
Self-Calibrating Conformal Prediction

Lars van der Laan
University of Washington
lvdlaan@uw.edu

Ahmed M. Alaa
UC Berkeley and UCSF
amaalaa@berkeley.edu

Abstract

In machine learning, model calibration and predictive inference are essential for producing reliable predictions and quantifying uncertainty to support decision-making. Recognizing the complementary roles of point and interval predictions, we introduce *Self-Calibrating Conformal Prediction*, a method that combines Venn-Abers calibration and conformal prediction to deliver calibrated point predictions alongside prediction intervals with finite-sample validity conditional on these predictions. To achieve this, we extend the original Venn-Abers procedure from binary classification to regression. Our theoretical framework supports analyzing conformal prediction methods that involve calibrating model predictions and subsequently constructing conditionally valid prediction intervals on the same data, where the conditioning set or conformity scores may depend on the calibrated predictions. Real-data experiments show that our method improves interval efficiency through model calibration and offers a practical alternative to feature-conditional validity.

1 Introduction

Particularly in safety-critical sectors, such as healthcare, it is important to ensure decisions inferred from machine learning models are reliable, under minimal assumptions [38, 57, 11, 46, 56]. As a response, there is growing interest in predictive inference methods that quantify uncertainty in model predictions via prediction intervals [44, 28]. Conformal prediction (CP) is a popular, model-agnostic, and distribution-free framework for predictive inference, which can be applied post-hoc to any prediction pipeline [59, 49, 3, 34]. Given a prediction issued by a black-box model, CP outputs a prediction interval that is guaranteed to contain the unseen outcome with a user-specified probability [34]. However, a limitation of CP is that this prediction interval only provides valid coverage marginally, when averaged across all possible contexts – with ‘context’ referring to the information available for decision-making. Constructing informative prediction intervals that offer context-conditional coverage is generally unattainable without making additional distributional assumptions [58, 35, 4]. Consequently, there has been an upsurge in research developing CP methods that offer weaker, yet practically useful, notions of conditional validity; see, e.g., [43, 31, 47, 32, 23, 21].

In prediction settings, *model calibration* is a desirable property of machine learning predictors that ensures that the predicted outcomes accurately reflect the true outcomes [39, 65, 66, 22]. Specifically, a predictor is calibrated for the outcome if the average outcome among individuals with identical predictions is close to their shared prediction value [25]. Such a predictor is more robust against the over-or-under estimation of the outcome in extremes of predicted values. It also has the property that the best prediction of the outcome conditional on the model’s prediction is the prediction itself, which facilitates transparent decision-making [51]. There is a rich literature studying post-hoc calibration of prediction algorithms using techniques such as Platt’s scaling [45, 14], histogram binning [65, 25, 26], isotonic calibration [66, 40, 51], and Venn-Abers calibration [60].

Given the roles of both point and interval predictions in decision-making, we introduce a dual calibration objective that aims to construct (i) calibrated point predictions and (ii) associated prediction intervals with valid coverage conditional on these point predictions. Marrying model calibration and

predictive inference, we propose a solution to this objective that combines two post-hoc approaches — Venn-Abers calibration [61, 60] and CP [59] — to simultaneously provide point predictions and prediction intervals that achieve our dual objective in finite samples. In doing so, we extend the original Venn-Abers procedure from binary classification to the regression setting. Our theoretical and experimental results support the integration of model calibration into predictive inference methods to improve interval efficiency and interpretability.

2 Problem setup

2.1 Notation

We consider a standard regression setup in which the input $X \in \mathcal{X} \subset \mathbb{R}^d$ corresponds to contextual information available for decision-making, and the output $Y \in \mathcal{Y} \subset \mathbb{R}$ is an outcome of interest. We assume that we have access to a calibration dataset $\mathcal{C}_n = \{(X_i, Y_i)\}_{i=1}^n$ comprising n *i.i.d.* data points drawn from an unknown distribution $P := P_X P_{Y|X}$. We assume access to a black-box predictor $f : \mathcal{X} \mapsto \mathcal{Y}$, obtained by training an ML model on a dataset that is independent of \mathcal{C}_n . Throughout this paper, we do not make any assumptions on the model f or the distribution P . For a quantile level $\alpha \in (0, 1)$, we denote the “pinball” quantile loss function ℓ_α by $\ell_\alpha(f(x), y) := 1(y \geq f(x)) \cdot \alpha(y - f(x)) + 1(y < f(x)) \cdot (1 - \alpha)(f(x) - y)$.

2.2 Conditional predictive inference and a curse of dimensionality

Let (X_{n+1}, Y_{n+1}) be a new data point drawn from P independently of the calibration data \mathcal{C}_n . **Our high-level aim** is to develop a *predictive inference* algorithm that constructs a prediction interval $\widehat{\mathcal{C}}_n(X_{n+1})$ around the point prediction issued by the black-box model, i.e., $f(X_{n+1})$. For this prediction interval to be deemed *valid*, it should *cover* the true outcome Y_{n+1} with a probability $1 - \alpha$. Conformal prediction (CP) is a method for predictive inference that can be applied in a post-hoc fashion to any black-box model [59]. The vanilla CP procedure issues prediction intervals that satisfy the marginal coverage condition:

$$\mathbb{P}(Y_{n+1} \in \widehat{\mathcal{C}}_n(X_{n+1})) \geq 1 - \alpha, \tag{1}$$

where the probability \mathbb{P} is taken with respect to the randomness in \mathcal{C}_n and (X_{n+1}, Y_{n+1}) . However, marginal coverage might lack utility in decision-making scenarios where decisions are context-dependent. A prediction band $\widehat{\mathcal{C}}_n(x)$ achieving 95% coverage may exhibit arbitrarily poor coverage for specific contexts x . Ideally, we would like this coverage condition to hold for each context $x \in \mathcal{X}$, i.e., the conventional notion of “conditional validity” requires

$$\mathbb{P}(Y_{n+1} \in \widehat{\mathcal{C}}_n(X_{n+1}) | X_{n+1} = x) \geq 1 - \alpha, \tag{2}$$

for all $x \in \mathcal{X}$. However, previous work has shown that it is impossible to achieve (2) without distributional assumptions [58, 35, 4].

While context-conditional validity as in (2) is generally unachievable, it is feasible to attain weaker forms of conditional validity. Given any finite set of groups \mathcal{G} and a grouping function $G : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{G}$, Mondrian-CP offers coverage conditional on group membership, that is, $\mathbb{P}(Y_{n+1} \in \widehat{\mathcal{C}}_n(X_{n+1}) | G(X_{n+1}, Y_{n+1}) = g) \geq 1 - \alpha, \forall g \in \mathcal{G}$ [59, 47]. Expanding upon group- and context-conditional coverage, A *multicalibration* objective was introduced in [17] that seeks to satisfy, for all h in an (infinite-dimensional) class \mathcal{F} of weighting functions (i.e., ‘covariate shifts’), the property:

$$\mathbb{E}[h(X_{n+1})\{(1 - \alpha) - 1\{Y_{n+1} \in \widehat{\mathcal{C}}_n(X_{n+1})\}\}] = 0. \tag{3}$$

Gibbs et al. [21] proposed a regularized CP framework for (approximately) achieving (3) that provides a means to trade off the efficiency (i.e., width) of prediction intervals and the degree of conditional coverage achieved. However, Barber et al. [4] and Gibbs et al. [21] establish the existence of a “curse of dimensionality”: as the dimension of the context increases, smaller classes of weighting functions must be considered to retain the same level of efficiency. For group-conditional coverage, this curse of dimensionality manifests in the size of the subgroup class \mathcal{G} [4] via its VC dimension [55]. Thus, especially in data-rich contexts, prediction intervals with meaningful multicalibration guarantees over the context space may be too wide for decision-making.

2.3 A dual calibration objective

In decision-making, both point predictions and prediction intervals play a role. For example, in scenarios with a low signal-to-noise ratio, prediction intervals may be too wide to directly inform decision-making, as their width is typically of the order of the standard deviation of the outcome. Point predictions might be used to guide decisions, while prediction intervals help quantify deviations of point predictions from unseen outcomes and assess the risk associated with these decisions.

Viewing the black-box model f as a scalar dimension reduction of the context x , a natural relaxation of the infeasible objective of context-conditional validity in (2) is prediction-conditional validity, i.e., $P(Y_{n+1} \in \hat{C}_n(X_{n+1}) | f(X_{n+1})) \geq 1 - \alpha$. Prediction-conditional validity ensures that the interval widths adapt to the outputs of the model $f(\cdot)$, so that the intervals can be reliably used to quantify the deviation of model predictions from unseen outcomes. Since prediction-conditional validity only requires coverage conditional on a one-dimensional random variable, it avoids the curse of dimensionality associated with context-conditional validity. In addition, as illustrated in our experiments in Section 5 and Appendix C, when the heteroscedasticity (e.g., variance) in the outcome is a function of its conditional mean, prediction-conditional validity can closely approximate context-conditional validity, so long as the predictor estimates the conditional mean of the outcome sufficiently well.

Given the roles of both point and interval predictions in decision-making, we introduce a novel dual calibration objective, *self-calibration*, that aims to construct (i) calibrated point predictions and (ii) associated prediction intervals with valid coverage conditional on these point predictions. Formally, given the model f and calibration data $C_n \cup \{X_{n+1}\}$, **our objective** is to post-hoc construct a calibrated point prediction $f_{n+1}(X_{n+1})$ and a compatible prediction interval $\hat{C}_{n+1}(X_{n+1})$ centered around $f_{n+1}(X_{n+1})$ that satisfies the following desiderata:

- (i) **Perfectly Calibrated Point Prediction:** $f_{n+1}(X_{n+1}) = \mathbb{E}[Y_{n+1} | f_{n+1}(X_{n+1})]$.
- (ii) **Prediction-Conditional Validity:** $\mathbb{P}(Y_{n+1} \in \hat{C}_{n+1}(X_{n+1}) | f_{n+1}(X_{n+1})) \geq 1 - \alpha$.

Desideratum (i) states that the point prediction $f_{n+1}(X_{n+1})$ should be *perfectly calibrated* — or self-consistent [19] — for the true outcome Y_{n+1} [37, 25, 51]. It is widely recognized that the calibration of model predictions is important to ensure their reliability, trustworthiness, and interpretability in decision-making [37, 65, 7, 24, 15]. Desideratum (i) also improves interval efficiency by ensuring $\hat{C}_{n+1}(X_{n+1})$ is centered around an unbiased prediction, meaning the interval’s width is driven by outcome variation rather than by prediction bias. Desideratum (ii) is a prediction interval variant of (i) that ensures the prediction interval $\hat{C}_{n+1}(X_{n+1})$ is calibrated with respect to the model f_{n+1} , providing valid coverage for Y_{n+1} within contexts with the same *calibrated* point prediction. We refer to a predictive inference algorithm simultaneously satisfying (i) and (ii) as *self-calibrating*, as such a procedure is automatically able to adapt to miscalibration in the model $f(\cdot)$ due to, e.g., model misspecification or distribution shifts, ensuring that the interval is constructed from a calibrated predictive model.

Self-calibration can also be motivated by a decision-making scenario where point predictions determine actions and prediction intervals are used to apply these actions selectively. When point predictions are sufficient statistics for actions, self-calibration implies that the point and interval predictions are accurate, on average, within the subset of all contexts receiving the same prediction and, therefore, the same action.

3 Self-Calibrating Conformal Prediction

A key advantage of CP is that it can be applied post-hoc to any black-box model f without disrupting its point predictions. However, desideratum (i) introduces a perfect calibration requirement for the point predictions of f , thereby interfering with the underlying model specification. In this section, we introduce Self-Calibrating CP (SC-CP), a modified version of CP that is self-calibrating in that it satisfies (i) and (ii), while preserving all the favorable properties of CP, including its finite-sample validity and post-hoc applicability. Before describing our complete procedure in Section 3.3, we provide background on point calibration and propose Venn-Abers calibration for regression.

3.1 Preliminaries on point calibration

Following the framing of van der Laan et al. [51] (see also [25]), a *point calibrator* is a post-hoc procedure that aims to learn a transformation $\theta_n : \mathbb{R} \rightarrow \mathbb{R}$ of the black-box model f such that: (1) $\theta_n(f(X_{n+1}))$ is well-calibrated for Y_{n+1} in the sense of Desideratum (i); and (2) $\theta_n \circ f$ is comparably predictive to f . Condition (2) ensures that in the process of achieving (1), the quality of the model f is not compromised, and excludes trivial calibrators such as $\theta_n(f(\cdot)) := \frac{1}{n} \sum_{i=1}^n Y_i$ [26]. To our knowledge, this notion of calibration traces back to Mincer and Zarnowitz (1969) [39], who introduced the idea of regressing outcomes on predictions to achieve calibration in forecasting.

Commonly-employed point calibrators include Platt’s scaling [45, 14], histogram (or quantile) binning [65], and isotonic calibration [66, 40]. Mechanistically, these point calibrators learn θ_n by regressing the outcomes $\{Y_i\}_{i=1}^n$ on the model predictions $\{f(X_i)\}_{i=1}^n$. Importantly, however, point calibration fundamentally differs from the regression task of learning $E_P[Y|f(X)]$, as calibration can be achieved without smoothness assumptions, allowing for misspecification of the regression task [25]. Histogram binning is a simple and distribution-free calibration procedure [25, 26] that learns θ_n via a histogram regression over a finite (outcome-agnostic) binning of the output space $f(\mathcal{X})$. Isotonic calibration is an outcome-adaptive binning method that uses isotonic regression [5] to learn θ_n by minimizing the empirical mean square error over all 1D piece-wise constant, monotone nondecreasing transformations. Isotonic calibration is distribution-free — it does not rely on monotonicity assumptions — and, in contrast with histogram binning, it is tuning parameter-free and naturally preserves the mean-square error of the original predictor (as the identity transform is monotonic) [51]. A key limitation of histogram binning and isotonic calibration is that their calibration guarantees are only approximate, and desideratum (i) only holds asymptotically.

3.2 Venn-Abers calibration

For binary classification, Vovk and Petej [60] proposed Venn-Abers calibration, which iterates isotonic calibration over imputations $y \in \mathcal{Y}$ of the unseen outcome Y_{n+1} to provide calibrated multi-probabilistic predictions in finite samples. In this section, we generalize the Venn-Abers calibration procedure to regression, offering finite-sample calibration guarantees for non-binary outcomes.

