zeni et al., 2020, Ndiaye and Takeuchi, 2019]. The recent work of [Barber et al., 2022] provided an extension of conformal prediction and the jackknife+ to cases beyond exchangeability that includes a weighted version of the jackknife+. However, it assumes that the weights are fixed and not data-dependent, which limits its applicability in more general dataset shift problems.

In this work, we develop a predictor auditing method with a coverage guarantee under covariate shift. **Our contributions can summarized as follows**:

1. We develop JAW: a weighted jackknife+ method for uncertainty quantification under covarite shift. To the best of our knowledge, this is the first in the literature.
2. We provide rigorous coverage guarantees for the proposed method and show that we achieve the same coverage guarantee as [Barber et al., 2021] without assuming the exchangability of the data.
3. We demonstrate superior empirical performance of the proposed method over the state-of-the-art baselines on a variety of benchmark datasets under covariate shift.

2 Preliminaries and Related Work

2.1 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 are allowed to change.

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

This is a strong but common assumption for many dataset shift problems. It is also closed related to data missingness and sample selection bias [Bickel et al., 2009]. Rich literature exist in this domain. The most prevalent method for correcting the shift is by applying important weights [Sugiyama et al., 2007, Shimodaira, 2000]. Density ratio estimation is then a key subproblem of covariate shift correction [Sugiyama et al., 2012]. Other methods dealing with covariate shift include matching the (kernel) representation between the two distributions [Gretton et al., 2009, Yu and Szepesvári, 2012, Zhang et al., 2013, Zhao et al., 2021] and robust optimization [Liu and Ziebart, 2014, Chen et al., 2016, Duchi et al., 2019, Rezaei et al., 2021]. However, uncertainty quantification is relatively less explored under covariate shift. Recent work [Ovadia et al., 2019, Zhou and Levine, 2021, Chan et al.,

2020] has emphasized the importance of a reliable uncertainty estimation under data shift, especially in the deep learning cases. Tibshirani et al. [2019] opens the possibility of conducting conformal prediction under covariate shift. We follow this line of work and focus on enabling the guaranteed predictive inference under covariate shift based on the jackknife+ method.

2.2 Standard conformal prediction

We assume a standard regression setup with training data $\{Z_1, ..., Z_n\} = \{(X_1, Y_1), ..., (X_n, Y_n)\}$, a test point $Z_{n+1} = (X_{n+1}, Y_{n+1})$, and a predictor $\hat{\mu} = \mathcal{A}(\{(X_1, Y_1), ..., (X_n, Y_n)\})$, where $\mathcal{A}$ is the algorithm for fitting the model. Conformal prediction aims to generate a predictive interval or confidence interval (or region) $\widehat{C}^{\text{audit}}_{n,\alpha} : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}\}$ around the prediction $\widehat{\mu}(X_{n+1})$ that contains the true label Y_{n+1} with high probability: a small predictive interval corresponds to high confidence in the original prediction, while a large interval corresponds to low confidence. A central requirement for the predictive interval is the *coverage*: for a target coverage rate $1 - \alpha \in (0, 1)$, we would like the true label to reside in the confidence interval with high probability, $\widehat{C}^{\text{audit}}_{n,\alpha}$ to satisfy

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

with flexible assumptions about the distribution of the training data and the test data $(X_1, Y_1), ..., (X_{n+1}, Y_{n+1})$.

Since its initial development in the 1990s, the study of conformal prediction for producing predictive regions (in regression) or predictive sets (in classification) with guaranteed coverage has grown into a broader research field [Vovk et al., 2005], [Shafer and Vovk, 2008], [Vovk et al., 2009], [Vovk, 2013], [Burnaev and Vovk, 2014]. Traditional conformal prediction crucially relies on the notion of "exchangeability" of the data in two ways:

- The data $Z_i = (X_i, Y_i)$ are assumed to be exchangeable (for example, iid)
- The algorithm $\mathcal{A}$ that maps a data set to a fitte model $\widehat{\mu}$ must treat the data symmetrically

Given these assumptions, conformal prediction methods use a score function $\mathcal{S} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ on the data, such as the residual score $\mathcal{S}(X_i, Y_i) = |Y_i - \widehat{\mu}(X_i)|$ to generate a distribution of "nonconformity" scores (or "conformity" scores) from which predictive regions can be derived. This kinds of estimates are also called distribution-free uncertainty estimates [Lei and Wasserman, 2014, Lei et al., 2018] to emphasize their applicability irrespective of distribution family, though this is not to be confused with independence from distribution *shift*. We consider the distribution-free uncertainty quantification under the covariate shift setting when the standard exchangability condition for standard conformal prediction is violated.

