Journal of the Royal Statistical Society Series B: Statistical Methodology, 2023, **85**, 1680–1705
https://doi.org/10.1093/jrsssb/qkad069
Advance access publication 17 July 2023
Original Article

Prediction sets adaptive to unknown covariate shift

Hongxiang Qiu, Edgar Dobriban and Eric Tchetgen Tchetgen

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

Address for correspondence: Hongxiang Qiu, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: qiuhx@wharton.upenn.edu

Abstract

Predicting sets of outcomes—instead of unique outcomes—is a promising solution to uncertainty quantification in statistical learning. Despite a rich literature on constructing prediction sets with statistical guarantees, adapting to unknown covariate shift—a prevalent issue in practice—poses a serious unsolved challenge. In this article, we show that prediction sets with finite-sample coverage guarantee are uninformative and propose a novel flexible distribution-free method, PredSet-1Step, to efficiently construct prediction sets with an asymptotic coverage guarantee under unknown covariate shift. We formally show that our method is *asymptotically probably approximately correct*, having well-calibrated coverage error with high confidence for large samples. We illustrate that it achieves nominal coverage in a number of experiments and a data set concerning HIV risk prediction in a South African cohort study. Our theory hinges on a new bound for the convergence rate of the coverage of Wald confidence intervals based on general asymptotically linear estimators.

Keywords: covariate shift, machine learning, nonparametric inference, nonparametric model, PAC guarantee, prediction set

Abbreviations: APAC, asymptotically probably approximately correct; CI, confidence interval; CUB, confidence upper bound; PAAC, probably asymptotically approximately correct; PAC, probably approximately correct.

1 Introduction

With recent advances in data acquisition, computing, and fitting algorithms, modern statistical machine learning methods can often produce accurate predictions. However, a key statistical challenge is to accurately quantify the uncertainty of the predictions. At the moment, it remains a subject of active research how to properly quantify uncertainty for the most powerful algorithms, such as deep neural nets and random forests. The difficulty is salient because in many applications, there are some instances whose outcomes are intrinsically difficult to predict accurately. In a classification problem, for such objects, it may be more desirable to produce a small prediction set that covers the truth with high probability, instead of outputting a single prediction. Reliable prediction sets can be especially important in safety-critical applications, such as in medicine (Berkenkamp et al., 2017; Bojarski et al., 2016; Gal et al., 2017; Kitani et al., 2012; Malik et al., 2019; Moja et al., 2014; Ren et al., 2017). The idea of such prediction sets has a rich statistical history dating back at least to the pioneering works of Wilks (1941), Wald (1943), Scheffe and Tukey (1945), and Tukey (1947, 1948).

To address this challenge, there is an emerging body of work on constructing prediction sets with coverage guarantees under various assumptions (see e.g. Bates et al., 2021; Chernozhukov et al., 2018b; Dunn et al., 2018; Lei et al., 2013, 2015, 2018; Lei & Wasserman, 2014; Park et al., 2020; Sadinle et al., 2019). Most of these methods have theoretical coverage guarantees when the data distribution for which the predictions are constructed matches that from which

Received: August 22, 2022. **Revised:** June 17, 2023. **Accepted:** June 20, 2023

the predictive model was generated. Among these, one of the best known methods is conformal prediction (CP) (see e.g. Chernozhukov et al., 2018b; Dunn et al., 2018; Saunders et al., 1999; Vovk et al., 1999, 2005; Lei et al., 2013, 2018; Lei & Wasserman, 2014). Conformal prediction can guarantee a high probability of covering a new observation, where the probability is marginal over the entire data set and the new observation.

Moreover, inductive conformal prediction (Papadopoulos et al., 2002)—where the data at hand are split into a training set and a calibration set, satisfies a *training-set conditional*, or *probably approximately correct* (PAC) guarantee (Park et al., 2020; Vovk, 2013). A prediction set learned from data is PAC if, over the randomness in the data, there is a high probability that its coverage error is low for new observations. This guarantee decouples the randomness in data at hand and the randomness in new observations. This allows a more fine-tuned control over the probability of error. This guarantee is a generalisation of the notion of tolerance regions of Wilks (1941) and Wald (1943) to the setting of supervised learning. As a generalisation in another direction, the method in Bates et al. (2021) provides risk-controlling prediction sets, which have low prediction risk with high probability over the randomness in the data.

The aforementioned methods are valid when the new observation and the data at hand are drawn from the same population, but this condition might fail to hold in applications. This phenomenon has been referred to in statistical machine learning as *dataset shift* (see e.g. Quiñonero-Candela et al., 2009; Shimodaira, 2000; Sugiyama & Kawanabe, 2012). More specifically, an important form of data set shift is *covariate shift*: a change of only the distribution of input covariates (or features), with an unchanged distribution of the outcome given covariates. For example, the shift may arise due to a change in the sampling probabilities of different subpopulations or individuals in surveys or designed experiments. Another setting is the assessment of future risks based on current data, such as predicting an individual's risk of a disease based on the patient's features. Here, the features can shift (e.g. as the conditions of the patient change), but the distribution of the outcome given the features may be unchanged (Quiñonero-Candela et al., 2009). Other examples of covariate shift include changes in the colour and lighting of image data (Hendrycks & Dietterich, 2019), or even adversarial attacks that slightly perturb the data points (Szegedy et al., 2014). In both cases, the distribution of labels given input covariates is unchanged.

A concrete example of covariate shift arises in a data set concerning HIV risk prediction in a South African cohort that was analysed in Tanser et al. (2013) and is also studied in our article. The empirical distribution of HIV prevalence across communities in the source (urban and rural communities) and target populations (peri-urban communities) are presented in Figure 1, and a severe covariate shift is present. The distributions of the outcome given covariates in the two populations appear to be similar.

Another example of covariate shift arises in causal inference. As discussed in Lei and Candès (2021) and studied in this article, under standard causal assumptions, predicting counterfactual (hypothetical) outcomes can be formulated as a prediction problem under covariate shift. In this setup, the two covariate distributions are those in the two treatment groups (treated and untreated), which may be different in observational data due to confounding.

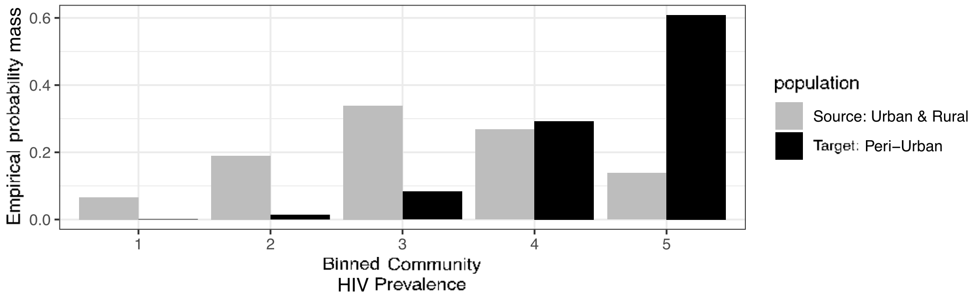

Figure 1. Empirical distribution of a covariate (binned community HIV prevalence with categories encoded by 1–6) in the two populations of the data concerning HIV risk prediction in a South African cohort.

In presence of covariate shift, prediction coverage guarantees may not hold if one assumes no covariate shift. Possible solutions have only recently formally been studied. Tibshirani et al. (2019) studied conformal prediction under covariate shift, assuming that the likelihood ratio is known *a priori*. Park et al. (2022) studied the PAC property of inductive conformal prediction (or, PAC prediction sets) under covariate shift. Their methods rely on knowing the covariate shift, i.e. the likelihood ratio of the covariate distribution in the target population to that in the source population, or on bounding its smoothness, which may not always be practical.

Cauchois et al. (2020) studied conformal prediction that is robust to a specified level of deviation of the target population from the source population. On the other hand, Lei and Candès (2021) studied conformal prediction under covariate shift without assuming that the likelihood ratio is known and allowed estimation of this ratio instead. In this article, we focus on the PAC property. In inductive conformal prediction under no covariate shift or known covariate shift, the PAC property can be obtained even though this method was developed to obtain marginal validity (see Vovk, 2013 for the case without covariate shift and Park et al., 2022 for the case with known covariate shift). However, to our knowledge, PAC property results for inductive conformal prediction under completely unknown covariate shift have not yet been obtained.

In this article, we focus on achieving a PAC guarantee and show that PAC prediction sets under unknown covariate shift are uninformative. We next propose novel methods to construct prediction sets that are *asymptotically PAC* (APAC) as the sample size grows to infinity, with a convergence rate that we unravel. Our main method, *PredSet-1Step*, is based on asymptotically efficient one-step corrected estimators of the true coverage error and the associated Wald confidence intervals. The procedure to construct the estimator is illustrated in Figure 2, and the procedure to construct prediction sets afterwards is illustrated in Online Supplementary Material, Figure S2 (see notations in the rest of this article). PredSet-1Step heavily relies on semiparametric efficiency theory (see e.g. Bickel et al., 1993; Chernozhukov et al., 2018a; Kennedy, 2022; Levit, 1974; Newey, 1990; Pfanzagl, 1985, 1990; Van Der Vaart, 1991; van der Vaart & Wellner, 1996) to obtain improved convergence rates. PredSet-1Step may also be used to construct asymptotically risk-controlling prediction sets (Bates et al., 2021).

This article is organised as follows. We introduce the problem setup, present a negative result on PAC prediction sets, and present identification results under unknown covariate shift in Section 2. In Section 3, we provide an overview of our proposed methods. We describe our method to estimate the likelihood ratio, and present pathwise differentiability results of the miscoverage in the target population; akin to those of Hahn (1998). These form the basis of our proposed methods. We next describe our proposed PredSet-1Step method, which builds on cross-fitting/double machine learning (Chernozhukov et al., 2018a; Schick, 1986), along with its theoretical properties, in Section 4. We show in Corollary 7 that PredSet-1Step yields APAC prediction sets with an error in the confidence level that is typically of order $n^{1/4}$ multiplied by the square root of the product of the convergence rates of estimators of two nuisance functions. These results are based on a novel bound on the difference between the realised and nominal coverage for Wald confidence intervals based on general asymptotically linear estimators (Theorem 4). We then present simulation studies in Section 5 and data analysis results in Section 6.

We present further results in the Online Supplementary Material. In Online Supplementary Material, Section S1, we propose an extension, PredSet-RS, of the rejection sampling method from Park et al. (2022). In Online Supplementary Material, Section S2, we present an alternative approach to PredSet-1Step, PredSet-TMLE, a targeted maximum likelihood estimation (TMLE) implementation of our efficient influence function based approach (Van der Laan & Rubin, 2006). We describe two methods to construct asymptotically risk-controlling prediction sets (ARCPS) in Online Supplementary Material, Section S3. These methods are slightly modified versions of PredSet-1Step and PredSet-TMLE. The proofs of our theoretical results can be found in Online Supplementary Material, Section S7. We discuss the PAC property, comparing it with marginal validity, in Online Supplementary Material, Section S8. We finally clarify a connection between causal inference and covariate shift, based on which we may apply methods for covariate shift to obtain well-calibrated prediction sets for individual treatment effects (ITEs), in Online Supplementary Material, Section S9. Our proposed methods are implemented in an R package available at https://github.com/QIU-Hongxiang-David/APACpredset.

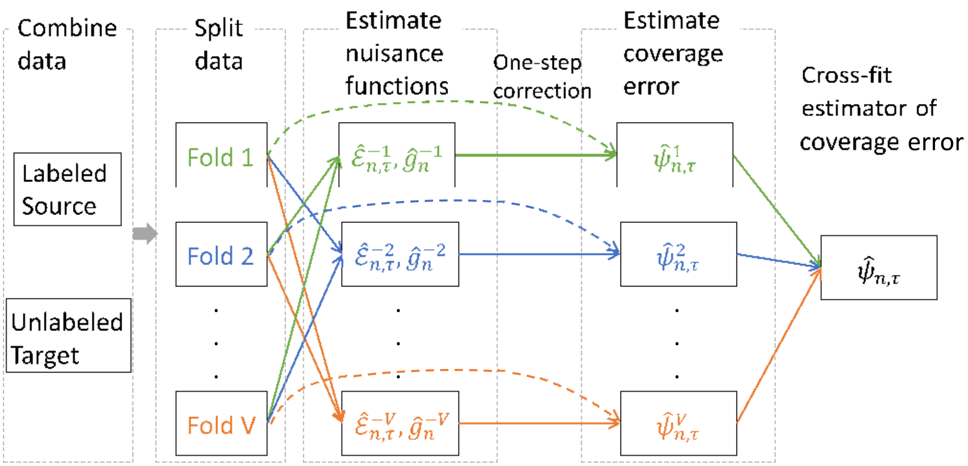
Combine data
Split data
Estimate nuisance functions
One-step correction
Estimate coverage error
Cross-fit estimator of coverage error
Labeled Source
Unlabeled Target
Fold 1
Fold 2
Fold V
$\hat{\mathcal{E}}_{n,\tau}^{-1}, \hat{g}_n^{-1}$
$\hat{\mathcal{E}}_{n,\tau}^{-2}, \hat{g}_n^{-2}$
$\hat{\mathcal{E}}_{n,\tau}^{-V}, \hat{g}_n^{-V}$
$\hat{\psi}_{n,\tau}^{1}$
$\hat{\psi}_{n,\tau}^{2}$
$\hat{\psi}_{n,\tau}^{V}$
$\hat{\psi}_{n,\tau}$

Figure 2. Overall procedure of the cross-fit one-step corrected estimator $\psi_{n,\tau}$ of the coverage error corresponding to the prediction set with threshold τ, which forms the basis of our PredSet-1Step method.

2 Problem setup and assumptions

2.1 Basic setting

Suppose one has observed labelled data from a *source population*, and unlabelled data from a *target population*. Denote a prototypical full (but unobserved) data point as $O := (A, X, Y) \sim \bar{P}^0$, where $A \in \{0, 1\}$ is the indicator of the data point being drawn from the source population ($A = 1$) or the target population ($A = 0$), $X \in \mathcal{X}$ are the covariates, and $Y \in \mathcal{Y}$ is the outcome, label or dependent variable to be predicted. The observed data points are of the form $O := (A, X, AY) \sim P^0$. In other words, in the observed data, outcomes (dependent variables) are observed only from the source population, and are missing from the target population (encoded as zero for notational convenience).

