Journal of Machine Learning Research 27 (2026) 1-68 Submitted 11/24; Revised 2/26; Published 2/26

Transfer Conformal Predictive Inference for Regression

Ce Zhang* CE5@UALBERTA.CA
Department of Mathematical and Statistical Sciences
University of Alberta
Edmonton, AB T6G 2G1, Canada

Ting Li* TINGLI@MAIL.SHUFE.EDU.CN
School of Statistics and Data Science
Shanghai University of Finance and Economics
Shanghai, Shanghai 200433, China

Jinhan Xie* JINHANXIE@YNU.EDU.CN
Yunnan Key Laboratory of Statistical Modeling and Data Analysis
Yunnan University
Kunming, Yunnan 650091, China

Linglong Kong† LKONG@UALBERTA.CA
Department of Mathematical and Statistical Sciences
University of Alberta
Edmonton, AB T6G 2G1, Canada

Bei Jiang† BEI1@UALBERTA.CA
Department of Mathematical and Statistical Sciences
University of Alberta
Edmonton, AB T6G 2G1, Canada

Editor: Chris Oates

Abstract

Conformal prediction, a powerful framework for constructing prediction intervals for response variables using any regression function estimators, often faces the challenge of producing overly broad intervals with limited target data. In this paper, we study the transfer learning problem in conformal prediction, aiming to improve the precision of the prediction interval of the target data with insufficient data by leveraging related auxiliary source datasets. Allowing for the potential non-exchangeability between source and target datasets, we propose two transfer conformal prediction algorithms designed for scenarios where knowledge of informative source data is either present or absent. Our approach uses conditional Kullback-Leibler divergence to effectively identify relevant source datasets for transfer. A comprehensive theoretical analysis of the non-asymptotic properties of the proposed algorithms is provided, including lower and upper bounds, and the prediction interval width. These results illustrate the potential to achieve more efficient, narrower intervals without compromising coverage accuracy. Empirical results from extensive simulations and real-world data confirm the efficacy of our methods, demonstrating significant improvements in prediction interval precision by leveraging source data, achieving narrower intervals while maintaining desired coverage levels.

* Co-first author
† Corresponding author

Keywords: Conditional Kullback-Leibler divergence; Non-exchangeability; Posterior drift; Weighted conformal prediction; Transferability.

1. Introduction

Conformal prediction provides a powerful and flexible tool for generating prediction intervals in classification and regression problems without relying on distributional assumptions on the data (Vovk et al., 2005; Shafer and Vovk, 2008; Vovk et al., 2009; Lei et al., 2018; Angelopoulos and Bates, 2021). It provides a straightforward way to generate prediction sets for any model. Formally, consider a set of n observations $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where $X_i \in \mathbb{R}^p$ represents the covariate and $Y_i \in \mathbb{R}$ denotes the response. For a new data point (X_{n+1}, Y_{n+1}), the primary objective of conformal prediction is to construct a prediction interval $C_n(X_{n+1})$ that covers Y_{n+1} with a confidence level of at least $1 - \alpha$ such that $\Pr\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1 - \alpha$ for any distribution of the data. Without specific assumptions on the distribution and the model, conformal prediction makes it a valuable and general tool for learning the prediction interval of the new response.

One fundamental challenge in conformal prediction lies in securing the desired coverage level without imposing assumptions on the underlying data distribution. Based on the exchangeable assumption, where the data-generating distribution is invariant under permutations of sample points, split conformal prediction (Vovk et al., 2005; Solari and Djordjilović, 2022) employs empirical quantiles derived from the residuals of a holdout set to ensure predictive coverage. This strategy incorporates data splitting to avoid multiple refitting of the predictor (Shafer and Vovk, 2008). However, real-world applications frequently encounter scenarios where the exchangeability assumption does not hold, such as distribution drift between the observations and the new data point and correlations between data points. Most existing works crucially rely on the assumption of data exchangeability, with a few exceptions. In particular, Barber et al. (2023) modified the existing conformal prediction methods to retain predictive coverage in the presence of violations of exchangeability. Meanwhile, several conformal prediction methods have been proposed for dependent data, such as split-conformal based method (Oliveira et al., 2022), block-based construction (Chernozhukov et al., 2018) and online adaptive method (Gibbs and Candes, 2021).

Another challenge in conformal prediction is the limited sample size of the data, leading to inaccurate estimates of uncertainty. When the sample size of the target data is notably small, the resulting prediction intervals may become excessively wide, thereby diminishing their coverage probability (Linusson et al., 2014). The often limited sample size of a single dataset prompts the need to augment it with supplementary external datasets to enhance the precision of prediction. Transfer learning has been demonstrated as an effective tool in utilizing information gained from related source data to improve performance in target data (Cai and Pu, 2022). It has received significant attention in statistics, including transfer learning in classification (Cai and Wei, 2021; Reeve et al., 2021), high-dimensional linear regression (Li et al., 2022), high-dimensional generalized linear regression (Bastani, 2021; Tian and Feng, 2023), and nonparametric regression (Cai and Pu, 2022). Despite the considerable focus on utilizing transfer learning for point estimates, its application in constructing conformal prediction intervals remains relatively unexplored. A notable exception is Fisch et al. (2021), however, it requires the exchangeability of related source data,

an assumption often invalidated in practical scenarios by factors like distribution drift and correlations among data points.

The aim of this paper is to develop a new transfer conformal prediction (TCP) method that leverages auxiliary source datasets to improve the prediction precision of the target dataset. Specifically, we aim to overcome the challenge of limited sample size by leveraging information from source datasets. We seek to achieve the desired coverage level without imposing the exchangeability assumption between the source data and the target data while maintaining a narrower length of predictive interval compared to solely utilizing the target dataset. The non-exchangeability between the target and source data, coupled with potential distributional shifts, complicates the TCP method. Unlike recent studies that develop conformal prediction methods for hierarchical or grouped data (Dunn et al., 2023; Lee et al., 2023; Duchi et al., 2024; Liu et al., 2024), our proposed TCP framework addresses a fundamentally different problem setting. Hierarchical methods (Dunn et al., 2023; Lee et al., 2023; Duchi et al., 2024) focused on group-structured environments, where observations are arranged into exchangeable clusters (e.g., students within classrooms, patients within hospitals), and their validity depends on both between-group exchangeability (interchangeability of groups) and within-group exchangeability (i.i.d. observations within each group). By contrast, our TCP framework targets transfer learning scenarios with non-exchangeable source domains and no inherent hierarchical structure, and thus makes no exchangeability assumptions. Although our TCP framework can also be applied in settings where hierarchical methods are valid, the reverse is not true: hierarchical methods cannot accommodate non-exchangeable sources.

As for reducing the predictive band's length, various conformal prediction methods have been proposed. Recent work by Stutz et al. (2021) on the conformal training method focuses primarily on classification tasks and does not explicitly handle non-exchangeable datasets or distributional shifts. Dheur et al. (2024) provides a systematic extension of conformal highest density regions (HDRs) to multivariate responses, including joint prediction for mixed continuous responses and categorical responses with numerous categories, while maintaining finite-sample marginal coverage guarantees under exchangeability. Their approach builds on the HPD-split framework of Izbicki et al. (2022) and leverages joint predictive densities to construct sharp, distribution-free prediction regions. In contrast, our work focuses on transfer learning settings with non-exchangeable source and target domains, where such finite-sample guarantees are no longer attainable, and new methods are required to control coverage under posterior drift. The approach by Huang et al. (2024) with CF-GNN is designed specifically for graph-structured data, limiting its broader applicability. Xie et al. (2024) propose a boosted conformal method that, while effective, relies on post-hoc adjustments, which may lead to inefficiencies if the underlying model is sub-optimal. Different from existing literature that relied on the conditional quantile regression on the outcome (Romano et al., 2019; Kivaranovic et al., 2020), adapted to skewed data by estimating the conditional histograms (Sesia and Romano, 2021), or estimating the conditional density function to produce non-convex predictive bands (Izbicki et al., 2019; Hoff, 2023), we propose to leverage the information from external source datasets to reduce the length of the predictive band.