Let Θ_{iso} consist of all univariate, piecewise constant functions that are monotonically nondecreasing. Our Venn-Abers calibration procedure, outlined in Alg. 1, is derived from an *oracle* variant of isotonic calibration that provides a perfectly calibrated point prediction *in finite samples*, but requires knowledge of the true outcome Y_{n+1} . Specifically, the Venn-Abers calibration algorithm iterates over imputed outcomes $y \in \mathcal{Y}$ for Y_{n+1} and applies isotonic calibration to the augmented dataset $C_n \cup \{(X_{n+1}, y)\}$ to produce a set of point predictions $f_{n, X_{n+1}}(X_{n+1}) := \{f_n^{(X_{n+1}, y)}(X_{n+1}) : y \in \mathcal{Y}\}$. When the outcome space \mathcal{Y} is non-discrete, Alg. 1 may be infeasible to compute exactly and can be approximated by discretizing \mathcal{Y} . Nonetheless, the range of the Venn-Abers multi-prediction can be feasibly computed as $[f_n^{(x, y_{\min})}(x), f_n^{(x, y_{\max})}(x)]$ where $[y_{\min}, y_{\max}] := \text{range}(\mathcal{Y})$, in light of the min-max representation of isotonic regression [33].

Unlike point calibrators, Venn-Abers calibration generates a set of calibrated predictions for each context X_{n+1} , indexed by $y \in \mathcal{Y}$. As we demonstrate later, this set prediction is guaranteed *in finite samples* to include a perfectly calibrated point prediction, namely, the oracle prediction $f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1})$ corresponding to the true outcome Y_{n+1} . Moreover, each prediction in the set, being obtained via isotonic calibration, still enjoys the same large-sample calibration guarantees as isotonic calibration [51]. By the stability of isotonic regression, as the size of the calibration set n increases, the width of this set of predictions rapidly narrows, eventually converging to a single, perfectly calibrated point prediction [60]. Venn-Abers calibration thus provides a measure of epistemic uncertainty by producing a range of values for a perfectly calibrated point prediction. In cases with small sample sizes, standard isotonic calibration can overfit, leading to poorly calibrated point predictions. When this overfitting occurs, the Venn-Abers set prediction widens, reflecting greater uncertainty in the perfectly calibrated point prediction within the set [30].

For the binary classification case, [60] derived a (large-sample) calibrated point prediction from the Venn-Abers multi-prediction using a shrinkage approach. We can similarly construct, for each $x \in \mathcal{X}$,

Algorithm 1 Venn-Abers Calibration

Input: Calibration data $\mathcal{C}_n = \{(X_i, Y_i)\}_{i=1}^n$, model f , context $x \in \mathcal{X}$.

- 1: **for** each $y \in \mathcal{Y}$ **do**
- 2: Set augmented dataset $\mathcal{C}_n^{(x,y)} := \mathcal{C}_n \cup \{(x, y)\}$;
- 3: Apply isotonic calibration to f using $\mathcal{C}_n^{(x,y)}$:

$$\theta_n^{(x,y)} := \operatorname{argmin}_{\theta \in \Theta_{\text{iso}}} \sum_{i \in \mathcal{C}_n^{(x,y)}} \{Y_i - \theta \circ f(X_i)\}^2.$$

$$f_n^{(x,y)} := \theta_n^{(x,y)} \circ f.$$
- 4: **end for**

Output: Multi-prediction $\{f_n^{(x,y)}(x) : y \in \mathcal{Y}\}$.

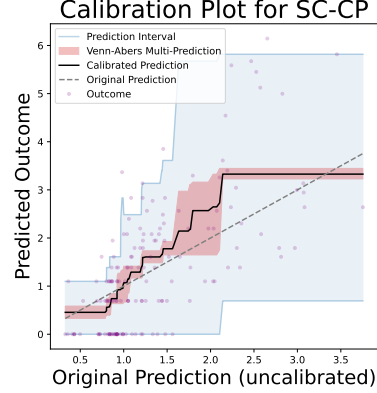


Figure 1: Example SC-CP output with small \mathcal{C}_n ($n = 200$).

Algorithm 2 Self-Calibrating Conformal Prediction

Input: Calibration data $\mathcal{C}_n = \{(X_i, Y_i)\}_{i=1}^n$, model f , context $x \in \mathcal{X}$, miscoverage level $\alpha \in (0, 1)$

- 1: **for** each $y \in \mathcal{Y}$ **do**
- 2: Obtain calibrated model $f_n^{(x,y)}$ by isotonic calibrating f on $\mathcal{C}_n \cup \{(x, y)\}$ as in Alg. 1;
- 3: Set self-calibrating conformity scores $S_i^{(x,y)} = |Y_i - f_n^{(x,y)}(X_i)|, \forall i \in [n]$ and $S_{n+1}^{(x,y)} = |y - f_n^{(x,y)}(x)|$;
- 4: Calculate $1 - \alpha$ empirical quantile $\rho_n^{(x,y)}(x)$ of conformity scores with same calibrated prediction as x :

$$\operatorname{argmin}_{q \in \mathbb{R}} \sum_{i=1}^n \mathbf{1}\{f_n^{(x,y)}(X_i) = f_n^{(x,y)}(x)\} \cdot \ell_\alpha(q, S_i^{(x,y)}) + \ell_\alpha(q, S_{n+1}^{(x,y)});$$

- 5: **end for**
- 6: Set $f_{n+1}(x) := \{f_n^{(x,y)}(x) : y \in \mathcal{Y}\}$.
- 7: Set $\widehat{\mathcal{C}}_{n+1}(x) := \{y \in \mathcal{Y} : |y - f_n^{(x,y)}(x)| \leq \rho_n^{(x,y)}(x)\}$.

Output: $f_{n+1}(x) \subset \operatorname{conv}(\mathcal{Y}), \widehat{\mathcal{C}}_{n+1}(x) \subset \mathcal{Y}$

a point prediction as follows:

$$\widetilde{f}_{n+1,x}(x) := f_{n+1,x}^{\text{mid}}(x) + \frac{f_n^{(x,y_{\max})}(x) - f_n^{(x,y_{\min})}(x)}{y_{\max} - y_{\min}} \{\bar{y}_n - f_{n+1,x}^{\text{mid}}(x)\}, \quad (4)$$

where $f_{n+1,x}^{\text{mid}}(x) := \frac{1}{2} \{f_n^{(x,y_{\max})}(x) + f_n^{(x,y_{\min})}(x)\}$ is the midpoint of the multi-prediction and $\bar{y}_n := \frac{1}{n} \sum_{i=1}^n Y_i$. The behavior of $\widetilde{f}_{n+1,x}(x)$ is natural; it shrinks the point prediction $f_{n+1,x}^{\text{mid}}(x)$ towards the average outcome \bar{y}_n (a well-calibrated prediction) proportional to how uncertain we are in the calibration of $f_{n+1,x}^{\text{mid}}(x)$. The ratio $\frac{1}{y_{\max} - y_{\min}} (f_n^{(x,y_{\max})}(x) - f_n^{(x,y_{\min})}(x))$ measures the sensitivity of isotonic regression to the addition of a single data point to \mathcal{C}_n , and a value closer to 1 corresponds to a higher degree of overfitting. In the extreme case where the calibration dataset is very large, we have $f_n^{(x,y_{\max})}(x) \approx f_n^{(x,y_{\min})}(x)$, implying that $\widetilde{f}_{n+1,x}(x) \approx f_{n+1,x}^{\text{mid}}(x)$. Conversely, in the opposite extreme where the calibration dataset is very small and isotonic regression overfits, we have $f_n^{(x,y_{\max})}(x) \approx y_{\max}$ and $f_n^{(x,y_{\min})}(x) \approx y_{\min}$, in which case $\widetilde{f}_{n+1,x}(x) \approx \bar{y}_n$. We could replace \bar{y}_n in (4) with any reference predictor, such as one calibrated using Platt's scaling or quantile-binning.

3.3 Conformalizing Venn-Abers Calibration

In this section, we propose Self-Calibrating Conformal Prediction, which conformalizes the Venn-Abers calibration procedure to provide prediction intervals centered around the Venn-Abers multi-prediction that are self-calibrated in the sense of desiderata (i) and (ii).

A simple, albeit naive, strategy for achieving (i) and (ii) without finite-sample guarantees involves using the dataset \mathcal{C}_n to calibrate point predictions of $f(\cdot)$ through isotonic regression, and then constructing prediction intervals from the $1 - \alpha$ empirical quantiles of prediction errors within subgroups defined by unique values of the calibrated point predictions. To motivate our SC-CP algorithm, we introduce an infeasible variant of this seemingly naive procedure that is valid in finite samples, but can only be computed by an oracle that knows the unseen outcome Y_{n+1} . In

this oracle procedure, we compute a perfectly calibrated prediction $f_{n+1}^*(X_{n+1}) := \theta_{n+1}^*(f(X_{n+1}))$ by isotonic calibrating f using the oracle-augmented calibration set $\{(X_i, Y_i)\}_{i=1}^{n+1}$, where $\theta_{n+1}^* \in \operatorname{argmin}_{\theta \in \Theta_{\text{iso}}} \sum_{i=1}^{n+1} \{Y_i - \theta(f(X_i))\}^2$. Next, we compute the conformity scores $S_i^* := |Y_i - f_{n+1}^*(X_i)|$ as the absolute residuals of the calibrated predictions. An oracle prediction interval is then given by $C_{n+1}^*(X_{n+1}) := f_{n+1}^*(X_{n+1}) \pm \rho_{n+1}^*(X_{n+1})$, where $\rho_{n+1}^*(X_{n+1})$ is the empirical $1 - \alpha$ quantile of conformity scores with calibrated predictions identical to X_{n+1} , that is, scores in the set $\{S_i^* : f_{n+1}^*(X_i) = f_{n+1}^*(X_{n+1}), i \in [n+1]\}$. Importantly, isotonic regression, which is an outcome-adaptive histogram binning method, ensures that the calibrated model f_{n+1}^* is piece-wise constant, with a sufficiently large number of observations averaged within each constant segment—typically on the order of $n^{2/3}$ [16]. Consequently, the empirical quantile $\rho_{n+1}^*(X_{n+1})$ is generally stable with relatively low variability across realizations of \mathcal{C}_n . In our proofs, using the first-order conditions characterizing the optimizer θ_{n+1} and exchangeability, we show that $f_{n+1}^*(X_{n+1}) = \mathbb{E}[Y_{n+1} | f_{n+1}^*(X_{n+1})]$, so that desideratum (i) is satisfied. Furthermore, we establish that the interval $C_{n+1}^*(X_{n+1})$ achieves desideratum (ii), i.e., $\mathbb{P}(Y_{n+1} \in C_{n+1}^*(X_{n+1}) | f_{n+1}^*(X_{n+1})) \geq 1 - \alpha$. To do so, our key insight is that $\rho_{n+1}^*(X_{n+1})$ corresponds to the evaluation of the function ρ_{n+1}^* computed via prediction-conditional quantile regression as:

$$\rho_{n+1}^* \in \operatorname{argmin}_{\theta \circ f_{n+1}^*; \theta: \mathbb{R} \rightarrow \mathbb{R}} \frac{1}{n+1} \sum_{i=1}^{n+1} \ell_\alpha(\theta \circ f_{n+1}^*(X_i), S_i).$$

The first-order conditions characterizing the optimizer ρ_{n+1}^* combined with the exchangeability between \mathcal{C}_n and (X_{n+1}, Y_{n+1}) can be used to show that $C_{n+1}^*(X_{n+1})$ is *multi-calibrated* against the class of weighting functions $\mathcal{F}_{n+1} := \{\theta \circ f_{n+1}^*; \theta: \mathbb{R} \rightarrow \mathbb{R}\}$ in the sense of (3). Using first-order conditions to establish the theoretical validity of conformal prediction was also applied by Gibbs et al. [21] to demonstrate the multi-calibration of oracle prediction intervals obtained from quantile regression over a fixed class \mathcal{F} . In our case, quantile regression is performed over a data-dependent function class \mathcal{F}_{n+1} , learned from the calibration data, which introduces additional challenges in our proofs.

Our SC-CP method, which is outlined in Alg. 2, follows a similar procedure to the above oracle procedure. Since the new outcome Y_{n+1} is unobserved, we instead iterate the oracle procedure over all possible imputed values $y \in \mathcal{Y}$ for Y_{n+1} . As in Alg. 1., this yields a set of isotonic calibrated models $f_{n, X_{n+1}} := \{f_n^{(X_{n+1}, y)} : y \in \mathcal{Y}\}$, where $f_{n+1}(X_{n+1})$ is the Venn-Abers multi-prediction of Y_{n+1} . Then, for each $y \in \mathcal{Y}$ and $i \in [n]$, we define the *self-calibrating conformity scores* $S_i^{(X_{n+1}, y)} := |Y_i - f_n^{(X_{n+1}, y)}(X_i)|$ and $S_{n+1}^{(X_{n+1}, y)} := |y - f_n^{(X_{n+1}, y)}(X_{n+1})|$, where the dependency of our scores on the imputed outcome $y \in \mathcal{Y}$ is akin to Full (or transductive) CP [59]. Our SC-CP interval is then given by $\widehat{C}_{n+1}(X_{n+1}) := \{y \in \mathcal{Y} : S_{n+1}^{(X_{n+1}, y)} \leq \rho_n^{(X_{n+1}, y)}(X_{n+1})\}$, where $\rho_n^{(X_{n+1}, y)}(X_{n+1})$ is the empirical $1 - \alpha$ quantile of the level set $\{S_i^{(X_{n+1}, y)} : f_n^{(X_{n+1}, y)}(X_i) = f_n^{(X_{n+1}, y)}(X_{n+1}), i \in [n+1]\}$. By definition, $\widehat{C}_{n+1}(X_{n+1})$ covers Y_{n+1} if, and only if, the oracle interval $C_{n+1}^*(X_{n+1})$ covers Y_{n+1} , thereby inheriting the self-calibration of $C_{n+1}^*(X_{n+1})$. Formally, $\widehat{C}_{n+1}(X_{n+1})$ is a set, but it can be converted to an interval by taking its range, with little efficiency loss.