The observed data consist of n independently and identically distributed (i.i.d.) observed data points $O_i \sim P^0$ ($i \in [n] := \{1, 2, \ldots, n\}$). Let $s : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$ be a given scoring function. For example, when Y is a discrete variable, $s(x, y)$ may be an estimator of the probability of $Y = y$ given $X = x$ that has been trained from a held-out data set drawn from the source population. When Y is a continuous variable, $s(x, y)$ may be an estimator of the conditional density of Y at y given $X = x$ or $-|y - y(x)|$ for a given prediction model y. Tȟe function s can be arbitrary user-specified map-ping. We treat s as a fixed function throughout this article; as shown in the above examples, in practice s can be learned from a separate training set.

Let $\mathcal{B} \subseteq 2^{\mathcal{Y}}$ be the Borel σ-algebra of $\mathcal{Y}$, which is assumed to be a topological space. We refer to a map $C : \mathcal{X}$! $\mathcal{B}$ that assigns to each input $x \in \mathcal{X}$ a prediction set simply as a *prediction set*. Our goal is to construct a prediction set that is *asymptotically probably approximately correct* (APAC) in the target population. In other words, the prediction set should be asymptotically training-set-conditionally valid in the target population.

To be more precise, we first review a few related concepts. A prediction set C is *approximately correct* in the target population if the true coverage error in the target population, $\Pr_{\bar{P}^0}(Y \notin C(X) \mid A = 0)$, is less than or equal to a given target upper bound $\alpha_{\text{error}} \in (0, 1)$.

An estimated prediction set $\hat{C}$ constructed from the data is *probably approximately correct* (PAC) in the target population, with miscoverage level (also termed *content*) α_{error} and confidence level $1 - \alpha_{\text{conf}}$ ($\alpha_{\text{conf}} \in (0, 1)$), if, for a $\hat{C}$-independent draw (X, Y) from the target population,

$$\Pr_{P^0}(\Pr_{\bar{P}^0}(Y \notin \hat{C}(X) \mid A = 0, \hat{C}) \le \alpha_{\text{error}}) \ge 1 - \alpha_{\text{conf}}.$$

In other words, $\hat{C}$ is PAC if we have confidence at least $1 - \alpha_{\text{conf}}$ that the true coverage error of the estimated prediction set $\hat{C}$ in the target population is below the desired level α_{error}.

Remark 1 For conciseness in notations, in the rest of the article, we may drop the distribution over which a probability is taken over when this distribution is clear (e.g. $\bar{P}^0$ or P^0) from the context. For example, we may write the above PAC guarantee as $\Pr(\Pr(Y \notin \hat{C}(X) \mid A = 0, \hat{C}) \leq \alpha_{\text{error}}) \geq 1 - \alpha_{\text{conf}}$.

Methods to construct PAC prediction sets under covariate shift have been proposed when Y is observed in data points drawn from the target population, or when the distribution shift from the source to the target population is known (Park et al., 2020, 2022; Vovk, 2013). Risk-controlling prediction set (RCPS) have also been constructed, without considering covariate shift (Bates et al., 2021), while the problem with covariate shift has not been addressed, to our knowledge.

However, in our setting, neither the outcomes from the target population nor the distribution shift is known. Due to these unknown nuisance parameters, we have the following negative result on nontrivial prediction sets with a finite-sample marginal or PAC coverage guarantee.

Lemma 1 Suppose that $\mathcal{X}$ and $\mathcal{Y}$ are Euclidean spaces. Let $\bar{\mathcal{M}}^{*}$ be the set of all distributions $\bar{P}^0$ on the full data point $\bar{O}$ such that unknown covariate shift is present (namely Conditions 1–3 in Section 2.2 hold), and the joint distribution of (X, Y) is absolutely continuous with respect to the Lebesgue measure on $\mathcal{X} \times \mathcal{Y}$. Suppose that a (possibly randomised) prediction set $\hat{C}$ is PAC in the target population, that is,

$$\Pr\left\{\Pr(Y \notin \hat{C}(X) \mid A = 0, \hat{C}) \leq \alpha_{\text{error}}\right\} \geq 1 - \alpha_{\text{conf}}$$

for all $\bar{P}^0 \in \bar{\mathcal{M}}^{*}$. Then, for any $\bar{P}^0 \in \bar{\mathcal{M}}^{*}$ and a.e. $y \in \mathcal{Y}$ with respect to the Lebesgue measure,

$$\Pr(y \notin \hat{C}(X) \mid A = 0) \leq \alpha_{\text{error}} + \alpha_{\text{conf}}.$$

If $\alpha_{\text{error}} + \alpha_{\text{conf}} < 1$, Lemma 1 indicates that any PAC prediction set $\hat{C}$ in the target population under unknown covariate shift is essentially uninformative since it will contain almost any possible outcome with a nonzero probability for any data-generating mechanism. This lack of information can be clearly seen in the simple illustrative case where the support of Y is $\mathbb{R}$ and $Y \perp\!\!\!\perp X$. In this case, it would be desirable to obtain a PAC prediction set that outputs, for example, an esti-mated central or highest-density $1 - \alpha_{\text{error}}$ probability region of the distribution $Y \mid A = 0$. However, Lemma 1 implies that such a PAC prediction set does not exist, and that a PAC predic-tion set would instead cover almost every $y \in \mathbb{R}$ with probability at least $1 - (\alpha_{\text{error}} + \alpha_{\text{conf}})$ with respect to X. A similar negative result holds when Y is discrete. We prove this lemma by (a) obtain-ing a similar negative result for prediction sets with finite-sample marginal coverage guarantees (Online Supplementary Material, Lemma S1) and (b) using Theorem 2 and Remark 4 in Shah and Peters (2020). The proof can be found in Online Supplementary Material, Section S7.1.

Because of this negative result on finite-sample coverage guarantee, in this article, we choose to relax the validity criterion to an asymptotic one. It turns out that this way, we can account for unknown covariate shift. Recall that n is the sample size used to estimate the prediction set.

Definition A sequence of estimated prediction sets $(\hat{C}_n)_{n \geq 1}$ is asymptotically probably approximately correct (APAC) if

$$\Pr(\Pr(Y \notin \hat{C}_n(X) \mid A = 0, \hat{C}_n) \leq \alpha_{\text{error}}) \geq 1 - \alpha_{\text{conf}} + \mathrm{o}(1) \tag{1}$$

as $n \longrightarrow \infty$, where the $\mathrm{o}(1)$ term tends to zero as $n \longrightarrow \infty$.

In other words, a sequence of APAC prediction sets $\hat{C}_n$ is almost PAC for sufficiently large n. We will further quantify the magnitude of the $o(1)$ error in the confidence level. We use an estimated prediction set $\hat{C}_n$ and a sequence $(\hat{C}_n)_{n \geq 1}$ interchangeably in this article and may say that $\hat{C}_n$ is APAC. Further, we treat α_{error} and α_{conf} as fixed.

Remark 2 *Risk-controlling prediction set* (RCPS) (Bates et al., 2021) is more general than but similar to PAC. Our proposed PredSet-1Step method can be readily applied to constructing asymptotic RCPS (ARCPS) with a slight modification. We introduce the concept of ARCPS and describe the modified method in Online Supplementary Material, Section S3.

Vovk (2013) derived the PAC property of inductive conformal predictors (Papadopoulos et al., 2002), and Park et al. (2020) presented a nested perspective (Vovk et al., 2005). While the formulations of Vovk (2013) and Park et al. (2020) are equivalent, we follow Park et al. (2020). For thresholds $\tau \in \bar{\mathbb{R}}$, where $\bar{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$, we consider nested prediction sets (Vovk et al., 2005) of the form

$$C_\tau : x \mapsto \{y \in \mathcal{Y} : s(x, y) \geq \tau\}.$$

Since $C_{\tau_1}(x) \subseteq C_{\tau_2}(x)$ for any x and any $\tau_1 \geq \tau_2$, typical measures of the size of C_τ (such as the cardinality or Lebesgue measure) are nonincreasing functions of τ. Therefore, to obtain an APAC prediction set with a small size, given a finite set $\mathcal{T}_n \subseteq \bar{\mathbb{R}}$ of candidate thresholds, we select a threshold $\hat{\tau}_n \in \mathcal{T}_n$ such that $\Pr(\Pr(Y \notin C_{\hat{\tau}_n}(X) \mid A = 0) \leq \alpha_{\text{error}}) \geq 1 - \alpha_{\text{conf}} + o(1)$ and $\hat{\tau}_n$ is as large as pos-sible. Our methods under this setting are our main contribution.

We summarise one of our main results informally. Formal results can be found in later sections. The algorithm to estimate the coverage error $\Psi_\tau(P^0) = \Pr(Y \notin C_\tau(X) \mid A = 0)$ corresponding to the threshold τ used in PredSet-1Step is Algorithm 1. We will show in Corollary 7 that, under certain conditions, the prediction set $C_{\hat{\tau}_n^{\text{1Step}}}$ with the threshold $\hat{\tau}_n^{\text{1Step}}$ selected by PredSet-1Step is APAC:

$$\Pr(\Psi_{\hat{\tau}_n^{\text{1Step}}}(P^0) \leq \alpha_{\text{error}}) \geq 1 - \alpha_{\text{conf}} - \mathcal{C}\Delta_{n,\epsilon},$$

where $\mathcal{C}$ is an absolute positive constant and $\Delta_{n,\epsilon}$ is typically of order $n^{1/4}$ multiplied by the square root of the product of the convergence rates of two nuisance function estimators. This corollary relies on a novel result on the convergence rate of Wald confidence interval coverage for general asymptotically linear estimators (Theorem 4). The result bounds the difference between the true and the nominal coverage by three error terms: (a) the difference between the estimator and a sample mean, (b) the estimation error of the asymptotic variance, and (c) the difference of the distribution of the sample mean from its limiting normal distribution.

Remark 3 Beyond the APAC criterion, an alternative approach is to find a prediction set C that approximately solves

$$\min \quad \text{size}(C) \quad \text{subject to} \quad \widehat{\Pr}(Y \notin C(X) \mid A = 0) \leq \alpha_{\text{error}},$$

where $\widehat{\Pr}(Y \notin C(X) \mid A = 0)$ is an estimator of $\Pr(Y \notin C(X) \mid A = 0)$ and size(C) is a measure of the size of the prediction set C. This approach has been considered in Yang et al. (2022), and generally results in smaller prediction sets than the APAC ones we consider in this article. The reason is that the APAC guarantee requires approximately controlling the confidence level $1 - \alpha_{\text{conf}}$ to achieve the desired coverage error level α_{error} over the data, which leads to some conservativeness. This difference can also been seen from the PAC guarantee in Yang et al. (2022) taking the form

$$\Pr(\Pr(Y \notin \hat{C}_n(X) \mid A = 0, \hat{C}_n) \leq \alpha_{\text{error}} + o_p(1)) \geq 1 - \alpha_{\text{conf}}. \tag{2}$$

To distinguish from the APAC guarantee in (1), we call the guarantee in (2) a *probably asymptotically approximately correct* (PAAC) guarantee. The difference between APAC (1) and PAAC (2) is in the asymptotically vanishing approximation error: in APAC (1), the approximation is on the confidence level; in PAAC (2), the approximation is on the coverage error. This difference may seem subtle but has substantial impact on the performance of prediction sets satisfying these guarantees. We illustrate this difference by interpreting APAC (1) and PAAC (2) in words: APAC (1) states that, with confidence approaching the desired level $1 - \alpha_{\text{conf}}$, the true coverage error does not exceed the desired level α_{error}, but may frequently be a little conservative; PAAC (2) states that, with confidence at least $1 - \alpha_{\text{conf}}$, the true coverage error does not exceed the desired level α_{error} by much, but may frequently exceed α_{error} by a little. We also illustrate the difference in Figure 3. In some applications, having a high confidence guarantee on the desired level α_{error} of true coverage error at a price of slight conservativeness may be desirable, for example, for safety purposes.

We conclude this section by introducing a few more notations. We use C to denote an absolute positive constant that may vary line by line. For two scalar sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, we use $a_n \lesssim b_n$ to denote that for some constant $C > 0$ and all $n \geq 1$, $a_n \leq Cb_n$, and we define $\gtrsim$ similarly. We use $a_n \asymp b_n$ to denote that both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. We also adopt the little-o and big-O notations. For a probability distribution P and a scalar $p \geq 1$, we use $\| \cdot \|_{P,p}$ to denote the $L^p(P)$-norm of a function.

2.2 Identification

Without any further assumptions, it is impossible to estimate $\Pr(Y \notin C(X) \mid A = 0)$ for an arbitrary prediction set C, since the joint distribution $(X, Y) \mid A = 0$ of (X, Y) in the target population cannot be identified due to Y missing in the data. We make a few assumptions, following the standard setting in the covariate shift literature (see e.g. Quiñonero-Candela et al., 2009; Shimodaira, 2000; Sugiyama & Kawanabe, 2012), so that $\Pr(Y \notin C(X) \mid A = 0)$ can be identified as a functional of the true distribution P^0 on the observed data.

Let P^0_A denote the marginal distribution of A under P^0, $P^0_{X\mid a}$ denote the distribution of $X \mid A = a$ under P^0 for $a \in \{0, 1\}$, $\bar{P}^0_{Y\mid x,a}$ the distribution of $Y \mid X = x$, $A = a$ under the full data distribution $\bar{P}^0$, and $P^0_{Y\mid x} := \bar{P}^0_{Y\mid x,1}$. For any distribution P of the observed data point O, we define these marginal and conditional distributions similarly and denote them with similar notations except that the superscript 0 denoting P^0 is dropped. For example, $P_{X\mid a}$ stands for the distribution of $X \mid A = a$ under P. It will also be convenient to define the loss function

$$Z_\tau : (x, y) \mapsto 1(y \notin C_\tau(x))$$

for any $\tau \in \bar{\mathbb{R}}$. Our first condition is:

Condition 1 (Data available from both populations). $0 < \Pr(A = 1) < 1$.

This condition ensures that data points from both source and target populations are collected in sufficient quantity, and that the conditional distributions introduced above are well defined. In practice, this condition requires that a reasonable amount of data from both populations is collected.

Next, we state the key *covariate shift* assumption (see e.g. Quiñonero-Candela et al., 2009; Shimodaira, 2000; Sugiyama & Kawanabe, 2012), which is central to our article.

Condition 2 (Covariate shift: Identical conditional outcome distribution). The conditional distribution of $Y \mid X = x$ in the target population is identical to that in the source population for all $x \in \mathcal{X}$.[1] Mathematically, $\bar{P}^0_{Y\mid x,1} = \bar{P}^0_{Y\mid x,0} = P^0_{Y\mid x}$.