Meanwhile, to mitigate the risk of negative transfer due to substantial distributional differences between the target data and the source data, two prevalent frameworks have been

utilized to delineate these differences and facilitate the transfer of source data that exhibits a certain degree of similarity. In the case of covariate shift, the conditional distributions of the response given the covariate are the same, but different distributions of the covariates are allowed across the target data and the source data (Hanneke and Kpotufe, 2019; Schneider et al., 2020). In the case of posterior drift, the distributions of the covariates exhibit consistency while variations emerge in the conditional distribution of the response given the covariates (Liu et al., 2020; Cai and Wei, 2021; Maity et al., 2024). This commonly seen posterior drift model is also adopted in this article. We focus on posterior shifts rather than covariate shifts because posterior shifts directly affect the conditional mean $\mathbb{E}(Y \mid X)$, which is essential for conditional mean regression. Accurate estimation $\mathbb{E}(Y \mid X)$ is critical for constructing valid and efficient conformal prediction intervals (Zhang and Candès, 2024). Moreover, transfer learning can be highly beneficial by using source datasets with small posterior drift differences. Intuitively, this can be achieved by controlling the estimated difference of the posterior drifts similar to existing works. However, the specific influence of this difference on the predictive interval in conformal prediction is not well understood, which is one of the focuses of this paper.

In this paper, we propose a novel TCP method that leverages multiple non-exchangeable source datasets to enhance the prediction efficiency for target data, based on the posterior drift model. In contrast to the existing literature discussed above, we make several major contributions, as outlined below.

- We propose two novel transfer prediction algorithms designed for scenarios with and without prior knowledge of information sources, accommodating non-exchangeability between source and target data. For cases where information sources are predetermined, the algorithm integrates a transferring step, a debiasing step, and a conformal step. The conditional Kullback-Leibler (cKL) divergence is employed to identify the most informative transfer set for cases without knowledge of information sources.
- We thoroughly investigate the theoretical properties of the proposed TCP method, including lower and upper bounds as well as the bounds on the width of the prediction bands. The theoretical results not only affirm the method's capability to provide valid asymptotic coverage but also demonstrate its efficiency in producing prediction bands of narrower widths compared to conventional conformal methods using a single target data (Lei et al., 2018; Barber et al., 2023). Furthermore, the proposed algorithm for source data detection is proven to accurately select the informative set with high probability.
- We consider three prevalent regression settings: high-dimensional linear, generalized linear models, and nonparametric regression. For each setting, we present the explicit configurations of the corresponding TCP bands and detail the specific convergence rates for both coverage and the width of the prediction intervals.
- Extensive studies and real data analysis of an election data set are conducted to examine the finite-sample performance of the proposed algorithm. These results show that the proposed algorithms can improve the quality of the prediction intervals, achieving narrower widths while preserving the desired coverage level, by effectively leveraging information from source data while avoiding negative transfer.

The remainder of the article is organized as follows. In Section 2, we introduce the basics of the two different split conformal models. Section 3 presents the proposal TCP

approaches with and without prior knowledge of useful sources. Section 4 investigates theoretical properties, the lower and upper bounds, and the width of the prediction interval. Section 5 and 6 assess the finite sample performance through simulations and an election dataset, respectively. All technical details are included in the Appendix.

2. Preliminaries

In this section, we present a brief review of split conformal prediction and weighted conformal prediction, which lays the groundwork for the proposed method.

2.1 Split conformal prediction

Split conformal prediction begins by fitting a pre-trained model to estimate the conditional mean $\mu = \mathbb{E}(Y|X)$, denoted as $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$, using an initial training dataset. Subsequently, residuals are computed from the model $\widehat{\mu}$, using a separate holdout dataset. From these residuals, the corresponding quantile is determined. Finally, the prediction interval is constructed by integrating both the model $\widehat{\mu}$ and the calculated quantile.

Specifically, suppose a model $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$ that has been fitted on an original training dataset. Residuals for n subsequent exchangeable observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ can be derived as $R_i = |Y_i - \widehat{\mu}(X_i)|,\ i = 1, \ldots, n$. The prediction interval for a new testing feature vector X_{n+1} can be expressed as

$$\widehat{C}_n(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{n} \frac{1}{n+1} \cdot \delta_{R_i} + \frac{1}{n+1} \cdot \delta_{+\infty}\right), \tag{1}$$

with $\mathrm{Q}_\tau(\cdot)$ signifying the τ-quantile of its input and δ_a representing the point mass at a. Notice that the confidence radius is the $\lceil (1-\alpha)(n+1) \rceil$-th order statistic of the residuals $\{R_1, \ldots, R_n\}$. Notably, the split conformal prediction technique has been demonstrated to ensure finite sample validity, with predictive coverage at the desired $1-\alpha$ level, regardless of the distributional assumptions of the observations (Papadopoulos et al., 2002; Vovk et al., 2005; Lei et al., 2015).

To validate (1), a fundamental assumption of the split conformal method is the exchangeability of the n observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ with the subsequent observation (X_{n+1}, Y_{n+1}) (Lei et al., 2018). However, this assumption is often compromised by factors such as distribution drift, correlations among data points, and other related phenomena (Barber et al., 2023). This challenge is particularly pronounced in datasets collected from multiple sources, as frequently observed in transfer learning scenarios. In such contexts, distribution drift is likely to occur, leading to disparate distributions across datasets and thus violating the exchangeability assumption. Consequently, to effectively leverage information from the source data, it is imperative to acknowledge and account for non-exchangeability when constructing prediction intervals for the target data.

2.2 Weighted conformal prediction

It is essential to emphasize that the standard split conformal prediction method uniformly assigns a weight of $1/(n+1)$ to the residuals in (1). To address the limitations posed by the exchangeability assumption, weighted conformal prediction provides an alternative

approach. This method assigns different weights to the residuals based on the similarity between the testing and training data (Tibshirani et al., 2019; Lei and Candès, 2021; Fannjiang et al., 2022; Barber et al., 2023). In scenarios where there are no distributional shifts, characterized by a density ratio across data points equal to one, the weighted conformal prediction method simplifies to the conventional split conformal prediction approach.

In the weighted conformal prediction framework, each ith data point $Z_i = (X_i, Y_i)$ is assigned a weight w_i. This weighting is based on the principle that a higher value w_i, indicates a greater degree of reliability associated with the data point Z_i. This reliability is derived from the similarity of its distribution to that of the new point $Z_{n+1} = (X_{n+1}, Y_{n+1})$. Common choices for the weights include the density ratio of the covariate in the context of covariate shift (Tibshirani et al., 2019; Fannjiang et al., 2022) or fixed weights based on the ordering of the data points (Barber et al., 2023). Consequently, the non-exchangeable weighted split conformal prediction, articulated within a symmetric algorithm, is defined as follows:

$$\widehat{C}_n\left(X_{n+1}\right) = \widehat{\mu}\left(X_{n+1}\right) \pm \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{n} \widetilde{w}_i \cdot \delta_{R_i} + \widetilde{w}_{n+1} \cdot \delta_{+\infty}\right),$$

where the normalized weights $\widetilde{w}_i = w_i/(w_1 + \cdots + w_n + 1), i = 1, \ldots, n$, and $\widetilde{w}_{n+1} = 1/(w_1 + \cdots + w_n + 1)$ given $w_i \in [0, 1]$, $i = 1, \ldots, n$. Barber et al. (2023) demonstrated that this weighted conformal model extends a coverage guarantee, applicable even in the absence of exchangeability:

$$\Pr\left\{Y_{n+1} \in \widehat{C}_n\left(X_{n+1}\right)\right\} \geq 1 - \alpha - \sum_{i=1}^{n} \widetilde{w}_i d_{\mathrm{TV}}\left(Z, Z^i\right), \tag{2}$$

where d_{TV} denotes the total variation distance between distribution, $Z = (Z_1, \ldots, Z_{n+1})$ and $Z^i = (Z_1, \ldots, Z_{i-1}, Z_{n+1}, Z_{i+1}, \ldots, Z_i)$ that posts the substitution of (X_{n+1}, Y_{n+1}) with (X_i, Y_i). The coverage gap between $1 - \alpha$ and $\mathbb{P}\{Y_{n+1} \in \widehat{C}_n\left(X_{n+1}\right)\}$ is bounded by $d_{\mathrm{TV}}\left(Z, Z^i\right)$. Notably, this method aligns with the conventional split conformal prediction (1) when the weights $\{w_i\}_{i=1}^n$ are uniformly set to 1. Moreover, in the case of exchangeable data, the distribution of Z equals that of Z^i, ensuring that $d_{\mathrm{TV}}\left(Z, Z^i\right) = 0$ for all i, and thus preserving the coverage.