3.4 Computational considerations

The main computational cost of Alg. 1 and Alg. 2 is in the isotonic calibration step, executed for each $y \in \mathcal{Y}$. Isotonic regression [5] can be scalably and efficiently computed using implementations of `xgboost` [12] for univariate regression trees with monotonicity constraints. Similar to Full (or transductive) CP [59], Alg. 2 may be computationally infeasible for non-discrete outcomes, and can be approximated by iterating over a finite subset of \mathcal{Y} . In our implementation, we iterate over a grid of \mathcal{Y} and use linear interpolation to impute the threshold $\rho_n^{(x, y)}(x)$ and score $S_{n+1}^{(x, y)}$ for each $y \in \mathcal{Y}$. As with Full CP and multicalibrated CP [21], Alg. 1 and Alg. 2 must be separately applied for each context $x \in \mathcal{X}$. The algorithms depend solely on $x \in \mathcal{X}$ through its prediction $f(x)$, so we can approximate the outputs for all $x \in \mathcal{X}$ by running each algorithm for a finite number of $x \in \mathcal{X}$ corresponding to a finite grid of the 1D output space $f(\mathcal{X}) = \{f(x) : x \in \mathcal{X}\} \subset \mathbb{R}$. In addition, both algorithms are fully parallelizable across both the input context $x \in \mathcal{X}$ and the imputed outcome $y \in \mathcal{Y}$. In our implementation, we use nearest neighbor interpolation in the prediction space to impute outputs for each $x \in \mathcal{X}$. In our experiments with sample sizes ranging from $n = 5000$

to 40000, quantile binning of both $f(\mathcal{X})$ and \mathcal{Y} into 200 equal-frequency bins enables execution of Alg. 1 and Alg. 2 across all contexts in minutes with negligible approximation error.

4 Theoretical guarantees

In this section, under exchangeability of the data, we establish that the Venn-Abers multi-prediction $f_{n, X_{n+1}}(X_{n+1}) := \{f_n^{(X_{n+1}, y)}(X_{n+1}) : y \in \mathcal{Y}\}$ and SC-CP interval $\widehat{C}_{n+1}(X_{n+1})$ output by Alg. 2 satisfy desiderata (i) and (ii) in finite samples and without distributional assumptions. Under an *iid* condition, we further establish that, asymptotically, the Venn-Abers calibration step within the SC-CP algorithm results in better point predictions and, consequently, more efficient prediction intervals.

The following theorem establishes that the Venn-Abers multi-prediction is perfectly calibrated in the sense of [61], containing a perfectly calibrated point prediction of Y_{n+1} in finite samples.

C1) Exchangeability: $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable.

C2) Finite second moment: $E_P[Y^2] < \infty$.

Theorem 4.1 (Perfect calibration of Venn-Abers multi-prediction). *Under Conditions C1 and C2, the Venn-Abers multi-prediction $f_{n, X_{n+1}}(X_{n+1})$ almost surely satisfies the condition $f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) = \mathbb{E}[Y_{n+1} | f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1})]$.*

Theorem 4.1 generalizes an analogous result by [60] for the special case of binary classification. Even in this special case, our proof is novel and elucidates how Venn-Abers calibration uses exchangeability with the least-squares loss in a manner analogous to how CP uses exchangeability with the quantile loss [21].

The following theorem establishes desideratum (ii) for the interval $\widehat{C}_{n+1}(X_{n+1})$ with respect to the oracle prediction $f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1})$ of Theorem 4.1. In what follows, let $\text{polylog}(n)$ be a given sequence that grows polynomially logarithmically in n .

C3) The conformity scores $|Y_i - f_n^{(X_{n+1}, Y_{n+1})}(X_i)|, \forall i \in [n+1]$, are almost surely distinct.

C4) The number of constant segments for $f_n^{(X_{n+1}, Y_{n+1})}$ is at most $n^{1/3} \text{polylog}(n)$.

Theorem 4.2 (Self-calibration of prediction interval). *Under C1, it holds almost surely that*

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

If also C3 and C4 hold, then $\mathbb{E}\left|\alpha - \mathbb{P}\left(Y_{n+1} \notin \widehat{C}_{n+1}(X_{n+1}) \mid f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1})\right)\right| \leq \frac{\text{polylog}(n)}{n^{2/3}}$.

Theorem 4.2 says that $\widehat{C}_{n+1}(X_{n+1})$ satisfies desideratum (ii) with coverage that is, on average, nearly exact up to a factor $\frac{\text{polylog}(n)}{n^{2/3}}$. Notably, the deviation error from exact coverage tends to zero at a fast dimensionless rate and, therefore, does not suffer from a ‘‘curse of dimensionality’’. Condition C3 is only required to establish the upper coverage bound and is standard in CP - see, e.g., [36, 21]. Although it may fail for non-continuous outcomes, this condition can be avoided by adding a small amount of noise to all outcomes [36]. The constant segment number of $n^{1/3} \text{polylog}(n)$ in C4 is motivated by the theoretical properties of isotonic regression; Assuming C2 and continuous differentiability of $t \mapsto E_P[Y | f(X) = t]$, it is shown in [16] that the number of observations in a given constant segment of an isotonic regression solution concentrates in probability around $n^{2/3}$. In general, without C4, our proof establishes a miscoverage upper bound of $\frac{1}{n+1} \mathbb{E}[N_{n+1}]$, where $\mathbb{E}[N_{n+1}]$ is the expected number of constant segments of $f_n^{(X_{n+1}, Y_{n+1})}$.

The next theorem examines the interaction between calibration and CP within SC-CP in terms of efficiency of the self-calibrating conformity scores. In the following, let $x \in \mathcal{X}, y \in \mathcal{Y}$ be arbitrary. For each $\theta \in \Theta_{iso}$, define the θ -transformed conformity scoring function $S_\theta : (x', y') \mapsto |y' - \theta \circ f(x')|$. Let $\theta_0 := \text{argmin}_{\theta \in \Theta_{iso}} \int \{S_\theta(x', y')\}^2 dP(x', y')$ be the optimal isotonic transformation of $f(\cdot)$ that minimizes the population mean-square error. Define the self-calibrating conformity scoring function as $S_n^{(x, y)}(x', y') := |y' - f_n^{(x, y)}(x')|$, where $f_n^{(x, y)}$ is obtained as in Alg. 1.

C5) Independent data: $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are iid.

C6) Bounded outcomes: \mathcal{Y} is a uniformly bounded set.

Theorem 4.3. Under C5 and C6, we have $\int \{S_n^{(x,y)}(x', y') - S_{\theta_0}(x', y')\}^2 dP(x', y') = O_p(n^{-2/3})$.

The above theorem indicates that the self-calibrating scoring function $S_n^{(x,y)}$ used in Alg. 2 asymptotically converges in mean-square error to the oracle scoring function S_{θ_0} at a rate of $n^{-2/3}$. Since the oracle scoring function S_{θ_0} corresponds to a model $\theta_0 \circ f$ with better mean square error than f , we heuristically expect that the Venn-Abers scoring function $S_n^{(x,y)}$ will translate to narrower CP intervals, at least asymptotically. We provide experimental evidence for this heuristic in Section 5.

Limitations. The perfectly calibrated prediction $f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1})$, guaranteed to lie by Theorem 4.1 in the Venn-Abers multi-prediction, typically cannot be determined precisely without knowledge of Y_{n+1} . However, the stability of isotonic regression implies that the width of multi-prediction $f_{n+1}(X_{n+1})$ shrinks towards zero very quickly as the size of the calibration set increases [10]. Moreover, the large-sample theory for isotonic calibration in [51] demonstrates that the ℓ^2 -calibration error of each model $f_n^{(X_{n+1}, y)}$ with $y \in \mathcal{Y}$ is $O_p(n^{-2/3})$. One caveat of SC-CP intervals is that desideratum (ii) is satisfied with respect to the unknown, oracle point prediction $f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1})$. However, we know that this oracle prediction lies within the Venn-Abers multi-prediction by Theorem 4.1, and its value can be determined with high precision with relatively small calibration sets [60] — see, e.g., Figure 1. These limitations appear to be unavoidable as perfectly calibrated point predictions can generally not be constructed in finite samples without oracle knowledge [61, 60].

4.1 Related work

The work of [42] proposes a regression extension of Venn-Abers calibration that differs from ours, both algorithmically and in its objective. While our extension constructs a calibrated point prediction $f(X)$ of Y such that $f(X) = \mathbb{E}[Y | f(X)]$, their approach uses the original Venn-Abers calibration procedure of [60] to construct a distributional prediction $f_t(X)$ of $1(Y \leq t)$ that satisfies $f_t(X) = \mathbb{P}(Y \leq t | f_t(X))$ for $t \in \mathcal{Y}$.

The impossibility results of [25] imply that any universal procedure providing prediction-conditionally calibrated intervals must explicitly or implicitly discretize the output of the model $f(\cdot)$. The works of [31] and [29] apply Mondrian CP [59] within leaves of a regression tree f to construct prediction intervals with prediction-conditional validity. However, this approach is restricted to tree-based predictors and does not guarantee calibrated point predictions and self-calibrated intervals. Mondrian conformal predictive distributions were applied within bins of model predictions in [9] to satisfy a coarser, distributional form of prediction-conditional validity. A limitation of Mondrian-CP approaches to prediction-conditional validity is that they require pre-specification of a binning scheme for the predictor $f(\cdot)$, which introduces a trade-off between model performance and the width of prediction intervals, and they do not perform point calibration (desideratum (i)) and, thereby, do not guarantee self-calibration. In contrast, SC-CP data-adaptively discretizes the predictor $f(\cdot)$ using isotonic calibration and, in doing so, provides calibrated predictions, improved conformity scores, and self-calibrated intervals.

Other notions of conditional validity have been proposed that, like prediction-conditional validity and self-calibration, avoid the curse of dimensionality of context-conditional validity. In the multiclassification setup, [50] and [18] use Mondrian CP to provide prediction intervals with valid coverage conditional on the class label (i.e., outcome). In [8], Mondrian CP is applied within bins categorized by context-specific difficulty estimates, such as conditional variance estimates. Multivald-CP [32, 6] offers coverage based on a threshold defining the prediction interval. For multiclassification, [41] propose a procedure for attaining valid coverage conditional on the prediction set size [2].

5 Real-Data Experiments: predicting utilization of medical services

5.1 Experimental setup

In this experiment, we illustrate how prediction-conditional validity can approximate context-conditional validity when the heteroscedasticity in outcomes is strongly associated with model

predictions, thereby ensuring validity across critical subgroups without their pre-specification. We analyze the Medical Expenditure Panel Survey (MEPS) dataset [1], supplied by the Agency for Healthcare Research and Quality [13], which was used in [47] for Mondrian CP with fairness applications. We use the preprocessed dataset acquired using the Python package `cqr`, also associated with [47]. This dataset contains $n = 15,656$ observations and $d = 139$ features, and includes information such as age, marital status, race, and poverty status, alongside medical service utilization. Our objective is to predict each individual’s healthcare system utilization, represented by a score that reflects visits to doctors’ offices, hospital visits, etc. Following [47], we designate race as the sensitive attribute A , aiming for *equalized coverage*, where $A = 0$ represents non-white individuals ($n_0 = 9640$) and $A = 1$ represents white individuals ($n_1 = 6016$). The outcome variable Y is transformed by $Y = \log(1 + \text{utilization score})$ to address the skewness of the raw score. In Appendix B, we present additional experimental results for the *Concrete*, *Community*, *STAR*, *Bike*, and *Bio* datasets used in [48] and publicly available in the Python package `cqr`, associated with [48] and [47].

We randomly partition the dataset into three segments: a training set (50%) for model training, a calibration set (30%) for CP, and a test set (20%) for evaluation. For training the initial model $f(\cdot)$, we use the `xgboost` [12] implementation of gradient boosted regression trees [20], where maximum tree depth, boosting rounds, and learning rate are tuned using 5-fold cross-validation. We consider two settings for training the model. In **Setting A**, we train the initial model on the untransformed outcomes and then transform the predictions as $\hat{y} \mapsto \log(1 + \hat{y})$, which makes the model predictive but poorly calibrated because it overestimates the true outcomes, in light of Jensen’s inequality. In **Setting B**, we train the initial model on the transformed outcomes, leading to fairly well-calibrated predictions. In both settings, calibration and evaluation are applied to the transformed outcomes.

For direct comparison, we compare **SC-CP** with baselines that leverage the standard absolute residual scoring function $S(x, y) := |y - f(x)|$ and target either marginal validity or prediction-conditional validity. The baselines are: **Marginal CP** [34], **Mondrian CP** with categories defined by bins of model predictions [59, 9], **CQR** [48] with model predictions used as features, and the **Kernel-smoothed conditional CP** approach of [21] with model predictions $\{f(X_i)\}_{i=1}^n$ used as features and bandwidth tuned with cross-validation. Due to the slow computing time of the implementation provided by [21], we apply **Kernel** on a subset of the calibration data of size $n_{cal} = 500$. **SC-CP** is implemented as described in Alg. 2, using isotonic regression constrained to have at least 20 observations averaged within each constant segment to mitigate overfitting (via the minimum child weight argument of `xgboost`). The miscoverage level is taken to be $\alpha = 0.1$. **SC-CP** provides calibrated point predictions and self-calibrated intervals, while the **Mondrian** and **Kernel** baselines offer approximate prediction-conditional validity, and **Marginal** and **CQR** only guarantee marginal coverage. We report empirical coverage, average interval width, and calibration error of model predictions in the test set within the sensitive attribute. Calibration error is defined as the mean error of the point predictions, $\mathbb{E}[\hat{Y} - Y \mid A]$ within the sensitive attribute A , which measures model over- or under-confidence. For **SC-CP**, we use the calibrated point predictions from (4), while the original point predictions are used for **Marginal**, **Mondrian**, and **Kernel**. For **CQR**, we use an estimate of the conditional median, obtained from a separate `xgboost` quantile regression model, as the point prediction. We note that, since quantiles are preserved under monotone transformations of the outcome, we expect the conditional median model of **CQR** to be well-calibrated, at least in a median sense, in both **Setting A** and **Setting B**. We include the baseline **Mondrian*** for direct comparison with **SC-CP**, in which **Mondrian** is applied with the same number of prediction bins as data-adaptively selected by **SC-CP**.