2.3 Weighted conformal prediction

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 called 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, that is, $g(v_{\sigma(1)}, ..., v_{\sigma(n)}) = g(v_1, ..., v_n)$ for any permutation σ of $1, ..., n$. In the case of covariate shift as in (1), Tibshirani et al. [2019] notes that if the training distribution P_X and the test distribution $\tilde{P}_X$ are known, then the data $\{Z_i\}$ are weighted exchangeable with weights given by the likelihood ratio $w(X_i) = \mathrm{d}\tilde{P}_X(X_i)/\mathrm{d}P_X(X_i)$.

Note that in standard conformal prediction, if $\{V_i\}$ represent a set of nonconformity scores, then the empirical distribution of $\{V_i\}$ can equivalently be represented as a sum of uniform weighted point masses on the augmented real line, that is $\frac{1}{n+1}\sum_{i=1}^{n} \delta_{V_i} + \frac{1}{n+1}\delta_\infty$ (where δ_{V_i} denotes a point mass at V_i). Standard conformal predictive intervals are then generated from the quantiles of this uniform-weighted empirical distribution. By extension, weighted conformal prediction (for split or full conformal variants) generates confidence intervals from the *weighted* empirical distribution

described as

$$\sum_{i=1}^{n} p_i^w(x)\delta_{V_i} + p_{n+1}^w(x)\delta_\infty$$

with weights defined as

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

, where the $p_i^w(x)$ can be thought of as normalized likelihood ratio weights (normalized with respect to the data $1, ..., n+1$).

Corollary 1 in [Tibshirani et al., 2019] provides the coverage guarantee of weighted conformal prediction: Assuming data from (2), that $\tilde{P}_X$ is absolutely continuous with respect to P_X, $w = \mathrm{d}\tilde{P}_X/\mathrm{d}P_X$, then for any score function $\mathcal{S}$, and any $\alpha \in (0, 1)$ and $x \in \mathbb{R}^d$, if

$$\widehat{C}_{n,\alpha}^w(x) = \Big\{ y \in \mathbb{R} : V_{n+1} \leq \text{Quantile}\Big(1 - \alpha; \sum_{i=1}^{n} p_i^w(x)\delta_{V_i} + p_{n+1}^w(x)\delta_\infty\Big)\Big\} \tag{4}$$

then $\widehat{C}_{n,\alpha}^w$ satisfies

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}_{n,\alpha}^w(X_{n+1})\} \geq 1 - \alpha \tag{5}$$

However, weighted standard conformal prediction methods inherit the same statistical and computational issues as nonweighted standard conformal prediction. In particular, conformal prediction methods largely fall into two main groups, each with its own pitfall: split conformal and full conformal prediction [Vovk et al., 2005, Shafer and Vovk, 2008]. Split conformal prediction generates nonconformity scores on a labeled holdout dataset and is computationally efficient in that it does not require model retraining, but the requirement of having a holdout set can be impractical or lead to a drop in model accuracy if obtained from sample splitting Papadopoulos [2008], Lei et al. [2018], Vovk [2012]. On the other hand, full conformal prediction avoids sample splitting, but at the expense of a heavy computational cost due to the requirement 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]. Outside of standard conformal prediction, the recent jackknife+ [Barber et al., 2021] achieves a tradeoff between these statistical and computational considerations, which motivates our current work that extends the jackknife+ coverage guarantee to hold under covariate shift.