Remark 1 *The significance of utilizing prior knowledge about the data distribution is particularly evident in the selection of appropriate weights w_i aimed at minimizing the final term in Equation* (2)*. By assigning larger weights to calibration points (X_i, Y_i) that exhibit similar distribution characteristics to Z_{n+1}, while allocating smaller weights to others, one can attain more stringent bounds.*

By moving away from the exchangeability assumption, the weighted conformal method demonstrates greater precision in prediction intervals within a non-exchangeable context compared to the standard split conformal approach (Tibshirani et al., 2019). However, challenges arise when training data are significantly limited, which can lead to excessively broad prediction intervals that may lack practical utility. We will provide a more detailed discussion of this phenomenon in the simulation and application sections. To mitigate these

limitations, we propose the transfer conformal model. This innovative approach seeks to enhance the accuracy of empirical quantile estimations by incorporating relevant information from source data and employing the non-exchangeable weighted conformal method, thereby yielding more precise prediction intervals.

3. Methodology

In this section, we propose the TCP method designed to address the limitations of conventional conformal prediction for single target data. We begin by describing the challenges associated with the conventional method in Section 3.1 and then introduce a novel data structure that integrates both the source and target data. Section 3.2 details a general TCP algorithm applicable to scenarios with known informative source data. Additionally, in Section 3.3, we introduce a transferable source detection algorithm that employs cKL divergence to identify transferable source data that can enhance prediction accuracy.

3.1 Problem set-up and transferability

In this subsection, we formally define the problem underlying the proposed TCP method. The target domain consists of an independent and identically distributed (i.i.d.) dataset $(X_i^{(0)}, Y_i^{(0)}) \sim Q$, for $i = 1, \ldots, n_0$, where $X^{(0)} \in \mathbb{R}^{n_0 \times p}$ and $Y^{(0)} \in \mathbb{R}^{n_0}$. In the realm of transfer learning, additional samples are available from K auxiliary source domains, denoted as $(X_i^{(k)}, Y_i^{(k)}) \sim P^{(k)}$, for $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$. Typically, it is often the case that the K source data are non-exchangeable with the target data. The objective is to construct a conformal prediction for a new data point (X_{new}, Y_{new}) from the target domain by effectively leveraging not only the target dataset but also the source datasets.

When constructing the prediction interval for Y_{new} solely using the target data, employing either standard split conformal prediction or weighted conformal prediction, a limited sample size n_0 often leads to an expanded prediction interval. Intervals with excessive width can significantly reduce the practical utility of conventional conformal prediction methods, posing a challenge in achieving the desired balance between interval precision and coverage. To address the above challenge of data scarcity, we propose a novel approach that incorporates conformal prediction within a transfer learning framework. The proposed method utilizes K nonexchangeable source datasets with similar distribution structures to improve the data efficiency of conformal prediction, resulting in narrower and more precise prediction intervals for the target dataset.

In transfer learning, the source domain's joint distributions $P^{(k)}, k = 1, \ldots, K$, often differ from the target domain's distribution Q. Directly combining these datasets for conformal prediction can introduce substantial bias when their distributions vary significantly. It is more effective to utilize an auxiliary dataset that closely resembles the target dataset, often termed the "information set" (Li et al., 2022). This study concentrates on scenarios with posterior drift across the K source datasets and the target dataset. We assess the similarity between different distributions by examining the divergence in their conditional distributions. Specifically, if the conditional distribution $P^{(k)}(Y^{(k)}|X^{(k)})$ of the kth source dataset $(X^{(k)}, Y^{(k)})$ closely aligns with $Q(Y^{(k)}|X^{(k)})$ of the target, valuable information can be transferred from the kth source to enhance the TCP interval for the target data. A

crucial aspect of our methodology is the identification of similar source data to target data. To quantify this similarity, we introduce the "level-h" transferring set, denoted as $\mathcal{A}_h$:

$$\mathcal{A}_h = \left\{ k \in \{1, \ldots, K\} : \left\| P^{(k)}(Y^{(k)}|X^{(k)}) - Q(Y^{(k)}|X^{(k)}) \right\|_\infty \leq h \right\}, \tag{3}$$

where $\|\cdot\|_\infty$ is defined as the infinity norm such that $\left\|P^{(k)}(Y^{(k)}|X^{(k)}) - Q(Y^{(k)}|X^{(k)})\right\|_\infty = \sup\{|P^{(k)}(Y^{(k)}|X^{(k)}) - Q(Y^{(k)}|X^{(k)})|\}$. This set includes source data where the transferability level is controlled by h. It's important to note that h can be any positive value, and adjusting h defines different sets $\mathcal{A}_h$. A smaller h indicates that the source data within $\mathcal{A}_h$ are more closely aligned with the target, and a larger cardinality of $\mathcal{A}_h$ ($|\mathcal{A}_h|$) signifies a greater number of informative auxiliary samples. When h is small but $|\mathcal{A}_h|$ is large, transfer learning can offer significant advantages.

3.2 TCP with known information set

In this section, we consider the proposed TCP method when the information set $\mathcal{A}_h$ is known and propose a novel TCP algorithm. The conditional mean function of the target population is denoted by $\mu_0(X^{(0)}) = \mathbb{E}\{Y^{(0)}|X = X^{(0)}\}$, while the conditional mean function for the kth source population is represented by $\mu_k(X^{(k)}) = \mathbb{E}\{Y^{(k)}|X = X^{(k)}\}$.

Given a new test feature X_{new} from the target population, we propose constructing the prediction interval for its corresponding response Y_{new} using all the available observed data,

$$\widehat{C}_n\left(X_{new}\right) = \widehat{\mu}\left(X_{new}\right) \pm \left(q_1 + \Lambda\right), \tag{4}$$

where $\widehat{\mu}$ is constructed using both the target data and source data, q_1 represents the estimated $1-\alpha$ quantile of the fitted residuals from source data, and Λ denotes the bias correction term that accounts for potential distributional shifts between the source and target data. This correction is defined as $\Lambda = q_2 - q_1$, where q_2 is the population $1-\alpha$ quantile of the residual distribution for the target model. Importantly, when the target and source distributions are identical, we have $\Lambda = 0$.

Remark 2 *The proposed prediction interval* (4) *is built on two key components: an enhanced transfer learning-based regression function estimator and an improved* $1-\alpha$ *empirical quantile of the fitted residuals, both tailored for the target data. First, the precision of the regression function estimate* $\widehat{\mu}$ *is enhanced by applying transfer learning techniques in regression (Li et al., 2022; Cai and Pu, 2022; Tian and Feng, 2023). Second, the* $1-\alpha$ *empirical quantile* q_2 *for the target data is improved by initially calculating a weighted quantile* q_1 *from the source datasets. This preliminary* q_1 *is then debiased and refined using* Λ*, derived from the target data, to better align the quantile with the specific characteristics of the target data.*

Our proposed TCP algorithm consists mainly of three steps: the transferring step, the debiasing step, and the conformal step. To avoid the overfitting problem, we partition the target data $(X^{(0)}, Y^{(0)})$ into two separate subsets, $\mathcal{I}_1$ and $\mathcal{I}_2$. Similarly, for each k-th source data $(X^{(k)}, Y^{(k)})$, ranging from $k = 1$ to K, we divide them into two non-overlapping subsets, $\mathcal{S}_1^k$ and $\mathcal{S}_2^k$. By amalgamating all individual subsets $\mathcal{S}_1^k$ and $\mathcal{S}_2^k$ for $k = 1$ to K, we

generate two independent, equally-sized sets: $\mathcal{S}_1 = \cup_{k=1}^{K} \mathcal{S}_1^k$ and $\mathcal{S}_2 = \cup_{k=1}^{K} \mathcal{S}_2^k$. Specifically, in the transfer step, we use data from $\mathcal{I}_1$ and $\mathcal{S}_1$ to estimate a preliminary regression function and a weighted $1-\alpha$ quantile of the residuals. The debiasing step employs a bootstrap method to compute the correction parameter Λ, refining this preliminary quantile. In the conformal step, $\mathcal{S}_1, \mathcal{I}_1$, and Λ are used to obtain the debiased weighted $1-\alpha$ quantile, while $\mathcal{S}_2$ and $\mathcal{I}_2$ are used to fit a transfer-learned regression function. These components are then combined to construct the TCP interval.