5.2 Results and discussion

The experimental results for each setting are depicted in Figure 2. Each panel’s left-most plot showcases a calibration plot [62] for **SC-CP**, illustrating original and calibrated predictions alongside prediction bands. On the right, the panels display prediction bands of our baselines as a function of the original model predictions. Visually, as expected by Theorem 4.2, the **SC-CP** bands adapt to outcome heteroscedasticity within model predictions, while **Marginal** lacks adaptation, **Mondrian** under-adapts due to insufficient bins, and **Kernel** adapts but offers wider intervals for large predictions where observations are sparse. The bands of **CQR** appear adaptive and similar to those of **SC-CP**, however, do not guarantee finite-sample prediction-conditional validity. The calibration plots reveal that heteroscedasticity in outcomes is primarily driven by their mean, suggesting that prediction-conditional validity may approximate context-conditional validity. This heuristic is supported by the

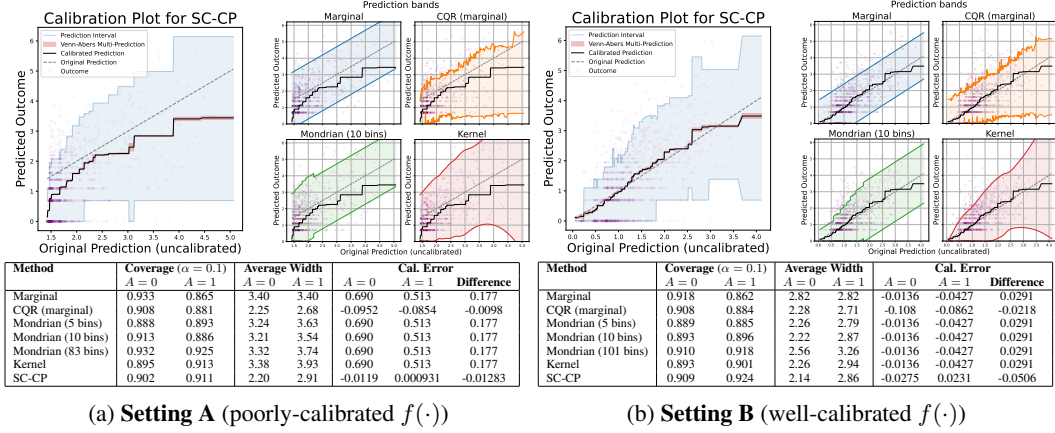


Figure 2: **MEPS-21 dataset**: Calibration plot for SC-CP, prediction bands for SC-CP and baselines, and empirical coverage, width, and calibration error within sensitive subgroup.

tables in Figure 2, which display empirical coverage, average interval width, and calibration error within the sensitive attribute (A) for all methods. In **Setting A**, the base regression model $f(\cdot)$ are poorly calibrated, i.e., $\mathbb{E}[f(X) - Y | A]$ is not close to 0, resulting in wider intervals, overconfidence in point predictions, and decreased interpretability for the baselines, as their intervals center around biased point predictions. In contrast, being self-calibrated, **SC-CP** corrects the calibration error in f , achieving the smallest interval widths and well-calibrated point predictions in both settings, as guaranteed by Theorem 4.1. In both settings, the quantile regression model of **CQR** appears to have worse calibration than **SC-CP**, which may be due to the median differing from the mean because of the skewness of the outcomes. Additionally, **SC-CP** predictions achieves a smaller difference in calibration error between the two subgroups than **Marginal**, **Mondrian**, and **Kernel**, suggesting they are less discriminatory and more fair [47]. **SC-CP** and **Kernel** achieve the desired coverage level of $1 - \alpha = 0.9$ in each subgroup and setting, whereas **Marginal** exhibits over- or under-coverage in each subgroup. **Mondrian** tends to under-cover with 5 and 10 bins and only attains good coverage when using the same binning number data-adaptively selected by **SC-CP**, highlighting its sensitivity to the pre-specified binning scheme. **CQR** attains good coverage in the $A = 0$ group but undercovers in the $A = 1$ group, which may be explained by **CQR** only guaranteeing marginal coverage in finite samples. Even with **SC-CP** having higher coverage, the intervals of **SC-CP** are narrower than those of **Kernel** and **Mondrian***. This provides experimental evidence that calibration improves conformity scores and translates into greater interval efficiency, as suggested by Theorem 4.3.

6 Extensions

Our theoretical techniques can be used to analyze conformal prediction methods that involve the calibration of model predictions followed by the construction of conditionally valid prediction intervals. Our analysis can be adapted to the general case where either the conformity score or the conditioning variable depends on the calibrated model prediction. While we use the absolute residual conformity score in our work, **SC-CP** can be applied to other conformity scores, such as the normalized absolute residual scoring function [43], allowing for the inclusion of context-specific difficulty estimates in the **SC-CP** procedure. Although we use Venn-Abers calibration in **SC-CP**, our analysis also applies to other binning calibration methods, such as Venn calibration [61, 60]. Thus, we can replace the isotonic calibration step in Alg. 1 and 2, for example, with histogram binning [25]. Additionally, a group-valid form of **SC-CP** can be achieved by applying Alg. 2 separately within subgroups, similar to Multivald CP [32]. Interesting areas for future work involve integrating point calibration with conformal prediction methods for predictive models beyond regression, such as the isotonic calibration of quantile predictions in conformal quantile regression [48].

References

- [1] Medical expenditure panel survey, panel 21. Accessed: May, 2024.
- [2] A. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.
- [3] V. Balasubramanian, S.-S. Ho, and V. Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.
- [4] R. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.
- [5] R. E. Barlow and H. D. Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.
- [6] O. Bastani, V. Gupta, C. Jung, G. Noarov, R. Ramalingam, and A. Roth. Practical adversarial multivald conformal prediction. *Advances in Neural Information Processing Systems*, 35:29362–29373, 2022.
- [7] A. Bella, C. Ferri, J. Hernández-Orallo, and M. J. Ramírez-Quintana. Calibration of machine learning models. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 128–146. IGI Global, 2010.
- [8] H. Boström and U. Johansson. Mondrian conformal regressors. In *Conformal and Probabilistic Prediction and Applications*, pages 114–133. PMLR, 2020.
- [9] H. Boström, U. Johansson, and T. Löfström. Mondrian conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*, pages 24–38. PMLR, 2021.
- [10] A. Caponnetto and A. Rakhlin. Stability properties of empirical risk minimization over donsker classes. *Journal of Machine Learning Research*, 7(12), 2006.
- [11] D. S. Char, N. H. Shah, and D. Magnus. Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine*, 378(11):981, 2018.
- [12] T. Chen and C. 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.
- [13] J. W. Cohen, S. B. Cohen, and J. S. Banthin. The medical expenditure panel survey: a national information resource to support healthcare cost research and inform policy and practice. *Medical care*, 47(7_Supplement_1):S44–S50, 2009.
- [14] D. R. Cox. Two further applications of a model for binary regression. *Biometrika*, 45(3/4):562–565, 1958.
- [15] S. E. Davis, T. A. Lasko, G. Chen, E. D. Siew, and M. E. Matheny. Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association*, 24(6):1052–1061, 2017.
- [16] H. Deng, Q. Han, and C.-H. Zhang. Confidence intervals for multiple isotonic regression and other monotone models. *The Annals of Statistics*, 49(4):2021–2052, 2021.
- [17] Z. Deng, C. Dwork, and L. Zhang. Happymap: A generalized multi-calibration method. *arXiv preprint arXiv:2303.04379*, 2023.
- [18] T. Ding, A. N. Angelopoulos, S. Bates, M. I. Jordan, and R. J. Tibshirani. Class-conditional conformal prediction with many classes. *arXiv preprint arXiv:2306.09335*, 2023.
- [19] B. Flury and T. Tarpey. Self-consistency: A fundamental concept in statistics. *Statistical Science*, 11(3):229–243, 1996.

- [20] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.
- [21] I. Gibbs, J. J. Cherian, and E. J. Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.
- [22] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268, 2007.
- [23] L. Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.
- [24] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.
- [25] C. Gupta, A. Podkopaev, and A. Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33:3711–3723, 2020.
- [26] C. Gupta and A. Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International Conference on Machine Learning*, pages 3942–3952. PMLR, 2021.
- [27] T. Hastie and R. Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.
- [28] T. Heskes. Practical confidence and prediction intervals. *Advances in neural information processing systems*, 9, 1996.
- [29] U. Johansson, H. Linusson, T. Löfström, and H. Boström. Interpretable regression trees using conformal prediction. *Expert systems with applications*, 97:394–404, 2018.
- [30] U. Johansson, T. Löfström, and C. Sönströd. Well-calibrated probabilistic predictive maintenance using venn-abers. *arXiv preprint arXiv:2306.06642*, 2023.
- [31] U. Johansson, C. Sönströd, H. Linusson, and H. Boström. Regression trees for streaming data with local performance guarantees. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 461–470. IEEE, 2014.
- [32] C. Jung, G. Noarov, R. Ramalingam, and A. Roth. Batch multivald conformal prediction. *arXiv preprint arXiv:2209.15145*, 2022.
- [33] C.-I. C. Lee. The min-max algorithm and isotonic regression. *The Annals of Statistics*, 11(2):467–477, 1983.
- [34] J. Lei, M. G’Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.
- [35] J. Lei and L. Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.
- [36] Y. Li, L. Qi, and Y. Sun. Semiparametric varying-coefficient regression analysis of recurrent events with applications to treatment switching. *Statistics in Medicine*, 37:3959–3974, 2018. doi: 10.1002/sim.7856. PubMed PMID: 29992591. NIHMSID: NIHMS1033642.
- [37] S. Lichtenstein, B. Fischhoff, and L. D. Phillips. Calibration of probabilities: The state of the art. In *Decision Making and Change in Human Affairs: Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision Making, Darmstadt, 1–4 September, 1975*, pages 275–324. Springer, 1977.
- [38] E. B. Mandinach, M. Honey, and D. Light. A theoretical framework for data-driven decision making. In *annual meeting of the American Educational Research Association, San Francisco, CA*, 2006.

- [39] J. A. Mincer and V. Zarnowitz. The evaluation of economic forecasts. In *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, pages 3–46. NBER, 1969.
- [40] A. Niculescu-Mizil and R. Caruana. Obtaining calibrated probabilities from boosting. In *UAI*, volume 5, pages 413–20, 2005.
- [41] G. Noarov, R. Ramalingam, A. Roth, and S. Xie. High-dimensional prediction for sequential decision making. *arXiv preprint arXiv:2310.17651*, 2023.
- [42] I. Nouretdinov, D. Volkhonskiy, P. Lim, P. Toccaceli, and A. Gammerman. Inductive venn-abers predictive distribution. In *Conformal and Probabilistic Prediction and Applications*, pages 15–36. PMLR, 2018.
- [43] H. Papadopoulos, A. Gammerman, and V. Vovk. Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, pages 64–69, 2008.
- [44] J. Patel. Prediction intervals-a review. *Communications in Statistics-Theory and Methods*, 18(7):2393–2465, 1989.
- [45] J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.
- [46] L. Quest, A. Charrie, L. du Croo de Jongh, and S. Roy. The risks and benefits of using ai to detect crime. *Harv. Bus. Rev. Digit. Artic*, 8:2–5, 2018.
- [47] Y. Romano, R. F. Barber, C. Sabatti, and E. Candès. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2(2):4, 2020.
- [48] Y. Romano, E. Patterson, and E. Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.
- [49] G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.
- [50] F. Shi, C. S. Ong, and C. Leckie. Applications of class-conditional conformal predictor in multi-class classification. In *2013 12th International Conference on Machine Learning and Applications*, volume 1, pages 235–239. IEEE, 2013.
- [51] L. van der Laan, E. Ulloa-Pérez, M. Carone, and A. Luedtke. Causal isotonic calibration for heterogeneous treatment effects. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, Honolulu, Hawaii, USA, 2023. PMLR.
- [52] L. van der Laan, E. Ulloa-Pérez, M. Carone, and A. Luedtke. Causal isotonic calibration for heterogeneous treatment effects. *arXiv preprint arXiv:2302.14011*, 2023.
- [53] A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.
- [54] A. Van Der Vaart and J. A. Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5(2011):192, 2011.
- [55] V. Vapnik, E. Levin, and Y. Le Cun. Measuring the vc-dimension of a learning machine. *Neural computation*, 6(5):851–876, 1994.
- [56] J. Vazquez and J. C. Facelli. Conformal prediction in clinical medical sciences. *Journal of Healthcare Informatics Research*, 6(3):241–252, 2022.
- [57] M. Veale, M. Van Kleek, and R. Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.
- [58] V. Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

- [59] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.
- [60] V. Vovk and I. Petej. Venn-abers predictors. *arXiv preprint arXiv:1211.0025*, 2012.
- [61] V. Vovk, G. Shafer, and I. Nouretdinov. Self-calibrating probability forecasting. *Advances in neural information processing systems*, 16, 2003.
- [62] M. Vuk and T. Curk. Roc curve, lift chart and calibration plot. *Metodoloski zvezki*, 3(1):89, 2006.
- [63] M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017.
- [64] Y. Xu and S. Yadlowsky. Calibration error for heterogeneous treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 9280–9303. PMLR, 2022.
- [65] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616. Citeseer, 2001.
- [66] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

A Code

The methods implemented in this paper are not computationally intensive and were run in a Jupyter notebook environment on a MacBook Pro with 16GB RAM and an M1 chip. A Python implementation of SC-CP is provided in the package `SelfCalibratingConformal`, available via `pip`. Code implementing SC-CP and reproducing our experiments is available in the GitHub repository `SelfCalibratingConformal`, which can be accessed at the following link: <https://github.com/Larsvanderlaan/SelfCalibratingConformal>.

B Supplementary real data experiments

B.1 Additional results

In this section, we present the experimental results for the *concrete*, *STAR*, *bike*, *community*, and *bio* datasets used in [48] and publicly available in the Python package `cqr`, associated with [48] and [47]. For the *STAR* dataset, the sensitive attribute A was set to “gender.” For the *Bike* dataset, the sensitive attribute A was set to “workingday,” and for *Community*, it was set to “race_binary.” For the remaining datasets, the sensitive attribute A was set to a dichotomization of the final column in the feature matrix, as above or below its median value.

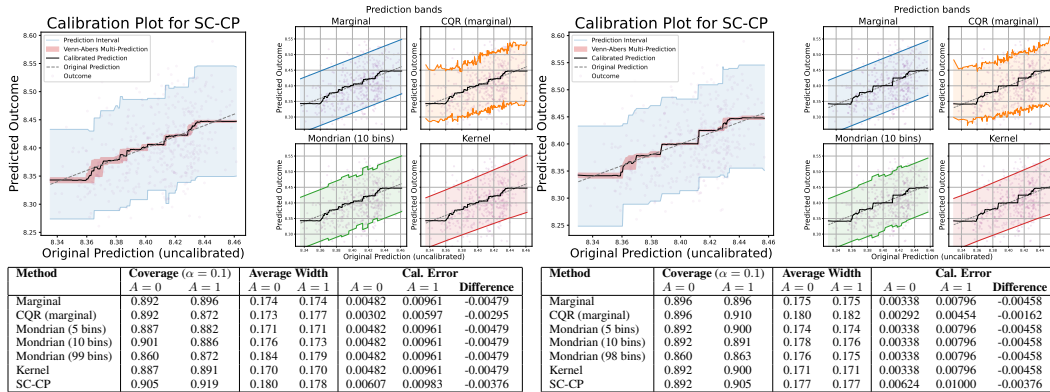


Figure 3: **STAR dataset:** Calibration plot for SC-CP, prediction bands for SC-CP and baselines, and empirical coverage, width, and calibration error within sensitive subgroup.