Moreover, Barber et al. [2022] recently proposed a weighted version of the jackknife+, their theoretical and empirical results assume that the weights are fixed (for example to account for time series data). To the best of our knowledge, a weighted version of the jackknife+ with data-dependent, likelihood-ratio weights has not yet been investigated theoretically or empirically.

2.4 Jackknife+

The jackknife+ [Barber et al., 2021] is a recently developed alternative to split and full conformal prediction that offers a compromise between those methods' statistical and computational drawbacks at the cost of a slightly weaker coverage guarantee. It incorporates the leave-one-out methods for assessing the predictive accuracy 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 $1 - \alpha$ quantile on the empirical distribution of some values $v_1, ..., v_n$ as $\widehat{q}_{n,\alpha}^+\{v_i\}$, and define the level α quantile on the values as $\widehat{q}_{n,\alpha}^-\{v_i\} = -\widehat{q}_{n,\alpha}^+\{v_i\}$. Then, letting $R_i^{LOO} = |Y_i - \widehat{\mu}_{-i}(X_i)|$ denote the leave-one-out residual for sample $Z_i = (X_i, Y_i)$, a predictive interval generated from classic jackknife resampling is defined as

$$\begin{aligned}\widehat{C}_{n,\alpha}^{\text{jackknife}}(X_{n+1}) &= \widehat{\mu}(X_{n+1}) \pm \widehat{q}_{n,\alpha}^+\{R_i^{LOO}\} \\ &= \Big[\widehat{q}_{n,\alpha}^-\{\widehat{\mu}(X_{n+1}) - R_i^{LOO}\}, \widehat{q}_{n,\alpha}^+\{\widehat{\mu}(X_{n+1}) + R_i^{LOO}\}\Big]\end{aligned} \tag{6}$$

In contrast, the jackknife+ predictive interval as defined in [Barber et al., 2021] is obtained by replacing the full model prediction $\widehat{\mu}(X_{n+1})$ with $\widehat{\mu}_{-i}(X_{n+1})$ in (6), where $\widehat{\mu}_{-i} = \mathcal{A}\big((X_1, Y_1), ..., (X_{n+1}, Y_{n+1})\big)$, the model trained with the ith datapoint removed:

$$\widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1}) = \Big[\widehat{q}^{-}_{n,\alpha}\{\widehat{\mu}_{-i}(X_{n+1}) - R_i^{LOO}\}, \widehat{q}^{+}_{n,\alpha}\{\widehat{\mu}_{-i}(X_{n+1}) + R_i^{LOO}\}\Big] \tag{7}$$

[Barber et al., 2021] prove that, whereas no coverage guarantee is possible for the classic jackknife predictive interval (6), the jackknife+ prediction interval satisfies

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})\} \geq 1 - 2\alpha \tag{8}$$

with the same assumptions of exchangeable data and a symmetric algorithm as in standard conformal prediction. Accordingly, the coverage guarantee (8) for the jackknife+ does not hold under dataset shift such as covariate shift. In this paper, we apply importance weights from density ratios to Jackknife+ and achieve the same coverage guarantee under covariate shift.

3 Proposed method

3.1 JAW: Jackknife+ Weighted with data-dependent weights

We consider adapting the likelihood ratio weighting as in [Tibshirani et al., 2019] to the jackknife+ from [Barber et al., 2021]. While the coverage results of [Tibshirani et al., 2019] extend split and full conformal prediction to their corresponding weighted versions, it is not trivial to see that the same is true for the jackknife+. We first present the main result before discussing the implications.

We present the following weighted jackknife+ predictive interval with data-dependent, likelihood-ratio weights:

$$\widehat{C}^{\text{weighted jackknife+}}_{n,\alpha}(X_{n+1}) = \Big[\widehat{q}^{-}_{n,\alpha}\big\{p_i^w(X_{n+1}) \cdot \delta_{\widehat{\mu}_{-i}(X_{n+1}) - R_i^{LOO}}\big\}, \widehat{q}^{+}_{n,\alpha}\big\{p_i^w(X_{n+1}) \cdot \delta_{\widehat{\mu}_{-i}(X_{n+1}) + R_i^{LOO}}\big\}\Big] \tag{9}$$