3.2.1 Transferring step

In this subsection, we apply transfer learning techniques to integrate information from both target and source datasets. Specifically, we use data from $\mathcal{I}_1$ and $\mathcal{S}_1$ to estimate the preliminary conditional mean and a preliminary weighted $1-\alpha$ quantile of the residuals. This approach addresses the limitations of limited target data by leveraging auxiliary information from source datasets. Under exchangeability conditions, Fisch et al. (2021) utilized multiple preliminary conditional mean estimates $\widetilde{\mu}_t$ and preliminary weighted quantiles of residuals $\widetilde{q}_{1-\alpha}^t$, $t = 1, \cdots, T$, derived from T independent and non-overlapping source datasets, to calculate the correction parameter Λ. However, the assumption of exchangeability between source and target datasets is often unrealistic in practical applications. In this paper, we extend the bias correction framework to accommodate non-exchangeable cases.

To generate multiple sets of $\{\widetilde{\mu}_t, \widetilde{q}_{1-\alpha}^t\}$ for bias correction, we begin by using the bootstrap resampling method to create T distinct random copies from $\mathcal{I}_1$, denoted as $\mathcal{I}_1^t$ for $t = 1, \ldots, T$. We define the ensemble of these samples as $\mathcal{B} = \left\{\mathcal{I}_1^t, t = 1, \ldots, T\right\}$, and the complementary set $\mathcal{B}_{-t}$ as the collection with the t-th subset $\mathcal{I}_1^t$ removed. Since the bootstrap samples are drawn with replacement from the same empirical distribution as the target data, they remain exchangeable with the target data, ensuring their distribution remains unchanged regardless of sample order.

For each subsample $\mathcal{I}_1^t$ drawn from $\mathcal{I}_1$, we estimate a fitted regression function $\widetilde{\mu}_t$ by combining the data from $\mathcal{S}_1$ and $\mathcal{I}_1^t$, formulated as follows:

$$\widetilde{\mu}_t = \mathcal{D}_1\left(\{(X_i, Y_i)\}\right), \tag{5}$$

where $(X_i, Y_i) \in \mathcal{S}_1 \cup \mathcal{I}_1^t$ and $\mathcal{D}_1$ denotes a general regression algorithm. For the data within $\mathcal{S}_1$, we compute the absolute fitted residuals $\widetilde{R}_i^t$, given by $\widetilde{R}_i^t = |Y_i - \widetilde{\mu}_t(X_i)|$ with $(X_i, Y_i) \in \mathcal{S}_1$, and the weighted $1-\alpha$ quantile of the empirical distribution of $\widetilde{R}_i^t$ is expressed as $\widetilde{q}_{1-\alpha}^t = \mathrm{Q}_{1-\alpha}(\sum_{i=1}^{|\mathcal{S}_1|+1} \widetilde{w}_i^t \cdot \delta_{\widetilde{R}_i^t})$, where $\widetilde{w}_i^t$ denotes the weight corresponding to $\mathcal{I}_1^t$. As highlighted in Barber et al. (2023), these weights are held constant, with higher weights assigned to data points in $\mathcal{S}_1$ believed to be drawn from a distribution similar to the test data (X_{new}, Y_{new}). The necessity of this bootstrap procedure will be further discussed in the next subsection. With these steps in place, we are now prepared to estimate the correction parameter Λ.

3.2.2 Debiasing step

In the previous subsection, we introduce the transfer step. This step inherently introduces bias into both the fitted regression function $\widetilde{\mu}_t$ and the empirical quantile $\widetilde{q}_{1-\alpha}^t$ due to

distributional differences between the source and target data. To correct for these biases, we implement a bootstrap-based debiasing method, which adjusts the estimates to more closely align with the target distribution as follows.

1. **Correction of the fitted regression function**: To correct for the bias in $\widetilde{\mu}_t$, we use target data $\mathcal{I}_1^t$. The resulting corrected fitted regression function is denoted as $\widehat{\mu}_t$ and is defined as

$$\widehat{\mu}_t = \mathcal{D}_2\left(\{(X_i, Y_i)\}; \widetilde{\mu}_t\right), \text{ for all } (X_i, Y_i) \in \mathcal{I}_1^t, \tag{6}$$

   where $\mathcal{D}_2$ denotes a general transfer learning algorithm and $\widehat{\mu}_t$ is the debiased regression estimate specifically adjusted for the target data.

2. **Correction of the empirical quantile**: To correct for the bias in $\widetilde{q}_{1-\alpha}^t$, we need to accurately estimate the correction parameter Λ. This requires estimating the true distribution function (denoted by F_t) of $\widehat{R}_i^{-t}$, where $\widehat{R}_i^{-t} = |Y_i - \widehat{\mu}_t(X_i)|$, for all $(X_i, Y_i) \in \mathcal{B}_{-t}$ and $t = 1, \cdots, T$. A natural estimate of F_t is the empirical distribution function, defined as:

$$\widehat{F}_t\left\{\widetilde{q}_{1-\alpha}^t + \Lambda\left(\mathcal{I}_1\right)\right\} = \frac{\left|\left\{\widehat{R}_i^{-t} \leq \widetilde{q}_{1-\alpha}^t + \Lambda\left(\mathcal{I}_1\right)\right\}\right|}{|\mathcal{B}_{-t}| + 1}.$$

   An intuitive approach to estimating $\Lambda(\mathcal{I}_1)$ is to adjust it such that $\widehat{F}_t\{\widetilde{q}_{1-\alpha}^t + \Lambda(\mathcal{I}_1)\}$ closely approximates $1-\alpha$. In other words, for each t, we aim for $\widehat{F}_t\{\widetilde{q}_{1-\alpha}^t + \Lambda(\mathcal{I}_1)\}$ to be as close as possible to $1-\alpha$. However, relying on a single estimate for $\widehat{F}_t\{\widetilde{q}_{1-\alpha}^t + \Lambda(\mathcal{I}_1)\}$ can lead to instability. To address this, we apply a bootstrap procedure to generate a sequence of estimates $\{\widehat{F}_t\{\widetilde{q}_{1-\alpha}^t + \Lambda(\mathcal{I}_1)\}; t = 1, \ldots, T\}$. We then select $\Lambda\left(\mathcal{I}_1\right)$ such that the majority of these estimates are close to $1 - \alpha$, using a mode-based estimation approach.

