Robust Conformal Prediction Using Privileged Information

Shai Feldman
Department of Computer Science
Technion, Israel
shai.feldman@cs.technion.ac.il

Yaniv Romano
Departments of Electrical and Computer Engineering and of Computer Science
Technion, Israel
yromano@cs.technion.ac.il

Abstract

We develop a method to generate prediction sets with a guaranteed coverage rate that is robust to corruptions in the training data, such as missing or noisy variables. Our approach builds on conformal prediction, a powerful framework to construct prediction sets that are valid under the i.i.d assumption. Importantly, naively applying conformal prediction does not provide reliable predictions in this setting, due to the distribution shift induced by the corruptions. To account for the distribution shift, we assume access to privileged information (PI). The PI is formulated as additional features that explain the distribution shift, however, they are only available during training and absent at test time. We approach this problem by introducing a novel generalization of weighted conformal prediction and support our method with theoretical coverage guarantees. Empirical experiments on both real and synthetic datasets indicate that our approach achieves a valid coverage rate and constructs more informative predictions compared to existing methods, which are not supported by theoretical guarantees.

1 Introduction

1.1 Motivation

Uncertainty quantification plays a pivotal role in increasing the reliability of machine learning models. In this paper, we focus on situations where the training data is corrupted, e.g., due to missing variables or noisy labels. These corruptions are ubiquitous in high-stakes applications—such as diagnosing diseases, predicting financial outcomes, or personalizing treatment plans for patients—in which the data-collection process is complex, resource-intensive, or time-consuming [1–5].

One way to enhance the trustworthiness of data-driven predictions is to provide an uncertainty set containing the correct outcome at a user-specified coverage rate, e.g., 90%. Conformal prediction (CP) [6] is a general framework for constructing such reliable prediction sets, however, it assumes that the training and test data samples are drawn i.i.d from the same distribution. This assumption does not hold in the problem setting we consider in this work, in which the training data is a corrupted or a biased version of the ground truth. For instance, consider a medical application in which training data have missing or incorrect labels for some patients in a non-random pattern. Another example is a situation where we have missing feature values in the training data but, at test-time, we observe the full set of features. These examples illustrate common sources for a distribution shift between

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

the training and test data, which breaks the coverage guarantee of traditional CP techniques. In this work, we address this gap and propose a novel calibration technique, called *privileged conformal prediction* (PCP), which constructs provably valid uncertainty sets despite being employed with corrupted samples. Technically, we achieve this by utilizing privileged information—additional data available only during training time—to account for the distribution shift induced by the corruptions.

1.2 Problem setup

Suppose we are given n training samples $\{(X_i(M_i), Y_i(M_i), Z_i, M_i)\}_{i=1}^n$, where $X_i^{\text{obs}} = X_i(M_i) \in \mathcal{X}$ is the observed covariates, $Y_i^{\text{obs}} = Y_i(M_i) \in \mathcal{Y}$ is the observed response, $Z_i \in \mathcal{Z}$ is the privileged information (PI), and $M_i \in \{0, 1\}$ is the corruption indicator. Specifically, if $M_i = 1$ then either X_i^{obs} or Y_i^{obs} are corrupted, and if $M_i = 0$, then X_i^{obs} and Y_i^{obs} correctly reflect the ground truth. In our setup, we require that the privileged information Z_i explains the corruption occurences M_i. Formally, we assume that the clean data variables are independent of the corruption indicator given the privileged information, i.e., $(X(0), Y(0)) \perp\!\!\!\perp M \mid Z$.

At inference time, we aim to provide reliable predictions for the clean test response $Y^{\text{test}} = Y_{n+1}(0)$ given the clean version of the features: $X^{\text{test}} = X_{n+1}(0)$. That is, even though the observed $X^{\text{obs}}, Y^{\text{obs}}$ might be corrupted, the test $X^{\text{test}}, Y^{\text{test}}$ are always uncorrupted. Crucially, at test-time, we do not have access to the test privileged information $Z^{\text{test}} = Z_{n+1}$, nor the clean test label Y^{test}. Moreover, we assume that the PI Z is always clean and correctly reflects the ground truth. We now emphasize the importance of this problem setup by providing several examples.

Example 1 (Noisy response)**.** *Here, we refer to $Y_i(1)$ as a noisy version of the ground truth response $Y_i(0)$. For instance, $Y_i^{obs} = Y_i(M_i)$ could be a label obtained either by a non-expert annotator or an expert annotator, and Z_i could be information about the annotator, such as their level of expertise. In this case, $M_i = 0$ ($M_i = 1$) indicates that the annotator chose the correct (wrong) label $Y_i^{\text{obs}} = Y_i(0)$ ($Y_i^{\text{obs}} = Y_i(1)$). In contrast, the features are always uncorrupted: $X_i(0) = X_i(1)$. Notice that the test $Y^{\text{test}} = Y_{n+1}(0)$ always refers to the clean response. In this setup, since the PI, Z_i, is the information about the annotator, it is likely to explain the corruption appearances M_i. That is, it is sensible to believe that the assumption $(X(0), Y(0)) \perp\!\!\!\perp M \mid Z$ is approximately satisfied.*

Example 2 (Missing features)**.** *Here, $X_i(0)$ is the full clean feature vector, and $X_i(1)$ is a partial version of it, i.e., $X_{i,j}(1) =$ 'NA' for some entries j. For example, consider an application where participants are requested to fill a user experience (UX) questionnaire, in which the goal is to predict user engagement. This trial consists of expert participants, who tend to fully answer the questionnaire, resulting in a full $X_i(0)$, and non-experts, who tend to partially answer it, resulting in the incomplete $X_i(1)$. The PI Z_i could be the level of expertise of the participant. Also, the response is always uncorrupted: $Y_i(0) = Y_i(1)$. We remark that at test time, the full feature vector $X^{\text{test}} = X_{n+1}(0)$ is completely available. Since the PI Z_i is the information about the participant, it is likely to explain the missing indices M_i. Therefore, for this choice of PI, we have a good reason to believe that the conditional independence requirement, $X(0), Y(0) \perp\!\!\!\perp M \mid Z$, is approximately satisfied.*

Example 3 (Missing response)**.** *Consider a medical setup where patients are being selected for a costly diagnosis, such as an MRI scan. Here, $X_i(0) = X_i(1)$ is the more standard medical measurements of the i-th patient, such as age, gender, medical history, and disease-specific measurements. The PI Z_i is the information manually collected by the doctor to choose whether the patient should be examined by an MRI scan. This information is obtained through, e.g., a discussion of the doctor with the patient, or a physical examination, and could include, for instance, shortness of breath, swelling, blurred vision, etc. The response $Y_i(0)$ is the disease diagnosis obtained by the MRI scan, and $Y_i(1) =$ 'NA$'$. The missingness indicator M_i equals 0 if the doctor decides to conduct an MRI scan, and 1 otherwise. At test time, our goal is to assist the doctors in future decisions before examining the patients, and hence the test PI Z_{test} is unavailable. This task is relevant in situations where the number of available doctors is insufficient to examine all patients. Here, Z_i explains the missingness M_i, and M_i does not depend on X_i or Y_i given Z_i.*

With the above use cases in mind, our goal is to construct an uncertainty set $C(X^{\text{test}}) \subseteq \mathcal{Y}$ for the unknown clean test variable $Y^{\text{test}} = Y_{n+1}(0)$ given the clean features $X^{\text{test}} = X_{n+1}(0)$. Importantly, this uncertainty set should be statistically valid and satisfy the following coverage requirement:

$$\mathbb{P}(Y^{\text{test}} \in C(X^{\text{test}})) \geq 1 - \alpha, \tag{1}$$

where $1 - \alpha \in (0, 1)$ is a pre-specified coverage rate, e.g., 90%. This property is called *marginal coverage*, as the probability is taken over all samples $\{(X_i(0), X_i(1), Y_i(0), Y_i(1), Z_i, M_i)\}_{i=1}^{n+1}$, which are assumed to be drawn exchangeably (e.g., i.i.d.) from $P_{X(0),X(1),Y(0),Y(1),Z,M}$. The challenge in achieving (1) lies in the fact that there is a distribution shift between the training data $\{(X_i(M_i), Y_i(M_i))\}_{i=1}^n$ and the test data $(X_{n+1}(0), Y_{n+1}(0))$. Indeed, naively calibrating the model with the corrupted data may produce invalid uncertainty estimates [7]. Also, calibrating using only the clean data would result in biased predictions as the clean training samples are drawn from $P_{X(0),Y(0)|M=0}$, while the test samples are drawn from $P_{X(0),Y(0)}$.

To bypass this bias, we assume that the privileged information explains away the corruption appearances. Formally, we require that the corruption indicator is independent of the clean data conditional on the value of the privileged information, $(X(0), Y(0)) \perp\!\!\!\perp M \mid Z$. This assumption implies that our setting is a special case of covariate shift, with the covariates being the privileged information Z. Since the test PI $Z^{\text{test}} = Z_{n+1}$ is unknown at test time, conformal methods that account for covariate shift, such as *weighted conformal prediction* (WCP) [8] cannot be applied directly in the setup. In this paper, we re-formulate weighted conformal prediction and show how to construct uncertainty sets that satisfy the coverage requirement in (1) although Z^{test} is unavailable.

1.3 Our contribution

We introduce *privileged conformal prediction* (PCP)—a novel calibration scheme that effectively handles corrupted data, and constructs provably valid uncertainty sets in the sense of (1). Our key assumption is that the corruption indicator does not depend on the observed clean data given the privileged information, namely, $(Y(0), X(0)) \perp\!\!\!\perp M \mid Z$. This assumption implies that the privileged information explains away the corruption appearances. Building on WCP, we offer a specialized calibration scheme that carefully utilizes only the observed training privileged data $\{Z_i\}_{i=1}^n$ to attain a valid predictive inference at test time. To enhance statistical efficiency, we further adapt PCP for scarce data, building on leave-one-out arguments [9, 10]. Importantly, all methods we offer are supported by a theoretical valid coverage rate guarantee. Numerical experiments on both synthetic and real data show that naive conformal prediction techniques do not provide reliable uncertainty estimates, in contrast with the proposed PCP. To the best of our knowledge, this work is the first to propose a calibration scheme that generates statistically valid prediction sets, assuming that the privileged information explains away the corruption appearances. Software implementing the proposed method and reproducing our experiments is available at `https://github.com/Shai128/pcp`.

2 Background and related work

2.1 Conformal prediction

Conformal Prediction (CP) [6] is a powerful framework for constructing prediction sets that hold a marginal coverage rate guarantee, in the sense of (1). The general recipe to construct such prediction sets is as follows. First, split the data into a proper training set, indexed by $\mathcal{I}_1$, and a calibration set, indexed by $\mathcal{I}_2$. Then, fit a given learning model $\hat{f}$, e.g., a random forest or a neural network, on the training data. Next, evaluate the holdout prediction error of $\hat{f}$ by applying a non-conformity score function $\mathcal{S}(\cdot) \in \mathbb{R}$ to the calibration samples: $S_i = \mathcal{S}(X_i, Y_i; \hat{f}), \forall i \in \mathcal{I}_2$. Popular score functions include the absolute residual $\mathcal{S}(x, y; \hat{f}) = |\hat{f}(x) - y|$ in regression cases, where $\hat{f}$ is a mean estimator, or $1 - \hat{f}(x)_y$ in classification settings, where $\hat{f}(x)_y$ is the estimated probability of the y label given $X = x$. The latter is known as the homogeneous prediction sets (HPS) score [6]. Other score functions include the CQR score [11] for regression tasks and the APS score [12] for classification tasks. Armed with the non-conformity scores, the conformal procedure proceeds by computing the $(1 + 1/|\mathcal{I}_2|)(1 - \alpha)$-th empirical quantile of the calibration scores:

$$Q^{\text{CP}} = (1 + 1/|\mathcal{I}_2|)\,(1 - \alpha)\text{-th empirical quantile of the scores } \{S_i\}_{i \in \mathcal{I}_2}, \tag{2}$$

where $1 - \alpha$ is a user-specified coverage level. Lastly, the prediction set for the test point is given by

$$C^{\text{CP}}(X^{\text{test}}) = \{y : \mathcal{S}(X^{\text{test}}, y; \hat{f}) \leq Q^{\text{CP}}\}.$$

The above procedure is guaranteed to generate predictive sets with a valid marginal coverage (1) under the assumption that the calibration and test samples are exchangeable. We now turn to describe *weighted conformal prediction* (WCP) which is designed to handle exchangeability violations that arise from covariate shifts.

2.2 Weighted conformal prediction

Weighted Conformal Prediction (WCP) [8] extends the conformal prediction framework to handle covariate shifts. The key idea behind WCP is to weight the distribution of the calibration scores when taking their quantile in (2), so that the weighted scores 'look exchangeable' with the test non-conformity score. For the interest of space, we will not present the general form of WCP, and instead focus on the setup presented in this work, in which $(X(0), Y(0)) \perp\!\!\!\perp M \mid Z$. Under this assumption, the corruption indicator induces a covariate shift between the observed clean calibration samples and the test sample, which is explained by Z. That is, the clean calibration samples are drawn from $P_{X(0),Y(0)|M=0}$, while the test sample is drawn from $P_{X(0),Y(0)}$. Nevertheless, their distributions are equal conditionally on Z: $P_{X(0),Y(0)|Z=z,M=0} = P_{X(0),Y(0)|Z=z}$. With this in place, we follow the recipe of WCP and construct a prediction set as follows. First, we compute the ratio of likelihoods between the test and train data:

$$w(z) = \frac{dP_Z^{\text{test}}(z)}{dP_Z^{\text{train}}(z)} = \frac{f_Z^{\text{test}}(z)}{f_{Z|M=0}^{\text{test}}(z)} = \frac{f_Z^{\text{test}}(z)}{f_Z^{\text{test}}(z)\frac{\mathbb{P}(M=0|Z=z)}{\mathbb{P}(M=0)}} = \frac{\mathbb{P}(M=0)}{\mathbb{P}(M=0 \mid Z=z)}. \tag{3}$$

We define the set of uncorrupted calibration indexes as: $\mathcal{I}_2^{\text{uc}} = \{j :\in \mathcal{I}_2, M_j = 0\}$. The normalized weights are formulated as:

$$p_i(Z^{\text{test}}) = \frac{w(Z_i)}{\sum_{k\in\mathcal{I}_2^{\text{uc}}} w(Z_k) + w(Z^{\text{test}})}, \;\; p_{\text{test}}(Z^{\text{test}}) = \frac{w(Z^{\text{test}})}{\sum_{k\in\mathcal{I}_2^{\text{uc}}} w(Z_k) + w(Z^{\text{test}})} \tag{4}$$

Then, the calibration threshold for the test point is defined as the $1-\alpha$ empirical quantile of the weighted distribution of the scores:

$$Q^{\texttt{WCP}}(Z^{\text{test}}) := \text{Quantile}\left(1-\alpha; \sum_{i\in\mathcal{I}_2^{\text{uc}}} p_i(Z^{\text{test}})\delta_{S_i} + p_{\text{test}}(Z^{\text{test}})\delta_{\infty}\right), \tag{5}$$

and, the prediction set for the test sample is defined similarly to CP:

$$C^{\texttt{WCP}}(X^{\text{test}}, Z^{\text{test}}) = \{y : \mathcal{S}(X^{\text{test}}, y; \hat{f}) \leq Q^{\texttt{WCP}}(Z^{\text{test}})\}.$$

Remarkably, WCP produces uncertainty sets that achieve the desired marginal coverage rate (1) despite the induced covariate shift. Nonetheless, to implement this method, we must have access to Z^{test}, which is required to obtain $w(Z^{\text{test}})$. In our problem setup, however, we assume that Z^{test} is unavailable, and thereby WCP cannot be directly applied. This highlights the key challenge we aim to tackle in this paper, but before describing our method we first outline additional related work.

2.3 Additional related work

The concept of learning from privileged information was introduced by [13], which proposes techniques to leverage additional knowledge available during training to improve the prediction accuracy and accelerate algorithm convergence rate. This idea has been further explored to train models that are more robust to distribution shifts in the context of domain adaptation [14–16]. The method proposed in [17] utilizes PI to handle datasets containing weak labels and to obtain more accurate predictions. Furthermore, [18] combined model distillation with privileged information as a way to enhance learning from multiple models and data representations. The integration of PI with traditional conformal prediction to generate more informative uncertainty estimates was explored in [19, 20]. This line of work stands in striking contrast with our proposal, as we present a novel robust conformal calibration procedure based on PI. More broadly, there have been developed conformal methods that advance beyond the exchangeability assumption, such as WCP, among other contributions [7, 21–26]. However, none of these works utilize PI to ensure the validity of the constructed prediction sets.