where we use $\widehat{q}^{+}_{n,\alpha}\big\{p_i^w(X_{n+1}) \cdot \delta_{\widehat{\mu}_{-i}(X_{n+1}) + R_i^{LOO}}\big\}$ and $\widehat{q}^{-}_{n,\alpha}\big\{p_i^w(X_{n+1}) \cdot \delta_{\widehat{\mu}_{-i}(X_{n+1}) - R_i^{LOO}}\big\}$ to denote the level $1-\alpha$ and level α quantiles of the empirical distribution $\sum_{i=1}^{n}\big[p_i^w(X_{n+1}) \cdot \delta_{\widehat{\mu}_{-i}(X_{n+1}) + R_i^{LOO}} + p_{n+1}^w(X_{n+1}) \cdot \delta_{\infty}\big]$ respectively. Here, the $p_i^w(X)$ is the normalized likelihood ratio weights as defined in (3).

3.2 JAW predictive interval coverage

We show that $\widehat{q}^{+}_{n,\alpha}\big\{p_i^w(X_{n+1})$ satisfies a coverage guarantee identical to that of the unweighted jackknife+ except with covariate shift allowed, which we state formally in the following theorem:

Theorem 1. *The weighted jackknife+ interval in (9) satisfies*

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}^{\textit{weighted jackknife+}}_{n,\alpha}(X_{n+1})\} \geq 1 - 2\alpha \tag{10}$$

Remarks: We leave the proof to the Appendix **??**, but emphasize the implication of the theorem. Even though this guarantee is similar to Jackknife+ as how weighted conformal prediction framework [Tibshirani et al., 2019] achieves the similar guarantee as the standard conformal prediction. We highlight the significant difference in deriving this theorem for JAW, when comparing to weighted standard conformal prediction.

1. Non-existence of nonconformity scores: The standard conformal prediction depends on a score function $\mathcal{S}$ and corresponding nonconformity scores $\{V_i\}$, but the Jackknife+ has neither. For split and full conformal prediction, the results from [Tibshirani et al., 2019] directly apply because weighted exchangeability of the nonconformity scores $\{V_1, ..., V_n, V_{n+1}\}$ allow us to assume that V_{n+1} is as likely to lie outside of $q^{+}_{n,\alpha}\{p_i^w(x)\delta_{V_i}\}$ as any other nonconformity score V_i. However, the jackknife+ does not have nonconformity scores that we can directly apply this result to.
2. Non-existence of weighted exchangeability for the leave-one-out models: Even without nonconformity scores, the analogous approach to the jackknife+ would be to assume that $\{\widehat{\mu}_{-i}(X_{n+1}) + R_i^{LOO}\}$ is weighted exchangeable for $i \in \{1, ..., n+1\}$ (and same for $\{\widehat{\mu}_{-i}(X_{n+1}) + R_i^{LOO}\}$). However, 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: $\widehat{\mu}_{-i}$ and $\widehat{\mu}_{-(n+1)}$ differ beyond a mere reweighting, and thus $\widehat{\mu}_{-i}(X_{n+1}) + R_i^{LOO}$ and $\widehat{\mu}_{-(n+1)}(X_{n+1}) + R_{n+1}^{LOO}$ are not weighted exchangeable. As we will see in the proof sketch, extending weighted conformal prediction to the jackknife+ requires that we treat each $\widehat{\mu}_{-i}$ as the leave-*two*-out model $\widehat{\mu}_{-(i,n+1)}$.

Proof sketch: Our proof technique follows the approach of Barber et al. [2021], and it adapts weighted conformal prediction as described in Tibshirani et al. [2019] to the jackknife+ method. It can also be viewed as generalizing the 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 suppose access to the test point (X_{n+1}, Y_{n+1}) and define a set of leave-*two*-out models $\{\tilde{\mu}_{-(i,j)}\}$ (where $\tilde{\mu}$ rather than $\widehat{\mu}$ reminds us that the former may be fit with access to the test point). We then generalize the notion of "strange" or hard-to-predict points described in Barber et al. [2021] to the covariate shift setting.
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 are able to 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.