Remark 3 *We illustrate the core concept behind the correction. Let $\Lambda\left(\mathcal{I}_1\right)$ denote the estimate of the correction parameter Λ based on $\mathcal{I}_1$. If $\Lambda\left(\mathcal{I}_1\right)$ is valid, then for each t, the expression $\widetilde{q}_{1-\alpha}^t + \Lambda\left(\mathcal{I}_1\right)$ should closely approximate the $1 - \alpha$ empirical quantile of $\widehat{R}_i^{-t}$, $t = 1, \cdots, T$. Therefore, the correction can be estimated by selecting $\Lambda(\mathcal{I}_1)$ such that $\Pr(\widehat{R}_i^{-t} \leq \widetilde{q}_{1-\alpha}^t + \Lambda(\mathcal{I}_1)) \approx 1 - \alpha$. Since $\{\mathcal{I}_1^t\}$'s are bootstrap subsamples of $\mathcal{I}_1$, and thus exchangeable with $\mathcal{I}_1$, or the target data, $\Lambda(\mathcal{I}_1)$ can be regarded as the correction for the empirical quantile estimate in the target data.*

As described above in the correction of the empirical quantile $\Lambda(\mathcal{I}_1)$ for $t = 1, \ldots, T$, we aim to compute $\widehat{F}_t\{\widetilde{q}_{1-\alpha}^t + \Lambda\left(\mathcal{I}_1\right)\}$. To further improve the accuracy of $\widehat{F}_t\{\widetilde{q}_{1-\alpha}^t + \Lambda\left(\mathcal{I}_1\right)\}$, we can adopt the calibration-conditional empirical distribution function estimation method (Bates et al., 2023), denoted as $\widehat{F}_t^{(\mathrm{ccv})}$, which is given by

$$\widehat{F}_t^{(\mathrm{ccv})}\left\{\widetilde{q}_{1-\alpha}^t + \Lambda\left(\mathcal{I}_1\right)\right\} = \min\left\{\frac{\gamma}{|\mathcal{B}_{-t}|} + c(\delta)\frac{\sqrt{\gamma\left(|\mathcal{B}_{-t}| - \gamma\right)}}{|\mathcal{B}_{-t}|\sqrt{|\mathcal{B}_{-t}|}}, 1\right\}, \tag{7}$$

where $\gamma = \lceil (|\mathcal{B}_{-t}| + 1) * \widehat{F}_t(\widetilde{q}^t_{1-\alpha} + \Lambda(\mathcal{I}_1)) \rceil + 1$, and

$$c(\delta) = \frac{-\log[-\log(1-\delta)] + 2\log\log|\mathcal{B}_{-t}| + \frac{1}{2}\log\log\log|\mathcal{B}_{-t}| - \frac{1}{2}\log\pi}{\sqrt{2\log\log|\mathcal{B}_{-t}|}},$$

with δ being a user-specified value between 0 and 1. To compute $\widehat{F}_t^{(\mathrm{ccv})}$, a value for the parameter δ must be selected. Any value of δ within the range of 0 to 1 is acceptable. In Section 5.2, we will perform additional simulations to test the robustness of our approach with respect to different values of δ. As illustrated in Figure 4, both the coverage probability and the width of the TCP prediction intervals remain stable across various values of δ. After calibration of the empirical distribution function, $\widehat{F}_t^{(\mathrm{ccv})}$ remains mutually independent across different estimates $\{\widetilde{q}^t_{1\alpha} + \Lambda(\mathcal{I}_1), t = 1, \cdots, T\}$.

We now proceed to estimate the correction parameter Λ as follows. First, we consider a predefined grid of potential values for $\Lambda(\mathcal{I}_1)$ and, for each value, compute $\{\widehat{F}_t^{(\mathrm{ccv})}\{\widetilde{q}^t_{1-\alpha} + \Lambda(\mathcal{I}_1)\}; t = 1, \ldots, T\}$. The optimal value of $\Lambda(\mathcal{I}_1)$ is then determined by:

$$\Lambda(\mathcal{I}_1) = \inf\left[\Lambda(\mathcal{I}_1) : \widehat{F}^{(\mathrm{ccv})}_{(\lfloor T*0.05\rfloor+1)}\left\{\widetilde{q}^{(\lfloor T*0.05\rfloor+1)}_{1-\alpha} + \Lambda(\mathcal{I}_1)\right\} \geq 1-\alpha\right], \tag{8}$$

where $\widehat{F}^{(\mathrm{ccv})}_{(\lfloor T*0.05\rfloor+1)}$ is $(\lfloor T*0.05\rfloor+1)$-th order statistic of $\{\widehat{F}_t^{(\mathrm{ccv})}\{\widetilde{q}^t_{1-\alpha}+\Lambda(\mathcal{I}_1)\}; t = 1, \ldots, T\}$. This approach ensures that most values of $\widehat{F}_t^{(\mathrm{ccv})}\left\{\widetilde{q}^t_{1-\alpha} + \Lambda(\mathcal{I}_1)\right\}$ are close to $1-\alpha$, providing a robust estimate of Λ while reducing the influence of outliers, which usually correspond to lower $\widehat{F}_t^{(\mathrm{ccv})}$ values.

Figure 1 illustrates the process of estimating $\Lambda(\mathcal{I}_1)$, showing how the bootstrap samples $\mathcal{I}_1^1, \mathcal{I}_1^2, \ldots, \mathcal{I}_1^T$ are used to estimate $\widetilde{\mu}_t$ and $\widetilde{q}^t_{1-\alpha}$, leading to the calculation of $\widehat{F}_t^{(\mathrm{ccv})}$, $t = 1, \cdots, T$, and the final value of $\Lambda(\mathcal{I}_1)$.

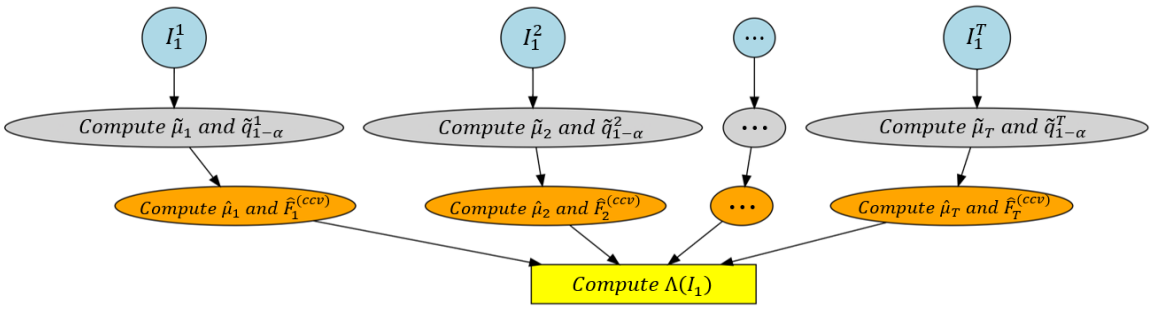

Figure 1: Flowchart of the computation process for $\Lambda(\mathcal{I}_1)$. The estimates $\widehat{F}_t^{(\mathrm{ccv})}$ are derived from the data $\mathcal{B}_{-t}$, $t = 1, \cdots, T$.

3.2.3 Conformal step

In this subsection, we implement the proposed TCP interval for the target data. First, we use data from $\mathcal{S}_1$ and $\mathcal{I}_1$ to estimate the preliminary $1-\alpha$ quantile of the residuals. By incorporating the information from $\mathcal{S}_1$ and combining it with $\mathcal{I}_1$, we define the preliminary regression function $\widetilde{\mu}_{\mathcal{I}_1}$ as $\widetilde{\mu}_{\mathcal{I}_1} = \mathcal{D}_1(\{(X_i, Y_i)\})$, where $(X_i, Y_i) \in \mathcal{S}_1 \cup \mathcal{I}_1$. The absolute

fitted residuals computed on $\mathcal{S}_1$, are given by $\widetilde{R}_i^{\mathcal{I}_1} = |y_i - \widetilde{\mu}_{\mathcal{I}_1}(X_i)|$, for all $(X_i, Y_i) \in \mathcal{S}_1$. Pooling the data from $\mathcal{S}_1$, the weighted $1-\alpha$ quantile of the empirical distribution of $\widetilde{R}_i^{\mathcal{I}_1}$ is defined as $\widetilde{q}_{1-\alpha}^{\mathcal{I}_1} = Q_{1-\alpha}(\sum_{i=1}^{|\mathcal{S}_1|+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}})$, where $\widetilde{w}_i^{\mathcal{I}_1}$ represents the normalized weights for $\mathcal{I}_1$, calculated as $\widetilde{w}_i^{\mathcal{I}_1} = w_i^{\mathcal{I}_1}/(w_1^{\mathcal{I}_1} + \cdots + w_N^{\mathcal{I}_1} + 1)$ with $w_1^{\mathcal{I}_1}, \ldots, w_N^{\mathcal{I}_1}$ being fixed values within the range $(0, 1)$. As in previous methods, higher weights are assigned to data points in $\mathcal{S}_1$ that are most likely drawn from a distribution similar to the test data $(X_{\text{new}}, Y_{\text{new}})$ (Barber et al., 2023).