Condition 2 is similar to the missing at random assumption in the missing data literature (see e.g. Little & Rubin, 2019). It holds, for instance, if in the target we observe the same Y (e.g.

[1] Formally, this has to hold almost surely with respect to a given probability measure over $\mathcal{X}$, with respect to which all distributions of X considered are absolutely continuous; however, we simplify the statement for clarity.

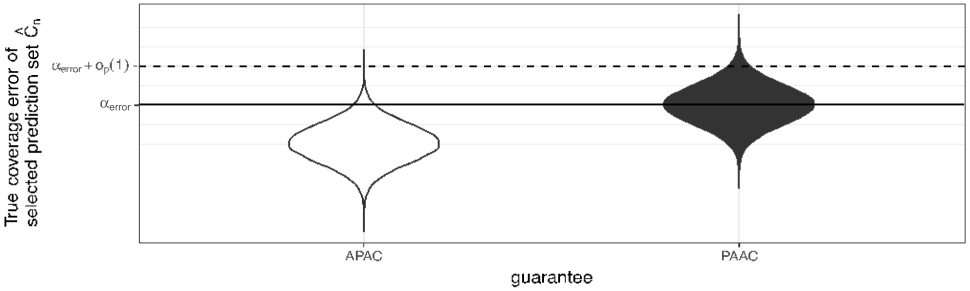

Figure 3. Exemplar sampling distributions of the true coverage error $\Pr(Y \notin \hat{C}_n(X) \mid A = 0, \hat{C}_n)$ of prediction sets $\hat{C}_n$ satisfying APAC and PAAC guarantees, respectively.

does a car face left or right), with X from a different distribution (images from cities A vs. B). Finally, we have an assumption to ensure that the target population overlaps with the source population.

Condition 3 (Dominance of covariate distributions). The marginal distribution of X in the target population, $P^0_{X|0}$, is dominated by that in the source population, $P^0_{X|1}$; that is, the Radon–Nikodym derivative $w_0 := \mathrm{d}P^0_{X|0}/\mathrm{d}P^0_{X|1}$ is well defined.

We assume that Conditions 1–3 hold throughout this article. For any distribution P of the observed data point O satisfying Condition 3 and any $\tau \in \mathbb{R}$, we define the functionals

$$w_P := \mathrm{d}P_{X|0}/\mathrm{d}P_{X|1} \quad \text{and} \quad \mathcal{E}_{P,\tau} : x \mapsto \Pr_P(Y \notin C_\tau(X) \mid X = x, A = 1). \tag{3}$$

We will also replace P in the subscripts of these and other quantities with 0 when referring to the functional components of P^0. Here, w_P is the likelihood ratio between target and source covariate distributions under P, and $\mathcal{E}_{P,\tau}$ is the covariate-conditional coverage error of the prediction set C_τ in the source population. It is not hard to show that, under the distribution $\bar{P}$ of the complete but unobserved data, we can also express $\mathcal{E}_{P,\tau}$ in terms of the target population as $\mathcal{E}_{P,\tau}(x) = \Pr^{-}(Y \notin C_{\bar{\beta}}(X) \mid X = x, A = 0)$. Further, we define

$$\Psi^{\text{Gcomp}}_\tau : P \mapsto E_P[\mathcal{E}_{P,\tau}(X) \mid A = 0] \quad \text{and} \quad \Psi^{\text{weight}}_\tau : P \mapsto E_P[w_P(X) Z_\tau(X, Y) \mid A = 1].$$

One can verify that for $j \in \{\text{Gcomp, weight}\}$,

$$\Psi^j_\tau(P^0) = \Pr_{\bar{P}^0}(Y \notin C_\tau(X) \mid A = 0). \tag{4}$$

In other words, although $\Psi^{\text{Gcomp}}_\tau(P^0)$ and $\Psi^{\text{weight}}_\tau(P^0)$ take as inputs different components of P^0, both correspond to the same functional of P^0, the coverage error of the prediction set C_τ in the target population. We will use Ψ_τ to denote these two functionals when we need not distinguish their mathematical expressions. In other words, $\Psi_\tau(P^0)$ equals the probability that $Y \notin C_\tau(X)$ in the 'covariate shifted' population where $A = 0$:

$$\Psi_\tau(P^0) = \Pr_{\bar{P}^0}(Y \notin C_\tau(X) \mid A = 0). \tag{5}$$

Borrowing terminology from causal inference and missing data, we refer to $\Psi^{\text{Gcomp}}_\tau(P^0)$ and $\Psi^{\text{weight}}_\tau(P^0)$ as the G-computation formula and the weighted formula, respectively.

Remark 4 Both Ψ^{Gcomp}_τ and $\Psi^{\text{weight}}_\tau$ take as inputs only certain components of the distribution rather than the entire distribution and hence may be computed as long as the relevant components are defined. For example, $\Psi^{\text{Gcomp}}_\tau(P)$ is defined if the distribution $P_{X|0}$ and $\mathcal{E}_{P,\tau}$ are defined. We will specify only the required components when defining our estimators.

Remark 5 There is a connection between counterfactuals in causal inference and covariate shift, as pointed out in Lei and Candès (2021). We discuss this connection in more detail in Online Supplementary Material, Section S9.

3 Overview and preliminaries of proposed methods

In all methods we propose, we assume that a finite set $\mathcal{T}_n$ of candidate thresholds is given, with a cardinality that may grow to infinity with n. Since, as a function of τ, there are at most $n + 1$ versions of the observed miscoverage indicators $\{Z_\tau(X_i, Y_i) = 1(s(X_i, Y_i) < \tau), i \in [n]\}$ in any data set, each corresponding to a threshold in the set $\{s(X_i, Y_i) : i \in [n]\} \cup \infty$, this assumption is not stringent.

Our general strategy is to construct an asymptotically valid $(1 - \alpha_{\text{conf}})$-confidence upper bound (CUB) for $\Psi_\tau(P^0)$ for each threshold $\tau \in \mathcal{T}_n$, and select the largest threshold $\hat{\tau}_n \in \mathcal{T}_n$ such that, for any candidate threshold less than or equal to $\hat{\tau}_n$, the corresponding CUB is less than α_{error}. This procedure is illustrated in Online Supplementary Material, Figure S2. To construct accurate approximate confidence intervals (CIs), we rely on semiparametric efficiency theory (Bickel, 1982; Bickel & Doksum, 2015; Bickel et al., 1993; Newey, 1990; Pfanzagl, 1985, 1990; Van Der Vaart, 1991; van der Vaart, 1998).

3.1 Estimation of nuisance functions

For a given threshold τ, we will see in Section 3.2 that it is helpful to estimate nuisance functions corresponding to the pointwise coverage error

$$\mathrm{E}_{0,\tau} = \mathrm{E}_{P^0,\tau} : x \mapsto \Pr_{P^0}(Y \notin C_\tau(X) \mid X = x, A = 1) \tag{6}$$

and the covariate shift likelihood ratio w_0 from Condition (3). An estimator $\mathrm{E}_{n,\tau}$ of $\mathrm{E}_{0,\tau}$ may be obtained with standard classification or regression algorithms in the subsample from the source population with dependent variable $Z_\tau(X, Y)$ and covariate X.

However, for the estimation of the likelihood ratio w_0, we opt for a re-parametrization to a classification problem. Inspired by Friedman (2004), Bickel et al. (2007), Sugiyama et al. (2008), and Menon and Ong (2016), we use the following observation from Bayes' Theorem. For any distribution P of the observed data point O satisfying Condition 3, define $g_P : x \mapsto [0, 1]$ and $\gamma_P \in (0, 1)$ via

$$g_P(x) := \Pr_P(A = 1 \mid X = x), \quad \gamma_P := \Pr_P(A = 1). \tag{7}$$

We define g_0 and γ_0 similarly for P^0:

$$g_0(x) := \Pr_{P^0}(A = 1 \mid X = x), \quad \gamma_0 := \Pr_{P^0}(A = 1). \tag{8}$$

Following terminology in causal inference, we call g_0 the *propensity score function* (Rosenbaum & Rubin, 1983). When referring to generic propensity score functions and probabilities, we will write g and γ instead of g_P and γ_P. For any propensity score function g and any probability $\gamma \in (0, 1)$, we define $\mathcal{W}(g, \gamma) : \mathcal{X} \to [0, \infty)$ as

$$\mathcal{W}(g, \gamma)(x) := \frac{1 - g(x)}{g(x)} \frac{\gamma}{1 - \gamma}. \tag{9}$$

Bayes' theorem directly shows that

$$w_0(x) = \mathcal{W}(g_0, \gamma_0)(x). \tag{10}$$

We will use this reparameterization through the rest of this article. We can estimate γ_0 by γ_n obtained from the empirical distribution. Further, we may estimate g_0 by g_n obtained with standard classification or regression algorithms with dependent variable A and covariate X. In our experience, existing classification techniques are more flexible in our setting than density estimation methods. For instance, density estimation procedures might need adjustment according to the support of variables in X (e.g. bounded continuous, unbounded continuous, discrete, a mixture, etc.), while most classification methods need not make this distinction.

3.2 Pathwise differentiability

We next present results on pathwise differentiability of the error rate parameter Ψ_τ with respect to $\mathcal{M}$, a model that is nonparametric at P^0 (Bickel, 1982; Bickel & Doksum, 2015; Bickel et al., 1993; Levit, 1974; Newey, 1990; Pfanzagl, 1985, 1990; Van Der Vaart, 1991; van der Vaart, 1998). We first briefly describe the intuition behind these terminologies. Consider a generic one-dimensional parametric submodel $(P_H^\epsilon)_{\epsilon \in \mathbb{R}}$ satisfying $\mathrm{d}P_H^\epsilon/\mathrm{d}P^0(o) \approx 1 + \epsilon H(o)$ for some function H with $\mathrm{E}_{P^0}[H(O)] = 0$ and finite variance. We only consider *regular parametric submodels* (see e.g. Newey, 1990, for more details). The function H is called the *score function* of this submodel. We say that a model $\mathcal{M}$ is nonparametric at P^0 if, for any function H with mean zero and finite variance, $P_H^\epsilon \in \mathcal{M}$ for ϵ sufficiently close to zero. Roughly speaking, H encodes the direction of local perturbations of P^0 in the submodel, and a nonparametric model allows any perturbation of P^0.

We focus on nonparametric models in the main text, in which case no information about P^0 is known. In particular, the estimator $(\mathcal{E}_{n,\tau}, w_n)$ of $(\mathcal{E}_{0,\tau}, w_0)$ may converge in probability in an $L^2(P^0)$ sense at a rate slower than or equal to $n^{-1/2}$. This rate is typically slower than the parametric rate $n^{-1/2}$ as long as the covariate X has continuous components.

A parameter $\Psi : \mathcal{M} \longrightarrow \mathbb{R}$ is pathwise differentiable if $\mathrm{d}\Psi(P_H^\epsilon)/\mathrm{d}\epsilon|_{\epsilon=0} = \mathrm{E}_{P^0}[H(O)D(O)]$ for some function D with $\mathrm{E}_{P^0}[D(O)] = 0$ and finite variance. This function D is called a *gradient* of the parameter Ψ at P^0, since it characterises the local change in the value of the parameter corresponding to a perturbation of P^0. We can then heuristically expand $\Psi(P_H^\epsilon)$ around $\Psi(P^0)$:

$$
\begin{aligned}
\Psi(P_H^\epsilon) - \Psi(P^0) &\approx \epsilon \int H(o)D(o)P^0(\mathrm{d}o) \\
&= \int (1 + \epsilon H(o))D(o)P^0(\mathrm{d}o) - \int D(o)P^0(\mathrm{d}o) \approx \int D(o)(P_H^\epsilon - P^0)(\mathrm{d}o).
\end{aligned} \tag{11}
$$

In nonparametric models, the gradient D is unique and also called the *canonical gradient*. The above explanation is informal, and we refer the readers to Online Supplementary Material, Section S7.2 and to Levit (1974) and Pfanzagl (1985, 1990) for more details. The pathwise differentiability of an estimator is closely related to efficiency and is crucial for the construction of a root-n-consistent and asymptotically normal estimator.

An estimator is *asymptotically efficient* under a nonparametric model if it equals the estimand plus the sample mean of the canonical gradient, up to an error $\mathrm{o}_p(n^{-1/2})$. An asymptotically efficient estimator is root-n consistent and has the smallest possible asymptotic variance among a large class of estimators called *regular estimators* (see e.g. Section 8.5 in van der Vaart, 1998). Hence, the result on pathwise differentiability of Ψ_τ forms the basis of constructing efficient estimators of $\Psi_\tau(P^0)$, based on which approximate CUBs can be constructed under a nonparametric model.

Now, we return to our prediction set problem. Consider an arbitrary function $\mathcal{E}$ defined on $\mathcal{X}$ with range contained in [0, 1], any scalar $\gamma \in (0, 1)$, and any positive scalar π. For each $\tau \in \mathbb{R}$, with $o := (a, x, y)$, we define the function

$$
D_\tau(P, g, \gamma) : o \mapsto \frac{a}{\gamma_P} \mathcal{W}(g, \gamma)(x) \left\{ Z_\tau(x, y) - \mathcal{E}_{P,\tau}(x) \right\} + \frac{1 - a}{1 - \gamma_P} [\mathcal{E}_{P,\tau}(x) - \Psi_\tau^{\mathrm{Gcomp}}(P)]. \tag{12}
$$

For notational convenience, we suppress the dependence of this gradient function on the target parameter $\Psi_\tau(P)$; this dependence is implicit through the dependence on P.

We require an additional bounded likelihood ratio condition, which is standard in the literature on covariate shift (Quiñonero-Candela et al., 2009; Shimodaira, 2000; Sugiyama & Kawanabe, 2012).

Condition 4 (Bounded likelihood ratio). There exists a constant $B < \infty$ such that $\sup_{x \in \mathcal{X}} w_0(x) < B$. Equivalently, there exists a constant $\delta \in (0, 1)$ such that the propensity score is bounded away from zero, namely $\inf_{x \in \mathcal{X}} g_0(x) > \delta$. Here, δ may be taken as $\gamma_0/(B(1 - \gamma_0) + 1)$.