3 Proposed method

In this section, we present our main contribution, the *privileged conformal prediction* (PCP) method. Since the setup we study in this paper has not been explored in the literature of conformal prediction, we start by suggesting a naive approach to achieve (1). Beyond serving as a baseline method for our PCP, this naive approach also reveals the challenges involved in constructing valid prediction sets when the calibration data is corrupted.

3.1 A naive approach: Two-Staged Conformal

Recall that `WCP` cannot be directly applied in our setup, since Z^{test} is unknown. To overcome this, the naive approach presented below consists of two stages: (i) estimate the unknown Z^{test} from the feature vector X^{test}, and (ii) employ `WCP` with the estimated privileged information.

While this approach is intuitive, the estimation of Z^{test} must be done in care: if the prediction of Z is incorrect, then `WCP` would not provide us the desired coverage guarantee. As a way out, instead of providing a point estimate, we will construct an interval $C^Z(X^{\text{test}})$ for Z^{test} given X^{test} using conformal prediction. This interval is guaranteed to contain the true PI Z^{test} with probability $1-\beta$, where β is a miscoverage rate of our choice, e.g., $\beta = 0.01$. Since we do not know which $z \in C^Z(X^{\text{test}})$ is the correct Z^{test}, we sweep over all possible elements $z \in C^Z(X^{\text{test}})$, compute their weights, $w(z)$, and take the largest weight:

$$w^{\text{conservative}}(X^{\text{test}}) := \max_{z \in C^Z(X^{\text{test}})} w(z). \tag{6}$$

The intuition behind taking the largest weight lies in Lemma 1, which states that the larger the test weight is, the larger the threshold $Q^{\texttt{WCP}}$ produced by `WCP`, which, in turn, increases the size of the prediction set. Armed with $w^{\text{conservative}}(X^{\text{test}})$, we can run `WCP` with a nominal coverage level $1-\alpha+\beta$ using the clean calibration samples and their weights $\{(X_i^{\text{obs}}, Y_i^{\text{obs}}, w(Z_i))\}_{i\in\mathcal{I}^{\text{uc}}}$, and the conservative test weight, $w^{\text{conservative}}(X^{\text{test}})$. We denote the weighted score quantile provided by `WCP` in (5) with this conservative test weight by $Q^{\texttt{WCP}}_{\text{conservative}}$. The prediction set is therefore defined as:

$$C^{\texttt{Two-Staged}}(X^{\text{test}}) := \left\{ y : \mathcal{S}(X^{\text{test}}, y; \hat{f}) \le Q^{\texttt{WCP}}_{\text{conservative}} \right\}. \tag{7}$$

An outline of this procedure is given in Algorithm 2 in Appendix B.1. The proposition below states that the uncertainty set generated by this naive approach is guaranteed to contain the test label Y^{test}, despite the presence of corrupted labels in the calibration set.

Proposition 1. *Suppose that $\{(X_i(0), X_i(1), Y_i(0), Y_i(1), Z_i, M_i)\}_{i=1}^{n+1}$ are exchangeable, the observed covariates are clean, i.e., $\forall i : X_i^{\text{obs}} = X_i(0) = X_i(1)$, the covariate shift assumption holds, i.e., $(X(0), Y(0)) \perp\!\!\!\perp M \mid Z$, and P_Z is absolutely continuous with respect to $P_{Z|M=0}$. Then, the prediction set $C^{\texttt{Two-Staged}}(X^{\text{test}})$ from* (7) *achieves a valid coverage rate:*

$$\mathbb{P}(Y^{\text{test}} \in C^{\texttt{Two-Staged}}(X^{\text{test}})) \ge 1-\alpha.$$

The proof is given in Appendix A.1. While this two-staged approach constructs valid prediction sets, it has several limitations. First, it requires predicting not only Y^{test} but also Z^{test}, and the prediction of the latter is anticipated to increase the uncertainty encapsulated in the resulting prediction set for Y^{test}. This algorithm also requires iterating over all $z \in C^Z(X^{\text{test}})$, which can be computationally expensive, especially when Z is continuous or multi-dimensional. In addition, and perhaps more importantly, the prediction set $C^Z(X^{\text{test}})$ for Z^{test} might contain unlikely, or off-support values of Z. This can lead to an extreme $w^{\text{conservative}}(X^{\text{test}})$, which, in turn, results in unnecessarily large prediction sets for Y^{test}. Moreover, this naive method assumes that the calibration features X_i^{obs} reflect the ground truth, i.e., $X_i^{\text{obs}} = X_i(0)$, and thus the coverage guarantee does not hold in situations where the features are missing or noisy. This discussion emphasizes the challenges in designing a calibration scheme that not only provides robust coverage guarantees but is also computationally and statistically efficient. In the next section, we present our main proposal which fully resolves all limitations of this naive approach.

3.2 Our main proposal: Privileged Conformal Prediction

In this section, we introduce our procedure to construct prediction sets with a valid coverage rate under the setting of corrupted samples. We begin similarly to `CP`, as described in Section 2.1, and split the data into a training set, $\mathcal{I}_1$, and a calibration set, $\mathcal{I}_2$. Next, we fit a predictive model $\hat{f}$ on the training data, and compute a non-conformity score for each calibration sample:

$$S_i = \mathcal{S}(X_i^{\text{obs}}, Y_i^{\text{obs}}; \hat{f}), \forall i \in \mathcal{I}_2.$$

Similarly to `WCP` and `Two-Staged` methods, we rely on the likelihood ratio of the training and test distributions, and compute the weight of the i-th sample: $w_i := \frac{\mathbb{P}(M=0)}{\mathbb{P}(M=0|Z=Z_i)}$. The problem in

WCP is that the scores threshold $Q^{\texttt{WCP}}(Z^{\text{test}})$ from (5) depends on Z^{test}. Here, we follow the intuition behind the two-staged baseline and propose an algorithm that provides a fixed threshold $Q^{\texttt{PCP}}$ that is not a function of Z^{test}. This threshold can be thought of as a conservative estimate, or an upper bound of $Q^{\texttt{WCP}}(Z^{\text{test}})$, which is based on the calibration data, and does not require Z^{test}. To achieve this, we consider every calibration point $i \in \mathcal{I}_2$ as a test point, and run WCP as a subroutine to obtain the i-th score threshold Q_i. The final PCP test score threshold, $Q^{\texttt{PCP}}$, is defined as the $(1-\beta)$-th empirical quantile of the calibration thresholds $\{Q_i\}_{i\in\mathcal{I}_2}$, where $\beta\in(0,\alpha)$ is a level of our choice, e.g., $\beta = 0.01$.

Formally, we consider the i-th sample as a test point and compute the normalized weight of the j-th sample:

$$p_j^i = \frac{w_j}{\sum_{k\in\mathcal{I}_2^{\text{uc}}} w_k + w_i}, \ \forall i, j \in \mathcal{I}_2.$$

Notice that p_j^i extends the WCP weights, p_j, from (4), since $p_j = p_j^{n+1}$. Now, we compute the i-th threshold Q_i by applying WCP using the uncorrupted calibration data:

$$Q_i := \text{Quantile}\left(1-\alpha+\beta; \sum_{j\in\mathcal{I}_2^{\text{uc}}} p_j^i \delta_{S_j} + p_i^i \delta_\infty\right), \tag{8}$$

Next, we extract from $\{Q_i\}_{i\in\mathcal{I}_2}$ a conservative estimate of $Q^{\texttt{WCP}}(Z^{\text{test}}) = Q_{n+1}$, denoted by $Q^{\texttt{PCP}}$:

$$Q^{\texttt{PCP}} := \text{Quantile}\left(1-\beta; \sum_{i\in\mathcal{I}_2} \frac{1}{|\mathcal{I}_2|+1}\delta_{Q_i} + \frac{1}{|\mathcal{I}_2|+1}\delta_\infty\right). \tag{9}$$

Finally, for a new input data X^{test}, we construct the prediction set for Y^{test} as follows:

$$C^{\texttt{PCP}}(X^{\text{test}}) = \left\{ y : \mathcal{S}(X^{\text{test}}, y, \hat{f}) \le Q^{\texttt{PCP}} \right\}.$$

For convenience, Algorithm 1 summarizes the above procedure and Algorithm 3 details a more efficient version of this procedure. We now show that the prediction sets constructed by PCP achieve a valid marginal coverage rate.

Theorem 1. *Suppose that $\{(X_i(0), X_i(1), Y_i(0), Y_i(1), Z_i, M_i)\}_{i=1}^{n+1}$ are exchangeable, $(X(0), Y(0)) \perp\!\!\!\perp M \mid Z$, and P_Z is absolutely continuous with respect to $P_{Z|M=0}$. Then, the prediction set $C^{PCP}(X^{\text{test}})$ constructed according to Algorithm 1 achieves the desired coverage rate:*

$$\mathbb{P}(Y^{\text{test}} \in C^{PCP}(X^{\text{test}})) \ge 1-\alpha.$$

The proof is given in Appendix A.2. We remark that while Theorem 1 requires that the PI satisfies the conditional independence assumption, i.e., $(X(0), Y(0)) \perp\!\!\!\perp M \mid Z$, in Appendix A.5 we relax this assumption and provide a lower bound for the coverage rate for settings where the conditional independence assumption is not exactly satisfied. We pause here to emphasize the significance of Theorem 1. The key challenge in proving this result lies in the fact that the $\{Q_i\}_{i\in\mathcal{I}_2\cup\{n+1\}}$ are not exchangeable. This is attributed to the fact that for every $i \in \mathcal{I}_2^{\text{uc}}$, the threshold Q_i is defined using its own score S_i, while the test Q_{n+1} does not rely on its corresponding score $S_{n+1} = \mathcal{S}(X^{\text{test}}, Y^{\text{test}}, \hat{f})$. As a side comment, if the thresholds were exchangeable, the proof was much simpler, as $Q^{\texttt{PCP}}$ would be greater than Q_{n+1} with probability $1-\beta$. In this case, $C^{\texttt{PCP}}$ includes $C^{\texttt{WCP}}$ at a high probability, meaning that it achieves the desired coverage rate. Due to the lack of exchangeability, the argument above is incorrect. Indeed, the validity of PCP does not follow directly from the guarantee of WCP, and it requires additional technical steps.

We now turn to discuss the role of β. First, we emphasize that Theorem 1 holds for any choice of $\beta \in (0, \alpha)$. Therefore, β only affects the sizes of the uncertainty sets. Intuitively, as $\beta \to \alpha$, a higher quantile of the weighted distribution of the scores is taken, and a lower quantile of the Q_i's is taken. Similarly, as $\beta \to 0$ a lower quantile of the weighted distribution of the scores is taken, and a higher quantile of the Q_i's is taken. An optimal β can be considered as the β that leads to the narrowest intervals. Such optimal β can be practically computed with a grid of values for β in $(0, \alpha)$, using a validation set. Nonetheless, in our experiments, we directly chose β that is close to 0. In Appendix E.5 we conduct an ablation study analyzing the effect of β on a synthetic dataset.

Lastly, we note that while the real ratios of likelihoods, w_i, are required to provide the validity guarantee in Theorem 1, PCP can be employed with estimates of w_i. These weights can be estimated in the following way. The first step is estimating the conditional corruption probability given Z, i.e., $\mathbb{P}(M = 0 \mid Z = z)$, using the training and validation sets with any off-the-shelf classifier. We remark that this classifier can be fit on unlabeled data, as this classifier only requires the PI Z and the corruption indicator M. We denote the model outputs by $\hat{p}(M = 0 \mid Z = z)$. Next, we estimate the marginal corruption probability directly from the data: $\hat{p}(M = 0) = \frac{1}{n}\sum_{i=1}^{n} M_i$. Finally, the estimated weights are computed according to (3): $\hat{w}_i = \hat{w}(z_i) = \frac{\hat{p}(M=0)}{\hat{p}(M=0|Z=z_i)}$. Even though PCP is not guaranteed to attain the nominal coverage level when employed with the estimates $\hat{w}_i$, the experiments from Section 4.3 indicate that it does achieve a conservative coverage rate in this case. The effect of inaccurate estimates of w_i on the coverage rate attained by PCP could be an exciting future direction to explore, perhaps by drawing on ideas from [26].

```
**Algorithm 1:** Privileged Conformal Prediction (PCP)
**Input:**
    Data $(X_i^{\text{obs}}, Y_i^{\text{obs}}, Z_i, M_i) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \times \{0,1\}, 1 \leq i \leq n$, weights $\{w_i\}_{i=1}^{n}$, miscoverage
    level $\alpha \in (0,1)$, level $\beta \in (0,\alpha)$, an algorithm $\hat{f}$, a score function $\mathcal{S}$, and a test point $X^{\text{test}} = x$.
**Process:**
    Randomly split $\{1, ..., n\}$ into two disjoint sets $\mathcal{I}_1, \mathcal{I}_2$.
    Fit the base algorithm $\hat{f}$ on the training data $\{(X_i^{\text{obs}}, Y_i^{\text{obs}})\}_{i\in\mathcal{I}_1}$.
    Compute the scores $S_i = \mathcal{S}(X_i^{\text{obs}}, Y_i^{\text{obs}}; \hat{f})$ for the calibration samples, $i \in \mathcal{I}_2^{\text{uc}}$.
    Compute a threshold $Q_i$ for each calibration sample according to (8).
    Compute $Q^{\text{PCP}}$, the $(1-\beta)$ quantile of $\{Q_i\}_{i\in\mathcal{I}_2}$, according to (9).
**Output:**
    Prediction set $C^{\text{PCP}}(x) = \{y : \mathcal{S}(x, y; \hat{f}) \leq Q^{\text{PCP}}\}$.
```

3.3 Privileged Conformal Prediction for scarce data

In this section, we present an adaptation of PCP to handle situations where the sample size is small. While PCP is computationally light, it requires splitting the data into training and calibration sets. This restriction is significant for small datasets in which the reduction in computations from the data splitting comes at the expense of statistical efficiency. To avoid data splitting, we build on the leave-one-out jackknife+ method [9, 27], and, in particular, its weighted version JAW [21]. The method we propose, which we refer to as LOO-PCP, better utilizes the training data compared to PCP. For the interest of space, we refer to Appendix B.3 for the description of LOO-PCP. The following Theorem states that prediction set $C^{\text{LOO-PCP}}$ constructed by LOO-PCP is guaranteed to achieve a valid coverage rate under our setup. In Appendix A.3 we provide the proof, which relies on results from [21].

Theorem 2. *Suppose that $\{(X_i(0), X_i(1), Y_i(0), Y_i(1), Z_i, M_i)\}_{i=1}^{n+1}$ are exchangeable, $(X(0), Y(0)) \perp\!\!\!\perp M \mid Z$, and P_Z is absolutely continuous with respect to $P_{Z|M=0}$. Then, the prediction set $C^{\text{LOO-PCP}}(X^{\text{test}})$ constructed according to Algorithm 4 satisfies:*

$$\mathbb{P}(Y^{\text{test}} \in C^{\text{LOO-PCP}}(X^{\text{test}})) \geq 1 - 2\alpha.$$

We note that in contrast to split CP, here, the coverage guarantee appears with a factor 2 in α. We refer the reader to [9] for a detailed explanation of why the factor 2 is necessary and cannot be removed. Nonetheless, it is well-known that this jackknife approach greatly improves statistical efficiency compared to split conformal methods.

4 Applications

In this section, we exemplify our proposal in three real-life applications. In all experiments, we randomly split the data into training, validation, calibration, and test sets. We fit a base learning model on the training data and use the validation set to avoid overfitting. We calibrate the model using the calibration data with the proposed PCP or with a baseline technique, and evaluate the performance on the test set. In all experiments, the calibration schemes are applied to achieve a $1-\alpha = 90\%$ marginal coverage rate. We use the CQR [11] non-conformity score in regression tasks, and the *homogeneous prediction sets* (HPS) non-conformity score [6, 28] in classification tasks. Appendix D describes the full details about the network architecture, training strategy, datasets, corruption technique, and this experimental protocol. The specific formulation of the PI is described in each experiment. In this section, we focus on three use cases: causal inference, missing response, and noisy response. We demonstrate the applicability of PCP on more datasets, and under different corruptions, including additional causal inference tasks in Appendix E.1, more response corruptions in Appendix E.3 and in Appendix E.4.1, and missing features settings in Appendix E.4.2.