4 Experiments

4.1 Datasets and Creation of Covariate Shift

We conduct experiments on three UCI datasets Dua and Graff [2017] with a variety number of features. Table 1 shows their statistics.

Table 1: Dataset statistics for the UCI datasets

Dataset	# of samples	# of features	label range
Airfoil	1503	5	[103.38, 140.987]
Red Wine	1599	11	[3, 8]
Communities & Crime	1994	99	[0, 1]

For the creation of covariate shift, we use the same exponential tilting procedure as in [Tibshirani et al., 2019] to induce covariate shift on the test data. In particular, we first randomly split the data into two portions. The first portion is used for training. The test distribution is then sampled from the remaining datapoints that are not used for training with probabilities proportional to exponential tilting weights, such that the total number of test points is equal to half the number of points not used for training. The exponential tilting weights for the airfoil and wine datasets take the form of $w(x) = \exp(x^T\beta)$, with $\beta = (-\beta_1, 0, ..., 0, \beta_2)$ such that the weights depend on the first and last features of each dataset. We obtained the weights for the communities dataset using exponential tilting on the first two principal components representation of the data, that is $w(x) = \exp(x_{\text{PCA}}^T\beta)$ with $\beta = (-\beta_1, \beta_2)$, where x_{PCA} are the first two principal components of the data. We can tune β to generate different magnitudes of covariate shift. Figure 2 shows the distribution first and last features in the airfoil dataset before and after the exponential tilting is applied to induce covariate shift with parameter $\beta = (-1, 0, 0, 0, 1)$.

4.2 Baselines

We compare the proposed method with the following baselines:

1. **Naive** method uses the residuals on the training data, $|Y_i - \widehat{\mu}(X_i)|$, to estimate the typical prediction error on the new test point. This method usually suffers from overfitting, which may be even worse under covariate shift.
2. **Jackknife** uses the classic Jackknife resampling as in (6).

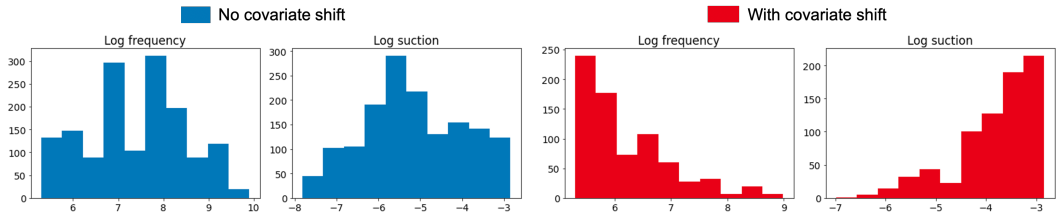
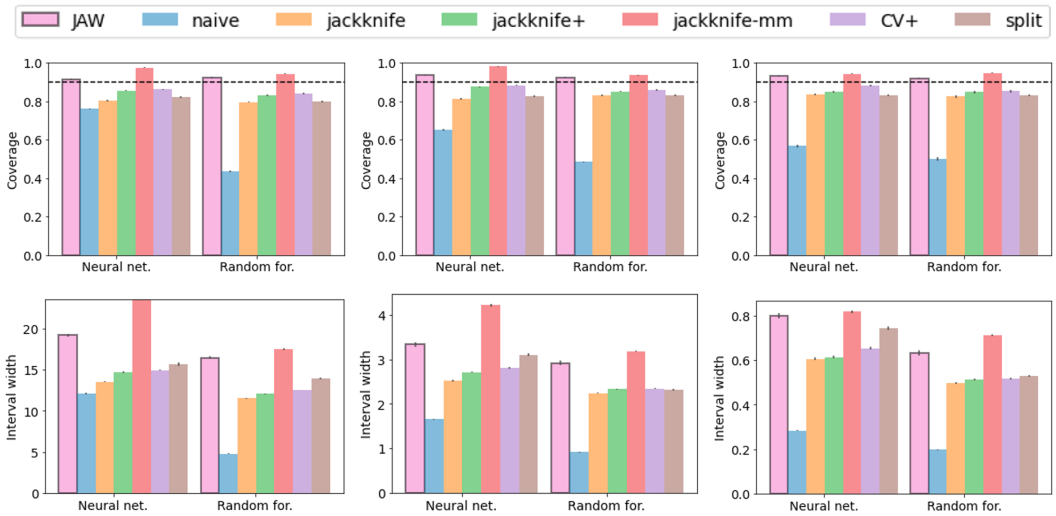


Figure 2: Distribution log frequency and log suction features of airfoil dataset before and after exponential tilting.