Under Condition 4, it holds that $\sup_{\tau \in \bar{\mathbb{R}}} \mathrm{E}_{P^0}[D_\tau(P^0, g_0, \gamma_0)(O)^2] < \infty$, because Z_τ is bounded. The pathwise differentiability of Ψ is presented in the following theorem, our first main result.

Theorem 2 (Pathwise differentiability of Ψ_τ). Under Conditions 1–4, for each $\tau \in \bar{\mathbb{R}}$, the functional Ψ_τ from (5), where $\Psi_\tau(P^0) = \Pr(Y \notin C_\tau(X) \mid A = 0)$ is the coverage error in the target, 'covariate shifted' population with $A = 0$, is pathwise differentiable at P^0 relative to $\mathcal{M}$ with canonical gradient $D_\tau(P^0, g_0, \gamma_0)$ from (12).

The proof of Theorem 2 is related to the proof of Theorem 1 in Hahn (1998): with $1 - A$ being the treatment indicator D in Hahn (1998), $1(Y \notin \hat{C}_\tau(X))$ being the counterfactual outcome Y_0 in Hahn (1998), $\Psi_\tau(P^0)$ can be written as the mean counterfactual outcome $\mathrm{E}[Y_0 \mid D = 1]$ in the treated group in Hahn (1998). Estimating this is the main challenge in estimating the average treatment effect $\mathrm{E}[Y_1 - Y_0 \mid D = 1]$ on the treated; the canonical gradient of $\mathrm{E}[1(Y \notin \hat{C}_\tau(X)) \mid A = 0]$ can be calculated using arguments similar to the proof of Theorem 1 in Hahn (1998). We provide the proof in Online Supplementary Material, Section S7.2. Since both nuisance functions $\mathcal{E}_{0,\tau}$ and w_0 appear in the canonical gradient, it is helpful to estimate both functions in order to construct asymptotically efficient estimators of $\Psi_\tau(P^0)$ as well as asymptotically valid CUBs. As is known in the sieve estimation literature (see e.g. Chen, 2007; Qiu et al., 2021; Shen, 1997) and other nonparametric inference literature (see e.g. Bickel & Ritov, 2003; Newey et al., 1998, 2004), it is possible to only estimate—for example—$\mathcal{E}_{0,\tau}$, with specific nonparametric methods, and still obtain an asymptotically efficient estimator of $\Psi_\tau(P^0)$. In this article, we do not take these approaches and propose methods that require estimating both nuisance functions in our procedures to allow for the most generality and flexibility in choosing estimators of nuisance functions.

Remark 6 The pathwise differentiability of Ψ_τ does not use that the loss function $Z_\tau(x, y) = 1(y \notin C_\tau(x))$ is binary. Therefore, our approach works for general loss functions, and we may construct asymptotically efficient estimators in that setting. Then, we can construct asymptotically efficient estimators for the true risk that corresponds to a general loss function for a prediction set under covariate shift. In particular, PredSet-1Step may be used with slight modifications for the estimation of the conditional risk function $\mathcal{E}_{0,\tau}$ for general losses. We present the corresponding results for constructing ARCPS in Online Supplementary Material, Section S3.

4 PredSet-1Step

In this section, we describe the PredSet-1Step method, based on an asymptotically efficient one-step corrected estimator of $\Psi_\tau(P^0)$, along with its main theoretical properties. For each candidate $\tau \in \mathcal{T}_n$, we first construct an asymptotically efficient estimator $\psi_{n,\tau}$ of $\Psi_\tau(P^0)$, and then obtain a consistent estimator $\hat{\sigma}^2_{n,\tau}$ of the asymptotic variance $\sigma^2_{0,\tau}$ of $\psi_{n,\tau}$. We finally construct a Wald CUB based on $\psi_{n,\tau}$ and $\hat{\sigma}^2_{n,\tau}$. We select thresholds $\tau \in \mathcal{T}_n$ for prediction sets based on the CUBs. We next describe each step in more detail.

4.1 Cross-fit one-step corrected estimator

In this section, we describe a cross-fit one-step corrected estimator of $\Psi_\tau(P^0)$ for a given τ. After obtaining an estimator $\hat{\mathcal{E}}_{n,\tau}$ of $\mathcal{E}_{0,\tau}$ via parametric (e.g. logistic regression, neural nets) or nonparametric methods, it might be tempting to estimate $\Psi_\tau(P^0)$ by the mean of $\hat{\mathcal{E}}_{n,\tau}(X)$ among observations from the target population. In other words, $\hat{\mathcal{E}}_{n,\tau}$ is plugged into Ψ^{Gcomp}. However, in general, this plug-in estimator may not be rate-optimal and may invalidate subsequent CUB construction and APAC guarantees. The reason is a bias term that may dominate the convergence of this estimator.

Fortunately, this excessive bias can be reduced by a one-step correction on the standard plug-in estimator (see e.g. Theorem 4 in Chapter 3 of Le Cam (1969) for early development of this idea for parametric models, and Pfanzagl (1985) and Chapter 25 in van der Vaart (1998) for more modern generalisations to semi/nonparametric models.) We further incorporate cross-fitting into the procedure to relax restrictions on the techniques used to estimate nuisance functions $\mathcal{E}_{0,\tau}$ and g_0. This technique improves performance in small to moderate samples. Such sample splitting ideas date back at least to Hajek (1962) and Bickel (1982).

Suppose that the data are split into V folds of approximately equal size completely at random. We assume that $V \geq 2$ is fixed. Common choices of V include 5 and 10. Let $I_v \subseteq [n]$ denote the index set of observations in fold $v \in [V]$. We use $P^{n,v}$ to denote the empirical distribution of data in fold v. We also use $P_A^{n,v}$ and $P_{X \mid a}^{n,v}$ to denote the empirical distribution of A and $X \mid A = a$ corresponding to data in fold v.

For each τ, let $\hat{\mathrm{E}}_{n,\tau}^{-v}$ denote a flexible estimator of $\mathrm{E}_{0,\tau}$ obtained from data points out of fold v via, for example, standard regression or supervised statistical learning tools. We also use $\hat{g}_n^{-v}$ to denote a flexible estimator of g_0 obtained from data points out of fold v. We define

$$\hat{\gamma}_n^v := \Pr_{P^{n,v}}(A = 1). \tag{13}$$

We construct a cross-fit one-step corrected estimator using Algorithm 1. Direct calculation shows that the one-step corrected estimator $\hat{\psi}_{n,\tau}^v$ for fold v from (15) equals

$$\Psi_\tau^{\mathrm{Gcomp}}(\hat{P}_\tau^{n,v}) + \frac{1}{|I_v|} \sum_{i \in I_v} D_\tau(\hat{P}_\tau^{n,v}, \hat{g}_n^{-v}, \hat{\gamma}_n^v)(O_i). \tag{14}$$

The key one-step correction in (14) based on the canonical gradient is similar to a correction based on a linear approximation using the gradient in (12) at the estimated distribution $\hat{P}_\tau^{n,v}$. We roughly describe the intuition below in an informal manner, and refer the readers to Online Supplementary Material, Section S6.3 for technical details. Following (11), we expand $\Psi(P^0)$ around $\Psi(\hat{P}_\tau^{n,v})$:

$$\begin{aligned}\Psi(P^0) &\approx \Psi(\hat{P}_\tau^{n,v}) + \int D_\tau(\hat{P}_\tau^{n,v}, \hat{g}_n^{-v}, \hat{\gamma}_n^v)(o)(P^0 - \hat{P}_\tau^{n,v})(do)\\ &= \Psi(\hat{P}_\tau^{n,v}) + \mathrm{E}_{P^0}[D_\tau(\hat{P}_\tau^{n,v}, \hat{g}_n^{-v}, \hat{\gamma}_n^v)(O)],\end{aligned}$$

where, in the expectation in the second line, we treat $(\hat{P}_\tau^{n,v}, \hat{g}_n^{-v}, \hat{\gamma}_n^v)$ as fixed. The second equality follows because a gradient at P has mean zero under P. Since the above correction term is unknown, we replace the expectation under P^0 with the empirical mean and thus find the one-step correction in (14). This idea is illustrated in Figure 4. This one-step correction is crucial to ensure root-n consistency and asymptotic normality of the estimator, as we illustrate in a simulation shown in Online Supplementary Material, Figure S3.

Algorithm 1 Cross-fit one-step estimator of coverage error $\Psi_\tau(P^0)$ used in PredSet-1Step

```
for $v \in [V]$ do Estimate $g_0$ by $\hat{g}_n^{-v}$ using data out of fold $v$.
for $v \in [V]$ and $\tau \in \mathrm{T}_n$ do Estimate $\mathrm{E}_{0,\tau}$ by $\hat{\mathrm{E}}_{n,\tau}^{-v}$ using data out of fold $v$.
for $v \in [V]$ and $\tau \in \mathrm{T}_n$ do    ▷ (Obtain a one-step corrected estimator for fold $v$)
    Let $\hat{P}_\tau^{n,v}$ be a distribution with the following components: (a) marginal distribution of $A$ being $P_A^{n,v}$, (b)
      conditional distribution of $X \mid A = 0$ being $P_{X \mid 0}^{n,v}$, and (c) distribution of $Z_\tau \mid X, A = 1$ defined by $\hat{\mathrm{E}}_{n,\tau}^{-v}$
    Let $|I_v|$ be the cardinality of the index set $I_v$. Set
```

$$\hat{\psi}_{n,\tau}^v := \frac{\sum_{i \in I_v} (1 - A_i)\hat{\mathrm{E}}_{n,\tau}^{-v}(X_i)}{\sum_{i \in I_v} (1 - A_i)} + \frac{1}{|I_v|} \sum_{i \in I_v} \frac{A_i}{\hat{\gamma}_n^v} W(\hat{g}_n^{-v}; \hat{\gamma}_n)[Z_\tau(X_i, Y_i) - \hat{\mathrm{E}}_{n,\tau}^{-v}(X_i)]. \tag{15}$$

```
for $\tau \in \mathrm{T}_n$ do Obtain the cross-fit one-step corrected estimator for threshold $\tau$:
```

$$\hat{\psi}_{n,\tau} := \frac{1}{n} \sum_{v=1}^{V} |I_v| \hat{\psi}_{n,\tau}^v. \tag{16}$$

We require two additional conditions on $\hat{g}_n^{-v}$ and $\hat{\mathrm{E}}_{n,\tau}^{-v}$ for $\hat{\psi}_{n,\tau}$ to be asymptotically efficient.

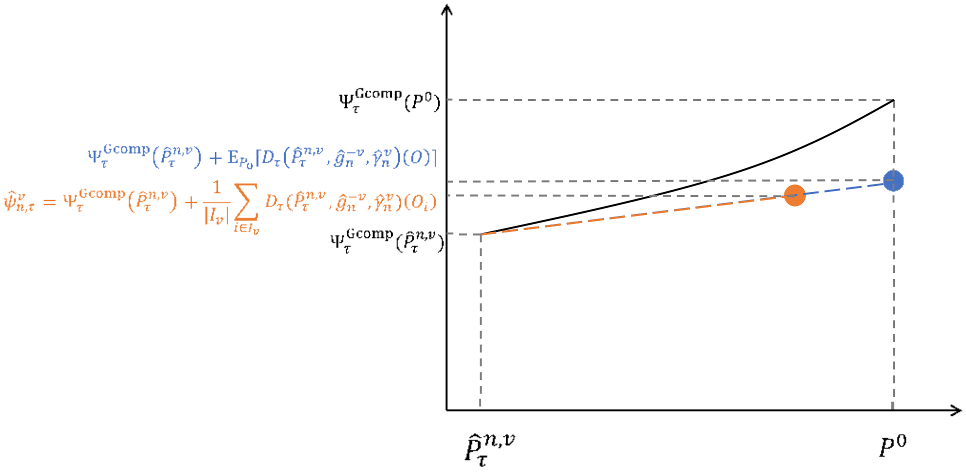
$\Psi_\tau^{\mathrm{Gcomp}}(P^0)$
$\Psi_\tau^{\mathrm{Gcomp}}(\hat{P}_\tau^{n,v}) + \mathrm{E}_{P_0}[D_\tau(\hat{P}_\tau^{n,v}, \hat{g}_n^{-v}, \hat{\gamma}_n^v)(O)]$
$\hat{\psi}_{n,\tau}^v = \Psi_\tau^{\mathrm{Gcomp}}(\hat{P}_\tau^{n,v}) + \frac{1}{|I_v|}\sum_{i \in I_v} D_\tau(\hat{P}_\tau^{n,v}, \hat{g}_n^{-v}, \hat{\gamma}_n^v)(O_i)$
$\Psi_\tau^{\mathrm{Gcomp}}(\hat{P}_\tau^{n,v})$
$\hat{P}_\tau^{n,v}$
P^0

Figure 4. Illustration of the idea behind the one-step correction in (14). In this figure, $\Psi_\tau^{\mathrm{Gcomp}}(\hat{P}_\tau^{n,v}) + \mathrm{E}_{P^0}[D_\tau(\hat{P}_\tau^{n,v}, \hat{g}_n^{-v}, \hat{\gamma}_n^v)(O)]$ is the ideal/oracle first-order approximation to the estimand $\Psi_\tau^{\mathrm{Gcomp}}(P^0)$ at the naïve plug-in estimator $\Psi_\tau^{\mathrm{Gcomp}}(\hat{P}_\tau^{n,v})$; $\hat{\psi}_{n,\tau}^v$ is the one-step corrected estimator.

Condition 5 (Bounded propensity score estimator). For some non-negative sequence $(q_n)_{n\geq 0}$ tending to zero as $n \to \infty$, with probability $1 - q_n$, $\inf_{v\in[V],x\in\mathcal{X}} \hat{g}_n^{-v}(x) > \delta$, where δ is the constant introduced in Condition 4.

Typically, with appropriate regularisation to avoid overfitting, the estimator $\hat{g}_n^{-v}$ of the propensity score is bounded away from zero except in extremely ill-posed data sets. These occur with extremely small probability (e.g. if the covariate X is discrete and for some x, all observations with $X = x$ are from the target population). Moreover, the user can always truncate the estimator to be bounded away from zero, in which case $q_n = 0$. Thus we often expect q_n in Condition 5 to decrease to zero at a much faster rate than the convergence rates of the nuisance function estimators in Condition 6. This does not have an effect on our second main result below, which is on the asymptotic efficiency of the cross-fit one-step estimator $\hat{\psi}_{n,\tau}$, but it will affect the Wald confidence interval coverage in the next subsection. Recall that $\|\cdot\|_{P,p}$ stands for the $L^p(P)$-norm of a function. Our condition for nuisance estimators is as follows.