4.1 Causal inference: semi-synthetic example

We begin with a causal inference example, in which the goal is to obtain inference for individual treatment effects [29]. In this setting, $X_i(0) = X_i(1) \in \mathcal{X}$ denotes the features, Z_i denotes the privileged information, $M_i \in \{0, 1\}$ denotes the binary treatment indicator, and $Y_i(0), Y_i(1) \in \mathbb{R}$ denote the counterfactual outcomes under control and treatment conditions, respectively. Recall that we only observe $Y_i^{\text{obs}} = Y_i(M_i)$ and that the PI explains the treatment pattern $(X(0), Y(0)) \perp\!\!\!\perp M \mid Z$. In this experimental setup, our goal is to construct a prediction set that covers the true test potential outcome under control conditions, i.e., $Y_{n+1}(0)$, at a user-specified level $1 - \alpha = 90\%$. Alternatively, we could also aim to predict the outcome under treatment $Y_{n+1}(1)$. However, in this experiment, we focus on $Y_{n+1}(0)$ since the dataset we use is highly imbalanced and there are few samples from the treatment group. This task is compelling since it can be used to generate a valid uncertainty interval for the individual treatment effect (ITE), $Y_i(1) - Y_i(0)$, which is a great interest for many problems [30–33]. For instance, the work in [34] shows how to construct a valid interval for the ITE by combining intervals for $Y_{n+1}(0)$ and $Y_{n+1}(1)$. We remark that providing statistically valid prediction intervals for $Y_{n+1}(0)$ is challenging due to the distribution shift between the observed control responses, which are drawn from $P_{Y(0)|M=0}$, whereas the test control response is drawn from $P_{Y(0)}$. Moreover, in this example, we intentionally design M_i to induce such a distribution shift; see Appendix D.1 for more details on the definition of M_i.

We test the applicability of our method on the semi-synthetic Infant Health and Development Program (IHDP) dataset [35], in which the objective is to find the effect of specialist home visits on a child's future cognitive test scores. That is, the feature vector X_i contains covariates describing the child's characteristics, the treatment M_i is the specialist home visits indicator, and the potential outcomes $Y_i(0), Y_i(1)$ are the future cognitive test scores. Since this dataset does not originally contain a privileged information variable, we artificially define it as the entry in X_i that correlates the most with $Y_i(0)$. This feature is then removed from X_i, so it is unavailable at inference time.

Since the IHDP dataset contains only 747 samples, we apply LOO-PCP in this example and compare it to the following calibration techniques. The first method is a naive jackkife+, which uses only the control samples and does not account for distribution shifts. The second and third techniques are two versions of JAW [21], which is a weighted conformal version of the jackknife+. The first version (Naive WCP) naively uses an estimate of $P_{M|X}$ as the likelihood ratio weights instead of $P_{M|Z}$, as Z^{test} is unknown. The second (Infeasible WCP) is an **infeasible** method which requires access to the unknown test privileged information Z^{test} for the computation of the likelihood ratio weights, which can be considered as an oracle calibration process. Importantly, the infeasible JAW and the proposed method use the true corruption probabilities when computing the weights $w(z)$ from (3).

Figure 1 reports the coverage rates and interval lengths of each calibration scheme. This figure shows that naive jackkife+ and the naive JAW achieve a lower coverage rate than desired. This is anticipated, as both schemes do not accurately account for the distribution shift. In contrast, the infeasible JAW and PCP achieve the desired coverage rate. This is not a surprise, as JAW is guaranteed to attain the nominal coverage level when applied with the correct weights [21, Theorem 1]. However, this method is infeasible to implement in contrast with our proposal, which, according to Theorem 2, is guaranteed to cover the response at the desired rate without using the test privileged information. Furthermore, Figure 1 reveals that PCP constructs intervals with approximately the same width as the ones generated by the infeasible JAW. This indicates that we do not lose much in terms of statistical efficiency by not having access to the test privileged information Z^{test}.

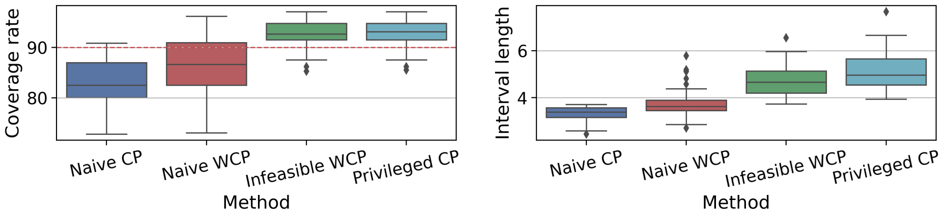

Figure 1: **Causal inference experiment: IHDP dataset.** The coverage rate and average interval length achieved by naive jackknife+ (`Naive CP`), naive `JAW` which considers only X to cope with the distribution shift (`Naive WCP`), an infeasible `JAW` which uses Z^{test} (`Infeasible WCP`), and the proposed method (`Privileged CP`). The metrics are evaluated over 50 random data splits.

4.2 Missing response variable: semi-synthetic example

In this section, we study the performance of `PCP` and compare it to baselines in a missing response setting using six real datasets: Facebook1,2 [36], Bio [37], House [38], Meps19 [39] and Blog [40]. Since these datasets do not originally contain privileged information, we artificially define Z_i as the feature from X_i that correlates the most with Y_i and then remove it from X_i. Furthermore, since all response variables are present in these datasets, we artificially remove the responses in 20% of the samples. We intentionally set the missing probability in a way that induces a distribution shift between the missing and observed variables. In Appendix D.1 we provide the full details about the corruption process and how we impute the missing data.

We compare the proposed method (`PCP`) to the following calibration schemes: a naive conformal prediction (`Naive CP`); a naive `WCP`, which considers only X to cope with the distribution shift; the two-staged baseline (`Two-Staged`); and an infeasible weighted conformal prediction (`Infeasible WCP`) which has access to the test privileged information Z^{test}. Importantly, the baseline `Two-Staged`, the infeasible `WCP`, and `PCP` use the real corruption probabilities when computing the weights $w(z)$ in (3). In contrast, `Naive WCP` estimates the corruption probability conditioned on X from the data. Figure 2 presents the performance of each calibration scheme, showing that the naive approach (`Naive CP`) consistently produces invalid prediction intervals. This is anticipated, as `Naive CP` does not provide guarantees under distribution shifts. Figure 2 also shows that `Two-Staged` generates too wide intervals, resulting in a conservative coverage rate of approximately 95%. By contrast, the infeasible `WCP` and the proposed `PCP` consistently achieve the desired 90% level. Crucially, `PCP` is comparable in the interval length to the infeasible `WCP`. In conclusion, this experiment demonstrates that `PCP` constructs intervals that are both reliable and informative.

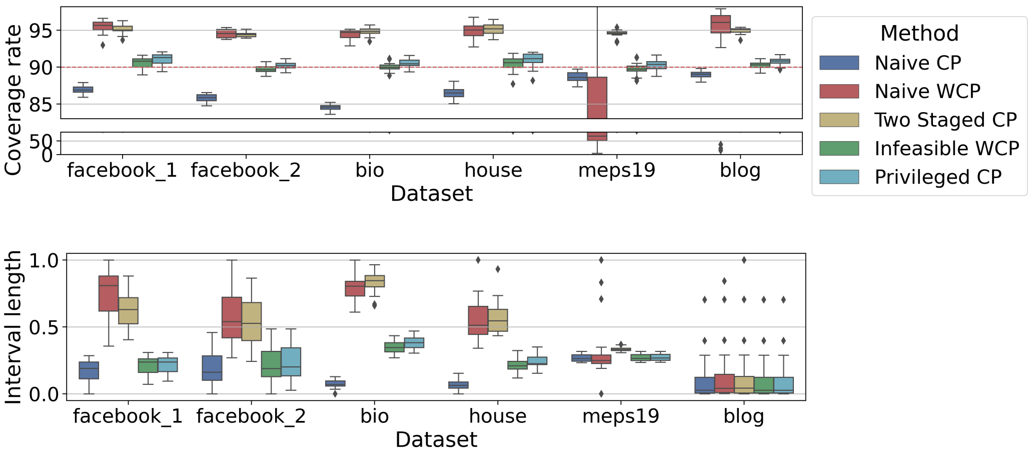

Figure 2: **Missing response experiment.** The coverage rate and average interval length obtained by various methods; see text for details. Performance metrics are evaluated over 20 random data splits.

4.3 Noisy response variable: real example

In what follows, we examine the performance of the proposed technique on the CIFAR-10N [41] image recognition dataset that contains noisy labels. Here, X is an image of one out of ten possible objects, and Y is its corresponding label. The noisy response, $Y(1)$, is the label annotated by one human annotator, while $Y(0)$ denotes the clean label obtained from CIFAR-10 [42]. That is, $M = 0$ indicates that the annotator correctly labeled the image. Similarly to [17], we define the privileged data as information about the annotators. Specifically, the variable Z_i contains two features: (i) the number of unique labels suggested by three annotators for the i-th sample, and (ii) the time took to annotate the corresponding sample batch, which contains ten images. In this experiment, we compare our method (`PCP`) to the following calibration schemes: a naive conformal prediction, applied either with the noisy labels `Naive CP (clean + noisy)` or ignoring them `Naive CP (only clean)`; the two-staged baseline (`Two Staged CP`); an infeasible `WCP` (`Infeasible WCP`) which assumes access to the unknown test privileged information Z^{test}. Additionally, since the corruption probabilities are not given in this dataset, we estimate them from the data and use these estimates to compute the weights w in (3).

Figure 3 presents each calibration scheme's coverage rate and uncertainty set size. This figure shows that `Naive CP` applied with noisy labels tends to overcover the clean label. This behavior is consistent with the work in [25, 43], which suggests that naive `CP` constructs conservative uncertainty sets when employed on data with dispersive label-noise. Figure 3 also indicates that calibrating the model only on the clean samples leads to invalid prediction sets that tend to undercover the clean test label. Observe also that the two-stage baseline is overly conservative, as it encapsulates the error in predicting both Z^{test} and the label. In contrast, the coverage rate of infeasible `WCP` and our proposed `PCP` is much closer to the desired level, yet slightly conservative. We suggest two possible explanations for this behavior: (i) the weights used are only estimates of the true likelihood ratios; (ii) in this real data we consider here, the PI may not fully explain the corruption mechanism. This highlights the robustness of our method to violations of our assumptions in the specific use-case studied here. Lastly, we remark that the prediction sets of `PCP` have a similar set size to the sets constructed by the infeasible `WCP`, which is in line with the results from Section 4.1 and Section 4.2.

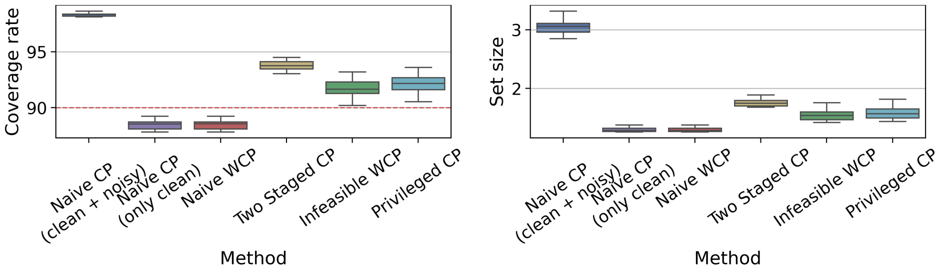

Figure 3: **Noisy response experiment: CIFAR-10N dataset.** Average coverage and set size obtained by various methods; see text for details. The metrics are evaluated over 20 random data splits.

5 Discussion and impact statement

In this paper, we introduced `PCP`, a novel calibration scheme to reliability quantify prediction uncertainty in situations where the training data is corrupted. The validity of our proposal is supported by theoretical guarantees and demonstrated in numerical experiments. The key assumption behind our method is that the features and responses are independent of the corruption indicator given the privileged information. This conditional independence resembles the strong ignorability assumption in causal inference [44–46]. While acquiring PI that satisfies this requirement can be challenging, our work relaxes the strong ignorability assumption, as the confounders are allowed to be absent during inference time. An additional restriction we make is that the true conditional corruption probability must be known to provide a theoretical coverage validity. However, our numerical experiments indicate that estimating these probabilities leads to reliable uncertainty estimates. A promising future direction would be to theoretically analyze the effect of inaccurate weights on the coverage guarantee, e.g., by borrowing ideas from [26]. Finally, we should note that there are potential social implications of our method, akin to many other works that aim to advance the ML field.

Acknowledgments and Disclosure of Funding

Y.R. and S.F. were supported by the ISRAEL SCIENCE FOUNDATION (grant No. 729/21). Y.R. thanks the Career Advancement Fellowship, Technion. Y.R. and S.F. thank Stephen Bates for insightful discussions and for providing feedback on this manuscript.

References

[1] H.A. Kahn and C.T. Sempos. *Statistical Methods in Epidemiology*. Monographs in epidemiology and biostatistics. Oxford University Press, 1989.

[2] David E Lilienfeld and Paul D Stolley. *Foundations of epidemiology*. Oxford University Press, USA, 1994.

[3] Steven Piantadosi. *Clinical trials: a methodologic perspective*. John Wiley & Sons, 1997.

[4] Steve Selvin. *Statistical analysis of epidemiologic data*, volume 35. Oxford University Press, 2004.

[5] Floyd J Fowler Jr. *Survey research methods*. Sage publications, 2013.

[6] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, NY, USA, 2005.

[7] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816 – 845, 2023.

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

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

[10] Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28, 2015.

[11] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[12] Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591, 2020.

[13] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

[14] Saeid Motiian. *Domain Adaptation and Privileged Information for Visual Recognition*. West Virginia University, 2019.

[15] Adam Breitholtz. *Towards practical and provable domain adaptation*. PhD thesis, Chalmers Tekniska Hogskola (Sweden), 2023.

[16] Judith Hoffman. *Adaptive learning algorithms for transferable visual recognition*. University of California, Berkeley, 2016.

[17] Yanshan Xiao, Zexin Ye, Liang Zhao, Xiangjun Kong, Bo Liu, Kemal Polat, and Adi Alhudhaif. Privileged information learning with weak labels. *Applied Soft Computing*, 142:110298, 2023.

[18] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *International Conference on Learning Representations*, 2016.

[19] Meng Yang, Ilia Nouretdinov, and Zhiyuan Luo. Learning by conformal predictors with additional information. In *Artificial Intelligence Applications and Innovations: 9th IFIP WG 12.5 International Conference, AIAI 2013, Paphos, Cyprus, September 30–October 2, 2013, Proceedings 9*, pages 394–400. Springer, 2013.

[20] Niharika Gauraha, Lars Carlsson, and Ola Spjuth. Conformal prediction in learning under privileged information paradigm with applications in drug discovery. In *Conformal and Probabilistic Prediction and Applications*, pages 147–156. PMLR, 2018.

[21] Drew Prinster, Anqi Liu, and Suchi Saria. Jaws: Auditing predictive uncertainty under covariate shift. *Advances in Neural Information Processing Systems*, 35:35907–35920, 2022.

[22] Maxime Cauchois, Suyash Gupta, Alnur Ali, and John Duchi. Predictive inference with weak supervision. *arXiv preprint arXiv:2201.08315*, 2022.

[23] Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, pages 1–66, 2024.

[24] Margaux Zaffran, Aymeric Dieuleveut, Julie Josse, and Yaniv Romano. Conformal prediction with missing values. In *International Conference on Machine Learning*, pages 40578–40604. PMLR, 2023.

[25] Matteo Sesia, YX Wang, and Xin Tong. Adaptive conformal classification with noisy labels. *arXiv preprint arXiv:2309.05092*, 2023.

[26] Yonghoon Lee, Edgar Dobriban, and Eric Tchetgen Tchetgen. Simultaneous conformal prediction of missing outcomes with propensity score ϵ-discretization. *arXiv preprint arXiv:2403.04613*, 2024.

[27] Chirag Gupta, Arun K. Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022.

[28] Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

[29] Miguel A Hernán and James M Robins. Causal inference, 2010.

[30] Jennie E Brand and Yu Xie. Who benefits most from college? evidence for negative selection in heterogeneous economic returns to higher education. *American sociological review*, 75(2):273–302, 2010.

[31] Stephen L Morgan. Counterfactuals, causal effect heterogeneity, and the catholic school effect on learning. *Sociology of education*, pages 341–374, 2001.

[32] Yu Xie, Jennie E Brand, and Ben Jann. Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347, 2012.

[33] Jean-Pierre Florens, James J Heckman, Costas Meghir, and Edward Vytlacil. Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica*, 76(5):1191–1206, 2008.

[34] Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

[35] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[36] Facebook comment volume data set. `https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset`. Accessed: January, 2019.

[37] bio. Physicochemical properties of protein tertiary structure data set. `https://archive.ics.uci.edu/ml/datasets/Physicochemical+Properties+of+Protein+Tertiary+Structure`. Accessed: January, 2019.