Next, we incorporate data from $\mathcal{S}_2$ by pooling it with $\mathcal{I}_2$ and estimate the preliminary regression function $\widetilde{\mu}_{\mathcal{I}_2}$ to mitigate the risk of overfitting:

$$\widetilde{\mu}_{\mathcal{I}_2} = \mathcal{D}_1\left(\{(X_i, Y_i)\}\right), \text{ where } (X_i, Y_i) \in \mathcal{S}_2 \cup \mathcal{I}_2,$$

After applying bias correction using data from $\mathcal{I}_2$, we obtain the debiased regression function $\widehat{\mu}_{\mathcal{I}_2}$,

$$\widehat{\mu}_{\mathcal{I}_2} = \mathcal{D}_2\left(\{(X_i, Y_i)\}; \widetilde{\mu}_{\mathcal{I}_2}\right), \text{ for all } (X_i, Y_i) \in \mathcal{I}_2.$$

Given a new data point X_{new}, the proposed TCP interval is constructed as:

$$\widehat{C}(X_{new}) = \widehat{\mu}_{\mathcal{I}_2}(X_{new}) \pm \left(\widetilde{q}_{1-\alpha}^{\mathcal{I}_1} + \Lambda(\mathcal{I}_1)\right), \tag{9}$$

with the computation process for $\widehat{C}(X_{new})$ shown in Figure 2.

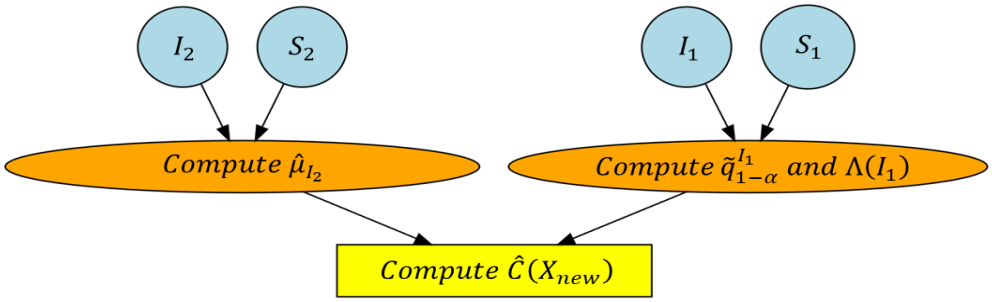

Figure 2: Flowchart of the computation process for $\widehat{C}(X_{new})$.

3.2.4 Summary

In this subsection, we present the proposed TCP method for constructing conformal prediction intervals with limited target data. The process begins by leveraging the source data through transfer learning to estimate both the initial conditional mean regression function and the empirical quantile of residuals. To address any bias in these estimates, we employ bootstrap-based debiasing techniques. This approach results in the final construction of the TCP interval, which ensures reliable coverage for the target data by incorporating the debiased estimates.

The details of the algorithm are provided in Algorithm 1. In Step 1, the dataset is defined, and in Step 2, we use the bootstrap method to generate T random replications of

$\mathcal{I}_1$, denoted as $\mathcal{I}_1^t, t = 1, \ldots, T$. In Steps 3-13, the correction parameter $\Lambda(\mathcal{I}_1)$ is estimated using $\mathcal{B}$ and $\mathcal{S}_1$. Steps 14-16 involve estimating $\widetilde{q}_{1-\alpha}^{\mathcal{I}_1}$ using $\mathcal{I}_1$ and $\mathcal{S}_1$. Finally, in Step 19, the proposed TCP interval $\widehat{C}(X_{\text{new}})$, is constructed using $\widehat{\mu}_{\mathcal{I}_2}$, $\widetilde{q}_{1-\alpha}^{\mathcal{I}_1}$, and $\Lambda(\mathcal{I}_1)$, where $\widehat{\mu}_{\mathcal{I}_2}$ is fitted with $\mathcal{I}_2$ and $\mathcal{S}_2$ in Steps 17-18.

Remark 4 *In contrast to Liu et al. (2024), our work addresses a different problem formulation. Liu et al. (2024) explicitly considered scenarios in which some target-site outcomes are missing, indicated by a missingness indicator $R = 0$. Their density ratio is defined as $\omega_{k,0}(\boldsymbol{X}) = p(\boldsymbol{X} \mid T = 0, R = 0)/p(\boldsymbol{X} \mid T = k, R = 1)$. If all training data satisfy $R = 1$, then the set $\{X : T = 0, R = 0\}$ is empty, and consequently the numerator term $p(X \mid T = 0, R = 0)$ is undefined. By contrast, our TCP framework assumes that all training samples have observed outcomes ($R = 1$ for all data points). Under this regime, the density ratio of Liu et al. (2024) cannot be computed, and their influence-function-based correction is therefore inapplicable. We emphasize that this distinction reflects fundamentally different problem settings rather than a limitation of either approach. We also conduct simulations against baselines inspired by Lee et al. (2023) and Duchi et al. (2024) (see Appendix H for details). The results show that even with limited target data, TCP consistently achieves better coverage and narrower prediction intervals.*

Remark 5 *A natural extension of TCP to time-series or dependent data is to treat TCP as the base forecaster in a one-step-ahead setting and then calibrate its prediction sets using a time-ordered adaptive conformal procedure, as in Zaffran et al. (2022). Specifically, we observe a dependent sequence $\{(X_t, Y_t)\}_{t\geq 1}$, where Y_t may depend on the past observations through lagged responses and historical covariates. A general time-series forecasting model can be written as:*

$$Y_t = m_t(\mathcal{H}_{t-1}) + \varepsilon_t, \qquad \mathcal{H}_{t-1} := \{(X_s, Y_s) : s \leq t-1\},$$

where $m_t(\cdot)$ is an unknown forecasting function and the innovations ε_t may exhibit temporal dependence (e.g., AR or ARMA-type structures) rather than being i.i.d.. The goal is to construct a prediction set $\widehat{C}_t(X_t)$ for Y_t at each time t using only the past history $\mathcal{H}_{t-1}$. This is the standard time-series forecasting regime considered in Zaffran et al. (2022), which allows for both temporal dependence and distributional drift. To apply the proposed TCP in this setting, at each time t, we refit the TCP forecaster using a moving window of recent target observations, augmented with auxiliary source data (e.g., earlier time periods, related environments, or parallel series) selected via our source detection procedure. We then calibrate the forecaster using the most recent target observations immediately following the training window. Prediction sets are constructed from the $(1-\alpha_t)$-quantile of the resulting nonconformity scores, with the proposed Λ correction incorporated in the same manner as in our main method. To accommodate temporal dependence and potential distributional drift, the nominal miscoverage level is updated online using the ACI (Zaffran et al., 2022) recursion:

$$\alpha_{t+1} \;=\; \alpha_t \;+\; \gamma\left(\alpha - \mathbf{1}\left\{Y_t \notin \widehat{C}_t(X_t)\right\}\right), \tag{10}$$

which adaptively enlarges or shrinks future prediction intervals in response to recent under- or over-coverage. When the auxiliary sources share a similar forecasting structure with the

target series, incorporating them through TCP improves predictive accuracy for the target. This extension therefore preserves the transfer learning benefits of TCP while providing a principled and practical mechanism for handling temporal dependence via adaptive calibration. In particular, improved forecasting accuracy typically leads to smaller nonconformity scores on recent target data, resulting in lower calibrated quantiles and shorter prediction sets for the same nominal coverage.