Condition 6 (Sufficient rates for nuisance estimators). The following conditions hold:

$$\mathrm{E}_{P^0} \sup_{v\in[V],\tau\in\mathrm{T}_n} \|\hat{\mathrm{E}}_{n,\tau}^{-v} - \mathrm{E}_{0,\tau}\|_{P^0_{X\mid 0},2} = \mathrm{o}(1),$$

$$\mathrm{E}_{P^0} \sup_{v\in[V]} \|(1 - \hat{g}_n^{-v})/\hat{g}_n^{-v} - (1 - g_0)/g_0\|_{P^0_{X\mid 0},2} = \mathrm{o}(1),$$

$$\mathrm{E}_{P^0} \sup_{v\in[V],\tau\in\mathrm{T}_n} \int \left|\left\{\frac{1 - \hat{g}_n^{-v}(x)}{\hat{g}_n^{-v}(x)} - \frac{1 - g_0(x)}{g_0(x)}\right\}\left\{\hat{\mathrm{E}}_{n,\tau}^{-v}(x) - \hat{\mathrm{E}}_{0,\tau}(x)\right\}\right| P^0_{X\mid 1}(\mathrm{d}x) = \mathrm{o}(n^{-1/2}).$$

By the Cauchy–Schwarz inequality, a sufficient condition for the last equation in Condition 6 is the following:

$$\mathrm{E}_{P^0} \sup_{v\in[V],\tau\in\mathrm{T}_n} \left\|\frac{1 - \hat{g}_n^{-v}}{\hat{g}_n^{-v}} - \frac{1 - g_0}{g_0}\right\|_{P^0_{X\mid 1},2} \left\|\mathrm{E}_{n,\tau}^{-v} - \hat{\mathrm{E}}_{0,\tau}\right\|_{P^0_{X\mid 1},2} = \mathrm{o}(n^{-1/2}).$$

Therefore, a sufficient condition is that both nuisance estimators converge at a rate faster than $n^{-1/4}$. Thus, we allow for much slower rates than the parametric root-n rate. This $\mathrm{o}(n^{-1/4})$ rate requirement is only a sufficient condition and is by no means necessary. Condition 6 is satisfied even if one nuisance estimator converges very slowly, as long as the other nuisance estimator converges sufficiently fast to compensate for this slow convergence. We require convergence of the estimator $\mathrm{E}_{n,\tau}^{-v}$ uniformly over $\tau \in \mathrm{T}_n$ to establish uniform asymptotic efficiency in Theorem 3 below. Though this assumption on convergence is stronger than assumptions typically needed to obtain efficient estimators, it may not be stringent. We illustrate this with an example in Online Supplementary Material, Section S5.

Remark 7 The aforementioned phenomenon that one convergence rate can compensate for the other is similar to the mixed bias property, which is frequently observed for semiparametrically or nonparametrically efficient estimators (Rotnitzky et al., 2021). The mixed bias property often leads to double robustness. An estimator is doubly robust if it is still consistent even when one nuisance function, but not the other, is estimated inconsistently (see e.g. the rejoinder to discussions of Scharfstein et al. (1999), and Bang and Robins (2005)). This double robustness property also holds for our estimator $\psi_{n,\tau}$ of coverage error $\Psi_\tau(P^0)$, similarly to the method in Yang et al. (2022). In other words, $\psi_{n,\tau}$ is consistent for $\Psi_\tau(P^0)$ even if either $\mathrm{E}_{0,\tau}$ or g_0 is estimated inconsistently, in which case Condition 6 fails. We do, however, generally require Condition 6 to hold for our proposed PredSet-1Step method except for special cases. The reason is that PredSet-1Step further relies on asymptotically valid CUBs, which rely on the asymptotic normality of $\psi_{n,\tau}$. If a nuisance function is estimated inconsistently and thus Condition 6 fails, even though $\hat{\psi}_{n,\tau}$ is still consistent for $\Psi_\tau(P^0)$, $\hat{\psi}_{n,\tau}$ is no longer asymptotically normal in general. In this case, it is challenging, if possible at all, to construct asymptotically valid CUB. We discuss special cases where PredSet-1Step is doubly robust under Online Supplementary Material, Condition S5 or S6 in Section S4. In particular, when one nuisance function is known, our proposed procedure remains valid with the known nuisance function plugged in.

This leads to our second main result.

Theorem 3 (Asymptotic efficiency of cross-fit one-step corrected estimator). Under Conditions 1–6, the one-step corrected cross-fit estimator $\psi_{n,\tau}$ from (16) is an asymptotically nonparametrically efficient estimator of the coverage error $\Psi_\tau(P^0) = \Pr(Y \notin C_\tau(X) \mid A = 0)$ from (5), in the target, 'covariate shifted' population where $A = 0$. Moreover, with the gradient D_τ from (12), the propensity score g_0 and the probability γ_0 of $A = 1$ from (8), and the conditional coverage error rate $\mathrm{E}_{0,\tau}$ from (6),

$$\sup_{\tau \in \mathrm{T}_n} \left| \hat{\psi}_{n,\tau} - \Psi_\tau(P^0) - \frac{1}{n} \sum_{i=1}^{n} D_\tau(P^0, g_0, \gamma_0)(O_i) \right| = o_p(n^{-1/2}).$$

Theorem 3 states the same asymptotic efficiency claim that is implied by the general result Theorem 3.1 and the more concrete result Theorem 5.1 in Chernozhukov et al. (2018a). See also Proposition 2 in Kennedy (2022). One difference is that Theorem 3 concerns a uniform efficiency claim over $\tau \in \mathrm{T}_n$, which is implied by pointwise efficiency and the uniform rate condition 6; another difference arises in the proof due to the different estimation strategies for the nuisance parameter γ_0. The proof of Theorem 3 can be found in Online Supplementary Material, Section S7.3. As explained in Section 3.2, this result implies that the one-step corrected estimator enjoys a desirable optimality property: it has the smallest possible asymptotic variance, among all regular estimators, under the nonparametric model $\mathcal{M}$.

Remark 8 Although $\psi_{n,\tau}$ is consistent for $\Psi_\tau(P^0) \in [0, 1]$, this estimator itself may fall outside of the interval [0, 1]. This possibility may harm the interpretation of $\psi_{n,\tau}$ as an estimator of a probability. We may project $\psi_{n,\tau}$ onto [0, 1], or instead use the targeted minimum-loss based estimator (TMLE) (Van der Laan & Rose, 2018; Van der Laan & Rubin, 2006). We present the method based on TMLE, PredSet-TMLE, in Online Supplementary Material, Section S2.

4.2 Wald CUB and selection of threshold

To construct Wald CUBs based on $\hat{\psi}_{n,\tau}$ using Theorem 3, we need to estimate the asymptotic variance $\sigma^2_{0,\tau} := \mathrm{E}_{P^0}[D_\tau(P^0, g_0, \gamma_0)(O)^2]$. We propose to use a plug-in estimator based on sample splitting. Let

$$(\hat{\sigma}^v_{n,\tau})^2 := \frac{1}{|I_v|}\sum_{i\in I_v} D_\tau(\hat{P}^{n,v}_\tau, \hat{g}^{-v}_n, \hat{\gamma}^v_n)(O_i)^2 \quad \text{and} \quad \hat{\sigma}_{n,\tau} := \left\{\frac{1}{n}\sum_{v=1}^{V} |I_v|(\hat{\sigma}^v_{n,\tau})^2\right\}^{1/2}. \tag{17}$$

We propose to use $\hat{\sigma}_{n,\tau}/\sqrt{n}$ as the standard error when constructing a $(1-\alpha_{\mathrm{conf}})$-Wald CUB of $\Psi_\tau(P^0)$. That is, we propose to use $\hat{\psi}_{n,\tau} + z_{\alpha_{\mathrm{conf}}}\hat{\sigma}_{n,\tau}/\sqrt{n}$ as an approximate $(1-\alpha_{\mathrm{conf}})$-CUB, where we use z_α to denote the $(1-\alpha)$-quantile of the standard normal distribution for any $\alpha \in (0, 1)$.

Our theoretical guarantees on the APAC property of PredSet-1Step rely on the following general result on the confidence interval coverage of Wald CIs based on asymptotically linear estimators. Recall that an estimator $\hat{\phi}_n$ of ϕ_0 is asymptotically linear with influence function IF if the expansion $\hat{\phi}_n = \phi_0 + \frac{1}{n}\sum_{i=1}^{n}\mathrm{IF}(O_i) + o_p(n^{-1/2})$ holds for $\hat{\phi}_n$ (see e.g. Chapter 25 in van der Vaart & Wellner, 1996). In this definition, it is implicitly assumed that $\mathrm{E}_{P^0}[\mathrm{IF}(O)] = 0$ and $\mathrm{E}_{P^0}[\mathrm{IF}(O)^2] < \infty$.

Theorem 4 (Coverage of Wald CIs). Suppose that $\hat{\phi}_n$ is an asymptotically linear estimator of ϕ_0 with influence function IF such that $\sigma_0^2 := \mathrm{E}_{P^0}[\mathrm{IF}(O)^2] > 0$. Let $\hat{\sigma}^2_n$ be a consistent estimator of the asymptotic variance σ_0^2. Consider the corresponding Wald $(1-\alpha)$-CUB $\hat{\phi}_n + z_\alpha\hat{\sigma}_n/\sqrt{n}$ for ϕ_0. Assume that $\mathrm{E}_{P^0}|\mathrm{IF}(O)|^3 = \rho_0 < \infty$. Then, for any fixed scalar $\eta > 0$, there exists a universal constant $\mathcal{C}$ such that

$$\left|\Pr_{P^0}(\phi_0 < \hat{\phi}_n + z_\alpha\hat{\sigma}_n/\sqrt{n}) - (1-\alpha)\right| \le \mathcal{C}\frac{n^{1/4}}{\sigma_0^{1/2}}\left\{\mathrm{E}_{P^0}\left[\left(\hat{\phi}_n - \phi_0 - \frac{1}{n}\sum_{i=1}^{n}\mathrm{IF}(O_i)\right)^2\right]\right\}^{1/2}$$
$$+ \left\{\mathcal{C}\frac{\mathrm{E}_{P^0}[1(|\hat{\sigma}_n - \sigma_0| \le \eta)|\hat{\sigma}_n - \sigma_0|]}{\sigma_0} + \Pr_{P^0}(|\hat{\sigma}_n - \sigma_0| > \eta)\right\} + \mathcal{C}\frac{\rho_0}{\sigma_0^3}n^{-1/2}.$$

The three terms on the right-hand side arise from three sources: (a) the deviation of $\hat{\phi}_n - \phi_0$ from the sample mean of the influence function, (b) the estimation error of the asymptotic variance, and (c) the deviation of a root-*n*-scaled centred sample mean from its limiting normal distribution. In nonparametric models, the above bound typically converges to zero slower than the root-*n*-rate that is standard for parametric models, which is a phenomenon observed in some semi/nonparametric problems (Han & Kato, 2019; Zhang & Liang, 2011). The above bound is likely to have room for improvement, but this result suffices to prove the desired APAC property of our procedure. The proof of Theorem 4 can be found in Online Supplementary Material, Section S7.4.

For our problem, Theorem 4 alone is insufficient for results on CUB coverage for all $\tau \in \mathcal{T}_n$. For extremely large or small τ, it is possible that $\Psi_\tau(P^0) = 0$ or 1 and $D_\tau(P^0, g_0, \gamma_0) \equiv 0$. Therefore, we need to consider this special case separately. For any $\epsilon \ge 0$, let

$$\mathcal{T}^{\epsilon} := \{\tau \in \bar{\mathbb{R}} : \sigma^2_{0,\tau} > \epsilon\} \quad \text{and} \quad \mathcal{T}^{-} := \bar{\mathbb{R}} \setminus \mathcal{T}^{0} = \{\tau \in \bar{\mathbb{R}} : \sigma^2_{0,\tau} = 0\}, \tag{18}$$

where $\sigma^2_{0,\tau}$ is defined at the beginning of Section 4.2. These sets of candidate thresholds depend on the true data-generating mechanism P^0 only, and are deterministic. For all $\tau \in \mathcal{T}^{-}$, one of the following scenarios occurs: either (a) $\mathcal{E}_{0,\tau} \equiv 0$ and $\Psi_\tau(P^0) = 0$ or (b) $\mathcal{E}_{0,\tau} \equiv 1$ and $\Psi_\tau(P^0) = 1$. We make the following assumption on $\hat{\ }_{n,\tau}\cdot\mathcal{E}^{-v}$

Condition 7 (Deterministic conditional coverage error estimator for extreme thresholds). For all $\tau \in \mathcal{T}^-$, it holds that $\hat{\mathrm{E}}^{-v}_{n,\tau} = \mathrm{E}_{0,\tau}$ for all $v \in [V]$ and all n.

Condition 7 is so mild that it is often automatically satisfied: for extremely small τ that lies in $\mathcal{T}^-$, the random variable $Z_\tau = 1(s(X, Y) > \tau)$ is a constant equal to one. Since one can only observe $Z_\tau(X_i, Y_i) = 1$ in any sample, it is natural to estimate $\mathrm{E}_{0,\tau}$ with $\hat{\mathrm{E}}^{-v}_{n,\tau} \equiv 1$, which equals $\mathrm{E}_{0,\tau}$. On the other hand, for extremely large τ that lies in $\mathcal{T}$, the random variable Z_τ is a constant equal to zero, and it is natural to estimate $\mathrm{E}_{0,\tau}$ with $\hat{\mathrm{E}}^{-v}_{n,\tau} \equiv 0$, which also equals $\mathrm{E}_{0,\tau}$.

We have the following convergence rate of the coverage to the nominal confidence $1 - \alpha_{\text{conf}}$, our third main result.