[38] house. House sales in king county, USA. `https://www.kaggle.com/harlfoxem/housesalesprediction/metadata`. Accessed: July, 2021.

[39] meps_19. Medical expenditure panel survey, panel 19. https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-181. Accessed: January, 2019.

[40] blog_data. Blogfeedback data set. https://archive.ics.uci.edu/ml/datasets/BlogFeedback. Accessed: January, 2019.

[41] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.

[42] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[43] Bat-Sheva Einbinder, Shai Feldman, Stephen Bates, Anastasios N Angelopoulos, Asaf Gendler, and Yaniv Romano. Conformal prediction is robust to dispersive label noise. In *Conformal and Probabilistic Prediction with Applications*, pages 624–626. PMLR, 2023.

[44] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

[45] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[46] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[47] Douglas Almond, Kenneth Y. Chay, and David S. Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

[48] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

[49] David S Yeager, Paul Hanselman, Gregory M Walton, Jared S Murray, Robert Crosnoe, Chandra Muller, Elizabeth Tipton, Barbara Schneider, Chris S Hulleman, Cintia P Hinojosa, et al. A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774):364–369, 2019.

[50] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[51] Carlos Carvalho, Avi Feller, Jared Murray, Spencer Woody, and David Yeager. Assessing treatment effect variation in observational studies: Results from a data challenge. *Observational Studies*, 5(2):21–35, 2019.

[52] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.

[53] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[55] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[56] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

A Theoretical results

A.1 Proof of Proposition 1

Proof. For the sake of this proof, we re-define the scores quantile $Q^{\texttt{WCP}}_{\text{conservative}}$ from Section 3.1 as a function of a test weight ω:

$$Q(\omega) := \text{Quantile}\left(1-\alpha+\beta; \sum_{j\in\mathcal{I}_2^{\text{uc}}} p_j(\omega)\delta_{S_j} + p_{n+1}(\omega)\delta_\infty\right), \tag{10}$$

where $p_j(\omega), p_{n+1}(\omega)$ are defined as:

$$p_j(\omega) = \frac{w_j}{\sum_{k\in\mathcal{I}_2^{\text{uc}}} w_k + \omega}, p_{n+1}(\omega) = \frac{\omega}{\sum_{k\in\mathcal{I}_2^{\text{uc}}} w_k + \omega}.$$

For ease of notation, we denote the test weight by $w_{n+1} = w(Z_{n+1})$ and its conservative counterpart by $\tilde{w}_{n+1} = w_{n+1}^{\text{conservative}}$, which is defined in (6). Note that by the definition of Q, we get $Q^{\texttt{WCP}}_{\text{conservative}} \equiv Q(\tilde{w}_{n+1})$. Since the observed calibration points $\{(X_i(M_i), Z_i)\}_{i\in\mathcal{I}_2}$ and the test point $(X_{n+1}(0), Z_{n+1})$ are exchangeable (we assume that $X_i(0) = X_i(1)$), then `CP` [6] guarantees that the prediction set $C^Z(X^{\text{test}})$ satisfies the coverage requirement:

$$\mathbb{P}(Z_{n+1} \in C^Z(X^{\text{test}})) \geq 1-\beta,$$

and therefore:

$$\mathbb{P}(w(Z_{n+1}) \in \{w(z) : z \in C^Z(X^{\text{test}})\}) \geq 1-\beta.$$

Following this, we get: $\mathbb{P}(w(Z_{n+1}) \leq \tilde{w}_{n+1}) \geq 1-\beta$. Assuming $(X(0), Y(0)) \perp\!\!\!\perp M \mid Z$, there is a covariate shift between the calibration and test samples, where the covariates are the privileged information Z. Specifically, the calibration covariates $\{Z_i\}_{i\in\mathcal{I}_2^{\text{uc}}}$ are drawn from $P_{Z|M=0}$ while the test covariates Z^{test} are drawn from P_Z. Importantly, both the calibration and response response variables have the same distribution conditional on Z: $P_{Y|Z}$. Thus, [8, Theorem 1] states that:

$$\mathbb{P}\left(Y^{\text{test}} \in \left\{y : \mathcal{S}(X^{\text{test}}, y) \leq Q(w_{n+1})\right\}\right) \geq 1-\alpha+\beta.$$

We note that Lemma 1 states that $Q(\omega)$ is non-decreasing, i.e., $\omega_1 \geq \omega_2 \Rightarrow Q(\omega_1) \geq Q(\omega_2)$. Finally, we combine everything together to get:

$$\begin{aligned}
\mathbb{P}\left(Y^{\text{test}} \in C^{\texttt{Two-Staged}}(X^{\text{test}})\right) &= \mathbb{P}\left(Y^{\text{test}} \in \left\{y : \mathcal{S}(X^{\text{test}}, y) \leq Q(\tilde{w}_{n+1})\right\}\right) \\
&= \mathbb{P}\left(\mathcal{S}(X^{\text{test}}, Y^{\text{test}}) \leq Q(\tilde{w}_{n+1})\right) \\
&\geq \mathbb{P}\left(\mathcal{S}(X^{\text{test}}, Y^{\text{test}}) \leq Q(\tilde{w}_{n+1}), \tilde{w}_{n+1} \geq w_{n+1}\right) \\
&\geq \mathbb{P}\left(\mathcal{S}(X^{\text{test}}, Y^{\text{test}}) \leq Q(w_{n+1}), \tilde{w}_{n+1} \geq w_{n+1}\right) \\
&= 1 - \mathbb{P}\left(\mathcal{S}(X^{\text{test}}, Y^{\text{test}}) > Q(w_{n+1}) \text{ or } \tilde{w}_{n+1} < w_{n+1}\right) \\
&\geq 1 - \mathbb{P}\left(\mathcal{S}(X^{\text{test}}, Y^{\text{test}}) > Q(w_{n+1})\right) - \mathbb{P}(\tilde{w}_{n+1} < w_{n+1}) \\
&\geq 1 - (\alpha-\beta) - (\beta) \\
&= 1-\alpha.
\end{aligned}$$

□

A.2 Proof of Theorem 1

Proof. For the sake of this proof, we re-define the scores quantile Q from (8) as a function of a test weight ω, similarly to (10):

$$Q(\omega) := \text{Quantile}\left(1-\alpha+\beta; \sum_{j\in\mathcal{I}_2^{\text{uc}}} p_j(\omega)\delta_{S_j} + p_{n+1}(\omega)\delta_\infty\right), \tag{11}$$

where $p_j(\omega), p_{n+1}(\omega)$ are defined as:

$$p_j(\omega) = \frac{w_j}{\sum_{k\in\mathcal{I}_2^{\text{uc}}} w_k + \omega}, \;\; p_{n+1}(\omega) = \frac{\omega}{\sum_{k\in\mathcal{I}_2^{\text{uc}}} w_k + \omega}.$$

Note that by the definition of Q, we get $Q_i \equiv Q(w_i)$. Furthermore:

$$Q^{\texttt{PCP}} \equiv \text{Quantile}\left(1-\beta; \sum_{i\in\mathcal{I}_2} \frac{1}{|\mathcal{I}_2|+1}\delta_{Q_i} + \frac{1}{|\mathcal{I}_2|+1}\delta_{\infty}\right)$$
$$= \text{Quantile}\left(1-\beta; \sum_{i\in\mathcal{I}_2} \frac{1}{|\mathcal{I}_2|+1}\delta_{Q(w_i)} + \frac{1}{|\mathcal{I}_2|+1}\delta_{\infty}\right).$$

Since $Q(\omega)$ is a non-decreasing function of ω, as proved in Lemma 1, we get:

$$Q^{\texttt{PCP}} = Q\left(\text{Quantile}\left(1-\beta; \sum_{i\in\mathcal{I}_2} \frac{1}{|\mathcal{I}_2|+1}\delta_{w_i} + \frac{1}{|\mathcal{I}_2|+1}\delta_{\infty}\right)\right).$$

Thus, $Q^{\texttt{PCP}}$ can be considered as if it was computed from the following weight:

$$\tilde{w}_{n+1} := \text{Quantile}\left(1-\beta; \sum_{i\in\mathcal{I}_2} \frac{1}{|\mathcal{I}_2|+1}\delta_{w_i} + \frac{1}{|\mathcal{I}_2|+1}\delta_{\infty}\right),$$

in the sense that $Q^{\texttt{PCP}} = Q(\tilde{w}_{n+1})$. Therefore:

$$C^{\texttt{PCP}}(X^{\text{test}}) := \left\{y : \mathcal{S}(X^{\text{test}}, y) \leq Q^{\texttt{PCP}}\right\} = \left\{y : \mathcal{S}(X^{\text{test}}, y) \leq Q(\tilde{w}_{n+1})\right\}.$$

The true weight of the $n+1$ sample is: $w_{n+1} = \frac{\mathbb{P}(M=0)}{\mathbb{P}(M=0|Z=Z_{n+1})}$. Now, since $\{Z_i\}_{i\in\mathcal{I}_2\cup\{n+1\}}$ are exchangeable, then $\{w_i\}_{i\in\mathcal{I}_2\cup\{n+1\}}$ are exchangeable and thus [11, Lemma 2] states that:

$$\mathbb{P}(w_{n+1} \leq \tilde{w}_{n+1}) \geq 1-\beta.$$

Assuming $(X(0), Y(0)) \perp\!\!\!\perp M \mid Z$, there is a covariate shift between the calibration and test samples, where the covariates are the privileged information Z. Specifically, the uncorrupted calibration covariates $\{Z_i\}_{i\in\mathcal{I}_2^{\text{uc}}}$ are drawn from $P_{Z|M=0}$ while the test covariates Z^{test} are drawn from P_Z. Importantly, both the calibration and response response variables have the same distribution conditional on Z: $P_{Y|Z}$. Thus, [8, Theorem 1] states that:

$$\mathbb{P}\left(Y^{\text{test}} \in \left\{y : \mathcal{S}(X^{\text{test}}, y) \leq Q(w_{n+1})\right\}\right) \geq 1-\alpha+\beta.$$

Note that Lemma 1 states that $Q(\omega)$ is non-decreasing, i.e., $\omega_1 \geq \omega_2 \Rightarrow Q(\omega_1) \geq Q(\omega_2)$. Finally, we combine all together and get:

$$\begin{aligned}
\mathbb{P}\left(Y^{\text{test}} \in C^{\texttt{PCP}}(X^{\text{test}})\right) &= \mathbb{P}\left(Y^{\text{test}} \in \left\{y : \mathcal{S}(X^{\text{test}}, y) \leq Q(\tilde{w}_{n+1})\right\}\right) \\
&= \mathbb{P}\left(\mathcal{S}(X^{\text{test}}, Y^{\text{test}}) \leq Q(\tilde{w}_{n+1})\right) \\
&\geq \mathbb{P}\left(\mathcal{S}(X^{\text{test}}, Y^{\text{test}}) \leq Q(\tilde{w}_{n+1}), \tilde{w}_{n+1} \geq w_{n+1}\right) \\
&\geq \mathbb{P}\left(\mathcal{S}(X^{\text{test}}, Y^{\text{test}}) \leq Q(w_{n+1}), \tilde{w}_{n+1} \geq w_{n+1}\right) \\
&= 1 - \mathbb{P}\left(\mathcal{S}(X^{\text{test}}, Y^{\text{test}}) > Q(w_{n+1}) \text{ or } \tilde{w}_{n+1} < w_{n+1}\right) \\
&\geq 1 - \mathbb{P}\left(\mathcal{S}(X^{\text{test}}, Y^{\text{test}}) > Q(w_{n+1})\right) - \mathbb{P}(\tilde{w}_{n+1} < w_{n+1}) \\
&\geq 1 - (\alpha-\beta) - (\beta) \\
&= 1-\alpha.
\end{aligned}$$

□

A.3 Proof of Theorem 2

Proof. For the sake of this proof, we re-define the prediction set as a function of the test weight $\omega \in [0,1]$:

$$C_{\omega}^{\texttt{LOO-PCP}}(x) = \left\{y \in \mathcal{Y} : \sum_{i\in\mathcal{I}^{\text{uc}}} p_i(\omega)\mathbb{1}\{S_i < \mathcal{S}(x, y; \hat{f}^{-i})\} < 1-\gamma\right\},$$

where $\gamma = \alpha - \frac{1}{2}\beta$, and $p_i(\omega), p_{n+1}(\omega)$ are defined as:

$$p_i(\omega) = \frac{w_i}{\sum_{k \in \mathcal{I}^{\text{uc}}} w_k + \omega}, p_{n+1}(\omega) = \frac{\omega}{\sum_{k \in \mathcal{I}^{\text{uc}}} w_k + \omega}.$$

The prediction set generated by `LOO-PCP` is: $C^{\texttt{LOO-PCP}}_{\tilde{w}_{n+1}}(x)$, where:

$$\tilde{w}_{n+1} := \text{Quantile}\left(1-\beta; \sum_{i=1}^{n} \frac{1}{n+1}\delta_{w_i} + \frac{1}{n+1}\delta_{\infty}\right).$$

Therefore, our goal is to show that this prediction set covers the response variable at the desired coverage rate:

$$\mathbb{P}(Y^{\text{test}} \in C^{\texttt{LOO-PCP}}_{\tilde{w}_{n+1}}(X^{\text{test}})) \geq 1 - 2\alpha. \tag{12}$$

The proof consists of two steps. In the first step, we show that $C^{\texttt{LOO-PCP}}_{w_{n+1}}(X^{\text{test}}))$ achieves a valid marginal coverage, namely:

$$\mathbb{P}(Y^{\text{test}} \in C^{\texttt{LOO-PCP}}_{w_{n+1}}(X^{\text{test}})) \geq 1 - 2\gamma.$$

In the second step, we show that $C^{\texttt{LOO-PCP}}_{\tilde{w}_{n+1}}(X^{\text{test}}))$ is a super set of $C^{\texttt{LOO-PCP}}_{w_{n+1}}(X^{\text{test}}))$ with a high probability. From these two steps, we will conclude (12).

We define the matrix of residuals, similarly to [9, 21], denoted by $R \in \mathbb{R}^{(n+1)\times(n+1)}$, with entries:

$$R_{i,j} = \begin{cases} +\infty & i = j, \\ \mathcal{S}(X_i(0), Y_i(0), \hat{f}^{-i,j}) & i \neq j, \end{cases}$$

where $\hat{f}^{-i,j}$ is the model $\hat{f}$ fitted on all samples except with the points i and j removed, namely, $\{1, ..., n+1\} - \{i, j\}$. For simplicity, we denote $\tilde{w}_i(\omega) := w_i$ for $i \in \{1, .., n\}$ and $\tilde{w}_{n+1}(\omega) = \omega$. We follow the definition in [21] of “strange” points $\mathcal{G}(\omega) \subseteq \mathcal{I}^{\text{uc}} \cup \{n+1\}$:

$$\mathcal{G}(\omega) = \left\{ i \in \mathcal{I}^{\text{uc}} \cup \{n+1\} : \tilde{w}_i(\omega) > 0, \sum_{j \in \mathcal{I}^{\text{uc}} \cup \{n+1\}} p_j(\omega) \mathbb{1}\left\{R_{ij} > R_{ji}\right\} \geq 1 - \gamma \right\}.$$

We begin by showing that $Y^{\text{test}} \in C^{\texttt{LOO-PCP}}_{w_{n+1}}(X^{\text{test}})) \Rightarrow n+1 \notin \mathcal{G}(w_{n+1})$. Suppose that $Y^{\text{test}} \in C^{\texttt{LOO-PCP}}_{w_{n+1}}(X^{\text{test}})$. Then, by definition of $C^{\texttt{LOO-PCP}}_{w_{n+1}}(X^{\text{test}}))$:

$$\begin{aligned} 1 - \gamma &> \sum_{j \in \mathcal{I}^{\text{uc}}} p_j(w_{n+1}) \mathbb{1}\{S_j < \mathcal{S}(X^{\text{test}}, Y^{\text{test}}; \hat{f}^{-j})\} \\ &= \sum_{j \in \mathcal{I}^{\text{uc}}} p_j(w_{n+1}) \mathbb{1}\{R_{j,n+1} < R_{n+1,j}\} \\ &= \sum_{j \in \mathcal{I}^{\text{uc}}} p_j(w_{n+1}) \mathbb{1}\{R_{n+1,j} > R_{j,n+1}\} \\ &= \sum_{j \in \mathcal{I}^{\text{uc}} \cup \{n+1\}} p_j(w_{n+1}) \mathbb{1}\{R_{n+1,j} > R_{j,n+1}\}. \end{aligned}$$