3. **Jackknife+** follows (7) that replaces the prediction $\widehat{\mu}(X_{n+1})$ with $\widehat{\mu}_{-i}(X_{n+1})$, which is the predictor trained using ith datapoint removed.
4. **CV+** (cross validation) splits the data into K portions and replaces the $\widehat{\mu}_{-i}(X_{n+1})$ with $\widehat{\mu}_{-k}(X_{n+1})$, which is the model trained with the kth subset removed. Jackknife+ can be viewed as the special case of CV+ when $K = n$.
5. **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.
6. **Jackknife-mm** method is a more conservative alternative to the Jackknife method. In the IID setting, it always achieves the target coverage rate $1 - \alpha$ but usually produces overly-wide confidence intervals.

$$\widehat{C}^{\text{jackknife-mm}}_{n,\alpha}(X_{n+1}) = \left[\min_{i=1,\ldots,n} \widehat{q}^{-}_{n,\alpha}\{\widehat{\mu}(X_{n+1}) - R_i^{LOO}\}, \max_{i=1,\ldots,n} \widehat{q}^{+}_{n,\alpha}\{\widehat{\mu}(X_{n+1}) + R_i^{LOO}\}\right] \quad (11)$$

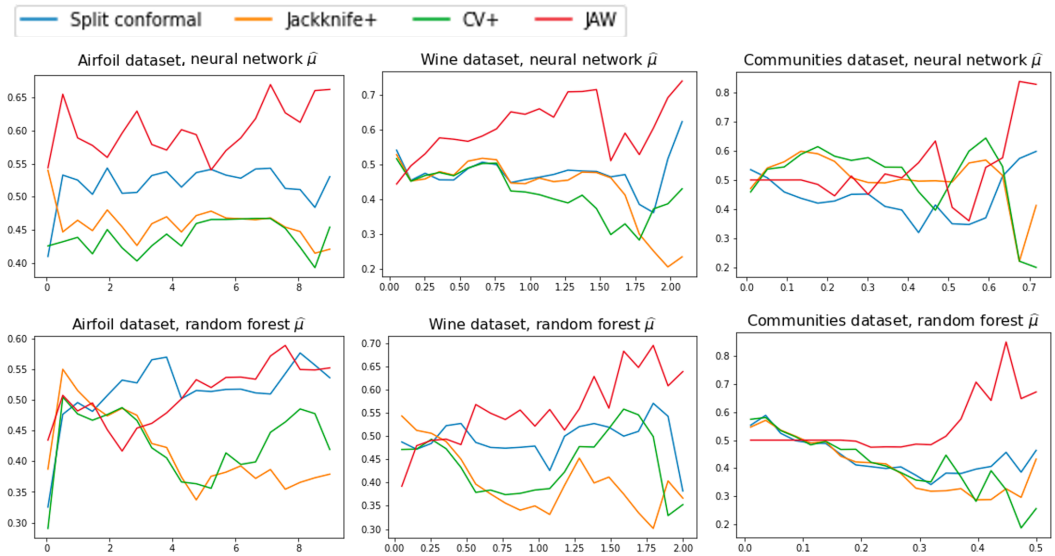


(a) Airfoil (b) Red Wine (c) Communities & Crime

Figure 3: Coverage (first row) and confidence interval width (second row) comparison for neural network and random forest predictors on the UCI datasets. Dashed line is the target coverage level $(1 - \alpha)$. JAW achieves the target coverage with much lower width of confidence interval than the other method that does the same (jackknife-mm). Error bar shows the standard error of 200 repeated experiments.