Theorem 5 (Convergence rate of Wald-CUB coverage based on cross-fit one-step corrected estimator). Consider the cross-fit one-step corrected estimator $\hat{\psi}_{n,\tau}$ from (16), for the standard error estimator $\hat{\sigma}_{n,\tau}$ from (17), and the coverage error $\Psi_\tau(P^0) = \Pr_{P^0}(Y \notin C_\tau(X) \mid A = 0)$ from (5), in the target, 'covariate shifted' population where $A = 0$. Under Conditions 1–6, for any fixed $\epsilon > 0$, with $\mathcal{T}^\epsilon$ from (18), it holds that

$$\sup_{\tau \in \mathcal{T}^\epsilon \cap \mathcal{T}_n} \left| \Pr\left(\Psi_\tau(P^0) < \hat{\psi}_{n,\tau} + z_{\alpha_{\text{conf}}}\hat{\sigma}_{n,\tau}/\sqrt{n}\right) - (1 - \alpha_{\text{conf}}) \right| \le \Delta_{n,\epsilon}$$

where, with $\hat{g}^{-v}_n$ from Line 2 of Algorithm 1, $\hat{\gamma}^v_n$ from (13), g_0, γ_0 from (8), $\hat{\mathrm{E}}^{-v}_{n,\tau}$ from Line 4 of Algorithm 1, $\mathrm{E}_{0,\tau}$ from (6), the marginal distribution $P^0_{X \mid 1}$ of X in the source population from Condition 3, and probability $1 - q_n$ of having a bounded nuisance estimator from Condition 5,

$$\begin{aligned}\Delta_{n,\epsilon} := n^{1/4}\epsilon^{-1/4} \sup_{v \in [V], \tau \in \mathcal{T}_n} \mathrm{E}_{P^0}\Bigg[\Bigg\{\int \left| \frac{1 - \hat{g}^{-v}_n(x)}{\hat{g}^{-v}_n(x)} - \frac{1 - g_0(x)}{g^0(x)} \right| \\ \cdot (\hat{\mathrm{E}}^{-v}_{n,\tau}(x) - \mathrm{E}_{0,\tau}(x)) P^0_{X \mid 1}(\mathrm{d}x)\Bigg\}^{1/2}\Bigg] + q_n \end{aligned} \tag{19}$$

converges to zero. In addition, with $\mathcal{T}^-$ from (18), under Condition 7, it holds that $\Pr(\Psi_\tau(P^0) \le \hat{\psi}_{n,\tau} + z_{\alpha_{\text{conf}}}\hat{\sigma}_{n,\tau}/\sqrt{n}) = 1$ for all $\tau \in \mathcal{T}^-$.

Theorem 5 is a consequence of Theorem 4, and the proof can be found in Online Supplementary Material, Section S7.4. The uniform bound only holds for thresholds in $\mathcal{T}^\epsilon$ for some $\epsilon > 0$ because, as $\sigma^2_{0,\tau}$ tends to zero, it becomes more difficult to estimate $\sigma^2_{0,\tau}$ with a small *relative* error. The error term $\Delta_{n,\epsilon}$ in Theorem 5 is essentially the square root of the product bias of the two nuisance function estimators $\hat{g}^{-v}_n$ and $\hat{\mathrm{E}}^{-v}_n$, scaled by $n^{1/4}$. This product bias term is the dominating term in the bound in Theorem 4. This dominance suggests that, when using flexible nonparametric nuisance estimators, the main challenge in improving the coverage of the Wald CI based on our proposed estimator $\hat{\psi}_{n,\tau}$ might be the product bias; improved estimators of the asymptotic variance $\sigma^2_{0,\tau}$ alone might not substantially improve the CI coverage. We conjecture that this phenomenon might hold for a variety of efficient estimators that are constructed using semiparametric efficiency theory and involve nuisance function estimation.

Based on the Wald-CUB, we select a threshold $\hat{\tau}^{\text{1Step}}_n$ to ensure that the size of prediction sets is small:

$$\hat{\tau}^{\text{1Step}}_n := \max\left\{\tau \in \mathcal{T}_n : \hat{\psi}_{n,\tau'} + z_{\alpha_{\text{conf}}}\hat{\sigma}_{n,\tau'}/\sqrt{n} < \alpha_{\text{error}} \text{ for all } \tau' \in \mathcal{T}_n \text{ such that } \tau' \le \tau\right\}. \tag{20}$$

This step is illustrated in Online Supplementary Material, Figure S2. This procedure for choosing a threshold based on CUBs is justified by the following general result on APAC prediction set construction based on pointwise CUBs, which is similar to Theorem 1 in Bates et al. (2021) with adaptations to finite candidate threshold sets, general distributions of the score $s(X, Y)$, and asymptotic CUBs.

Theorem 6 (Grid search threshold based on CUB). Given a finite set $\mathcal{T}_n$ of candidate thresholds and asymptotic $(1-\alpha_{\mathrm{conf}})$-level CUBs $\lambda_n(\tau)$ of $\Psi_\tau(P^0)$ valid pointwise for each $\tau \in \mathcal{T}_n$, define the selected threshold

$$\hat{\tau}_n := \max\{\tau \in \mathcal{T}_n : \lambda_n(\tau') < \alpha_{\mathrm{error}} \text{ for all } \tau' \in \mathcal{T}_n \text{ such that } \tau' \le \tau\}. \tag{21}$$

Then, the prediction set with threshold $\hat{\tau}_n$ satisfies the following:

$$\Pr_{P^0}(\Psi_{\hat{\tau}_n}(P^0) \le \alpha_{\mathrm{error}}) \ge \inf_{\tau \in \mathcal{T}_n} \Pr_{P^0}\{\lambda_n(\tau) \ge \Psi_\tau(P^0)\} \ge 1 - \alpha_{\mathrm{conf}} - \sup_{\tau \in \mathcal{T}_n} \left| \Pr_{P^0}\{\lambda_n(\tau) \ge \Psi_\tau(P^0)\} - (1 - \alpha_{\mathrm{conf}}) \right|.$$

Consequently, the prediction set with threshold $\hat{\tau}_n$ is APAC if the asymptotic validity of all $\lambda_n(\tau)$ $(\tau \in \mathcal{T}_n)$ is uniform; that is,

$$\inf_{\tau \in \mathcal{T}_n} \Pr_{P^0}\{\lambda_n(\tau) \ge \Psi_\tau(P^0)\} \ge 1 - \alpha_{\mathrm{conf}} - o(1),$$

which is implied by a uniform convergence of CUB coverage to the nominal level $1 - \alpha_{\mathrm{conf}}$:

$$\sup_{\tau \in \mathcal{T}_n} \left| \Pr_{P^0}\{\lambda_n(\tau) \ge \Psi_\tau(P^0)\} - (1 - \alpha_{\mathrm{conf}}) \right| = o(1).$$

If the coverage of the CUB $\lambda_n(\tau)$ is at least $1 - \alpha_{\mathrm{conf}}$ for all $\tau \in \mathcal{T}_n$, then the prediction set with threshold $\hat{\tau}_n$ is PAC.

In Theorem 6, pointwise valid CUBs, rather than uniform CUBs or confidence bands, are used. More general results on using pointwise valid tests to control risk can be found in Angelopoulos et al. (2021).

We require an additional condition to derive the APAC guarantee of PredSet-1Step from CUB coverage results in Theorem 5 and APAC results in Theorem 6.

Condition 8 (Asymptotic variance equal to, or bounded away from, zero). Define $\tau_n^\dagger := \min\{\tau \in \mathcal{T}_n : \Psi_\tau(P^0) > \alpha_{\mathrm{error}}\}$, where we define $\min \emptyset := \infty$. For some fixed $\epsilon > 0$, it holds that $\tau_n^\dagger \in \mathcal{T}^{-} \cup \mathcal{T}^{\epsilon}$.

We have dropped the dependence of $\tau_n^\dagger$ on P^0 from the notation for conciseness. Condition 8 is again often automatically satisfied as long as the set of candidate thresholds $\mathcal{T}_n$ is sufficiently dense. Indeed, we argue that this condition holds if the candidate set $\mathcal{T}_n$ increases with the sample size n. Since $\inf\{\Psi_{\tau_n^\dagger}(P^0) : n = 1, 2, \ldots\} \ge \alpha_{\mathrm{error}} > 0$ by definition, either $\inf\{\Psi_{\tau_n^\dagger}(P^0) : n = 1, 2, \ldots\} = 1$ or $\alpha_{\mathrm{error}} \le \inf\{\Psi_{\tau_n^\dagger}(P^0) : n = 1, 2, \ldots\} < 1$. In the first case, $\Psi_{\tau_n^\dagger}(P^0)$ is trivially equal to unity, and therefore $\tau_n^\dagger \in \mathcal{T}^{-}$, so Condition 8 holds. In the second case, since $\mathcal{T}_n$ is increasing with n, $\tau_n^\dagger$ is decreasing with n. Thus, for some $\delta > 0$ and N, $\alpha_{\mathrm{error}} < \Psi_{\tau_n^\dagger}(P^0) < 1 - \delta$ for all $n > N$ and thus $\tau_n^\dagger \in \mathcal{T}^{\epsilon'}$ for some $\epsilon' > 0$. For each $n \le N$, $\tau_n^\dagger \in \mathcal{T}^{\epsilon_n} \cup \mathcal{T}^{-}$ for some $\epsilon_n > 0$. Condition 8 hence holds with $\epsilon = \max\{\epsilon', \epsilon_1, \ldots, \epsilon_N\}$. Condition 8 may only fail if $\Psi_{\tau_n^\dagger}(P^0)$ can be arbitrarily close to—but not equal to—one. Even if $\mathcal{T}_n$ is not increasing, in all scenarios we can think of, Condition 8 only fails for extremely contrived sets $\mathcal{T}_n$.

We have the following corollary of Theorem 5, our final result showing the APAC property of PredSet-1Step.

Corollary 7 If Conditions 1–8 hold, then we have

$$\Pr_{P^0}(\Psi_{\hat{\tau}_n^{\mathrm{1Step}}}(P^0) \le \alpha_{\mathrm{error}}) \ge 1 - \alpha_{\mathrm{conf}} - \mathcal{C}\Delta_{n,\epsilon}, \tag{22}$$

where $\Delta_{n,\epsilon}$ is defined in (19). In other words, the prediction set with threshold $\hat{\tau}_n^{\mathrm{1Step}}$ is APAC.

Remark 9 It might be preferable to use another CUB—rather than the Wald CUB we propose. For example, it is well known that carefully constructed bootstrap procedures can lead to better coverage for certain problems (Hall, 2013). Another possibility is to efficiently estimate the asymptotic variance $\sigma^2_{0,\tau}$. However, this does not appear to improve the overall convergence rate, because the estimation error in the asymptotic variance is not the only term that dominates our bound on the convergence rate. The other dominating term is the deviation of the estimator from the sample mean of the influence function.

The empirical performance of the above methods has sometimes been observed to be comparable to the ones without efficient estimation of the asymptotic variance or bootstrap (see e.g. Chapter 28 in Van der Laan & Rose, 2018). To our knowledge, theory on the convergence rate of confidence interval coverage for general asymptotically linear estimators has not been developed in the literature. The bound we obtained in Theorem 4 requires the development of novel tools to propagate the difference between the estimator and the sample mean of the influence function to the difference between the true and the nominal coverage.

Remark 10 PredSet-1Step relies on an efficient estimator based on the G-computation formula $\Psi_\tau^{\mathrm{Gcomp}}$. An alternative approach is to use estimators based on the weighted formula $\Psi_\tau^{\mathrm{weight}}$. In this approach, a one-step correction is also crucial to achieving the same asymptotic efficiency. Furthermore, for each fold v, we have used $\hat{\gamma}_n^v$ based on data in fold v to estimate γ_0. Using the empirical estimator $\sum_{i \notin I_v} A_i/(n - |I_v|)$ based on data out of fold v—an approach that coincides with double/debiased machine learning (Chernozhukov et al., 2018a)—also leads to efficient estimators and APAC prediction sets under the same conditions. The proof is similar, with minor modifications. We have chosen to estimate γ_0 in the same fold because it leads to a remainder that aligns with the conventional definition of the mixed bias property (Rotnitzky et al., 2021).

5 Simulations

We conduct three simulation studies to investigate the performance of our methods. In the first simulation, we consider a moderate-to-high dimensional sparse setting; in the second simulation, we consider a relatively low dimensional setting; in the third simulation, we consider a relatively low dimensional setting without covariate shift. In all settings, we consider $\alpha_{\mathrm{error}} = \alpha_{\mathrm{conf}} = 0.05$ and the following methods: (a) PredSet-1Step; (b) PredSet-TMLE, described in Remark 8 and Online Supplementary Material, Section S2; (c) PredSet-RS, a method based on rejection sampling, described in Online Supplementary Material, Section S1; (d) plug-in, a naïve variant of PredSet-1Step based on a naïve cross-fit plug-in estimator of the true coverage error $\Psi_\tau(P^0)$; the same as PredSet-1Step except that the one-step correction in (14) is not included; (e) plug-in2, a method similar to PredSet-1Step based on a corss-fit estimator with the estimated likelihood ratio w_0 plugged into $\Psi_\tau^{\mathrm{weight}}$; (f) weighted CP, weighted Conformal Prediction (Tibshirani et al., 2019) with an estimated likelihood ratio and a target marginal coverage error at most α_{error}; and (g) inductive CP, inductive Conformal Prediction (Papadopoulos et al., 2002), tuned as in Park et al. (2022); Vovk (2013) to ensure training-set conditional validity (i.e. the PAC property), ignoring covariate shift. To our best knowledge, training-set conditional validity results are unknown for weighted Conformal Prediction. We still include this method for comparison and do not expect it to attain (approximate) training-set conditional validity. Whenever no threshold can be selected, that is, the CUB corresponding to $\tau = 0$ is above α_{error}, we set the selected threshold to zero. We consider a setting without covariate shift in the third simulation, because in this case, inductive CP has a finite sample PAC guarantee while our proposed methods do not. In this case, we focus on comparing our proposed methods PredSet-1Step and PredSet-TMLE with inductive CP.

For all methods incorporating covariate shift, we split the data into two folds of equal sizes ($V = 2$). When estimating the nuisance functions $\mathrm{E}_{0,\tau}$ and g_0, we use Super Learner (van der Laan et al., 2007) with the library consisting of logistic regression, generalised additive models (Hastie & Tibshirani, 1990), logistic LASSO regression (Hastie et al., 1995; Tibshirani, 1996), and gradient boosting (Friedman, 2001, 2002; Mason et al., 1999, 2000) with various combinations of tuning parameters (maximum number of boosting iterations being 100, 200, 400, 800, or 1,000; minimum sum of instance weights needed in a child being 1, 5, or 10). Super Learner is an ensemble learner that outputs a weighted average of the algorithms in the library to minimise the cross-validated prediction error. In all above methods except inductive CP, the candidate threshold set T_n is a fixed grid on the interval [0, 0.3] with distance between adjacent grid points being 0.05. PredSet-RS requires additional tuning parameters, and we present them in the Online Supplementary Material.