3.3 Transferable source data detection

The proposed TCP method above assumes knowledge of the transferable set. In practice, however, this set is typically unknown, and incorporating source data, while potentially beneficial, can conversely impair performance on the target task. This is often attributable to the disparities in the distributions of source and target data, a phenomenon termed "negative transfer" (Pan and Yang, 2009; Torrey and Shavlik, 2010; Weiss et al., 2016). The challenge of avoiding negative transfer has gained significant research interest in recent years. Consequently, the need for a data-driven approach to detect transferable source data is paramount.

Our TCP framework is developed for transfer learning in regression, where the validity and efficiency of conformal intervals hinge on accurate estimation of the regression function $\mu(X) = \mathbb{E}[Y|X]$ as discussed in Sections 1 and 3.3. To this end, we adopt the cKL divergence, which directly characterizes the conditional dependence of the response variable on the predictors, i.e., shifts in $P(X)$ are irrelevant as long as they do not alter $P(Y|X)$. We provide the steps of the proposed transferable source data detection algorithm as follows. Firstly, we employ nonparametric methods such as Kernel Density Estimation (KDE) (Sheather and Jones, 1991) to estimate the conditional distribution from k-th source data and target data, denoted as $\widehat{P}^{(k)}(Y|X)$ and $\widehat{Q}(Y|X)$, $k = 1, \cdots, K$. Subsequently, we calculate the estimated cKL divergence of k-th source data:

$$\widehat{cKL}^{(k)} = \sum_{i=1}^{n_k} \widehat{P}^{(k)}(Y_i|X_i) \log\left\{\frac{\widehat{P}^{(k)}(Y_i|X_i)}{\widehat{Q}(Y_i|X_i)}\right\},$$

where (X_i, Y_i) are the samples of k-th source data, $\widehat{P}^{(k)}(Y|X)$ is the estimated conditional probability of the k-th source data and $\widehat{Q}(Y|X)$ is that of the target data. Finally, we compute all $\widehat{cKL}^{(k)}$ values, for $k = 1, \cdots, K$, and compare them against a pre-determined threshold D_0. Source datasets where $\widehat{cKL}^{(k)}$ falls below D_0 are selected into the estimated transferring set, $\widehat{\mathcal{A}} = \{k : \widehat{cKL}^{(k)} \leq D_0\}$. This set serves as an empirical approximation of the level-h transferring set $\mathcal{A}_h$.

Notably, the proposed algorithm for detecting transferable source data does not require the specification of h. As explained in Section 4.3, under certain conditions, it is established that $\widehat{\mathcal{A}} = \mathcal{A}_h$ for a designated h, which means that using $\widehat{\mathcal{A}}$ for the transferable set can also improve the accuracy of the prediction interval's lower and upper bounds, as well as its width. This is in contrast to intervals based solely on the target data, especially when the target sample size, n_0, falls within a certain range.

Algorithm 1 Transfer conformal prediction.

Input: Data $\mathcal{I}_1$, $\mathcal{I}_2$, $\mathcal{S}_1$, $\mathcal{S}_2$, a grid of potential values for $\Lambda(\mathcal{I}_1)$, new feature X_{new}, coverage level $1-\alpha$, $\delta \in (0,1]$, any algorithms $\mathcal{D}$, transferring set $\mathcal{A}$.

Use the Bootstrap method to generate T random copies of $\mathcal{I}_1$, each of which are denoted by $\mathcal{I}_1^t, t = 1, \ldots, T$. Let $\mathcal{B} = \{\mathcal{I}_1^t, t = 1, \ldots, T\}$ and $\mathcal{B}_{-t} = \mathcal{B} \backslash \mathcal{I}_1^t$.

for all possible values of $\Lambda(\mathcal{I}_1)$ **do**

  for $t = 1$ to T **do**

    Compute a preliminary mean estimator using equation (5):

$$\widetilde{\mu}_t = \mathcal{D}\left(\{(X_i, Y_i)\}\right), \ \ (X_i, Y_i) \in \mathcal{S}_1 \cup \mathcal{I}_1^t.$$

    Compute the corresponding fitted residuals: $\widetilde{R}_i^t = |Y_i - \widetilde{\mu}_t(X_i)|, \ (X_i, Y_i) \in \mathcal{S}_1$.

    Compute a preliminary weighted empirical $1-\alpha$ quantile of $\widetilde{R}_i^t$:

$$\widetilde{q}_{1-\alpha}^t = \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{|\mathcal{S}_1|+1} \widetilde{w}_i^t \cdot \delta_{\widetilde{R}_i^t}\right).$$

    Compute a debiased mean estimator using equation (6):

$$\widehat{\mu}_t = \mathcal{D}\left(\{(X_i, Y_i)\}; \widetilde{\mu}_t\right), \ \ (X_i, Y_i) \in \mathcal{I}_1^t.$$

    Compute the corresponding fitted residuals: $\widehat{R}_i^{-t} = |Y_i - \widehat{\mu}_t(X_i)|, \ \ (X_i, Y_i) \in \mathcal{B}_{-t}$.

    Compute the calibration-conditional empirical distribution function using equation (7):

$$\widehat{F}_t^{(\mathrm{ccv})}\left\{\widetilde{q}_{1-\alpha}^t + \Lambda(\mathcal{I}_1)\right\}.$$

  end for

end for

Compute the correction parameter using eqation (8):

$$\Lambda(\mathcal{I}_1) := \inf\left\{\Lambda(\mathcal{I}_1) : \widehat{F}_{(\lfloor T*0.05 \rfloor+1)}^{(\mathrm{ccv})}\left\{\widetilde{q}_{1-\alpha}^{(\lfloor T*0.05 \rfloor+1)} + \Lambda(\mathcal{I}_1)\right\} \geq 1-\alpha\right\}.$$

Compute a preliminary mean estimator: $\widetilde{\mu}_{\mathcal{I}_1} = \mathcal{D}\left(\{(X_i, Y_i)\}\right), \ \ (X_i, Y_i) \in \mathcal{S}_1 \cup \mathcal{I}_1$.

Compute the corresponding fitted residuals: $\widetilde{R}_i^{\mathcal{I}_1} = |Y_i - \widetilde{\mu}_{\mathcal{I}_1}(X_i)|, \ \ (X_i, Y_i) \in \mathcal{S}_1$.

Compute a preliminary weighted empirical $1-\alpha$ quantile of $\widetilde{R}_i^{\mathcal{I}_1}$:

$$\widetilde{q}_{1-\alpha}^{\mathcal{I}_1} = \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{|\mathcal{S}_1|+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right).$$

Compute a preliminary mean estimator: $\widetilde{\mu}_{\mathcal{I}_2} = \mathcal{D}\left(\{(X_i, Y_i)\}\right), \ \ (X_i, Y_i) \in \mathcal{S}_2 \cup \mathcal{I}_2$.

Compute a debiased mean estimator: $\widehat{\mu}_{\mathcal{I}_2} = \mathcal{D}\left(\{(X_i, Y_i)\}; \widetilde{\mu}_{\mathcal{I}_2}\right), \ \ (X_i, Y_i) \in \mathcal{I}_2$.

Compute the TCP interval using equation (9):

$$\widehat{C}(X_{new}) = \widehat{\mu}_{\mathcal{I}_2}(X_{new}) \pm \left[\widetilde{q}_{1-\alpha}^{\mathcal{I}_1} + \Lambda(\mathcal{I}_1)\right].$$

Output: $\widehat{C}(X_{new})$.