4.3 Experimental details

Models We standardize the training data size across datasets by randomly sampling 200 data points from the full dataset to use as the training data. The test points are then sampled from the remaining points such that the total number of test points is equal to half the number of points not used for training. We run our experiment using two different regression algorithms: random forests (scikit-learn RandomForestRegressor), and neural networks (scikit-learn MLPRegressor with LBFGS optimizer, logistic activation, and default parameters otherwise) as our predictor $\widehat{\mu}$.



(a) Airfoil (b) Red Wine (c) Communities & Crime

Figure 4: AUROC values for different tolerance levels τ across the three datasets, and with neural network random forest predictors, averaged across 5 experiment replicates. JAW achieves the highest AUROC values for most tolerance levels in most of the conditions.

Experiments Our first set of experiments focus on evaluating the empirical coverage and predictive interval width JAW compared to Jackknife+ and other baselines on each dataset with covariate shift. The best performing models should have high coverage with minimal interval width. In our second set of experiments, we turn from evaluating the coverage of a method's predictive interval to instead evaluating each method's ability to estimate the coverage of a given interval: to detect errors in an interval around the original prediction $\widehat{\mu}$. That is, for a given tolerance τ, the goal is to use each method's predictive distribution to estimate the probability that the prediction is correct ($|Y_{n+1} - \widehat{\mu}(X_{n+1})| \leq \tau$) or incorrect ($|Y_{n+1} - \widehat{\mu}(X_{n+1})| > \tau$). These probabilities are estimated by each method's empirical predictive distribution—details are discussed in the Appendix **??**. In our third set of experiments we compare the coverage performance of JAW to the jackknife+ for different levels of covariate shift.

4.4 Result analysis

Coverage v.s. Confidence Width Figure 3 demonstrates the performance comparison of JAW and the baselines. We observe that even though coverage is the key metric we focus on, the confidence interval width represents the sharpness of the prediction confidence. An overly wide confidence interval has high coverage but is less informative and useful in practice. Therefore, we compare both the coverage and the width of the confidence interval. JAW achieves the high coverage with "sharper" confidence interval. Figure 5 also demonstrates an example histogram comparison between JAW and jackknife+. JAW maintains the high coverage under covariate shift.

Error detection Figure 4 illustrates the performance of the JAW predictive distribution in detecting errors in given intervals, that is in estimating the probability of the event $|Y_{n+1} - \widehat{\mu}(X_{n+1})| \leq \tau$ for a given tolerance τ. The x-axis of each plot in Figure 4 is different values of τ in between the 0.05 and 0.95 quantiles of the leave-one-out residual prediction errors. The y-axis for each plot is the area under the ROC curve for the given threshold value. Better performing methods have higher overall AUROC values for all values τ. On four of six the dataset x predictor conditions, JAW has higher AUROC values for nearly all tolerance levels, and the JAW AUROC values are among the highest of the baselines for the other two conditions (airfoil dataset with random forest $\widehat{\mu}$ and communities dataset with neural network $\widehat{\mu}$).

Effects of different magnitudes of covariate shift We now present experiment results for different amounts of covariate shift in Figure 6. The extent of covariate shift can be controlled by modifying a

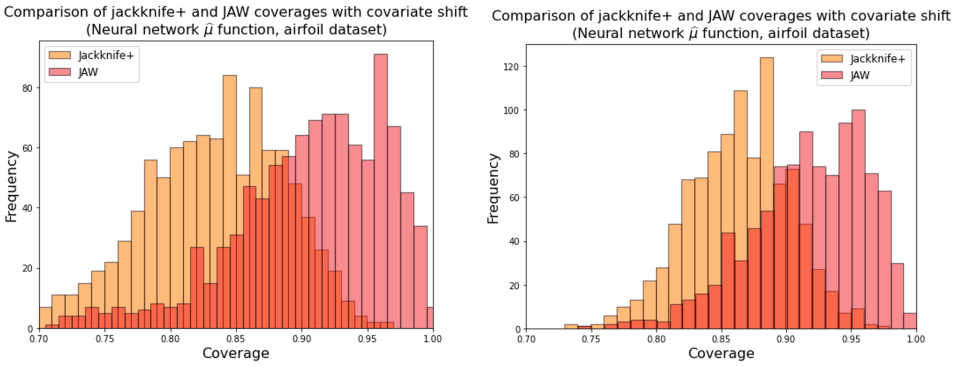


Figure 5: Comparison of the histogram of coverage on Airfoil under covariate shift, when $\beta = (-1, 0, 0, 0, 1)$. JAW still achieves high coverage while jackknife+ loses coverage significantly. This is 1000 replicates of the experiments.