We consider sample sizes $n =$ 500, 1,000, 2,000, and 4,000. For each sample size, we run all methods on 200 randomly generated data sets. We approximately calculate the true optimal threshold τ_0 by generating 10^6 samples from the target population and taking the α_{error}th quantile of $s(X, Y)$ in the sample. We next describe the data-generating mechanisms and the results of the three simulations.

5.1 Moderate-to-high dimensional sparse setting

To generate the data, we first generate the population indicator $A \sim \text{Bernoulli}(0.5)$. Given $A = a$, the covariate $X := (X_1, \ldots, X_{20})^\top$ is a 20-dimensional random vector generated from exponential distributions as follows:

$$X_1 \sim \text{Exp}(2^{1-a}), \qquad X_2 \sim \text{Exp}(2^{1-a}), \qquad X_k \sim \text{Exp}(1) \quad (k = 3, \ldots, 20),$$

where $X_1, \ldots, X_{20}$ are mutually independent. The outcome Y has three labels {0, 1, 2} and is generated according to the distribution implied by the following two equations:

$$\frac{\Pr(Y = 1 \mid X = x)}{\Pr(Y = 0 \mid X = x)} = \exp\left(2 + 2x_1 - 1.1x_2\right), \qquad \frac{\Pr(Y = 2 \mid X = x)}{\Pr(Y = 0 \mid X = x)} = \exp\left(-2.1 - 2x_1 + 1.2x_3\right).$$

Instead of the true conditional probability of Y defined above, we set the scoring function s to be the function satisfying the following three equations for all x: $s(x, 0) + s(x, 1) + s(x, 2) = 1$,

$$\frac{s(x, 1)}{s(x, 0)} = \exp\left(0.02 + 2.1x_1 - 0.91x_2 + 0.02x_4\right), \qquad \text{and}$$

$$\frac{s(x, 2)}{s(x, 0)} = \exp\left(-0.03 - 1.95x_1 + 1.25x_3 + 0.1x_5\right).$$

The empirical proportion that the true miscoverage is below α_{error} is presented in Figure 5. Since weighted CP was developed to achieve marginal coverage rather than training-set conditional coverage, its proportion of having a miscoverage exceeding α_{error} is much higher than the desired level α_{conf}. In this simulation, the optimal threshold for the source population is greater than the optimal threshold τ_0 for the target population. Hence, inductive CP performs considerably worse than all other methods—that incorporate covariate shift—in the sense that its miscoverage exceeds α_{error} much more often than the desired level α_{conf}, especially in large samples ($n =$ 4,000). As the sample size grows, the performance of inductive CP becomes worse.

The two plug-in methods appear not to be APAC because the Monte Carlo estimated actual confidence level is below 90% even in large samples ($n =$ 4,000) and the 95% confidence interval does not cover the desired 95% level. PredSet-RS performs much worse than other methods, including the invalid inductive CP and plug-in methods, when the sample size is not large ($n \leq$ 2,000). However, PredSet-RS might be APAC as its confidence level appears to approach 95% as the sample size grows. The other two methods, PredSet-1Step and PredSet-TMLE, appear to be APAC and have reasonable performance for moderate to large sample sizes ($n \geq$ 2,000).

As shown in Online Supplementary Material, Figure S4, the distribution of the threshold selected by PredSet-RS has a much wider spread than PredSet-1Step and PredSet-TMLE. We

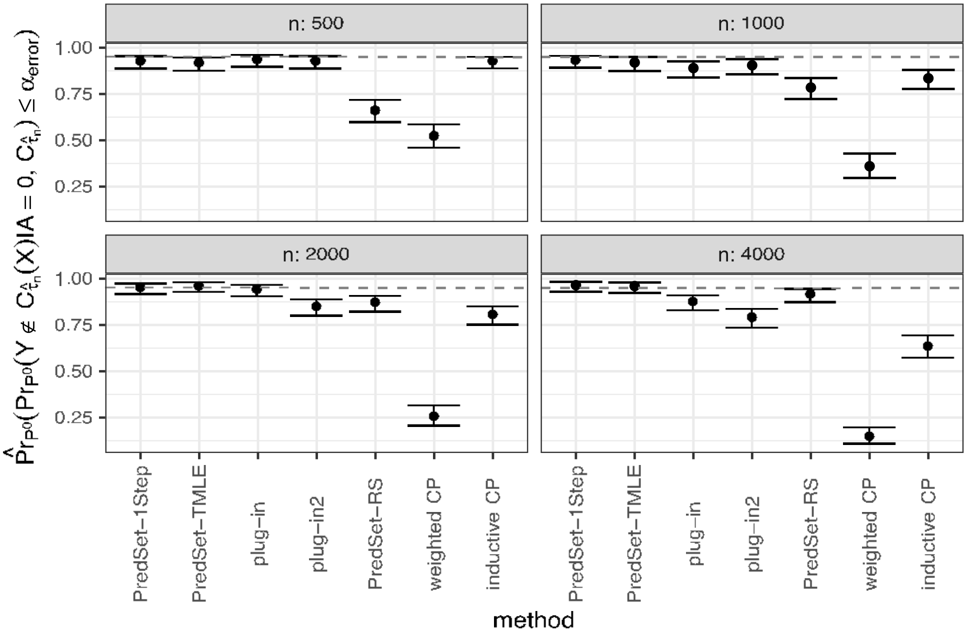

Figure 5. Empirical proportion of simulations where the estimated coverage error $\hat{\Pr}_{P^0}(Y \notin C_{\hat\tau_n}(X) \mid A = 0, C_{\hat\tau_n})$ does not exceed α_{error}, along with a 95% Wilson score confidence interval, in the moderate-to-high dimensional sparse setting. The grey horizontal dashed line is the desired confidence level $1 - \alpha_{\text{conf}}$.

therefore recommend PredSet-1Step and PredSet-TMLE rather than PredSet-RS, although all these methods appear to produce APAC prediction sets.

5.2 Low dimensional setting

The data-generating mechanism is similar to the previous simulation. We still generate A from a Bernoulli(0.5) random variable. We generate a three-dimensional covariate from a trivariate normal distribution:

$$X \mid A = a \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 & -0.2 \\ 0.2 & 1 & 0.2 \\ -0.2 & 0.2 & 1 \end{pmatrix} \times \left(\frac{1}{2}\right)^{1-a}\right).$$

The outcome Y also has three labels {0, 1, 2} and is generated according to the distribution implied by the following two equations:

$$\frac{\Pr(Y = 1 \mid X = x)}{\Pr(Y = 0 \mid X = x)} = \exp\left(1.4x_1 + 1.5x_2 - 1.5x_3 + 0.3(1 - x_1)^2 + 0.015x_2x_3\right),$$

$$\frac{\Pr(Y = 2 \mid X = x)}{\Pr(Y = 0 \mid X = x)} = \exp\left(-0.1 - 1.3x_1 - 2.2x_2 + 0.5x_3 + 0.5(1 - x_2)^2 + 0.03x_1x_3\right).$$

The scoring function is determined by the following three equations, valid for all $x : s(x, 0) + s(x, 1) + s(x, 2) = 1$,

$$\frac{s(x, 1)}{s(x, 0)} = \exp\left(0.02 + 1.2x_1 + 1.91x_2 - 1.6x_3\right), \quad \text{and}$$

$$\frac{s(x, 2)}{s(x, 0)} = \exp\left(-0.03 - 1.5x_1 - 2.4x_2 + 0.3x_3\right).$$

Unlike in the previous simulation where g_0 follows a logistic regression model, here neither g_0 nor $\mathrm{E}_{0,\tau}$ follows a parametric model that is correctly specified by an algorithm in the library of Super Learner due to interaction terms in the distribution of $Y \mid X$ and quadratic terms in the logit of w_0. The only exceptions are that $\mathrm{E}_{0,\tau}$ follows a logistic regression model with an infinite slope for an extremely large or small threshold τ. Thus, we do not expect our nuisance function estimators to generally converge at the parametric root-n rate.

The simulation results are presented in Online Supplementary Material, Figures S5 and S6. The performance of the methods is similar to the moderate-to-high dimensional sparse setting.

5.3 Low dimensional setting without covariate shift

The data-generating mechanism is identical to the previous simulation, except that

$$X \mid A = a \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 & -0.2 \\ 0.2 & 1 & 0.2 \\ -0.2 & 0.2 & 1 \end{pmatrix}\right).$$

In other words, covariate shift is not present.

The simulation results are presented in Online Supplementary Material, Figures S7 and S8. Inductive CP appears to perform the best for all sample sizes. This is not surprising, because inductive CP has a finite sample PAC guarantee in the no-covariate-shift setting. Our proposed methods PredSet-1Step and PredSet-TMLE also appear to be approximately PAC when the sample size is moderate to large ($n \geq 2{,}000$). The performance of our proposed methods appears to be comparable to that of inductive CP, even under no covariate shift. The performance of the other two methods—plug-in and PredSet-RS—is similar to the previous simulation.

We therefore conclude from our simulations that, when the sample size is reasonably large, our proposed methods PredSet-1Step and PredSet-TMLE empirically output approximately PAC prediction sets regardless of whether covariate shift is present or not. Even when no covariate shift is present, in which case inductive CP has a finite sample PAC guarantee, the performance of our methods is empirically comparable with inductive CP. Our proposed methods can be applied as a default method if the user suspects—but may be unsure—that covariate shift is present, and does not know the likelihood ratio w_0 of the shift.

6 Analysis of HIV risk prediction data in South Africa

We illustrate our methods with a data set concerning HIV risk prediction in a South African cohort study. Specifically, we use data from a large population-based prospective cohort study in KwaZulu-Natal, South Africa which was collected and analysed by Tanser et al. (2013) to evaluate the causal effect of community coverage of antiretroviral HIV treatment on community-level HIV incidence. The study followed a total of 16,667 individuals who were HIV-uninfected at baseline in order to observe individual HIV seroconversions over the period 2004 to 2011. In the present analysis, we aim to predict HIV seroconversion status over the follow-up period, for a target population of individuals living in a peri-urban community, using urban and rural communities as a source population.

Although the outcome is in fact available in both source and target samples, we deliberately treat the outcome in the target population as missing when constructing prediction sets and then use the observed outcome in the target population to evaluate empirical coverage of prediction sets. There are 12,385 and 5,136 participants from source and target populations, respectively. All participants are treated as independent draws from their corresponding populations. The covariates used to predict the outcome are the followings: (a) binned number partners in the past 12 months, (b) current marital status, (c) wealth quintile, (d) binned age and sex, (e) binned community antiretroviral therapy (ART) coverage, and (f) binned community HIV prevalence. For covariates that are time-varying, we use the last observed value as the covariate. All covariates are treated as categorical variables in the analysis. Missing data for each covariate are treated as a separate category, which is equivalent to the missing-indicator method (Groenwold et al., 2012). Covariate distributions are presented in Online Supplementary Material, Figure S1. We also perform Fisher's exact test via a Monte Carlo approximation with 2,000 runs to test the equality of

Table 1. Empirical coverage of prediction sets, 95% Wilson score confidence interval for coverage, and selected thresholds in the synthetic sample from the target population in the South Africa HIV trial data

Method	Empirical coverage (%)	Coverage CI (%)	Selected threshold $\hat{\tau}$
PredSet-1Step	95.98	94.83–96.89	0.095
PredSet-TMLE	95.42	94.20–96.39	0.100
Inductive conformal prediction	91.89	90.35–93.20	0.195

Note. The target coverage is at least $1 - \alpha_{\text{error}} = 95\%$, with probability 95% over the training data.

covariate distributions in the two populations, and we observe evidence of shift in covariate distribution with a *p*-value < 0.001.

For illustration, in this analysis, we create a severe shift on a covariate that we believe to be strongly related to the outcome. In the target population, we only include individuals with community ART coverage below 15% (binned community ART coverage being 1 or 2 in Online Supplementary Material, Figure S1). In other words, we set the target population to be the population in the peri-urban communities with ART coverage below 15%, this sub-population maybe of particular public health policy interest as likely to carry most of the burden of incident HIV cases. We present the analysis results for the full data analysis (target population being peri-urban communities) in Online Supplementary Material, Section S6. In this subset of the data, there are 1418 participants from the target population.

We randomly select 10,967 participants from the source population to train the scoring function *s*. We use Super Learner (van der Laan et al., 2007), with the same setup as in the simulations, to train a classifier of the outcome on this subsample, which is used as the scoring function *s*. We then construct prediction sets using the rest of the sample consisting of 1,418 participants from each of the source and the target populations. The target PAC criterion has miscoverage level $\alpha_{\text{error}} = 0.05$ and confidence level $1 - \alpha_{\text{conf}} = 0.95$. The methods we apply are a subset of the meth-ods investigated in the simulations: PredSet-1Step, PredSet-TMLE, and inductive CP (Papadopoulos et al., 2002; Park et al., 2022; Vovk, 2013), which ignores covariate shift. The tuning parameters of these methods, such as the number of folds and the algorithm to estimate nuisance functions, are identical to those in the simulations.

The empirical coverage of the above methods in the sample from the target population is presented in Table 1. The empirical coverage of both PredSet-1Step and PredSet-TMLE is close to the target coverage level $1 - \alpha_{\text{error}} = 95\%$, with the 95% confidence interval containing $1 - \alpha_{\text{error}}$. In contrast, the empirical coverage of inductive CP is lower than the target coverage lev-el. Thus, properly accounting for covariate shift as in PredSet-1Step and PredSet-TMLE is crucial for achieving the PAC property in the 'covariate shifted' target population in this subset of the data.

7 Conclusion

There has been extensive literature on (a) constructing prediction sets based on fitted machine learning models and (b) supervised learning under covariate shift. In this work, we study the intersection of these two problems in the challenging setting where the covariate shift needs to be estimated. We propose a distribution-free method, PredSet-1Step, to construct asymptotically probably approximately correct (APAC) prediction sets under unknown covariate shift. PredSet-1Step may also be used to construct ARCPS with a slight modification. Our method is flexible, taking as input an arbitrary given scoring function, produced by essentially any statistical or machine learning method.