Therefore: $n+1 \notin \mathcal{G}(w_{n+1})$, by the definition of $\mathcal{G}$. We deduce that: $\mathbb{P}(Y^{\text{test}} \in C^{\texttt{LOO-PCP}}_{w_{n+1}}(X^{\text{test}})) \geq \mathbb{P}(n+1 \notin \mathcal{G}(w_{n+1}))$. In [21] it is shown that:

$$\mathbb{P}(n+1 \notin \mathcal{G}(w_{n+1})) \geq 1 - 2\gamma.$$

Thus:

$$\mathbb{P}(Y^{\text{test}} \in C^{\texttt{LOO-PCP}}_{w_{n+1}}(X^{\text{test}})) \geq \mathbb{P}(n+1 \notin \mathcal{G}(w_{n+1})) \geq 1 - 2\gamma.$$

We now turn to the second step of the proof. Similarly to Lemma 1, we now show that $C^{\texttt{LOO-PCP}}_{\omega}(x)$ is a monotonic function of ω, i.e., if $\omega_1 \geq \omega_2$ then $C^{\texttt{LOO-PCP}}_{\omega_2}(x) \subseteq C^{\texttt{LOO-PCP}}_{\omega_1}(x)$. Suppose that $y \in C^{\texttt{LOO-PCP}}_{\omega_2}(x)$. Then:

$$\sum_{i \in \mathcal{I}^{\text{uc}}} p_i(\omega_2) \mathbb{1}\{S_i < \mathcal{S}(x, y; \hat{f}^{-i})\} < 1 - \gamma.$$

Since $\omega_1 \geq \omega_2$, we get $p_i(\omega_1) \leq p_i(\omega_2)$ for all $i \in \{1, ..., n\}$. Therefore:

$$\sum_{i \in \mathcal{I}^{\text{uc}}} p_i(\omega_1) \mathbb{1}\{S_i < \mathcal{S}(x, y; \hat{f}^{-i})\} \leq \sum_{i \in \mathcal{I}^{\text{uc}}} p_i(\omega_2) \mathbb{1}\{S_i < \mathcal{S}(x, y; \hat{f}^{-i})\} < 1 - \gamma.$$

Meaning that $y \in C_{\omega_1}^{\texttt{LOO-PCP}}(x)$ as well. Lastly, we recall that according to [11, Lemma 2] the weight $\tilde{w}_{n+1}$ satisfies:

$$\mathbb{P}(w_{n+1} \leq \tilde{w}_{n+1}) \geq 1 - \beta.$$

Finally, we combine everything together to get:

$$\begin{aligned}
\mathbb{P}\left(Y^{\text{test}} \in C_{\tilde{w}_{n+1}}^{\texttt{LOO-PCP}}(X^{\text{test}})\right) &\geq \mathbb{P}\left(Y^{\text{test}} \in C_{\tilde{w}_{n+1}}^{\texttt{LOO-PCP}}(X^{\text{test}}), \tilde{w}_{n+1} \geq w_{n+1}\right) \\
&\geq \mathbb{P}\left(Y^{\text{test}} \in C_{w_{n+1}}^{\texttt{LOO-PCP}}(X^{\text{test}}), \tilde{w}_{n+1} \geq w_{n+1}\right) \\
&= 1 - \mathbb{P}\left(Y^{\text{test}} \notin C_{w_{n+1}}^{\texttt{LOO-PCP}}(X^{\text{test}}) \text{ or } \tilde{w}_{n+1} < w_{n+1}\right) \\
&\geq 1 - \mathbb{P}\left(Y^{\text{test}} \notin C_{w_{n+1}}^{\texttt{LOO-PCP}}(X^{\text{test}})\right) - \mathbb{P}\left(\tilde{w}_{n+1} < w_{n+1}\right) \\
&\geq 1 - (2\gamma) - (\beta) \\
&\geq 1 - 2\left(\alpha - \frac{1}{2}\beta\right) - (\beta) \\
&= 1 - 2\alpha.
\end{aligned}$$

□

A.4 Lemma about weighted quantiles

Lemma 1. *Suppose that $w_i, S_i \in \mathbb{R}$ for $1 \leq i \leq n$. Further, suppose that $\gamma \in [0, 1]$, and $\mathcal{I} \subseteq \{1, ..., n\}$. Denote the normalized weights $p_i(\omega), p_{n+1}(\omega)$ as:*

$$p_i(\omega) = \frac{w_i}{\sum_{k \in \mathcal{I}} w_k + \omega}, p_{n+1}(\omega) = \frac{\omega}{\sum_{k \in \mathcal{I}} w_k + \omega}.$$

Then, the weighted quantile:

$$Q(\omega) := \textit{Quantile}\left(\gamma; \sum_{j \in \mathcal{I}} p_i(\omega)\delta_{S_j} + p_{n+1}(\omega)\delta_{\infty}\right),$$

is a non-decreasing function of ω, i.e.,

$$\omega_1 \geq \omega_2 \Rightarrow Q(\omega_1) \geq Q(\omega_2).$$

Proof. Without loss of generality, suppose that $\{S_i\}_{i=1}^n$ are sorted in an increasing order, i.e., $S_{i+1} \geq S_i$. For the sake of this proof, we define $S_{n+1} = \infty$. For ease of notation, denote $\mathcal{I}' = \mathcal{I} \cup \{n + 1\}$ and $\mathcal{I}_i = \mathcal{I}' \cap \{1, ..., i\}$. The formal definition of $Q(\omega)$ is:

$$Q(\omega) = S_{\min\left\{i \in \mathcal{I}' : \sum_{k \in \mathcal{I}_i} p_k(\omega) \geq \gamma\right\}}.$$

Suppose that $\omega_1 \geq \omega_2$. Then, by the definition of p_i, we get that for all $i \in \mathcal{I}$:

$$p_i(\omega_1) \leq p_i(\omega_2).$$

Therefore for all $i \in \mathcal{I}$:

$$\sum_{k \in \mathcal{I}_i} p_k(\omega_1) \leq \sum_{k \in \mathcal{I}_i} p_k(\omega_2).$$

For $i = n + 1$ the above equation is trivially satisfied as $\sum_{k \in \mathcal{I}_{n+1}} p_k(\omega_1) = \sum_{k \in \mathcal{I}_{n+1}} p_k(\omega_2) = 1$. Thus,

$$\min\left\{i \in \{1, ..., n + 1\} : \sum_{k \in \mathcal{I}_i} p_k(\omega_1) \geq \gamma\right\} \geq$$

$$\min\left\{i \in \{1, ..., n + 1\} : \sum_{k \in \mathcal{I}_i} p_k(\omega_2) \geq \gamma\right\}.$$

Therefore,

$$S_{\min\left\{i\in\{1,\ldots,n+1\}:\sum_{k\in\mathcal{I}_i} p_k(\omega_1)\geq\gamma\right\}} \geq S_{\min\left\{i\in\{1,\ldots,n+1\}:\sum_{k\in\mathcal{I}_i} p_k(\omega_2)\geq\gamma\right\}},$$

and finally,

$$Q(\omega_1) \geq Q(\omega_2).$$

□

A.5 Relaxing the conditional independence assumption

In this section, we present two relaxations for conditional independence assumption, $(X(0), Y(0)) \perp\perp M \mid Z$, in Theorem 1. The first relaxation allows $X(0)$ to depend on M given the PI at the expense of using the following weights $\mathbb{P}(M \mid X = x, Z = x)$. The second result relaxes the conditional independence to hold approximately, up to some error, denoted by ε. We begin with formalizing the first result and then turn to the second one.

Theorem 3 (Robustness of PCP to dependence of the features and corruption indicator)**.** *Suppose that* $\{(X_i(0), X_i(1), Y_i(0), Y_i(1), Z_i, M_i)\}_{i=1}^{n+1}$ *are exchangeable,* $(Y(0) \perp\!\!\!\perp M) \mid (X, Z)$*, the features are uncorrupted,* $X(0) = X(1)$*, and* $P_{X(0),Z}$ *is absolutely continuous with respect to* $P_{X(0),Z|M=0}$*. Denote by* $C^{PCP}(X^{\text{test}})$ *the prediction set constructed according to Algorithm 1 with the weights:*

$$w_i := \frac{\mathbb{P}(M = 0)}{\mathbb{P}(M = 0 \mid X(0) = X_i(0), Z = Z_i)}$$

Then, this prediction set achieves the desired coverage rate:

$$\mathbb{P}(Y^{\text{test}} \in C^{PCP}(X^{\text{test}})) \geq 1 - \alpha.$$

Proof. This result follows from the proof of Theorem 1, except for the following changes. Here, we get that then there is a covariate shift between the calibration and test samples, where the covariates are $X(0), Z$. Therefore, the following weight w_{n+1} of this setup:

$$w_{n+1} = \frac{\mathbb{P}(M = 0)}{\mathbb{P}(M = 0 \mid X(0) = X_{n+1}(0), Z = Z_{n+1})}$$

satisfies:

$$\mathbb{P}\left(Y^{\text{test}} \in \left\{y : \mathcal{S}(X^{\text{test}}, y) \leq Q(w_{n+1})\right\}\right) \geq 1 - \alpha + \beta,$$

where Q is defined as in Appendix A.2. The rest of the proof is as in Appendix A.2. □

We now turn to present an initial extension of Theorem 1 to a setting where the conditional independence assumption is not fully satisfied. For the simplicity of this extension, we assume $X(0) \perp\!\!\!\perp M \mid Z = z$.

We note that the independence assumption $Y(0) \perp\!\!\!\perp M \mid Z = z$ is equivalent to assuming that the density of $Y(0) \mid M = m, Z = z$ is the same for $m \in \{0, 1\}$, formally:

$$f_{Y(0)|M=0,X=x,Z=z}(y; 0, x, z) = f_{Y(0)|M=1,X=x,Z=z}(y; 1, x, z).$$

In this extension, we relax this assumption and instead require that $\forall x \in \mathcal{X}$, there exists $\varepsilon_x \in \mathbb{R}$ such that the difference between the two densities is bounded by ε_x:

$$\forall y \in \mathcal{Y}, z \in \mathcal{Z} : |f_{Y(0)|M=0,X=x,Z=z}(y; 0, x, z) - f_{Y(0)|M=1,X=x,Z=z}(y; 1, x, z)| \leq \varepsilon_x.$$

Theorem 4 (Robustness of PCP to conditional independence violation)**.** *Suppose that* $\{(X_i(0), X_i(1), Y_i(0), Y_i(1), Z_i, M_i)\}_{i=1}^{n+1}$ *are exchangeable, the features are independent of the corruption indicator given the PI,* $X(0) \perp\!\!\!\perp M \mid Z$*, the probability* P_Z *is absolutely continuous with respect to* $P_{Z|M=0}$*, and* $\forall x \in \mathcal{X}$ *there exists* $\varepsilon_x \in \mathbb{R}$ *such that:*

$$\forall y \in \mathcal{Y}, z \in \mathcal{Z} |f_{Y(0)|M=0,X=x,Z=z}(y; 0, x, z) - f_{Y(0)|M=1,X=x,Z=z}(y; 1, x, z)| \leq \varepsilon_x.$$

Then, the coverage rate of the prediction set $C^{PCP}(X^{\text{test}})$ *constructed according to Algorithm 1 is lower bounded by:*

$$\mathbb{P}(Y^{\text{test}} \in C^{PCP}(X^{\text{test}})) \geq 1 - \alpha - \mathbb{E}_{X,Z}[|C^{PCP}(X)|\varepsilon_X \mathbb{P}(M = 1 \mid X, Z)].$$

Proof. We define by V a variable that is drawn from:

$$(Z, V) \sim P_Z \times P_{Y(0)|Z=z,M=0}.$$

By the definition of V, and according to Theorem 1, `PCP` covers V with $1 - \alpha$ probability:

$$\mathbb{P}(V \in C^{\texttt{PCP}}(X)) \geq 1 - \alpha.$$

Notice that $V \mid X = x, Z = z$ equals in distribution to $Y(0) \mid M = 0, X = x, Z = z$ by definition. We denote the coverage rate of `PCP` over $Y(0) \mid M = 0, X = x, Z = z$ by $\beta_{0,x,z}$:

$$\beta_{0,x,z} := \mathbb{P}(Y(0) \in C^{\texttt{PCP}}(X) \mid X = x, Z = z, M = 0) = \mathbb{P}(V \in C^{\texttt{PCP}}(X) \mid X = x, Z = z).$$

Similarly, we denote the coverage rate of `PCP` over $Y(0) \mid M = 1, X = x, Z = z$ by $\beta_{1,x,z}$:

$$\beta_{1,x,z} := \mathbb{P}(Y(0) \in C^{\texttt{PCP}}(X) \mid M = 1, X = x, Z = z) = \int_{y \in C^{\texttt{PCP}}(x)} f_{Y(0)|M=1,X=x,Z=z}(y; 1, x, z)dy$$

This probability can be lower bounded by:

$$\begin{aligned}\int_{y \in C^{\texttt{PCP}}(x)} f_{Y(0)|M=1,X=x,Z=z}(y; 1, x, z)dy &\geq \int_{y \in C^{\texttt{PCP}}(x)} (f_{Y(0)|M=0,X=x,Z=z}(y; 0, x, z) - \varepsilon_x)dy \\ &= \beta_{0,x,z} - |C^{\texttt{PCP}}(x)|\varepsilon_x\end{aligned}$$

We now compute the conditional coverage rate of `PCP`:

$$\begin{aligned}\mathbb{P}(Y(0) \in C^{\texttt{PCP}}(X) \mid Z = z, X = x) &= \mathbb{P}(Y(0) \in C^{\texttt{PCP}}(X) \mid M = 0, Z = z, X = x)\mathbb{P}(M = 0 \mid X = x, Z = z) \\ &+ \mathbb{P}(Y(0) \in C^{\texttt{PCP}}(X) \mid M = 1, Z = z, X = x)\mathbb{P}(M = 1 \mid X = x, Z = z) \\ &\geq \beta_{0,x,z}\mathbb{P}(M = 0 \mid X = x, Z = z) \\ &+ (\beta_{0,x,z} - |C^{\texttt{PCP}}(x)|\varepsilon_x)\mathbb{P}(M = 1 \mid X = x, Z = z) \\ &= \beta_{0,x,z}\mathbb{P}(M = 0 \mid X = x, Z = z) \\ &+ \beta_{0,x,z}\mathbb{P}(M = 1 \mid X = x, Z = z) - |C^{\texttt{PCP}}(x)|\varepsilon_x\mathbb{P}(M = 1 \mid X = x, Z = z) \\ &= \beta_{0,x,z}(\mathbb{P}(M = 0, X = x, Z = z) + \mathbb{P}(M = 1 \mid X = x, Z = z)) \\ &- |C^{\texttt{PCP}}(x)|\varepsilon_x\mathbb{P}(M = 1 \mid X = x, Z = z) \\ &= \beta_{0,x,z} - |C^{\texttt{PCP}}(x)|\varepsilon_x\mathbb{P}(M = 1 \mid X = x, Z = z) \\ &= \mathbb{P}(V \in C^{\texttt{PCP}}(X) \mid X = x, Z = z) - |C^{\texttt{PCP}}(x)|\varepsilon_x\mathbb{P}(M = 1 \mid X = x, Z = z)\end{aligned}$$

By marginalizing this result we get:

$$\begin{aligned}\mathbb{P}(Y(0) \in C^{\texttt{PCP}}(X)) &= \int_{x \in \mathcal{X}, z \in \mathcal{Z}} \mathbb{P}(Y(0) \in C^{\texttt{PCP}}(X) \mid Z = z, X = x)f_{X,Z}(x, z)dxdz \\ &\geq \int_{x \in \mathcal{X}, z \in \mathcal{Z}} [\mathbb{P}(V \in C^{\texttt{PCP}}(X) \mid X = x, Z = z)]f_{X,Z}(x, z)dxdz \\ &- \int_{x \in \mathcal{X}, z \in \mathcal{Z}} [|C^{\texttt{PCP}}(x)|\varepsilon_x\mathbb{P}(M = 1 \mid X = x, Z = z)]f_{X,Z}(x, z)dxdz \\ &\geq 1 - \alpha - \mathbb{E}_{X,Z}[|C^{\texttt{PCP}}(X)|\varepsilon_X\mathbb{P}(M = 1 \mid X, Z)]\end{aligned}$$

□

This result provides a lower bound for the coverage rate of `PCP` in the setting where the conditional independence assumption is not exactly satisfied. Intuitively, as ε_x decreases, i.e., as the two distributions $Y(0) \mid M = m, Z = z$ for $m \in \{0, 1\}$ are closer to each other, the lower bound is tighter, and closer to the target level. Similarly, as the two distributions diverge, the lower bound becomes looser.

B Algorithms

B.1 Two-Staged Conformal

In this section, we outline the two-staged conformal prediction algorithm (`Two-Staged`).