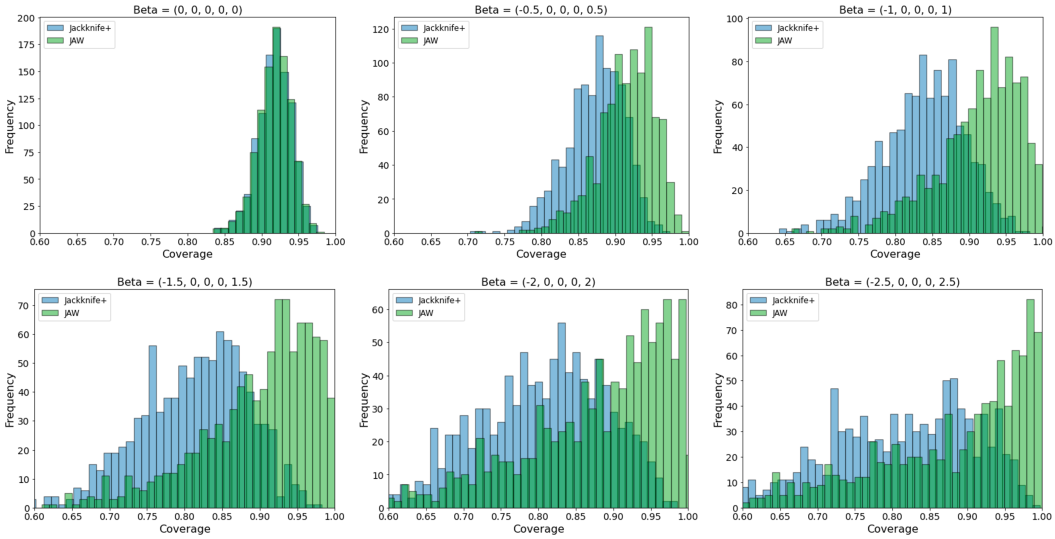


Figure 6: JAW performance compared to jackknife+ on the airfoil dataset with random forest $\widehat{\mu}$ function, under increasing magnitude of covariate shift, with 1000 replicates.

parameter in the exponential tilting weights so that weights are are more or less drastic. The bias level of 1 corresponds to the amount of bias in the previous experiments, and the bias of 0 corresponds to no bias (iid data). We can see that JAW is robust to different amounts of covariate shift, generating high coverage even under high level of shift. We also observe that the variance is higher when the level of shift is high. This is due to the high variance issue of important weighting methods that is well known [Reddi et al., 2015, Li et al., 2020] in the literature. We leave the variance reduction of our work to the future work.

5 Conclusion and Future Work

In this paper, we study the problem of uncertainty quantification under covariate shift. We develop JAW, a weighted jackknief+ method with guaranteed coverage, when the exchangebility assumption is absent. Our method achieves the target coverage with sharper confidence intervals than the baselines. We also show how our method can benefit the downstream safety-critical decision making. In practice, we do observe that JAW appears to have high variance in its coverage relative to the unweighted jackknife+. The retraining of JAW can also be computationally expensive. In the future, we are interested in developing more efficient variants of the method and further resolving the variance reduction problem.

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.

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.

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.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *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.

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

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

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

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

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.

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.

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.

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

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

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.

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 Section **??**.
- 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 4.4) and the conclusion. 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 in Section 3.2, and have the full proof of our theoretical results in the supplemental materials.
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.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.3.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Figure 3. We repeat our experiments 200 times and report the error bar.

(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.

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]