We use semiparametric efficiency theory when constructing prediction sets to obtain root-*n* convergence of the true miscoverage corresponding to the selected prediction sets, even if the estimators of the nuisance functions may converge slower than root-*n*. Our theoretical analysis of PredSet-1Step relies on a novel result on the convergence of Wald confidence intervals based on general asymptotically linear estimators, which is a technical tool of independent interest. We illustrate that our method has good coverage in a number of experiments and by analysing a data set concerning HIV risk prediction in a South African cohort. In experiments without covariate shift, PredSet-1Step performs similarly to inductive CP, which has finite-sample PAC properties. Thus, PredSet-1Step may be used in the common scenario if the user suspects—but may not be certain—that covariate shift is present, and does not know the form of the shift.

One interesting open question is the asymptotic behaviour of our selected threshold compared to the true optimal threshold. Our simulation results (Online Supplementary Material, Figures S4, S6, and S8) suggest that our selected threshold might converge in probability to the true optimal threshold. Our selected threshold also appears to have a vanishing negative bias that ensures the desired confidence level. Theoretical analysis is in need to confirm these conjectures.

Acknowledgments

We thank Arun Kumar Kuchibotla, Jing Lei, Lihua Lei, and Yachong Yang for helpful comments.

Conflict of interest: The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Funding

This work was supported in part by the NSF DMS 2046874 (CAREER) award, NIH grants R01AI27271, R01CA222147, R01AG065276, R01GM139926, and Analytics at Wharton.

Data availability

Data subject to third party restriction.

Supplementary material

Supplementary material are available at *Journal of the Royal Statistical Society: Series B* online.

References

Angelopoulos A. N., Bates S., Candès E. J., Jordan M. I., & Lei L. (2021). 'Learn then test: Calibrating predictive algorithms to achieve risk control', arXiv, arXiv:2110.01052v5, preprint: not peer reviewed.

Bang H., & Robins J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

Bates S., Angelopoulos A., Lei L., Malik J., & Jordan M. I. (2021). 'Distribution-free, risk-controlling prediction sets', arXiv, arXiv:2101.02703, preprint: not peer reviewed.

Berkenkamp F., Turchetta M., Schoellig A. P., & Krause A. (2017). Safe model-based reinforcement learning with stability guarantees. *Advances in Neural Information Processing Systems*, *30*, 909–919. https://proceedings.neurips.cc/paper_files/paper/2017/file/766ebcd59621e305170616ba3d3dac32-Paper.pdf

Bickel P. J. (1982). On adaptive estimation. *The Annals of Statistics*, *10*(3), 647–671. https://doi.org/10.1214/aos/1176345863

Bickel P. J., & Doksum K. A. (2015). *Mathematical statistics: Basic ideas and selected topics* (2nd ed., Vol. *1*). Chapman and Hall/CRC.

Bickel P. J., Klaassen C. A., Ritov Y., & Wellner J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press.

Bickel P. J., & Ritov Y. (2003). Nonparametric estimators which can be "plugged-in". *Annals of Statistics*, *31*(4), 1033–1053. https://doi.org/10.1214/aos/1059655904

Bickel S., Brückner M., & Scheffer T. (2007). Discriminative learning for differing training and test distributions. In *ACM International Conference Proceeding Series* (pp. 81–88). Association for Computing Machinery.

Bojarski M., Del Testa D., Dworakowski D., Firner B., Flepp B., Goyal P., Jackel L. D., Monfort M., Muller U., Zhang J., Zhang X., Zhao J., & Zieba K. (2016). 'End to end learning for self-driving cars', arXiv, arXiv:1604.07316v1, preprint: not peer reviewed.

Cauchois M., Gupta S., Ali A., & Duchi J. C. (2020). 'Robust validation: Confident predictions even when distributions shift', arXiv, arXiv:2008.04267v1, preprint: not peer reviewed.

Chen X. (2007). Chapter 76: Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, *6*(SUPPL. PART B), 5549–5632. https://doi.org/10.1016/S1573-4412(07)06076-X

Chernozhukov V., Chetverikov D., Demirer M., Duflo E., Hansen C., Newey W., & Robins J. (2018a). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Chernozhukov V., Wuthrich K., & Zhu Y. (2018b). Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Proceedings of the 31st Conference on Learning Theory, PMLR* (Vol. 75, pp. 732–749). PMLR. http://arxiv.org/abs/1802.06300

Dunn R., Wasserman L., & Ramdas A. (2018). 'Distribution-free prediction sets with random effects', arXiv, arXiv:1809.07441, preprint: not peer reviewed.

Friedman J. H. (2001). *Greedy function approximation: A gradient boosting machine* (Technical Report 5).

Friedman J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Friedman J. H. (2004). *On multivariate goodness-of-fit and two-sample testing* (Technical Report). Citeseer.

Gal Y., Islam R., & Ghahramani Z. (2017). Deep Bayesian active learning with image data. In *34th International Conference on Machine Learning, ICML 2017* (Vol. 3, pp. 1923–1932). PMLR.

Groenwold R. H., White I. R., Donders A. R. T., Carpenter J. R., Altman D. G., & Moons K. G. (2012). Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis. *Canadian Medical Association Journal*, *184*(11), 1265–1269. https://doi.org/10.1503/cmaj.110977

Hahn J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, *66*(2), 315. https://doi.org/10.2307/2998560

Hajek J. (1962). Asymptotically most powerful rank-order tests. *The Annals of Mathematical Statistics*, *33*(3), 1124–1147. https://doi.org/10.1214/aoms/1177704476

Hall P. (2013). *The bootstrap and Edgeworth expansion*. Springer Science & Business Media.

Han Q., & Kato K. (2019). 'Berry-Esseen bounds for Chernoff-type non-standard asymptotics in isotonic regression', arXiv, arXiv:1910.09662v2, preprint: not peer reviewed.

Hastie T., Buja A., & Tibshirani R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, *23*(1), 73–102. https://doi.org/10.1214/aos/1176324456

Hastie T. J., & Tibshirani R. J. (1990). *Generalized additive models*. Chapman and Hall/CRC.

Hendrycks D., & Dietterich T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations* ICLR.

Kennedy E. H. (2022). 'Semiparametric doubly robust targeted double machine learning: A review', arXiv, arXiv:2203.06469, preprint: not peer reviewed.

Kitani K. M., Ziebart B. D., Bagnell J. A., & Hebert M. (2012). Activity forecasting. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7575, pp. 201–214). LNCS.

Le Cam L. M. (1969). *Théorie asymptotique de la décision statistique* (Vol. *33*). Presses de l'Université de Montréal.

Lei J., G'Sell M., Rinaldo A., Tibshirani R. J., & Wasserman L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, *113*(523), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

Lei J., Rinaldo A., & Wasserman L. (2015). A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, *74*(1-2), 29–43. https://doi.org/10.1007/s10472-013-9366-6

Lei J., Robins J., & Wasserman L. (2013). Distribution-free prediction sets. *Journal of the American Statistical Association*, *108*(501), 278–287. https://doi.org/10.1080/01621459.2012.751873

Lei J., & Wasserman L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *76*(1), 71–96. https://doi.org/10.1111/rssb.12021

Lei L., & Candès E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *83*(5), 911–938.http://arxiv.org/abs/2006.06138. https://doi.org/10.1111/rssb.12445

Levit B. Y. (1974). On optimality of some statistical estimates. In *Proceedings of the Prague symposium on asymptotic statistics* (Vol. 2, pp. 215–238). Charles University Prague.

Little R. J., & Rubin D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons, Ltd.

Malik A., Kuleshov V., Song J., Nemer D., Seymour H., & Ermon S. (2019). Calibrated model-based deep reinforcement learning. In *36th International Conference on Machine Learning, ICML 2019* (pp. 4314–4323). PMLR.

Mason L., Baxter J., Bartlett P., & Frean M. (1999). *Boosting algorithms as gradient descent in function space* (Technical Report).

Mason L., Baxter J., Bartlett P. L., & Frean M. (2000). *Boosting algorithms as gradient descent*. (Technical Report).

Menon A. K., & Ong C. S. (2016). Linking losses for density ratio and class-probability estimation. In *33rd International Conference on Machine Learning, ICML 2016* (Vol. 1, pp. 484–504). PMLR.

Moja L., Kwag K. H., Lytras T., Bertizzolo L., Brandt L., Pecoraro V., Rigon G., Vaona A., Ruggiero F., Mangia M., Iorio A., Kunnamo I., & Bonovas S. (2014). Effectiveness of computerized decision support systems linked to electronic health records: A systematic review and meta-analysis. *American Journal of Public Health*, *104*(12), e12–e22. https://doi.org/10.2105/AJPH.2014.302164

Newey W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, *5*(2), 99–135. https://doi.org/10.1002/jae.3950050202

Newey W. K., Hsieh F., & Robins J. (1998). Undersmoothing and bias corrected functional estimation.

Newey W. K., Hsieh F., & Robins J. M. (2004). Twicing kernels and a small bias property of semiparametric estimators. *Econometrica*, *72*(3), 947–962. https://doi.org/10.1111/j.1468-0262.2004.00518.x

Papadopoulos H., Proedrou K., Vovk V., & Gammerman A. (2002). Inductive confidence machines for regression. In *European Conference on Machine Learning* (pp. 345–356). Springer.

Park S., Dobriban E., Lee I., & Bastani O. (2022). PAC prediction sets under covariate shift. *In International Conference on Learning Representations*. https://doi.org/10.48550/arxiv.2106.09848

Park S., Li S., Lee I., & Bastani O. (2020). 'PAC confidence predictions for deep neural network classifiers', arXiv, arXiv:2011.00716, preprint: not peer reviewed.

Pfanzagl J. (1985). *Contributions to a general asymptotic statistical theory* (Vol. *3*). Lecture Notes in Statistics. Springer New York.

Pfanzagl J. (1990). *Estimation in semiparametric models* (Vol. *63*). Lecture Notes in Statistics. Springer.

Qiu H., Luedtke A., & Carone M. (2021). Universal sieve-based strategies for efficient estimation using machine learning tools. *Bernoulli*, *27*(4), 2300–2336. https://arxiv.org/abs/2003.01856. https://doi.org/10.3150/20-BEJ1309

Quiñonero-Candela J., Sugiyama M., Lawrence N. D., & Schwaighofer A. (2009). *Dataset shift in machine learning*. MIT Press.

Ren S., He K., Girshick R., & Sun J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Rosenbaum P. R., & Rubin D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rotnitzky A., Smucler E., & Robins J. M. (2021). Characterization of parameters with a mixed bias property. *Biometrika*, *108*(1), 231–238. https://doi.org/10.1093/biomet/asaa054

Sadinle M., Lei J., & Wasserman L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, *114*(525), 223–234. https://doi.org/10.1080/01621459.2017.1395341

Saunders C., Gammerman A., & Vovk V. (1999). Transduction with confidence and credibility. In *IJCAI*.

Scharfstein D. O., Rotnitzky A., & Robins J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, *94*(448), 1096–1120. https://doi.org/10.1080/01621459.1999.10473862

Scheffe H., & Tukey J. W. (1945). Non-parametric estimation. I. Validation of order statistics. *The Annals of Mathematical Statistics*, *16*(2), 187–192. https://doi.org/10.1214/aoms/1177731119

Schick A. (1986). On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, *14*(3), 1139–1151. https://doi.org/10.1214/aos/1176350055

Shah R. D., & Peters J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, *48*(3), 1514–1538. https://doi.org/10.1214/19-AOS1857

Shen X. (1997). On methods of sieves and penalization. *The Annals of Statistics*, *25*(6), 2555–2591. https://doi.org/10.1214/aos/1030741085

Shimodaira H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, *90*(2), 227–244. https://doi.org/10.1016/S0378-3758(00)00115-4

Sugiyama M., & Kawanabe M. (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT Press.

Sugiyama M., Suzuki T., Nakajima S., Kashima H., Von Bünau P., & Kawanabe M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, *60*(4), 699–746. https://doi.org/10.1007/s10463-008-0197-x

Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., & Fergus R. (2014). Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*.

Tanser F., Bärnighausen T., Grapsa E., Zaidi J., & Newell M. L. (2013). High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-Natal, South Africa. *Science*, *339*(6122), 966–971. https://doi.org/10.1126/science.1228160

Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani R. J., Barber R. F., Candès E. J., & Ramdas A. (2019). Conformal prediction under covariate shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.). *Advances in Neural Information Processing Systems 32 (NIPS 2019)* (pp. 2526–2536). Curran Associates.

Tukey J. W. (1947). Non-parametric estimation II. Statistically equivalent blocks and tolerance regions–the continuous case. *The Annals of Mathematical Statistics*, *18*(4), 529–539. https://doi.org/10.1214/aoms/1177730343

Tukey J. W. (1948). Nonparametric estimation, III. Statistically equivalent blocks and multivariate tolerance regions–the discontinuous case. *The Annals of Mathematical Statistics*, *19*(1), 30–39. https://doi.org/10.1214/aoms/1177730287

van der Laan M. J., Polley E. C., & Hubbard A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1). https://doi.org/10.2202/1544-6115.1309

Van der Laan M. J., & Rose S. (2018). *Targeted learning in data science: Causal inference for complex longitudinal studies*. Springer.

Van der Laan M. J., & Rubin D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, *2*(1). https://doi.org/10.2202/1557-4679.1043

Van Der Vaart A. (1991). On differentiable functionals. *The Annals of Statistics*, *19*(1), 178–204. https://doi.org/10.1214/aos/1176347976

van der Vaart A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

van der Vaart A. W., & Wellner J. (1996). *Weak convergence and empirical processes: With applications to statistics*. Springer Science & Business Media.

Vovk V. (2013). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning* (Vol. 25, pp. 475–490). PMLR.

Vovk V., Gammerman A., & Saunders C. (1999). Machine-learning applications of algorithmic randomness. In *Sixteenth International Conference on Machine Learning (ICML-1999)* (pp. 444–453). Morgan Kaufmann Publishers Inc.

Vovk V., Gammerman A., & Shafer G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Wald A. (1943). An extension of Wilks' method for setting tolerance limits. *The Annals of Mathematical Statistics*, *14*(1), 45–55. https://doi.org/10.1214/aoms/1177731491

Wilks S. S. (1941). Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, *12*(1), 91–96. https://doi.org/10.1214/aoms/1177731788

Yang Y., Kuchibhotla A. K., & Tchetgen E. T. (2022). 'Doubly robust calibration of prediction sets under covariate shift', arXiv, arXiv:2203.01761, preprint: not peer reviewed.

Zhang J. J., & Liang H. Y. (2011). Berry-Esseen type bounds in heteroscedastic semi-parametric model. *Journal of Statistical Planning and Inference*, *141*(11), 3447–3462. https://doi.org/10.1016/j.jspi.2011.05.001