Algorithm 2: Two-Staged Conformal Prediction (`Two-Staged`)

```
Input:
  Data $(X_i^{\text{obs}}, Y_i^{\text{obs}}, Z_i, M_i) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \times \{0,1\}, 1 \leq i \leq n$.
  Weights $\{w_i\}_{i=1}^n$.
  Miscoverage level $\alpha \in (0,1)$.
  Level $\beta \in (0,\alpha)$.
  An algorithm $\hat{f}^Y$ for $Y$.
  An algorithm $\hat{f}^Z$ for $Z$.
  A score function $\mathcal{S}$.
  A test point $X^{\text{test}} = x$.
Process:
  Randomly split $\{1, ..., n\}$ into two disjoint sets $\mathcal{I}_1, \mathcal{I}_2$.
  Fit the base algorithm $\hat{f}^Y$ on $\{(X_i^{\text{obs}}, Y_i^{\text{obs}})\}_{i \in \mathcal{I}_1}$ and the algorithm $\hat{f}^Z$ on $\{(X_i^{\text{obs}}, Z_i)\}_{i \in \mathcal{I}_1}$.
  Compute the scores $S_i^Z = \mathcal{S}(X_i^{\text{obs}}, Z_i; \hat{f}^Z)$, $S_i = \mathcal{S}(X_i^{\text{obs}}, Y_i^{\text{obs}}; \hat{f}^Y)$ for the calibration samples, $i \in \mathcal{I}_2$.
  Compute a threshold $Q^Z$ as the $(1 - 1/|\mathcal{I}_2|)(1-\beta)$-th empirical quantile of the scores $\{S_i\}_{i \in \mathcal{I}_2}$.
  Construct a prediction set for $Z$: $C^Z(x) = \{z : \mathcal{S}(x, z, \hat{f}^Z) \leq Q^Z\}$.
  Compute the conservative test weight $w^{\text{conservative}}(x) := \max_{z \in C^Z(x)} w(z)$.
  Compute the test threshold $Q^{\texttt{WCP}}_{\text{conservative}}$ according to (5) with the clean calibration points $\{(X_i, Y_i, w(Z_i))\}_{i \in \mathcal{I}_2^{\text{uc}}}$, and the conservative test weight, $w^{\text{conservative}}(x)$, with a nominal coverage level of $1 - \alpha + \beta$.
Output:
  Prediction set $C^{\texttt{Two-Staged}}(x) = \{y : \mathcal{S}(x, y; \hat{f}^Y) \leq Q^{\texttt{WCP}}_{\text{conservative}}\}$.
```

B.2 Efficient Privileged Conformal Prediction

Since the complexity of Algorithm 1 is squared in the number of calibration samples, here, we provide a more efficient algorithm with a complexity that is linear in the number of calibration samples. The efficient version is detailed in Algorithm 3. The proof of Theorem 1 in Section A.2 shows that these two algorithms are identical, in the sense that they produce exactly the same outputs.

B.3 Privileged Conformal Prediction for scarce data

In this section, we describe the leave-one-out privileged conformal prediction algorithm (`LOO-PCP`), which is designed to handle scarce data more efficiently. This procedure is summarized in Algorithm 4.

C Datasets details

C.1 General real dataset details

Table 1 displays the size of each data set, the feature dimension, and the feature that is used as privileged information in the tabular data experiments.

C.2 CIFAR-10N dataset details

CIFAR-10N [41] is a variation of the CIFAR-10 [42] in which the labels are given by human annotators. In this task, X_i is an image from the CIFAR-10 dataset. The response $Y_i \in \{1, ..., 10\}$ is the noisy image label chosen by the first human annotator. The ratio of noisy samples is 17.23%. The

Algorithm 3: Efficient Privileged Conformal Prediction (`PCP`)

```
Input:
    Data $(X_i^{\text{obs}}, Y_i^{\text{obs}}, Z_i, M_i) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \times \{0,1\}, 1 \leq i \leq n$.
    Weights $\{w_i\}_{i=1}^n$.
    Miscoverage level $\alpha \in (0,1)$.
    Level $\beta \in (0, \alpha)$.
    An algorithm $\hat{f}$.
    A score function $\mathcal{S}$.
    A test point $X^{\text{test}} = x$.
Process:
    Randomly split $\{1, ..., n\}$ into two disjoint sets $\mathcal{I}_1, \mathcal{I}_2$.
    Fit the base algorithm $\hat{f}$ on $\{(X_i^{\text{obs}}, Y_i^{\text{obs}})\}_{i \in \mathcal{I}_1}$.
    Compute the scores $S_i = \mathcal{S}(X_i^{\text{obs}}, Y_i^{\text{obs}}; \hat{f})$ for the calibration samples, $i \in \mathcal{I}_2^{\text{uc}}$.
    Compute an estimated test weight based on the calibration samples:
    $\tilde{w}_{n+1} := \text{Quantile}\left(1 - \beta; \sum_{i \in \mathcal{I}_2} \frac{1}{n_2+1} \delta_{w_i} + \frac{1}{n_2+1} \delta_{\infty}\right)$
    Compute $Q^{\texttt{PCP}} := Q(\tilde{w}_{n+1})$, where $Q$ is defined in (11).
Output:
    Prediction set $C^{\texttt{PCP}}(x) = \{y : \mathcal{S}(x, y; \hat{f}) \leq Q^{\texttt{PCP}}\}$.
```

Algorithm 4: Privileged Conformal Prediction for scarce data (`LOO-PCP`)

```
Input:
    Data $(X_i^{\text{obs}}, Y_i^{\text{obs}}, Z_i, M_i) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \times \{0,1\}, 1 \leq i \leq n$, weights $\{w_i\}_{i=1}^n$, miscoverage
    level $\alpha \in (0,1)$, level $\beta \in (0, \alpha)$, an algorithm $\hat{f}$, a score function $\mathcal{S}$, and a test point $X^{\text{test}} = x$.
Process:
for $i \leftarrow 1$ to $n$ do
    Define the current training set as $\mathcal{I} = \{1, ..., i-1, i+1, ..., n\}$.
    Fit the base algorithm $\hat{f}$ on $(X_j, Y_j)_{j \in \mathcal{I}_1}$ to obtain $\hat{f}^{-i}$.
    Compute the score $S_i = \mathcal{S}(X_i^{\text{obs}}, Y_i^{\text{obs}}; \hat{f}^{-i})$.
Guess the weight of the $n+1$ sample:
$\tilde{w}_{n+1} = \text{Quantile}\left(1 - \beta; \sum_{i \in \{1,...,n\}} \frac{1}{n+1} \delta_{w_i} + \frac{1}{n+1} \delta_{\infty}\right)$.
Compute: $p_i = \frac{w_i}{\sum_{j \in \mathcal{I}^{\text{uc}}} w_j + \tilde{w}_{n+1}}$, for $i \in \{1, ..., n\}$.
Define the threshold $\gamma := \alpha - \frac{1}{2}\beta$.
Output:
    Prediction set $C^{\texttt{LOO-PCP}}(x) = \left\{y \in \mathcal{Y} : \sum_{i \in \mathcal{I}^{\text{uc}}} p_i \mathbb{1}\{S_i < \mathcal{S}(x, y; \hat{f}^{-i})\} < 1 - \gamma\right\}$.
```

privileged information Z_i has two features: the working time of the first annotator, and the number of different labels chosen by the three human annotators.

C.3 CIFAR-10C dataset details

CIFAR-10C [50] is a variation of the CIFAR-10 [42] dataset in which the images are contaminated by artificial corruptions. Here, we randomly apply one of the following corruptions for 15% of the images: snow, defocus blur, pixelate, or fog. The corruption severity level is either 4 or 5, chosen with equal probability. A label of a corrupted image is flipped according to the severity and corruption type. The severities 4,5 are assigned the values 0.92, and 0.95, respectively, and the snow, defocus blur, pixelate, and fog corruptions are assigned with 0.95, 0.93, 0.9, 0.93, 0.94, respectively. The flip probability is the multiplication of the value assigned with the corresponding severity and corruption type. In total, 13.09% of the labels are flipped. Importantly, in training time, both the image and the label are corrupted, while at inference time, only the image is corrupted and the performance is computed with respect to the clean label. In the dispersive noise setting, the label is uniformly flipped

Table 1: Information about the real data sets.

Dataset Name	# Samples	$X/Z/Y$ Dimensions	Z description
facebook1 [36]	40948	52/1/1	Number of posts comments
facebook2 [36]	81311	52/1/1	Number of posts comments
Bio [37]	45730	8/1/1	Fractional area of exposed non polar residue
House [38]	21613	17/1/1	Square footage of the apartments interior living space
Meps19 [39]	15785	138 /1/1	Overall rating of feelings
Blog [40]	52397	279/1/1	The time between the blog post publication and base-time
IHDP [35]	747	24/1/1	Birth weight
Twins [47, 48]	12042	50/1/1	The birth weight of the lighter twin
NSLM [49]	10391	10/1/1	Synthetic normally distributed random variable
CIFAR-10N [41]	50000	32x32x3/1/2	Annotation time & Label variability
CIFAR-10C [50]	40000	32x32x3/1/2	Corruption severity & type

into a wrong one. In the contractive noise setting, if the image is affected by either snow or defocus blur corruption, then the noisy label is deterministically set to 2, or to 1 if the original label was already 2. If the image is corrupted by either pixelate, or fog, then the noisy label is deterministically set to 7, or to 6 if the original label was already 7. The privileged information contains two features: the severity level and the corruption type. Since CIFAR-10C contains only 10000 samples, we add 30000 clean samples from CIFAR-10 to our dataset. This way, 60% of the 10000 CIFAR-10C images are corrupted, while the other 30000 CIFAR-10 samples are clean. In total, 15% of the images are corrupted.

C.4 IHDP dataset details

The Infant Health and Development Program (IHDP) dataset [35], is a semi-synthetic data containing in which the response variable is the future cognitive test scores. The feature vector includes information about the child such as child—birth weight, head circumference, weeks born preterm, birth order, first born indicator, neonatal health index, twin status, and more. The treatment M is the specialist home visits indicator. The privileged information is defined as the covariate in X_i with the highest correlation to Y_i. We then remove this feature from X_i, so it cannot be used at inference time. Since this dataset is semi-synthetic, every sample point includes both potential outcomes $Y_i(0)$, $Y_i(1)$ for every sample i. We therefore need to choose which samples are treated with $M = 0$ and which are treated with $M = 1$. In section D.1 we explain how M is chosen to intentionally induce a distribution shift between the observed and test distributions.

C.5 Twins dataset details

The Twins dataset contains information about twin births in the USA between 1989 and 1991. The raw data was provided by [47], and [48] introduced it as a new benchmark. In this dataset, the treatment $T = 1$ indicates for being born the heavier twin, the covariates X include features about the twins and their parents, and the outcome corresponds to the mortality of each of the twins in their first year of life. Since the records of the two twins are available, their mortality rates are considered as two potential outcomes, where the treatment is the indicator of being born heavier. We follow the protocol of [48] and focus only on twins with birth weight less than 2kg. The privileged information variable is set to the birth weight of the lighter twin. Since both potential outcomes are observed in the data, we must selectively hide one of them to simulate an observational study. We choose the treatment probability as explained in Section D.1.

C.6 NSLM dataset details

The National Study of Learning Mindsets (NSLM) dataset [49] is a semi-synthetic data that was analyzed in the 2018 Atlantic Causal Inference Conference workshop on heterogeneous treatment effects [51]. See [51, Section 2] for more information about this dataset. In our experiments, we follow the protocol introduced in [34, 51] to generate two synthetic potential outcomes and a synthetic PI variable. Specifically, we begin by scaling the data to have 0 mean and standard deviation of 1. We split the dataset into a training set and a validation set, containing 80% and 20% samples from the entire data, respectively. Using these training and validation sets, we fit a neural network with one hidden layer of size 32 to predict Y from X. The training parameters are as detailed in Section D.1. This network function is denoted by $\hat{\mu}_0(\cdot)$. Similarly, we fit an XGBoost classifier to predict the original treatment variable M given in the data from the feature vector X. We set the max_depth and n_estimators parameters to 2, 10, respectively. We then calibrate the estimated propensity score to have the same mean as the marginal treatment probability. This calibrated estimated propensity score is denoted by $\hat{e}(X_i)$. We generate a new treatment variable M_i, a synthetic PI variable Z_i, and a semi-synthetic target variable $Y_i(0)$ as follows.

$$\begin{aligned}
&Z_i \sim \mathcal{N}(0, 0.2^2)\\
&E_i = \mathbb{1}\{Z_i \geq \text{Quantile}(0.9, Z) \text{ or } Z_i \leq \text{Quantile}(0.1, Z)\}\\
&M_i \sim Ber(\min(0.8, (1 + E_i)\hat{e}(X_i)))\\
&\tau_i = 0.228 + 0.05\mathbb{1}\{X_{i,5} < 0.07) - 0.05\mathbb{1}\{X_{i,6} < -0.69\} - 0.08\mathbb{1}\{X_{i,1} \in \{1, 13, 14\}\}\\
&Y_i(0) = \hat{\mu}_0(X_i) + \tau_i + (1 + E_i)Z_i.
\end{aligned}$$

C.7 Synthetic dataset details

In this section, we present the synthetic dataset used in the ablation study of the parameter β in Appendix E.5.

The feature vectors are uniformly sampled as follows:

$$X_i \sim \text{Uni}(1, 5)^{10},$$

where Uni(a, b) is a unifrom distribution in the range (a, b). The PI is sampled as:

$$\begin{aligned}
E_i^1 &\sim \mathcal{N}(0, 1),\\
E_i^2 &\sim \text{Uni}(-1, 1),\\
E_i^3 &\sim \mathcal{N}(0, 1),\\
P_i &\sim \text{Pois}(\cos(E_i^2 + 0.1)) * E_i^2,\\
Z_i &\sim P_i + 2E_i^3.
\end{aligned}$$

Above, Pois(λ) is a poisson distribution with parameter λ, and $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2. Finally, the label is defined as:

$$\begin{aligned}
\beta &\sim \text{Uni}(0, 1)^5\\
\beta &= \beta / ||\beta||_1\\
U_i &= \mathbb{1}_{Z_i < -3} + 2 * \mathbb{1}_{-3 \leq Z_i \leq 1} + 8 * \mathbb{1}_{Z_i > 1}\\
E_i &\sim \mathcal{N}(0, 1)\\
Y_i &= 0.3X_i\beta + 0.8Z_i + 0.2 + U_iE_i.
\end{aligned}$$

D Experimental setup

D.1 General setup

In all experiments, except for scarce data experiments, we split the data into a training set (50%), calibration (20%), validation set (10%) used for early stopping, and a test set (20%) to evaluate performance. See Section D.2 for the specific details in the scrace data experiments. Then, we normalize the feature vectors and response variables to have a zero mean and unit variance. In experiments involving missing variables, we impute them with a linear model fitted on variables that are always observed from X, Y, Z. The linear model is trained on samples from the training and validation sets. For datasets that are not originally corrupted, the corruption probability is determined as follows. First, for IHDP and Twins datasets, we fit a linear model on the entire data to predict Y given X, Z, and use its predictions as an initial value. For other datasets, we take Z as the initial value.

We take the initial values, set the maximal value as the 85% quantile, and divide by the 90% quantile of the initial values. Then, we zero the lowest 75% values. Then, we raise all values to the exponent that achieves an average of 0.20. The result is the corruption probability. Therefore, by definition, the average corruption probability is 20%. In all experiments, we fit a base learning model and wrap its output with a calibration scheme. In regression tasks, the model is trained to learn the 5% and 95% conditional quantiles of $Y \mid X$. In Table 2 we summarize the model we used for each dataset for both tasks. For neural network models, we used an Adam optimizer [52] with 1e-4 learning rate, and batch size of 128. The network is composed of hidden layers of sizes: 32, 64, 64, 32, 0.1 dropout, and leaky relu as an activation function. For xgboost and random forest models, we used 100 estimators. We train the networks for 1000 epochs, but stop the training earlier if the validation loss does not improve for 200 epochs, and in this case, the model with the lowest validation loss is chosen. In our experiments, we use the xgboost package [53] and the scikit-learn package [54] from random forest. The neural networks were implemented with the pytorch package [55]. Regarding the hyper-parameters of the calibration schemes, in all experiments, we set the parameter β of `PCP` to $\beta = 0.005$ and the parameter β of `Two-Staged` to $\beta = 0.05$.

Table 2: The learning models used for each dataset.