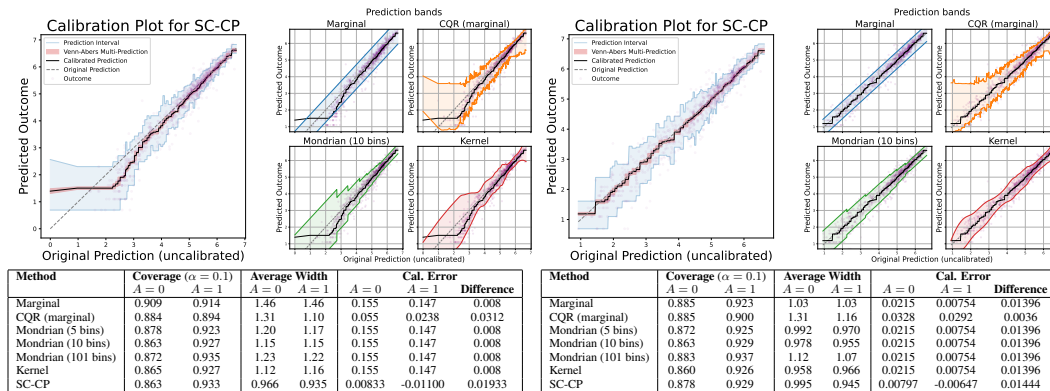


Figure 4: **Bike dataset:** Calibration plot for SC-CP, prediction bands for SC-CP and baselines, and empirical coverage, width, and calibration error within sensitive subgroup.

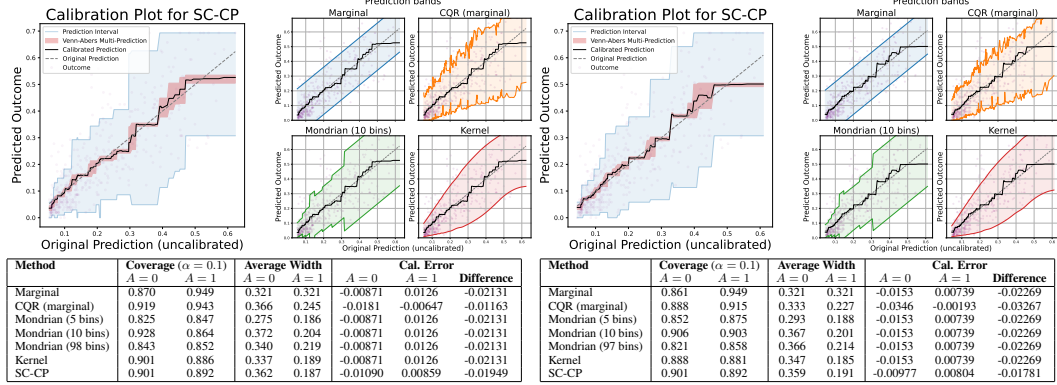


Figure 5: **Community dataset:** Calibration plot for SC-CP, prediction bands for SC-CP and baselines, and empirical coverage, width, and calibration error within sensitive subgroup.

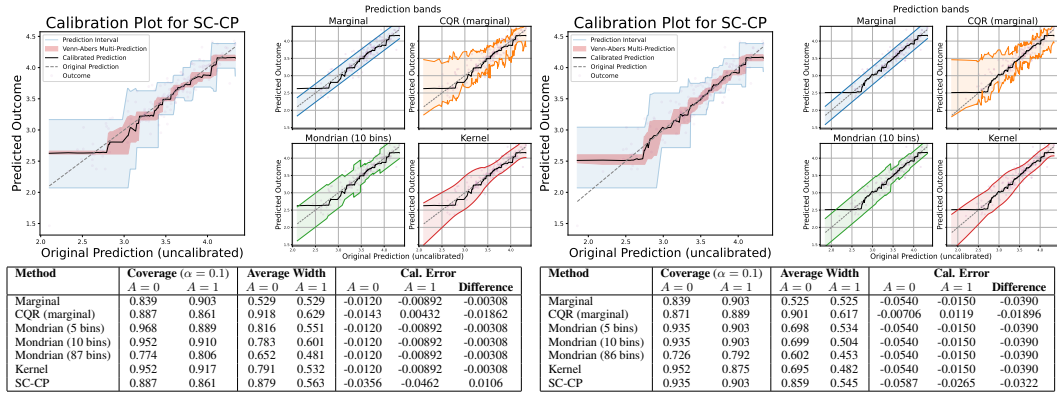


Figure 6: **Concrete dataset:** Calibration plot for SC-CP, prediction bands for SC-CP and baselines, and empirical coverage, width, and calibration error within sensitive subgroup.

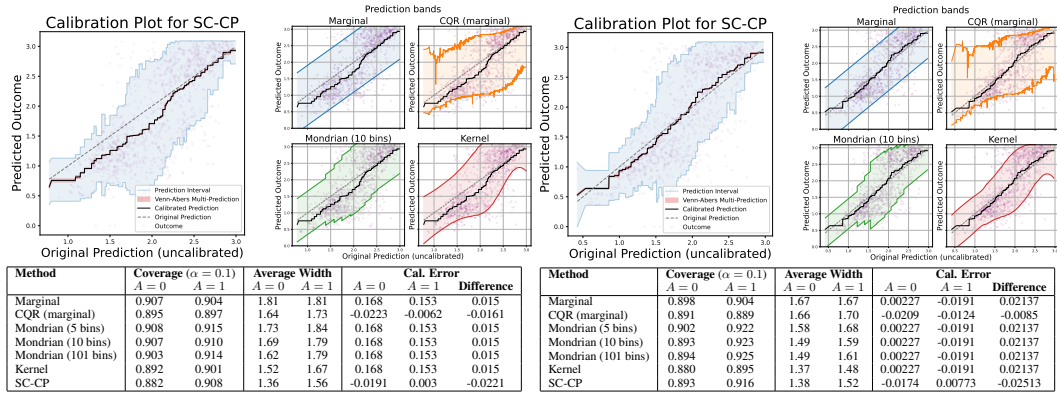


Figure 7: **Bio dataset:** Calibration plot for SC-CP, prediction bands for SC-CP and baselines, and empirical coverage, width, and calibration error within sensitive subgroup.

C Supplementary synthetic data experiments

C.1 Experimental setup

In this appendix, we perform additional synthetic experiments to evaluate the prediction-conditional validity of our method and how it translates to approximate context-conditional validity in certain cases.

Synthetic datasets. We construct synthetic training, calibration, and test datasets $\mathcal{D}_{train}, \mathcal{D}_{cal}, \mathcal{D}_{test}$ of sizes $n_{train}, n_{cal}, n_{test}$, which are respectively used to train f , apply CP, and evaluate performance. For parameters $d \in \mathbb{N}, \kappa > 0, a \geq 0, b \geq 0$, each dataset consists of *iid* observations of (X, Y) drawn as follows. The covariate vector $X := (X_1, \dots, X_d) \in [0, 1]^d$ is coordinate-wise independently drawn from a Beta(1, κ) distribution with shape parameter κ . Then, conditionally on $X = x$, the outcome Y is drawn normally distributed with conditional mean $\mu(x) := d^{-1/2} \sum_{j=1}^d \{x_j + \sin(4x_j)\}$ and conditional variance $\sigma^2(x) := \{0.035 - a \log(0.5 + 0.5x_1)/8 + b (|\mu_0(x)|^6/20 - 0.02) / 2\}^2$. Here, a and b control the heteroscedasticity and mean-variance relationship for the outcomes. For \mathcal{D}_{cal} and \mathcal{D}_{test} , we set $\kappa_{cal} = \kappa_{test} = 1$ and, for \mathcal{D}_{train} , we vary κ_{train} to introduce distribution shift and, thereby, calibration error in f . The parameters d, a , and b are fixed across the datasets.

Baselines. To mitigate overfitting, we implement SC-CP so that each function in Θ_{iso} has at least 20 observations averaged within each constant segment (implemented using the minimum leaf node size of argument `xgboost`). When appropriate, we will consider the following baseline CP algorithms for comparison. Unless stated otherwise, for all baselines, we use the scoring function $S(x, y) := |y - f(x)|$. The first baseline, uncond-CP, is split-CP [36], which provides only unconditional coverage guarantees. The second baseline, cond-CP, is adapted from [21] and provides conditional coverage over distribution shifts within a specified reproducing kernel Hilbert space. Following Section 5.1 of [21], we use the Gaussian kernel $K(X_i, X_j) := \exp(-4\|X_i - X_j\|_2^2)$ with euclidean norm $\|\cdot\|_2$ and select the regularization parameter λ using 5-fold cross-validation. The third baseline, Mondrian CP, applies the Mondrian CP method [59, 8] to categories formed by dividing f 's predictions into 20 equal-frequency bins based on \mathcal{D}_{cal} . As an optimal benchmark, we consider the oracle satisfying (2).

C.2 Experiment 1: Calibration and efficiency

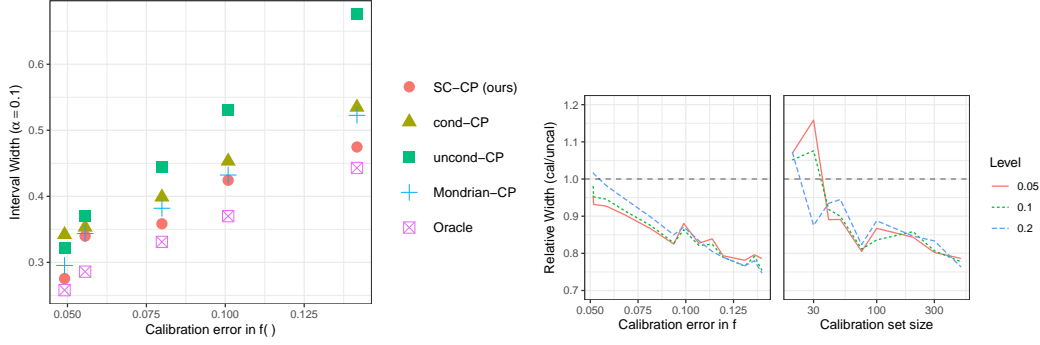
In this experiment, we illustrate how calibration of the predictor f can improve the efficiency (i.e., width) of the resulting prediction intervals. We consider the data-generating process of the previous section, with $n_{train} = n_{cal} = n_{test} = 1000, d = 5, a = 0$, and $b = 0.6$. The predictor f is trained on \mathcal{D}_{train} using the *ranger* [63] implementation of random forests with default settings. To control the calibration error in f , we vary the distribution shift parameter κ_{train} for \mathcal{D}_{train} over $\{1, 1.5, 2, 2.5, 3\}$.

Results. Figure 8a compares the average interval width across \mathcal{D}_{test} for SC-CP and baselines as the ℓ^2 -calibration error in f increases. Here, we estimate the calibration error using the approach of [64]. As calibration error increases, the average interval width for SC-SP appears smaller than those of uncond-CP, Mondrian-CP, and cond-CP, especially in comparison to uncond-CP. The observed efficiency gains in SC-CP are consistent with Theorem 4.3 and provides empirical evidence that self-calibrated conformity scores translate to tighter prediction intervals, given sufficient data. To test this hypothesis under controlled conditions, we compare the widths of prediction intervals obtained using vanilla (unconditional) CP for two scoring functions: $S : (x, y) \mapsto |y - f(x)|$ and the Venn-Abers (worst-case) scoring function $S_{cal} : (x, y) \mapsto \max_{y \in \mathcal{Y}} |y - f_n^{(x,y)}(x)|$. For miscoverage levels $\alpha \in \{0.05, 0.1, 0.2\}$, the left panel of Figure 8b illustrates the relative efficiency gain achieved by using S_{cal} , which we define as the ratio of the average interval widths for S_{cal} relative to S . The widths and calibration errors in Figure 8b are averaged across 100 data replicates.

Role of calibration set size. With too small calibration sets, the isotonic calibration step in Alg. 2 can lead to overfitting. In such cases, the self-calibrated conformity scores could be larger than their uncalibrated counterparts, potentially resulting in less efficient prediction intervals. In our experiments, overfitting is mitigated by constraining the minimum size of the leaf node in the isotonic regression tree to 20. For $\alpha \in \{0.05, 0.1, 0.2\}$, the right panel of Figure 8b displays the relationship between n_{cal} and the relative efficiency gain achieved by using S_{cal} , holding calibration error fixed ($\kappa_{train} = 3$). We find that calibration leads to a noticeable reduction in interval width as soon as $n_{cal} \geq 50$.

C.3 Experiment 2: Coverage and Adaptivity

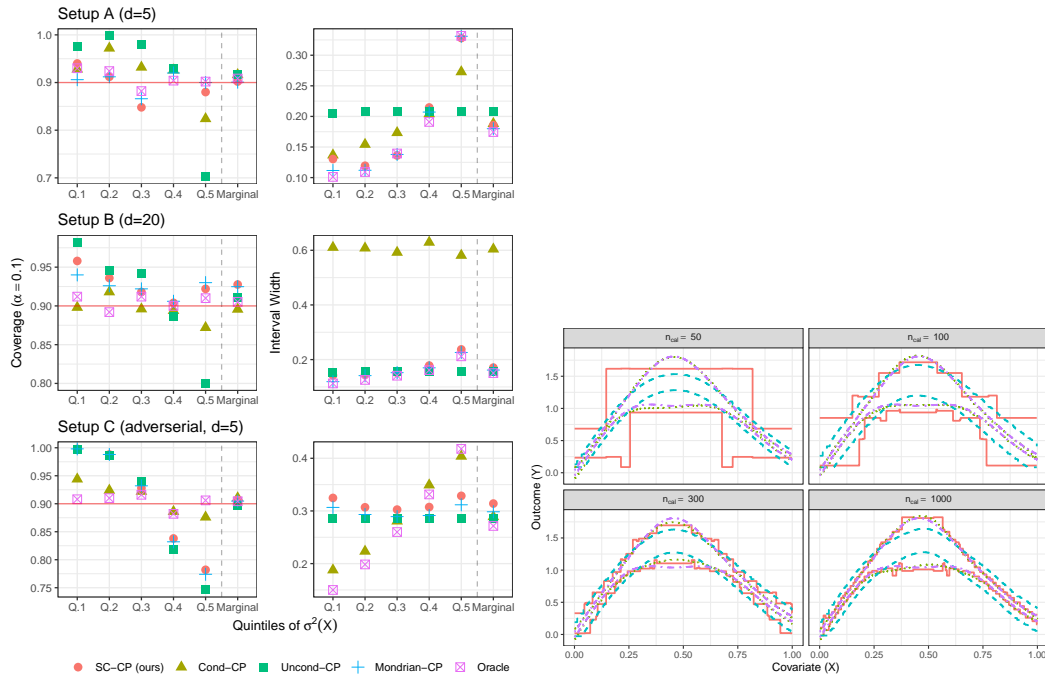
In this experiment, we illustrate how self-calibration of $\widehat{C}_{n+1}(X_{n+1})$ can, in some cases, translate to stronger conditional coverage guarantees. Here we take $n_{train} = n_{cal} = 1000$ and $n_{test} = 2500$,



(a) Avg. interval width with varying calibration error. (b) Relative efficiency change from calibration.

Figure 8: Figure 8a shows the average interval widths for varying ℓ^2 -calibration errors in f . Figure 8b shows the relative change in average interval width using Venn-Abers calibrated versus uncalibrated predictions, with varying ℓ^2 -calibration error in f (left), and calibration set size (right). Below the horizontal lines signifies efficiency gains for SC-CP.

and no distribution shift ($\kappa_{train} = 1$) in \mathcal{D}_{train} . We consider three setups: **Setup A** ($d = 5, a = 0, b = 0.6$) and **Setup B** ($d = 5, a = 0, b = 0.6$) which have a strong mean-variance relationship in the outcome process; and **Setup C** ($d = 5, a = 0.6, b = 0$) which has no such relationship. We obtain the predictor f from \mathcal{D}_{train} using a generalized additive model [27], so that it is well calibrated and accurately estimates μ . To assess the adaptivity of SC-CP to heteroscedasticity, we report the coverage and the average interval width within subgroups of \mathcal{D}_{test} defined by quintiles of the conditional variances $\{\sigma^2(X_i) : (X_i, Y_i) \in \mathcal{D}_{test}\}$.



(a) Coverage and interval width for $\sigma^2(X)$ quintiles

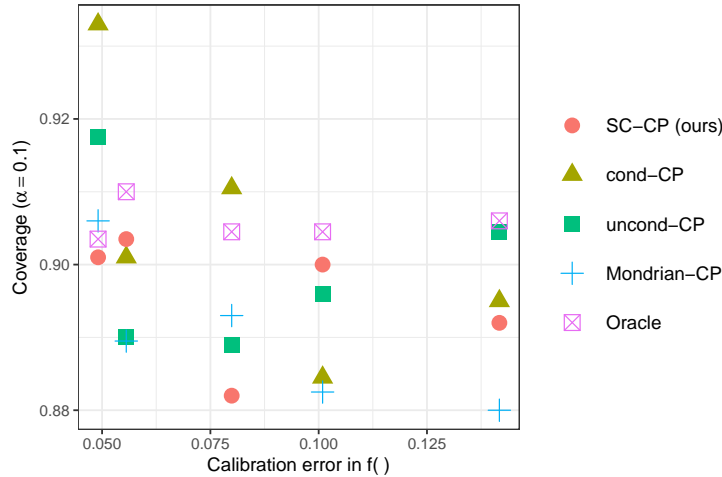
(b) Adaptivity of SC-CP with n_{cal}

Figure 9: Figure 9a displays the coverage and average interval width of SC-CP, marginally and within quintiles of the conditional outcome variance for setups A, B, and C. For a $d = 1$ example, Figure 9b shows the adaptivity of SC-CP prediction bands ($\alpha = 0.1$) across various calibration set sizes.

Results. The top two panels in Figure 9a displays the coverage and average interval width results for **Setup A** and **Setup B**. In both setups, SC-CP, cond-CP, and Mondrian-CP exhibit satisfactory coverage both marginally and within the quintile subgroups. In contrast, while uncond-CP attains the nominal level of marginal coverage, it exhibits noticeable overcoverage within the first three quintiles and significant undercoverage within the fifth quintile. The satisfactory performance of SC-CP with respect to heteroscedasticity in **Setups A and B** can be attributed to the strong mean-variance relationship in the outcome process. Regarding efficiency, the average interval widths of SC-CP are competitive with those of prediction-binned Mondrian-CP and the oracle intervals. The interval widths for cond-CP are wider than those for SC-CP and Mondrian-CP, especially for **Setup B**. This difference can be explained by cond-CP aiming for conditional validity in a 5D and 20D space, whereas SC-CP and Mondrian-CP target the 1D output space.

Limitation. If there is no mean-variance relationship in the outcomes, SC-CP is generally not expected to adapt to heteroscedasticity. The third (bottom) panel of Figure 9a displays SC-CP’s performance in **Setup C**, where there is no such relationship. In this scenario, it is evident that the conditional coverage of both uncond-CP and SC-CP are poor, while cond-CP maintains adequate coverage.

Adaptivity and calibration set size. SC-CP can be derived by applying CP within subgroups defined by a data-dependent binning of the output space $f(\mathcal{X})$, learned via Venn-Abers calibration, where the number of bins can grow with n_{cal} . In particular, if $t \mapsto E[Y | f(X) = t]$ is asymptotically monotone, which is plausible when f consistently estimates the outcome regression, then the number of bins selected by isotonic calibration will generally increase with n_{cal} , and the width of these bins will tend to zero. As a consequence, in such cases, the self-calibration result in Theorem 4.2 translates to conditional guarantees over finer partitions of $f(\mathcal{X})$ as n_{cal} increases. For $d = 1$ and a strong mean-variance relationship ($a = 0, b = 0.6$), Figure 9b demonstrates how the adaptivity of SC-CP bands improves as n_{cal} increases. In this case, for n_{cal} sufficiently large, we find that the SC-CP bands closely match those of cond-CP and the oracle.



(a) Avg. interval width with varying calibration error.

Figure 10: Figure 10a shows the marginal coverage corresponding to the interval widths of Figure 8a for varying ℓ^2 -calibration errors in f .

D Proofs

Proof of Theorem 4.1. It can be verified that an *isotonic calibrator class* Θ_{iso} satisfies the following properties: (a) It consists of univariate regression trees (i.e., piecewise constant functions) that are monotonically nondecreasing; (b) For all elements $\theta \in \Theta_{iso}$ and transformations $g : \mathbb{R} \rightarrow \mathbb{R}$, it holds that $g \circ \theta \in \Theta_{iso}$. Property (a) holds by definition. Property (b) is satisfied since constraining the maximum number of constant segments by $k(n)$ does not constrain the possible values that the

function may take in a given constant region. We note that the result of this theorem holds in general for any function class Θ_{iso} that satisfies (a) and (b).

Recall from Alg. 1 that $f_n^{(X_{n+1}, Y_{n+1})} = \theta_n^{(X_{n+1}, Y_{n+1})} \circ f$, where $\theta_n^{(X_{n+1}, Y_{n+1})} \in \Theta_{\text{iso}}$ is an isotonic calibrator satisfying $f_n^{(X_{n+1}, Y_{n+1})} = \theta_n^{(X_{n+1}, Y_{n+1})} \circ f$. By definition, recall that the isotonic calibrator class Θ_{iso} satisfies the invariance property that, for all $g : \mathbb{R} \rightarrow \mathbb{R}$, the inclusion $\theta \in \Theta_{\text{iso}}$ implies $g \circ \theta \in \Theta_{\text{iso}}$. Hence, for all $g : \mathbb{R} \rightarrow \mathbb{R}$ and $\varepsilon > 0$, it also holds that $(1 + \varepsilon g) \circ \theta_n^{(X_{n+1}, Y_{n+1})}$ lies in Θ_{iso} . Now, we use that $(1 + \varepsilon g) \circ \theta_n^{(X_{n+1}, Y_{n+1})} \circ f = (1 + \varepsilon g) \circ f_n^{(X_{n+1}, Y_{n+1})}$ and that $f_n^{(X_{n+1}, Y_{n+1})}$ is an empirical risk minimizer over the class $\{g \circ f : g \in \Theta_{\text{iso}}\}$. Using these two observations, the first order derivative equations characterizing the empirical risk minimizer $f_n^{(X_{n+1}, Y_{n+1})}$ imply, for all $g : \mathbb{R} \rightarrow \mathbb{R}$, that

$$\frac{1}{n+1} \sum_{i=1}^{n+1} (g \circ f_n^{(X_{n+1}, Y_{n+1})})(X_i) \left\{ Y_i - f_n^{(X_{n+1}, Y_{n+1})}(X_i) \right\} = \frac{d}{d\varepsilon} \left[\frac{1}{n+1} \sum_{i=1}^{n+1} \left\{ Y_i - (1 + \varepsilon g) \circ f_n^{(X_{n+1}, Y_{n+1})} \right\}^2 \right] \Big|_{\varepsilon=0} = 0. \quad (5)$$

Taking expectations of both sides of the above display, we conclude

$$\frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E} \left[(g \circ f_n^{(X_{n+1}, Y_{n+1})})(X_i) \left\{ Y_i - f_n^{(X_{n+1}, Y_{n+1})}(X_i) \right\} \right] = 0. \quad (6)$$

We now use the fact that $\{(X_i, Y_i, f_n^{(X_{n+1}, Y_{n+1})}(X_i)) : i \in [n+1]\}$ are exchangeable, since $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable by C1 and the function $f_n^{(X_{n+1}, Y_{n+1})}$ is invariant under permutations of $\{(X_i, Y_i)\}_{i=1}^{n+1}$. Consequently, Equation (6) remains true if we replace each $(X_i, Y_i, f_n^{(X_{n+1}, Y_{n+1})}(X_i))$ with $i \in [n]$ by $(X_{n+1}, Y_{n+1}, f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}))$. That is,

$$\begin{aligned} & \mathbb{E} \left[(g \circ f_n^{(X_{n+1}, Y_{n+1})})(X_{n+1}) \left\{ Y_{n+1} - f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right\} \right] \\ &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E} \left[(g \circ f_n^{(X_{n+1}, Y_{n+1})})(X_{n+1}) \left\{ Y_{n+1} - f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right\} \right] \\ &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E} \left[(g \circ f_n^{(X_{n+1}, Y_{n+1})})(X_i) \left\{ Y_i - f_n^{(X_{n+1}, Y_{n+1})}(X_i) \right\} \right] \\ &= 0. \end{aligned}$$

By the law of iterated conditional expectations, the preceding display further implies

$$\mathbb{E} \left[(g \circ f_n^{(X_{n+1}, Y_{n+1})})(X_{n+1}) \left\{ \mathbb{E}[Y_{n+1} | f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1})] - f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right\} \right] = 0.$$

Taking $g : \mathbb{R} \rightarrow \mathbb{R}$ to be defined by $(g \circ f_n^{(X_{n+1}, Y_{n+1})})(X_{n+1}) := \mathbb{E}[Y_{n+1} | f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1})] - f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1})$, we find

$$\mathbb{E} \left[\left\{ \mathbb{E}[Y_{n+1} | f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1})] - f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right\}^2 \right] = 0.$$

The above equality implies $\mathbb{E}[Y_{n+1} | f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1})] = f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1})$ almost surely, as desired. \square

Proof of Theorem 4.2. Recall, for a quantile level $\alpha \in (0, 1)$, the ‘‘pinball’’ quantile loss function ℓ_α is given by

$$\ell_\alpha(f(x), s) := \begin{cases} \alpha(s - f(x)) & \text{if } s \geq f(x), \\ (1 - \alpha)(f(x) - s) & \text{if } s < f(x). \end{cases}$$

As established in [21], each subgradient of $\ell_\alpha(\cdot, x)$ at f in the direction g , for some $\beta \in [\alpha - 1, \alpha]$, given by:

$$\partial_{\varepsilon[\beta]} \{ \ell_\alpha(f(x) + \varepsilon g(x), s) \} \Big|_{\varepsilon=0} := 1(f(x) \neq s) g(x) \{ \alpha - 1(f < s) \} + 1(f(x) = s) \beta g(x).$$

For $(x, y) \in \mathcal{X} \times \mathcal{Y}$, define the empirical risk minimizer:

$$\rho_n^{(x,y)} \in \operatorname{argmin}_{\theta \circ f_n^{(x,y)}; \theta: \mathbb{R} \rightarrow \mathbb{R}} \sum_{i=1}^n \ell_\alpha \left(\theta \circ f_n^{(x,y)}(X_i), S_i^{(x,y)} \right) + \ell_\alpha \left(\theta \circ f_n^{(x,y)}(x), S_{n+1}^{(x,y)} \right).$$

Then, since the isotonic calibrated predictor $f_n^{(x,y)}$ is piece-wise constant and the above optimization problem is unconstrained in the map $\theta : \mathbb{R} \rightarrow \mathbb{R}$, it holds that the evaluation $\rho_n^{(x,y)}(x)$ lies in the solution set:

$$\operatorname{argmin}_{q \in \mathbb{R}} \sum_{i=1}^n K_i(f_n^{(x,y)}, x) \cdot \ell_\alpha(q, S_i^{(x,y)}) + \ell_\alpha(q, S_{n+1}^{(x,y)}),$$

where $K_i(f_n^{(x,y)}, x) = \mathbf{1} \left\{ f_n^{(x,y)}(X_i) = f_n^{(x,y)}(x) \right\}$. Consequently, we see that the empirical quantile $\rho_n^{(x,y)}(x)$ defined in Alg. 2 coincides with the evaluation of the empirical risk minimizer $\rho_n^{(x,y)}(\cdot)$ at x , as suggested by our notation.

We will now theoretically analyze Alg. 2 by studying the empirical risk minimizer $x' \mapsto \rho_n^{(X_{n+1}, Y_{n+1})}(x')$. To do so, we modify the arguments used to establish Theorem 2 of [21]. As in [21], we begin by examining the first-order equations of the convex optimization problem defining $x' \mapsto \rho_n^{(X_{n+1}, Y_{n+1})}(x')$.