Dataset Name	Base learning model	Corruption probability estimator
Facebook1 [36]	Neural network	Neural network
Facebook2 [36]	Neural network	Neural network
Bio [37]	Neural network	Neural network
House [38]	Neural network	Neural network
Meps19 [39]	Neural network	Neural network
Blog [40]	Neural network	Neural network
IHDP [35]	XGBoost	XGBoost
Twins [47, 48]	XGBoost	XGBoost
NSLM [49]	XGBoost	XGBoost
CIFAR-10N [41]	Resnet-18	Resnet-18 when using x or Random forest when using only z
CIFAR-10C [50]	Resnet-18	Resnet-18 when using x or Random forest when using only z

D.2 Experimental setup for scarce data experiment

In this section, we detail the experimental setup employed in the experiment in Section 4.1. In this experiment, we split the data into a training set (30%), a validation set (10%), and a test set (60%). Furthermore, the nominal coverage level was set to $1 - 2\alpha = 90\%$. For each training sample i, we fit an XGBoost quantile regression model using the entire training set, except for the i-th sample, to obtain a leave-one-out model. Then, we calibrated the model outputs with each calibration scheme, using the leave-one-out models.

D.3 Experimental setup for CIFAR-10N and CIFAR-10C experiments

We use a ResNet-18 network [56] as a base model. If we used both X and Z as an input to the model, we forward X through the CNN, and then through a linear layer with an output dimension of 16. Then, we concatenate this result with Z and forward it through a network with hidden layers with sizes 32, 64, 64, 32. When fitting the CNN, we train the model with batches of size 32 for 50 epochs, and choose the model with the lowest validation loss. Additionally, we apply a random augment transform to improve performance. For the `Two-Staged` and `Naive WCP` calibration schemes, we use only X to estimate the corruption probability, and for the infeasible `WCP` and `PCP` we use only Z to estimate it.

D.4 Machine's spec

The resources used for the experiments are:

- **CPU**: Intel(R) Xeon(R) E5-2650 v4.

- **GPU**: Nvidia titanx, 1080ti, 2080ti.
- **OS**: Ubuntu 18.04.

D.5 Computational resources

The computation efficiency of the proposed algorithm is dominated by the efficiency of the base learning model. The reason is that our calibration scheme only requires one pass over all calibration samples, as explained in Section B.2.

E Additional experiments

In this section, we provide additional experiments and supply additional results. In all experiments, whenever possible, we display the performance of an uncalibrated model, `Naive CP` which uses clean and noisy samples, `Naive CP` which uses only the clean samples, `Weighted CP` with the following weights: (i) estimated using only from X, (ii) estimated using Z, and (iii) oracle weights as a function of Z. We further employ the `Two-Staged` algorithm and `PCP` with these three options for the weight function. It is important to note that it is not applicable in practice to apply `Weighted CP` or `Two-Staged CP` with weights estimated from Z, since they require the test privileged information Z^{test}. We conduct these experiments for demonstration purposes. Nevertheless, `PCP` is applicable with any of these weights, as it does not use Z^{test}.

E.1 Causal inference tasks experiments

We follow the protocol in Section 4.1, when our goal is to estimate the uncertainty of unknown response under no treatment $Y_{n+1}(0)$. As explained in Section 4.1, valid prediction intervals for $Y_{n+1}(0), Y_{n+1}(1)$ can be combined to construct a reliable interval for the individual treatment effect (ITE), which is valuable in many applications [30–33].

We begin with the semi-synthetic IHDP [35] dataset, in which the objective is to analyze the effect of specialist home visits on future cognitive test scores. See Section C.4 for more information about this dataset. Figure 4 displays the coverage rate and interval lengths of the prediction intervals constructed by the calibration schemes. This figure shows that the naive techniques: `Uncalibrated`, naive jackknife+ (`Naive CP`), and naive `JAW`, which uses weights estimated only from X, do not achieve the desired coverage level. This is not a surprise, as the naive approaches do not hold statistical guarantees. In contrast, `JAW` which uses Z^{test}, and the proposed `PCP` construct uncertainty intervals that achieve the desired coverage level. This is also anticipated since these methods are supported by theoretical guarantees. Nevertheless, the versions of `JAW` that achieve a valid coverage rate are infeasible in practice, as they require Z^{test}, which is unknown in our setting. In conclusion, Figure 4 suggests that `PCP` is the only applicable calibration scheme that achieves a valid coverage level in this experiment.

Next, we turn to the Twins dataset [47] which contains records about newborn twin babies, and the response variable is the mortality indicator. In Section C.5 we provide additional information about this dataset. Figure 5 presents the performance of each calibration scheme on the Twins dataset. This figure indicates that `Naive CP` tends to undercover the response variable while the two-staged baseline tends to overcover it. In contrast, observe that `WCP` and `PCP` achieve the desired 90% coverage rate when employed with oracle corruption probabilities, or with probabilities estimated from Z. However, we remark that these versions of `WCP` cannot be applied in practice, since they require Z^{test}, which is unavailable in our setup. The only practical version of `WCP` is with corruption probabilities estimated from X, which does not achieve the nominal coverage level. Notice, however, that the proposed `PCP` can be applied with all versions of corruption probabilities, as it does not require access to Z^{test}. To conclude, `PCP` and `Two-Staged` are the only calibration schemes that are both applicable in practice and guaranteed to generate uncertainty sets with a valid coverage rate. In addition, Figure 5 reveals that `PCP` achieves a comparable set size to the infeasible `WCP`, indicating that `PCP` does not lose much statistical efficiency by not using the test PI Z^{test}.

Lastly, we consider the semi-synthetic National Study of Learning Mindsets (NSLM) dataset [49], which examines behavioral interventions. See [51, Section 2] for information on the dataset, and Appendix C.6 for our adaptation for this dataset. The performance of each calibration scheme is provided in Figure 6. This figure shows the same trend: the naive methods undercover the response variable while the proposed PCP constructs valid uncertainty sets.

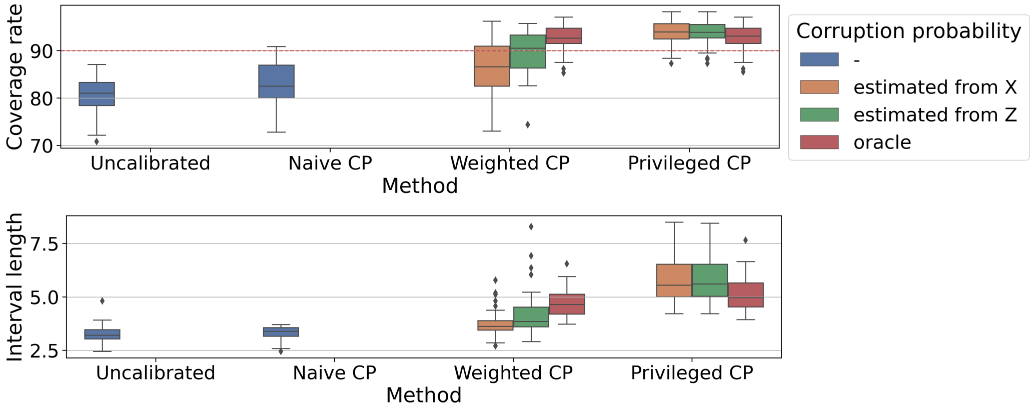

Figure 4: **IHDP dataset experiment.** The coverage rate and average interval length achieved by an uncalibrated quantile regression (Uncalibrated), a naive jackknife+ (Naive CP), JAW (Weighted CP) which estimates the corruption probability from either X (orange), Z (green), or uses the oracle probabilities (red), and the proposed method (Privileged CP) with the three options for the corruption probabilities. All methods are applied to attain a coverage rate at level $1 - 2\alpha = 90\%$. The metrics are evaluated over 50 random data splits.

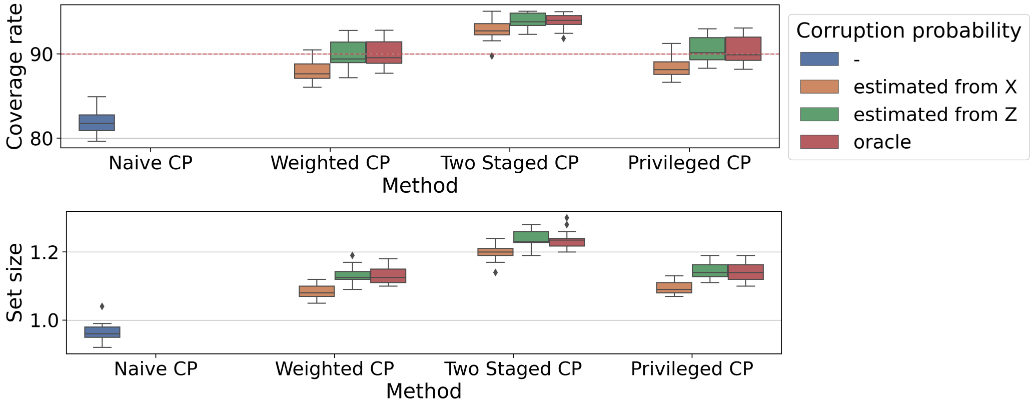

Figure 5: **Twins dataset experiment.** The coverage rate and average set size achieved by naive conformal prediction (Naive CP), Weighted CP which estimates the corruption probability from either X (orange), Z (green), or uses the oracle probabilities (red), the baseline Two Staged CP, and the proposed method (Privileged CP) with the three options for the corruption probabilities. All methods are applied to attain a coverage rate at level $1 - \alpha = 90\%$. The metrics are evaluated over 20 random data splits.

E.2 Noisy response variable: CIFAR-10N dataset

In this section, we provide additional results from the CIFAR-10N experiment conducted in Section 4.3. In Figure 7 we display the performance of naive CP, and WCP, Two-Staged, PCP, applied either with corruption probabilities estimated from X or from Z. This figure shows that WCP and PCP do not achieve the desired coverage rate when the corruption probabilities are estimated from X. In contrast, WCP and PCP attain a valid coverage rate when the corruption probabilities are estimated from Z. This is not a surprise, as it is guaranteed by [8, Theorem 1] and by Theorem 1. Lastly, this figure shows that Two-Staged constructs uncertainty sets that are too conservative. This is also anticipated since the prediction sets of Two-Staged encapsulate the uncertainty in both Z and Y. Importantly, we note that WCP cannot be used with corruption probabilities estimated from Z

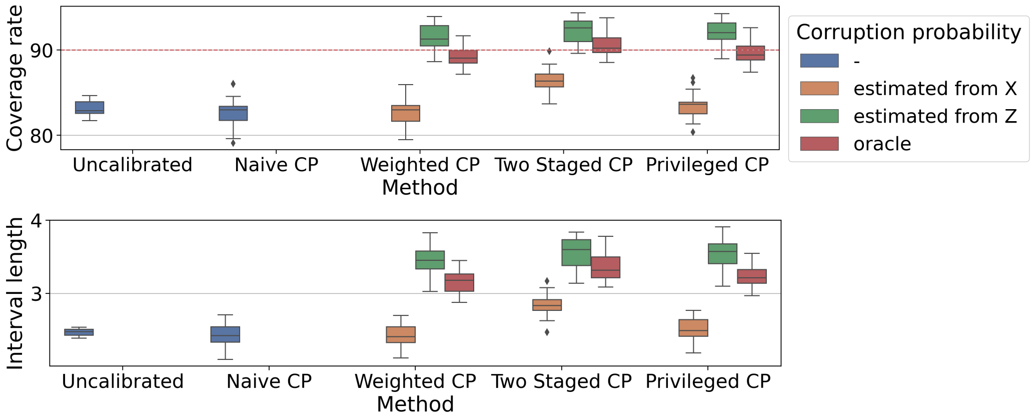

Figure 6: **NSLM dataset experiment.** The coverage rate and average interval length achieved by an uncalibrated quantile regression (`Uncalibrated`), naive conformal prediction (`Naive CP`), `Weighted CP` which estimates the corruption probability from either X (orange), Z (green), or uses the oracle probabilities (red), the baseline `Two Staged CP` and the proposed method (`Privileged CP`) with the three options for the corruption probabilities. All methods are applied to attain a coverage rate at level $1 - \alpha = 90\%$. The metrics are evaluated over 20 random data splits.

(`Weighted CP` in color green in Figure 7) since it requires access to Z^{test}, which is unknown in our setup.

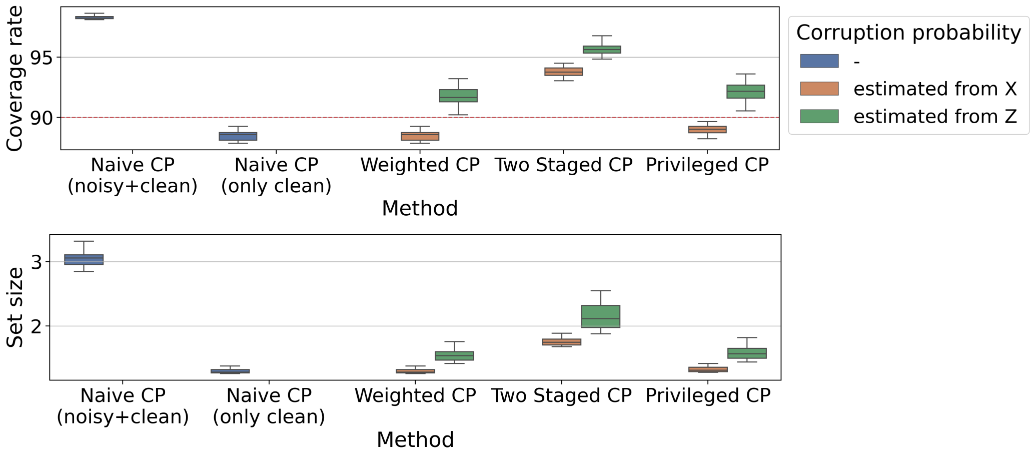

Figure 7: **Noisy response experiment: CIFAR-10N dataset.** The coverage rate and average set size length achieved by, naive conformal prediction (`Naive CP`), using either all calibration samples (noisy + clean) or only the uncorrupted ones (only clean), `Weighted CP` which estimates the corruption probability from either X (orange), Z (green), the baseline `Two Staged CP` and the proposed method (`Privileged CP`) with the two options for the corruption probabilities. All methods are applied to attain a coverage rate at level $1 - \alpha = 90\%$. The metrics are evaluated over 20 random data splits.

E.3 Noisy features and responses: CIFAR-10C dataset

We demonstrate our proposed method on the CIFAR-10C [50] dataset, which contains pairs of a noisy image with a noisy label. In short, some of the images are corrupted, e.g., with a defocus blur, and the responses of corrupted images are artificially corrupted. Nevertheless, we note that the test images are corrupted as well, in the sense that $X(0) = X(1)$ for all samples. In this experiment, we examine two options: dispersive noise, which randomly flips the label into a different one with uniform probability, and an adversarial noise, which deterministically flips the label into a different one. In Section C.3 we provide the full details about this dataset. Lastly, in the following experiments, we set the desired coverage rate to 80% since the model outputs extremely uncertain predictions for 10% of the samples.

Figure 8 displays the coverage rates and set sizes achieved by each calibration scheme, on the CIFAR-10C dataset corrupted with dispersive noise. This figure indicates that by considering the noisy labels, `Naive CP` achieves a conservative coverage rate, which is consistent with the work of [25, 43]. Nevertheless, when applied without the noisy labels, `Naive CP` tends to undercover the correct response, similarly to the two-staged baseline `Two-Staged`. Observe also that the feasible version of `WCP`, which uses weights estimated only from X achieves under coverage. In striking contrast, our proposed `PCP` covers the response at the desired coverage rate. This trend also applies for the CIFAR-10C dataset corrupted with adversarial noise, as presented in Figure 9. This figure illustrates that all feasible baselines suffer from undercoverage, except for the two-staged baseline which achieves an over-conservative coverage rate. In contrast, our proposal achieves the nominal coverage level.