Studying first-order equations of convex problem: Given any transformation $\theta : \mathbb{R} \rightarrow \mathbb{R}$, each subgradient of the map

$$\varepsilon \mapsto \ell_\alpha \left(\rho_n^{(X_{n+1}, Y_{n+1})}(X_i) + \varepsilon \theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_i), S_i^{(X_{n+1}, Y_{n+1})} \right)$$

is, for some $\beta \in [\alpha - 1, \alpha]^{n+1}$, of the following form:

$$\begin{aligned} \partial_{\varepsilon[\beta]} \left\{ \ell_\alpha \left(\rho_n^{(X_{n+1}, Y_{n+1})}(X_i) + \varepsilon \theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_i), S_i^{(X_{n+1}, Y_{n+1})} \right) \right\} \Big|_{\varepsilon=0} \\ := \begin{cases} \theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_i) \{ \alpha - 1(\rho_n^{(X_{n+1}, Y_{n+1})}(X_i) < S_i^{(X_{n+1}, Y_{n+1})}) \} & \text{if } S_i^{(X_{n+1}, Y_{n+1})} \neq \rho_n^{(X_{n+1}, Y_{n+1})}(X_i), \\ \beta_i (\theta \circ f_n^{(X_{n+1}, Y_{n+1})})(X_i) & \text{if } S_i^{(X_{n+1}, Y_{n+1})} = \rho_n^{(X_{n+1}, Y_{n+1})}(X_i). \end{cases} \end{aligned}$$

Now, since $\rho_n^{(X_{n+1}, Y_{n+1})}$ is an empirical risk minimizer of the quantile loss over the class $\{ \theta \circ f_n^{(X_{n+1}, Y_{n+1})}; \theta : \mathbb{R} \rightarrow \mathbb{R} \}$, there exists some vector $\beta^* = (\beta_1^*, \dots, \beta_{n+1}^*) \in [\alpha - 1, \alpha]^{n+1}$ such that

$$\begin{aligned} 0 &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1} \{ S_i^{(X_{n+1}, Y_{n+1})} \neq \rho_n^{(X_{n+1}, Y_{n+1})}(X_i) \} \left[\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_i) \{ \alpha - 1(\rho_n^{(X_{n+1}, Y_{n+1})}(X_i) < S_i^{(X_{n+1}, Y_{n+1})}) \} \right] \\ &\quad + \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1} \{ S_i^{(X_{n+1}, Y_{n+1})} = \rho_n^{(X_{n+1}, Y_{n+1})}(X_i) \} \left[\beta_i^* (\theta \circ f_n^{(X_{n+1}, Y_{n+1})})(X_i) \right] \end{aligned}$$

The above display can be rewritten as:

$$\begin{aligned} \frac{1}{n+1} \sum_{i=1}^{n+1} \left[\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_i) \{ \alpha - 1(\rho_n^{(X_{n+1}, Y_{n+1})}(X_i) < S_i^{(X_{n+1}, Y_{n+1})}) \} \right] \\ = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1} \{ S_i^{(X_{n+1}, Y_{n+1})} = \rho_n^{(X_{n+1}, Y_{n+1})}(X_i) \} \left[(1 - \beta_i^*) (\theta \circ f_n^{(X_{n+1}, Y_{n+1})})(X_i) \right]. \end{aligned} \tag{7}$$

Now, observe that the collection of random variables

$$\left\{ (f_n^{(X_{n+1}, Y_{n+1})}(X_i), S_i^{(X_{n+1}, Y_{n+1})}, \rho_n^{(X_{n+1}, Y_{n+1})}(X_i)) : i \in [n+1] \right\}$$

are exchangeable, since $\{(X_i, Y_i) : i \in [n+1]\}$ are exchangeable by C1 and the functions $f_n^{(X_{n+1}, Y_{n+1})}(\cdot)$ and $\rho_n^{(X_{n+1}, Y_{n+1})}(\cdot)$ are unchanged under permutations of the training data

$\{(X_i, Y_i)\}_{i \in [n+1]}$. Thus, the expectation of the left-hand side of Equation (7) can be expressed as:

$$\begin{aligned} & \mathbb{E} \left[\frac{1}{n+1} \sum_{i=1}^{n+1} \left[\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_i) \{ \alpha - 1(\rho_n^{(X_{n+1}, Y_{n+1})}(X_i) < S_i^{(X_{n+1}, Y_{n+1})}) \} \right] \right] \\ &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E} \left[\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_i) \{ \alpha - 1(\rho_n^{(X_{n+1}, Y_{n+1})}(X_i) < S_i^{(X_{n+1}, Y_{n+1})}) \} \right] \\ &= \mathbb{E} \left[\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \{ \alpha - 1(\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})}) \} \right], \end{aligned}$$

where the final inequality follows from exchangeability. Combining this with Equation (7), we find

$$\mathbb{E} \left[\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \{ \alpha - 1(\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})}) \} \right] \quad (8)$$

$$= \mathbb{E} \left[1 \{ S_i^{(X_{n+1}, Y_{n+1})} = \rho_n^{(X_{n+1}, Y_{n+1})}(X_i) \} \left[(1 - \beta_i^*) (\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_i)) \right] \right] \quad (9)$$

Lower bound on coverage: We first obtain the lower coverage bound in the theorem statement. Note, for any nonnegative $\theta : \mathbb{R} \rightarrow \mathbb{R}$, that (9) implies:

$$\mathbb{E} \left[\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \{ \alpha - 1(\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})}) \} \right] \geq 0.$$

This inequality holds since $(1 - \beta_i^*) \geq 0$ and $(\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_i)) \geq 0$ almost surely, for each $i \in [n+1]$.

By the law of iterated expectations, we then have

$$\mathbb{E} \left[\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \left\{ \alpha - \mathbb{P} \left(\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})} \mid f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right) \right\} \right] \geq 0.$$

Taking $\theta : \mathbb{R} \rightarrow \mathbb{R}$ as a nonnegative map that almost surely satisfies

$$\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) = 1 \left\{ \alpha \leq \mathbb{P} \left(\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})} \mid f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right) \right\},$$

we find

$$-\mathbb{E} \left[\left\{ \alpha - \mathbb{P} \left(\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})} \mid f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right) \right\}_- \right] \geq 0,$$

where the map $t \mapsto \{t\}_- := |t|1(t \leq 0)$ extracts the negative part of its input. Multiplying both sides of the previous inequality by -1 , we obtain

$$0 \leq \mathbb{E} \left[\left\{ \alpha - \mathbb{P} \left(\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})} \mid f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right) \right\}_- \right] \leq 0.$$

We conclude that the negative part of $\left\{ \alpha - \mathbb{P} \left(\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})} \mid f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right) \right\}$ is almost surely zero. Thus, it must be almost surely true that

$$\alpha \geq \mathbb{P} \left(\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})} \mid f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right).$$

Note that the event $Y_{n+1} \in \widehat{C}_{n+1}(X_{n+1}) = \{y \in \mathcal{Y} : S_{n+1}^{(X_{n+1}, y)} \leq \rho_n^{(X_{n+1}, y)}(X_{n+1})\}$ occurs if, and only if, $\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \geq S_{n+1}^{(X_{n+1}, Y_{n+1})}$. As a result, we obtain the desired lower coverage bound:

$$1 - \alpha \leq \mathbb{P} \left(Y_{n+1} \in \widehat{C}_{n+1}(X_{n+1}) \mid f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right).$$

Deviation bound for the coverage: We now bound the deviation of the coverage of SC-CP from the lower bound. Note, for any $f : \mathbb{R} \rightarrow \mathbb{R}$, that (9) implies:

$$\begin{aligned} & \mathbb{E} \left[\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \{ \alpha - 1(\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})}) \} \right] \\ & \leq \left| \mathbb{E} \left[1 \{ S_i^{(X_{n+1}, Y_{n+1})} = \rho_n^{(X_{n+1}, Y_{n+1})}(X_i) \} \left[(1 - \beta_i^*) (\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_i)) \right] \right] \right|. \quad (10) \end{aligned}$$

Using that $(1 - \beta_i^*) \in [0, 1]$ and exchangeability, we can bound the right-hand side of the above as

$$\begin{aligned} & \left| \mathbb{E} \left[\frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\{S_i^{(X_{n+1}, Y_{n+1})} = \rho_n^{(X_{n+1}, Y_{n+1})}(X_i)\} \left[(1 - \beta_i^*) (\theta \circ f_n^{(X_{n+1}, Y_{n+1})})(X_i) \right] \right] \right| \\ & \leq \frac{1}{n+1} \mathbb{E} \left[\left\{ \max_{i \in [n+1]} |(\theta \circ f_n^{(X_{n+1}, Y_{n+1})})(X_i)| \right\} \sum_{i=1}^{n+1} \mathbf{1}\{S_i^{(X_{n+1}, Y_{n+1})} = \rho_n^{(X_{n+1}, Y_{n+1})}(X_i)\} \right], \end{aligned}$$

Next, since there are no ties by C3, the event $\mathbf{1}\{S_i^{(X_{n+1}, Y_{n+1})} = \rho_n^{(X_{n+1}, Y_{n+1})}(X_i)\}$ for some index $i \in [n+1]$ can only occur once per piecewise constant segment of $\rho_n^{(X_{n+1}, Y_{n+1})}$, since, otherwise, $S_i^{(X_{n+1}, Y_{n+1})} = \rho_n^{(X_{n+1}, Y_{n+1})}(X_i) = \rho_n^{(X_{n+1}, Y_{n+1})}(X_j) = S_j^{(X_{n+1}, Y_{n+1})}$ for some $i \neq j$. However, $\rho_n^{(X_{n+1}, Y_{n+1})}$ is a transformation of $f_n^{(X_{n+1}, Y_{n+1})}$ and, therefore, has the same number of constant segments as $f_n^{(X_{n+1}, Y_{n+1})}$. Thus, it holds that

$$\sum_{i=1}^{n+1} \mathbf{1}\{S_i^{(X_{n+1}, Y_{n+1})} = \rho_n^{(X_{n+1}, Y_{n+1})}(X_i)\} \leq N^{(X_{n+1}, Y_{n+1})},$$

where $N^{(X_{n+1}, Y_{n+1})}$ is the (random) number of constant segments of $f_n^{(X_{n+1}, Y_{n+1})}$. This implies that

$$\begin{aligned} & \frac{1}{n+1} \mathbb{E} \left[\max_{i \in [n+1]} |(\theta \circ f_n^{(X_{n+1}, Y_{n+1})})(X_i)| \sum_{i=1}^{n+1} \mathbf{1}\{S_i^{(X_{n+1}, Y_{n+1})} = \rho_n^{(X_{n+1}, Y_{n+1})}(X_i)\} \right] \\ & \leq \frac{1}{n+1} \mathbb{E} \left[N^{(X_{n+1}, Y_{n+1})} \max_{i \in [n+1]} |(\theta \circ f_n^{(X_{n+1}, Y_{n+1})})(X_i)| \right]. \end{aligned}$$

Combining this bound with (10), we find

$$\begin{aligned} & \mathbb{E} \left[\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \left\{ \alpha - \mathbf{1}(\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})}) \right\} \right] \\ & \leq \frac{1}{n+1} \mathbb{E} \left[N^{(X_{n+1}, Y_{n+1})} \max_{i \in [n+1]} |(\theta \circ f_n^{(X_{n+1}, Y_{n+1})})(X_i)| \right]. \end{aligned}$$

By the law of iterated expectations, we then have

$$\begin{aligned} & \mathbb{E} \left[\theta \circ f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \left\{ \alpha - \mathbb{P} \left(\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})} \mid f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right) \right\} \right] \\ & \leq \frac{1}{n+1} \mathbb{E} \left[N^{(X_{n+1}, Y_{n+1})} \max_{i \in [n+1]} |(\theta \circ f_n^{(X_{n+1}, Y_{n+1})})(X_i)| \right]. \end{aligned}$$

Next, let $\mathcal{V} \subset \mathbb{R}$ denote the support of the random variable $f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1})$. Then, taking θ to be $t \mapsto \mathbf{1}(t \in \mathcal{V}) \text{sign} \left\{ \alpha - \mathbb{P} \left(\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})} \mid f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) = t \right) \right\}$, which falls almost surely in $\{-1, 1\}$, we obtain the mean absolute error bound:

$$\mathbb{E} \left| \alpha - \mathbb{P} \left(\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})} \mid f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right) \right| \leq \frac{1}{n+1} \mathbb{E} \left[N^{(X_{n+1}, Y_{n+1})} \right].$$

Since the event $Y_{n+1} \notin \widehat{C}_{n+1}(X_{n+1}) = \{y \in \mathcal{Y} : S_{n+1}^{(X_{n+1}, y)} \leq \rho_n^{(X_{n+1}, y)}(X_{n+1})\}$ occurs if, and only if, $\rho_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) < S_{n+1}^{(X_{n+1}, Y_{n+1})}$, we conclude that

$$\mathbb{E} \left| \alpha - \mathbb{P} \left(Y_{n+1} \notin \widehat{C}_{n+1}(X_{n+1}) \mid f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right) \right| \leq \frac{\mathbb{E}[N^{(X_{n+1}, Y_{n+1})}]}{n+1}.$$

Under C4, $\mathbb{E}[N^{(X_{n+1}, Y_{n+1})}] \leq n^{1/3} \text{polylog } n$, such that

$$\mathbb{E} \left| \alpha - \mathbb{P} \left(Y_{n+1} \notin \widehat{C}_{n+1}(X_{n+1}) \mid f_n^{(X_{n+1}, Y_{n+1})}(X_{n+1}) \right) \right| \leq \frac{n^{1/3} \text{polylog } n}{n+1} \leq \frac{\text{polylog } n}{n^{2/3}},$$

as desired. \square

Proof of Theorem 4.3. Let P_{n+1} denote the empirical distribution of $\{(X_i, Y_i)\}_{i=1}^{n+1}$ and let P_n denote the empirical distribution of $\{(X_i, Y_i)\}_{i=1}^n$. For any function $g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$: we use the following empirical process notation: $Pg := \int g(x, y)dP(x, y)$, $P_{n+1}g := \int g(x, y)dP_{n+1}(x, y)$, and $P_n g := \int g(x, y)dP_n(x, y)$.

Define the risk functions $R_n^{(x,y)}(\theta) := \frac{1}{n+1} \sum_{i=1}^n \{S_\theta(X_i, Y_i)\}^2 + \frac{1}{n+1} \{S_\theta(x, y)\}^2$, $R_{n+1}(\theta) := \frac{1}{n+1} \sum_{i=1}^{n+1} \{S_\theta(X_i, Y_i)\}^2$, and $R_0(\theta) := \int \{S_\theta(x, y)\}^2 dP(x, y)$. Moreover, define the risk minimizers as $\theta_n^{(x,y)} := \operatorname{argmin}_{\theta \in \Theta_{iso}} R_n^{(x,y)}(\theta)$ and $\theta_0 := \operatorname{argmin}_{\theta \in \Theta_{iso}} R_0(\theta)$. Observe that $R_n^{(x,y)}(\theta_n^{(x,y)}) - R_n^{(x,y)}(\theta_0) \leq 0$ since f_n minimizes R_n over Θ_{iso} . Using this inequality, it follows that

$$\begin{aligned} R_0(\theta_n^{(x,y)}) - R_0(\theta_0) &= R_0(\theta_n^{(x,y)}) - R_n^{(x,y)}(\theta_n^{(x,y)}) \\ &\quad + R_n^{(x,y)}(\theta_n^{(x,y)}) - R_n^{(x,y)}(\theta_0) + R_n^{(x,y)}(\theta_0) - R_0(\theta_0) \\ &\leq R_0(\theta_n^{(x,y)}) - R_n^{(x,y)}(\theta_n^{(x,y)}) - \{R_n^{(x,y)}(\theta_0) - R_0(\theta_0)\} \\ &\leq R_0(\theta_n^{(x,y)}) - R_{n+1}(\theta_n^{(x,y)}) - \{R_{n+1}(\theta_0) - R_0(\theta_0)\} \\ &\quad + R_n^{(x,y)}(\theta_n^{(x,y)}) - R_{n+1}(\theta_n^{(x,y)}) - \{R_n^{(x,y)}(\theta_0) - R_{n+1}(\theta_0)\}. \end{aligned}$$

The first term on the right-hand side of the above display can be written as

$$R_0(\theta_n^{(x,y)}) - R_{n+1}(\theta_n^{(x,y)}) - \{R_{n+1}(\theta_0) - R_0(\theta_0)\} = (P_{n+1} - P) \left[\{S_{\theta_n^{(x,y)}}\}^2 - \{S_{\theta_0}\}^2 \right].$$

We now bound the second term, $R_n^{(x,y)}(\theta_n^{(x,y)}) - R_{n+1}(\theta_n^{(x,y)}) - \{R_n^{(x,y)}(\theta_0) - R_{n+1}(\theta_0)\}$. For any $\theta \in \Theta_{iso}$, observe that

$$R_n^{(x,y)}(\theta) - R_{n+1}(\theta) = \frac{1}{n+1} \left[\{y - \theta \circ f(x)\}^2 - \{Y_{n+1} - \theta \circ f(X_{n+1})\}^2 \right].$$

We know that $\theta_n^{(x,y)}$ and θ_0 , being defined via isotonic regression, are uniformly bounded by $B := \sup_{y \in \mathcal{Y}} |y|$, which is finite by C6. Therefore,

$$\left| R_n^{(x,y)}(\theta_n^{(x,y)}) - R_{n+1}(\theta_n^{(x,y)}) - \{R_n^{(x,y)}(\theta_0) - R_{n+1}(\theta_0)\} \right| \leq \frac{8B^2}{n+1} = O(n^{-1}).$$

Combining the previous displays, we obtain the excess risk bound

$$R_0(\theta_n^{(x,y)}) - R_0(\theta_0) \leq (P_{n+1} - P) \left[\{S_{\theta_n^{(x,y)}}\}^2 - \{S_{\theta_0}\}^2 \right] + O(n^{-1}). \quad (11)$$

Next, we claim that $R_0(\theta_n^{(x,y)}) - R_0(\theta_0) \geq \|(\theta_n^{(x,y)} \circ f) - (\theta_0 \circ f)\|_P^2$. To show this, expanding the squares, note, pointwise for each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, that

$$\begin{aligned} \{S_{\theta_n^{(x,y)}}(x, y)\}^2 - \{S_{\theta_0}(x, y)\}^2 &= \{\theta_n^{(x,y)} \circ f(x)\}^2 - \{\theta_0 \circ f(x)\}^2 - 2y \left\{ (\theta_n^{(x,y)} \circ f)(x) - (\theta_0 \circ f)(x) \right\} \\ &= \left\{ (\theta_n^{(x,y)} \circ f)(x) + (\theta_0 \circ f)(x) - 2y \right\} \left\{ (\theta_n^{(x,y)} \circ f)(x) - (\theta_0 \circ f)(x) \right\}. \end{aligned}$$

Consequently,

$$R_0(\theta_n^{(x,y)}) - R_0(\theta_0) = \int \left\{ (\theta_n^{(x,y)} \circ f)(x) + (\theta_0 \circ f)(x) - 2y \right\} \left\{ (\theta_n^{(x,y)} \circ f)(x) - (\theta_0 \circ f)(x) \right\} dP(x). \quad (12)$$

The class Θ_{iso} consists of all isotonic functions and is, therefore, a convex space. Thus, the first-order derivative equations defining the population minimizer θ_0 imply that

$$\int \left\{ \theta_n^{(x,y)} \circ f(x) - \theta_0 \circ f(x) \right\} \{y - \theta_0 \circ f(x)\} dP(x, y) \leq 0. \quad (13)$$

Combining (12) and (13), we find

$$\begin{aligned} R_0(\theta_n^{(x,y)}) - R_0(\theta_0) &= \int \left\{ (\theta_n^{(x,y)} \circ f)(x) - (\theta_0 \circ f)(x) \right\}^2 dP(x) \\ &\quad + 2 \int \left\{ (\theta_0 \circ f)(x) - y \right\} \left\{ (\theta_n^{(x,y)} \circ f)(x) - (\theta_0 \circ f)(x) \right\} dP(x) \\ &\geq \int \left\{ (\theta_n^{(x,y)} \circ f)(x) - (\theta_0 \circ f)(x) \right\}^2 dP(x), \end{aligned}$$

as desired. Combining this lower bound with (11), we obtain the inequality

$$\int \left\{ (\theta_n^{(x,y)} \circ f)(x) - (\theta_0 \circ f)(x) \right\}^2 dP(x) \leq R_0(\theta_n^{(x,y)}) - R_0(\theta_0) \leq (P_n - P) \left[\{S_{\theta_n^{(x,y)}}\}^2 - \{S_{\theta_0}\}^2 \right] + O(1/n) \quad (14)$$

Define $\delta_n := \sqrt{\int \left\{ (\theta_n^{(x,y)} \circ f)(x) - (\theta_0 \circ f)(x) \right\}^2 dP(x)}$, the bound $B := \sup_{y \in \mathcal{Y}} |y|$, and the function class,

$$\Theta_{1,n} := \{(x, y) \mapsto \{(\theta_1 + \theta_2) \circ f - 2y\} \{(\theta_1 - \theta_2) \circ f\}\}.$$

Using this notation, (14) implies

$$\begin{aligned} \delta_n^2 &\leq \sup_{\theta_1, \theta_2 \in \Theta_{iso}: \|\theta_1 - \theta_2\| \leq \delta_n} \int \{(\theta_1 + \theta_2) \circ f(x) - 2y\} \{(\theta_1 - \theta_2) \circ f(x)\} d(P_n - P)(x, y) + O(1/n) \\ &\leq \sup_{h \in \Theta_{1,n}: \|h\| \leq 4B\delta_n} (P_n - P)h + O(1/n) \end{aligned}$$

Using the above inequality and C5, we will use an argument similar to the proof of Theorem 3 in [52] to establish that $\delta_n = O_p(n^{-1/3})$. This rate then implies the result of the theorem. To see this, note, by the reverse triangle inequality,

$$\begin{aligned} \left| S_n^{(x,y)}(y', x') - S_0(x', y') \right| &= \left| y' - \theta_n^{(x,y)}(x') - |y' - \theta_0(x')| \right| \\ &\leq \left| \theta_0(x') - \theta_n^{(x,y)}(x') \right|. \end{aligned}$$

Squaring and integrating the left- and right-hand sides, we find

$$\int \left| S_n^{(x,y)}(y', x') - S_0(x', y') \right|^2 dP(x', y') \leq \int \left| \theta_0(x') - \theta_n^{(x,y)}(x') \right|^2 dP(x') = \delta_n^2,$$

as desired.

We now establish that $\delta_n = O_p(n^{-1/3})$. For a function class \mathcal{F} , let $N(\epsilon, \mathcal{F}, L_2(P))$ denote the ϵ -covering number [53] of \mathcal{F} and define the uniform entropy integral of \mathcal{F} by

$$\mathcal{J}(\delta, \mathcal{F}) := \int_0^\delta \sup_Q \sqrt{\log N(\epsilon, \mathcal{F}, L_2(Q))} d\epsilon,$$

where the supremum is taken over all discrete probability distributions Q . We note that

$$\mathcal{J}(\delta, \Theta_{1,n}) = \int_0^\delta \sup_Q \sqrt{N(\epsilon, \Theta_{1,n}, \|\cdot\|_Q)} d\epsilon = \int_0^\delta \sup_Q \sqrt{N(\epsilon, \Theta_{iso}, \|\cdot\|_{Q \circ f^{-1}})} d\epsilon = \mathcal{J}(\delta, \Theta_{iso}),$$

where $Q \circ f^{-1}$ is the push-forward probability measure for the random variable $f(W)$. Additionally, the covering number bound for bounded monotone functions given in Theorem 2.7.5 of [53] implies that

$$\mathcal{J}(\delta, \Theta_{1,n}) = \mathcal{J}(\delta, \Theta_{iso}) \lesssim \sqrt{\delta}.$$

Recall that f is obtained from an external dataset, say \mathcal{E}_n , independent of the calibration data. Noting that f is deterministic conditional on a training dataset \mathcal{E}_n . Applying Theorem 2.1 of [54] conditional on \mathcal{E}_n , we obtain, for any $\delta > 0$, that

$$\begin{aligned} E \left[\sup_{h \in \Theta_{1,n}: \|h\| \leq 4B\delta} (P_n - P)h \mid \mathcal{E}_n \right] &\lesssim n^{-1/2} \mathcal{J}(\delta, \Theta_{1,n}) \left(1 + \frac{\mathcal{J}(\delta, \Theta_{1,n})}{\sqrt{n}\delta^2} \right) \\ &\lesssim n^{-1/2} \mathcal{J}(\delta, \Theta_{iso}) \left(1 + \frac{\mathcal{J}(\delta, \Theta_{iso})}{\sqrt{n}\delta^2} \right). \end{aligned}$$

Noting that the right-hand side of the above bound is deterministic, we conclude that

$$E \left[\sup_{h \in \Theta_{1,n}: \|h\| \leq 4B\delta} (P_n - P)h \right] \lesssim n^{-1/2} \mathcal{J}(\delta, \Theta_{iso}) \left(1 + \frac{\mathcal{J}(\delta, \Theta_{iso})}{\sqrt{n}\delta^2} \right).$$

We use the so-called ‘‘peeling’’ argument [53] to obtain our bound for δ_n . Note

$$\begin{aligned} P \left(\delta_n^2 \geq n^{-2/3} 2^M \right) &= \sum_{m=M}^{\infty} P \left(2^{m+1} \geq n^{2/3} \delta_n^2 \geq 2^m \right) \\ &= \sum_{m=M}^{\infty} P \left(2^{m+1} \geq n^{2/3} \delta_n^2 \geq 2^m, \delta_n^2 \leq \sup_{h \in \Theta_{1,n}: \|h\| \leq 4B\delta_n} (P_n - P)h + O(1/n) \right) \\ &= \sum_{m=M}^{\infty} P \left(2^{m+1} \geq n^{2/3} \delta_n^2 \geq 2^m, 2^{2m} n^{-2/3} \leq \sup_{h \in \Theta_{1,n}: \|h\| \leq 4B2^{m+1}n^{-1/3}} (P_n - P)h + O(1/n) \right) \\ &\leq \sum_{m=M}^{\infty} P \left(2^{2m} n^{-2/3} \leq \sup_{h \in \Theta_{1,n}: \|h\| \leq 4B2^{m+1}n^{-1/3}} (P_n - P)h + O(1/n) \right) \\ &\leq \sum_{m=M}^{\infty} \frac{E \left[\sup_{h \in \Theta_{1,n}: \|h\| \leq 4B2^{m+1}n^{-1/3}} (P_n - P)h \right] + O(1/n)}{2^{2m} n^{-2/3}} \\ &\leq \sum_{m=M}^{\infty} \frac{\mathcal{J}(2^{m+1}n^{-1/3}, \Theta_{iso}) \left(1 + \frac{\mathcal{J}(2^{2m+2}n^{-2/3}, \Theta_{iso})}{\sqrt{n}2^{m+1}n^{-1/3}} \right)}{\sqrt{n}2^{2m}n^{-2/3}} + \sum_{m=M}^{\infty} \frac{O(1/n)}{2^{2m}n^{-2/3}} \\ &\leq \sum_{m=M}^{\infty} \frac{2^{(m+1)/2}n^{-1/6}}{2^{2m}n^{-1/6}} + \sum_{m=M}^{\infty} \frac{o(1)}{2^{2m}} \\ &\lesssim \sum_{m=M}^{\infty} \frac{2^{(m+1)/2}}{2^{2m}}. \end{aligned}$$

Since $\sum_{m=1}^{\infty} \frac{2^{(m+1)/2}}{2^{2m}} < \infty$, we have that $\sum_{m=M}^{\infty} \frac{2^{(m+1)/2}}{2^{2m}} \rightarrow 0$ as $M \rightarrow \infty$. Therefore, for all $\varepsilon > 0$, we can choose $M > 0$ large enough so that

$$P \left(\delta_n^2 \geq n^{-2/3} 2^M \right) \leq \varepsilon.$$

We conclude that $\delta_n = O_p(n^{-1/3})$ as desired. □

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 abstract and introduction state our main contributions: a conformal prediction method that provides calibrated point predictions and prediction intervals with validity conditional on these point predictions. They also motivate our objective of prediction-conditional validity and self-calibration as a weaker, but feasible, notion of validity compared to context-conditional validity.

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: The objective of this paper is to construct calibrated point predictions and self-calibrated prediction intervals satisfying desiderata (i) and (ii). Below Theorem 4.1 and Theorem 4.2, we discuss two potential limitations: namely, that we construct a perfectly calibrated Venn-Abers multi- or set-prediction as opposed to a perfectly calibrated point prediction, and that our prediction interval is self-calibrated (achieving (ii)) concerning the unknown (oracle) perfectly calibrated point prediction. These limitations appear to be unavoidable as perfectly calibrated point predictions can generally not be constructed in finite samples without oracle knowledge [61, 60]. Other potential limitations regarding computation and implementation, as well as possible extensions are discussed in Section 3.4.

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: All assumptions for theoretical results are provided in the main text and formal and complete proofs are given in the appendix.

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: Important details of the experiment, including the dataset used for the analysis, the training, calibration, and test split, the definition of the evaluation metrics, the baseline methods for comparison, and how xgboost is trained, are all detailed in the text. Code to reproduce the experimental results is provided in the supplementary material.

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: A Python package implementing our method, as well as code for reproducing our experimental results, are provided in a supplementary zip file.

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: Important details such as the data split percentages, `xgboost` hyperparameters tuned using cross-validation, as well as parameters for SC-CP and baselines are all reported in the main text.

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: [No]

Justification: The experiment involved comparisons on real data, building on the experiment performed in [47]. As such, the ground truth is not known.

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: The methods implemented in this paper are not computationally intensive and were run in RStudio on a MacBook Pro with 16GB RAM and an M1 chip.

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: Our paper follows the NeurIPS Code of Ethics.

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: [NA]

Justification: Our work is foundational with no direct positive or negative societal impacts. We note, as alluded to in the introduction, that our work has the potential to improve the trustworthiness of machine learning methods by providing calibrated predictions and prediction intervals for uncertainty quantification.

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: Our paper does not involve releasing any data or models and, therefore, there is no risk of such 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: We use publicly available data and code, which are properly cited.

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: No new assets are introduced in this paper.

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: This paper does not involve crowdsourcing or 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: This 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.