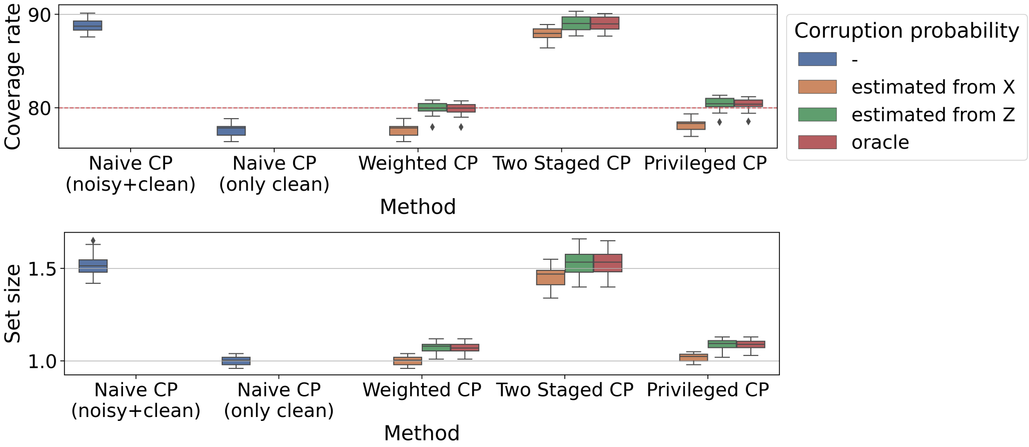

Figure 8: **Dispersive label-noise experiment: CIFAR-10C dataset.** The coverage rate and average set size length achieved by, naive conformal prediction (`Naive CP`), using either all calibration samples (noisy + clean) or only the uncorrupted ones (only clean), `Weighted CP` which estimates the corruption probability from either X (orange), Z (green), or uses the oracle probabilities (red), the baseline `Two Staged CP` and the proposed method (`Privileged CP`) with the three options for the corruption probabilities. All methods are applied to attain a coverage rate at level $1-\alpha = 90\%$. The metrics are evaluated over 20 random data splits.

E.4 Tabular data experiments

In this section, we conduct a series of experiments with the datasets: facebook1, facebook2, bio, house, meps19, and blog. In each experiment, we apply a different corruption to either the covariates or the response variable. Additionally, for the `Two-Staged`, `Infeasible WCP` and the proposed `PCP`, we use the oracle corruption probabilities when computing the weights $w(z)$. However, the `Naive WCP` method uses corruption probabilities estimated only from x using a neural network classifier. Section C.1 provides information about these datasets and Section D.1 describes the experimental setup.

E.4.1 Noisy response

We examine two artificial noise functions corrupting the response. The first noise is contractive, which reduces the variability by averaging the response with its mean: $Y(1) = \frac{1}{2}(Y(0) + \mathbb{E}[Y(0)])$. The second is dispersive, which adds to the response variable a normally distributed random noise with mean 0 and standard deviation 5 times the standard deviation of $Y(0)$.

We begin with the dispersive noise experiment. Figure 10 shows the performance of each calibration scheme in this setup. This figure indicates that the uncalibrated model and naive `CP` construct too conservative intervals. This result is consistent with the findings of [25, 43] which suggest that

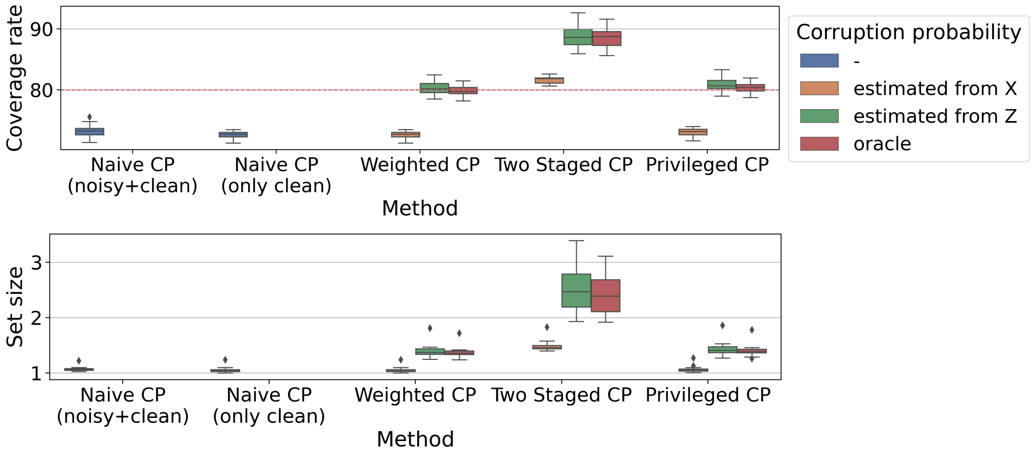

Figure 9: **Contractive label-noise experiment: CIFAR-10C dataset.** The coverage rate and average set size length achieved by, naive conformal prediction (`Naive CP`), using either all calibration samples (noisy + clean) or only the uncorrupted ones (only clean), `Weighted CP` which estimates the corruption probability from either X (orange), Z (green), or uses the oracle probabilities (red), the baseline `Two Staged CP` and the proposed method (`Privileged CP`) with the three options for the corruption probabilities. All methods are applied to attain a coverage rate at level $1-\alpha = 90\%$. The metrics are evaluated over 20 random data splits.

naively applying `CP` method on data with dispersive label-noise leads to conservative uncertainty sets. Nevertheless, `Naive WCP` achieves a coverage rate that is too low, possibly because the corruption probability estimates are not sufficiently accurate. Also, the two-staged baseline tends to output conservative intervals, as anticipated. Lastly, `Infeasible WCP` and the proposed `PCP` consistently achieve the desired coverage rate for all datasets. This is not a surprise, as a theoretical guarantee supports this result.

We now turn to the contractive noise experiment, and report in Figure 11 the coverage rate and interval lengths of the prediction intervals constructed by each calibration scheme. This figure shows that the uncalibrated model and naive `CP` generate intervals that undercover the correct outcome. This is anticipated, as the contractive noise confuses these techniques to 'think' that the underlying uncertainty is small, leading them to produce too small prediction intervals. In addition, the baselines `Naive WCP` and `Two Staged CP` construct too wide intervals, that tend to overcover the response. In contrast, `Infeasible WCP` and the proposed `PCP` achieve the desired coverage rate.

E.4.2 Missing features

In this section, we study the setting where the entries in the corrupted feature vector $X(1)$ are missing. Specifically, we artificially delete 20% of the features with the highest correlation to Y_i from $X_i(0)$ to obtain $X_i(1)$. The corruption indicator M_i is defined similarly to other experiments, as explained in Section D.1. Figure 12 shows the performance of each calibration scheme. This figure indicates that the uncalibrated model and naive `CP` tend to undercover the response variable. Also, the `Naive WCP` produces intervals with large variability. This behavior probably results from inaccurate estimates of the corruption probability $P_{M|X}$, as its oracle counterpart, `Infeasible WCP`, which uses the true corruption probabilities, precisely achieves the nominal coverage level. Furthermore, the baseline `Two Staged CP` generates too wide intervals that tend to overcover the response. In contrast, the proposed `PCP` consistently achieves the desired coverage rate $1-\alpha = 90\%$.

E.5 Ablation study of the effect of β

This section studies the effect of the parameter β on the prediction sets constructed by `PCP`. We apply `PCP` on a synthetic data, introduced in Appendix C.7 with different values of $\beta \in (0, \alpha)$. We follow the experimental protocol described in Appendix D.1, and report the coverage rate and average length achieved by `PCP` in Figure 13. This figure indicates that the smallest intervals are achieved for β that is close to 0, and the interval sizes are an increasing function of β. Yet, it is important to understand

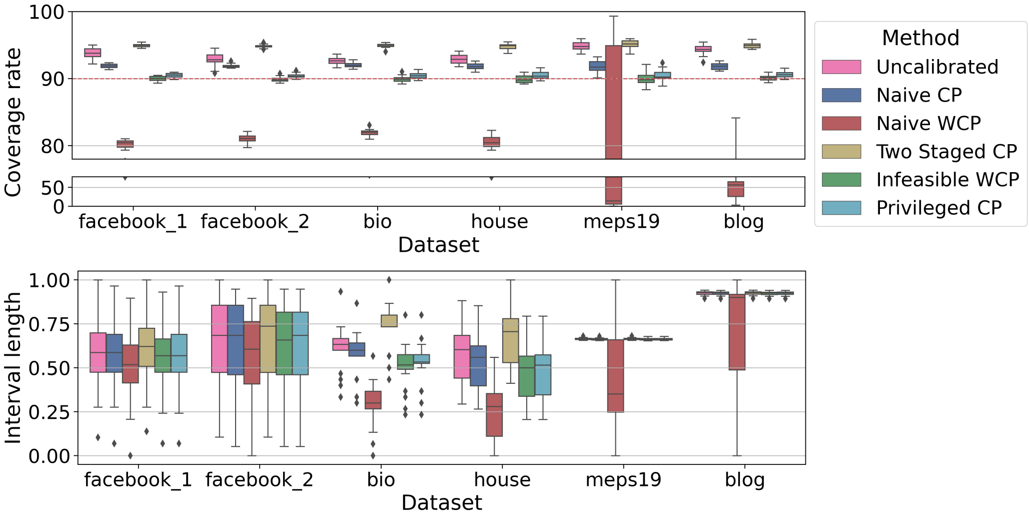

Figure 10: **Dispersive label-noise experiment: tabular datasets.** The coverage rate and average interval length achieved by an uncalibrated quantile regression (`Uncalibrated`), naive conformal prediction which uses both clean and noisy samples (`Naive CP`), `WCP` which estimates the corruption probability from X (`Naive WCP`), the baseline `Two Staged CP`, `WCP` which uses the oracle corruption probabilities (`Infeasible WCP`), and the proposed method (`Privileged CP`) that uses to the oracle corruption probabilities. All methods are applied to attain a coverage rate at level $1-\alpha = 90\%$. The metrics are evaluated over 20 random data splits.

that different results could be obtained for different datasets. Therefore, we recommend choosing β using a validation set, with a grid of values for β in $(0, \alpha)$.

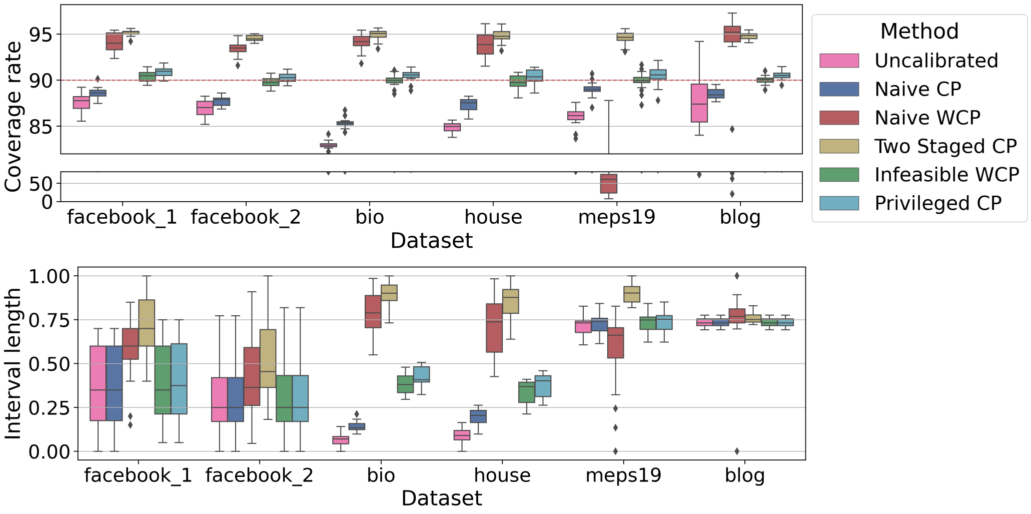

Figure 11: **Contractive label-noise experiment: tabular datasets.** The coverage rate and average interval length achieved by an uncalibrated quantile regression (`Uncalibrated`), naive conformal prediction which uses both clean and noisy samples (`Naive CP`), `WCP` which estimates the corruption probability from X (`Naive WCP`), the baseline `Two Staged CP`, `WCP` which uses the oracle corruption probabilities (`Infeasible WCP`), and the proposed method (`Privileged CP`) that uses the oracle corruption probabilities as well. All methods are applied to attain a coverage rate at level $1 - \alpha = 90\%$. The metrics are evaluated over 20 random data splits.

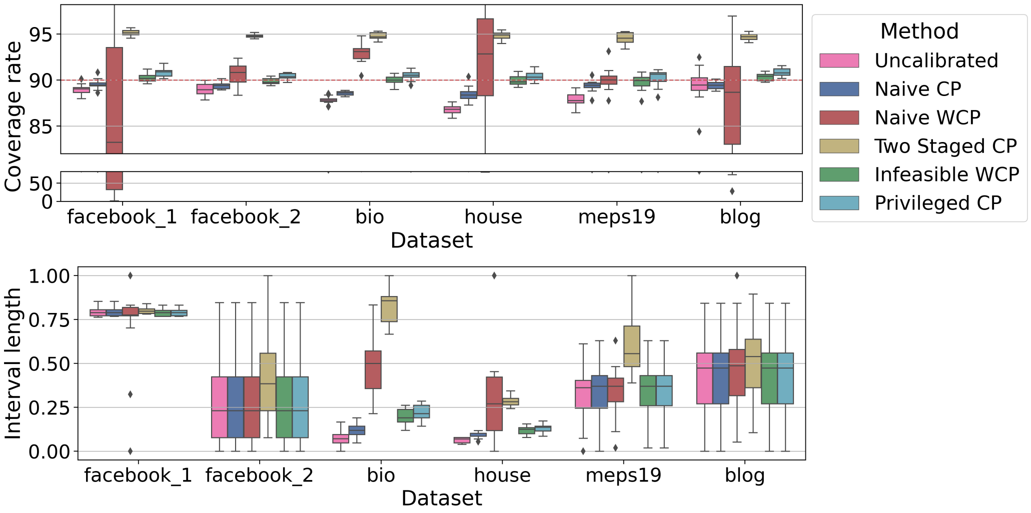

Figure 12: **Missing features experiment: tabular datasets.** The coverage rate and average interval length achieved by an uncalibrated quantile regression (`Uncalibrated`), naive conformal prediction which uses both clean and noisy samples (`Naive CP`), `WCP` which estimates the corruption probability from X (`Naive WCP`), the baseline `Two Staged CP`, `WCP` which uses the oracle corruption probabilities (`Infeasible WCP`), and the proposed method (`Privileged CP`) that uses to the oracle corruption probabilities. All methods are applied to attain a coverage rate at level $1 - \alpha = 90\%$. The metrics are evaluated over 20 random data splits.

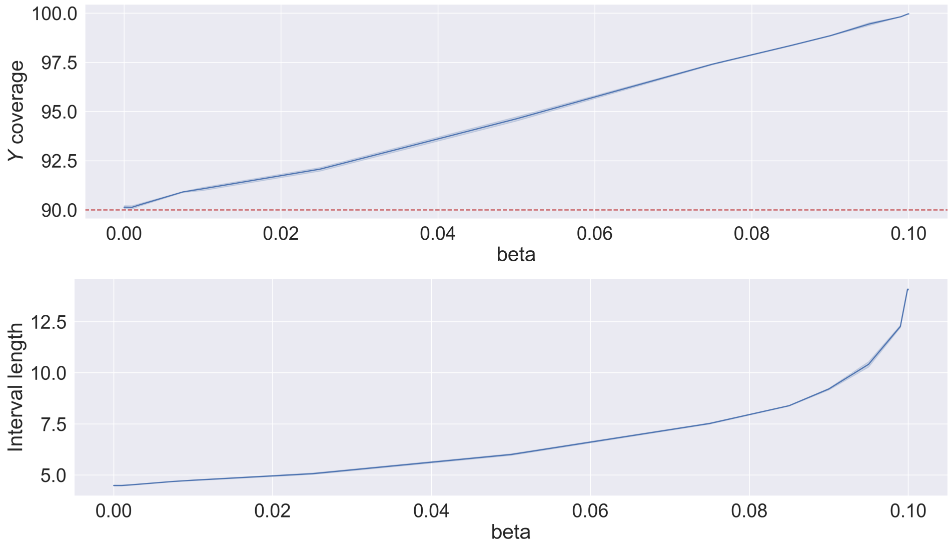

Figure 13: **Ablation study of β.** The coverage rate and average interval length achieved by `PCP` applied to attain a coverage rate at level $1 - \alpha = 90\%$. The metrics are evaluated over 20 random data splits.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The empirical experiments are provided in Section 4 and match the theoretical results in Appendix A. The problem setup is discussed in Section D.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 5 we discuss the limitations of our method, and in Sections 3.3 and 1.3 we review its computational efficiency.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theoretical claims are proved in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In Appendix D we detail the full experimental setup and in Appendix C we describe all information about the datasets we used.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The software package implementing our method and reproducing the experiments is attached to the supplementary material. The datasets we used are publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In Appendix D we detail the full experimental setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experiments are demonstrated with a boxplot which describes the variability of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Appendix D.4 we provide the spec of the machines we used and in Appendix D.5 we discuss the computation efficiency.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper does not involve human subjects or participants, and all datasets we used are publicly available.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Section 5 we discuss the potential positive and negative impacts of our method.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: In this paper, we introduce a calibration scheme that wraps predictive models. The datasets and models we use are standard, and thus we do not feel that our paper poses potential risks for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All sources we used are publicly available and we cited every asset we used in this work.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.