Remark 6 *We do not use the full KL divergence because it measures discrepancies in both the marginal covariate distribution $P(X)$ and the conditional distribution $P(Y \mid X)$. While this makes KL divergence suitable when both covariate and conditional shifts are of interest, in regression transfer problems, such generality can be misleading: it may cause the exclusion of informative sources only due to differences in feature distributions, even when the conditional structure is preserved. The cKL divergence addresses this limitation by isolating variations in $P(Y|X)$, making it more suitable for our transfer learning settings where accurate estimation of $\mu(X)$ is the primary goal and posterior drift is the main challenge. In particular, posterior drift, i.e., changes in $P(Y|X)$, directly undermines the estimation of $\mu(X)$, whereas covariate shift alters only $P(X)$ and leaves $\mu(X)$ unchanged. Consequently, cKL divergence serves as the correct selection criterion, as it identifies source datasets with similar conditional structures even when their covariate distributions differ substantially. To demonstrate this empirically, we conduct additional simulations under strong covariate shift with minimal posterior drift in Appendix D, where cKL consistently outperforms KL in identifying useful sources.*

Remark 7 *When transferability is limited by weak overlap in X, for example, due to large location and scale shifts, or by tail behavior that substantially affects predictive uncertainty, a purely conditional criterion may indeed become overly permissive. In this regime, predictive uncertainty depends critically on the joint behavior of (X, Y), specifically, on how much probability mass from a source lies outside the high-probability region of the target covariates and how frequently large tail events occur in the responses. As a result, sources that appear conditionally similar in local neighbourhoods may still be globally incompatible, leading to inflated calibration quantiles and wider prediction intervals after transfer. By contrast, a KL-based criterion on the joint distribution of (X, Y) penalizes both overlap failure in X and tail mismatch in Y more directly. These findings clarify that, while cKL is well suited to settings with adequate covariate overlap where conditional mismatch is the dominant source of heterogeneity, full KL divergence can be preferable when variance shifts, poor overlap, or heavy-tailed responses govern transfer performance. We include a dedicated simulation section to present this setting and to provide guidance on when KL-based detection can outperform cKL-based detection, as reported in the Appendix M.*

4. Theoretical properties

In this section, we present the theoretical results for the proposed TCP method. Specifically, we establish the coverage properties of general models when the information set is known and present three examples: high-dimensional linear, generalized linear models, and nonparametric regression, to demonstrate how TCP can outperform traditional conformal methods in Section 4.1. In Section 4.2, we discuss the conditional coverage properties of our TCP method. Section 4.3 introduces a theorem stating that the proposed transferable source detection algorithm can accurately identify the level-h transferring set $\mathcal{A}_h$ for a specified h. For proofs and further theoretical details, readers are referred to the Appendix.

4.1 Oracle TCP

In this section, we provide the coverage probability and interval width of the proposed TCP method, assuming the information set is known. Before proceeding, we introduce some notations and standard assumptions. Let $N = |\mathcal{S}_1| = |\mathcal{S}_2|$, and $n_0 = |\mathcal{I}_1| = |\mathcal{I}_1^t| = |\mathcal{I}_2|$. For theoretical purposes, we denote the population version of $\widetilde{\mu}_{\mathcal{I}_1}$, $\widetilde{\mu}_{\mathcal{I}_2}$, or $\widetilde{\mu}_t$ from Section 3 as $\mu_{\mathcal{A}_{h_0}}$, which combines μ_k for $k = 1, \ldots, K$. Correspondingly, the debiased regression estimators $\widehat{\mu}_{\mathcal{I}_1}$, $\widehat{\mu}_{\mathcal{I}_2}$, or $\widehat{\mu}_t$ are inferred as estimators of the population mean function of the target regression function μ_0. For simplicity, let $\widetilde{\mu}$ and $\widehat{\mu}$ respectively represent generic regression estimators from the sets $\widetilde{\mu}_{\mathcal{I}_1}, \widetilde{\mu}_{\mathcal{I}_2}, \widetilde{\mu}_t$ and $\widehat{\mu}_{\mathcal{I}_1}, \widehat{\mu}_{\mathcal{I}_2}, \widehat{\mu}_t$. Recall that $\mu_0(X^{(0)}) = \mathbb{E}\{Y^{(0)}|X = X^{(0)}\}$ and $\mu_k(X^{(k)}) = \mathbb{E}\{Y^{(k)}|X = X^{(k)}\}$ for the k-th source dataset.

Assumption 1 *The noise variables $\epsilon_i^{(0)} = Y_i^{(0)} - \mu_0(X_i^{(0)})$, $\epsilon_i^{(k)} = Y_i^{(k)} - \mu_0(X_i^{(k)})$, satisfy $\mathbb{E}\{\epsilon_i^{(0)}|X_i^{(0)}\} = 0$ and $\mathbb{E}\{\epsilon_i^{(k)}|X_i^{(k)}\} = 0$, for $k = 1, \ldots, K$ with bounded variance, respectively.*

Assumption 2 *$\|X^{(0)}\|_\infty$, $\|X^{(k)}\|_\infty$ for $k = 1, \ldots, K$ are bounded and we define a level-h_0 transferring set $\mathcal{A}_{h_0}$ as*

$$\mathcal{A}_{h_0} = \left\{k \in \{1, \ldots, K\} : \left\|\mu_k\left(X^{(k)}\right) - \mu_0\left(X^{(k)}\right)\right\|_\infty \leq h_0\right\}. \tag{11}$$

Let $\mu_{\mathcal{A}_{h_0}}$ denote the combined regression function over the transferring set $\mathcal{A}_{h_0}$, defined by

$$\mu_{\mathcal{A}_{h_0}}(x) := \mathbb{E}[\, Y \mid X = x, \ \text{data from } k \in \mathcal{A}_{h_0}]\,.$$

We assume that the same bound holds on the target support:

$$\left\|\mu_{\mathcal{A}_{h_0}}(X^{(0)}) - \mu_0(X^{(0)})\right\|_\infty \leq h_0.$$

Assumption 1 imposes less stringent conditions on the noise variables than those commonly found in the regression literature. This assumption is frequently employed in split conformal prediction; see (Vovk et al., 2009; Lei et al., 2018). Assumption 2 places specific conditions on the covariates and introduces the level-h_0 transferring set $\mathcal{A}_{h_0}$. This set encompasses source datasets whose conditional means μ_k are similar to the target dataset's conditional mean μ_0, specifically within a supremum norm difference of h_0. The rationale behind this transferring set definition instead of cKL divergence is that variations in conditional means often reflect differences in the underlying conditional probability distributions. This allows $\mathcal{A}_{h_0}$ to serve as a simplified and more structured alternative to $\mathcal{A}_h$ as specified in (3), facilitating easier analysis of the theoretical properties of the proposed TCP method.

Remark 8 *The transferring set at level h_0, denoted by $\mathcal{A}_{h_0}$ in* (11)*, varies depending on the specific regression models. For example, in high-dimensional transfer learning within linear regression, where $\mu_0(X_i^{(0)}) = (X_i^{(0)})^\top \boldsymbol{\beta}$ and $\mu_k(X_i^{(k)}) = (X_i^{(k)})^\top \boldsymbol{\theta}^{(k)}$, the level-$h_0$ transferring set reduces to $\mathcal{A}_{h_0} = \{k : \|\boldsymbol{\beta} - \boldsymbol{\theta}^{(k)}\|_1 \leq h_0\}$, as seen in Li et al. (2022). In the setting of high-dimensional generalized linear regression, this definition aligns with that described by Tian and Feng (2023), where $\mu_0(X_i^{(0)}) = G(X_i^{(0)})^\top \boldsymbol{\beta}$ and $\mu_k(X_i^{(k)}) = G(X_i^{(k)})^\top \boldsymbol{\theta}^{(k)}$, and G is a known link function. Additionally, the definition of* (11) *corresponds to the level-h_0 transferring set under a nonparametric regression setting in (Cai and Pu, 2022).*

Assumption 3 *(ℓ_∞-estimation error bound). There exist some sequences $\rho_{1,n} \to 0$ and $\eta_{1,n} \to 0$ such that $\Pr\{\|\widetilde{\mu} - \mu_{\mathcal{A}_{h_0}}\|_\infty \geq \eta_{1,n}\} \leq \rho_{1,n}$, as $N \to \infty$.*

Assumption 4 *(ℓ_∞-estimation error bound). There exist some sequences $\rho_{2,n} \to 0$ and $\eta_{2,n} \to 0$ such that $\Pr\{\|\widehat{\mu} - \mu_0\|_\infty \geq \eta_{2,n}\} \leq \rho_{2,n}$, as $N \to \infty$.*

Assumptions 3 and 4 ensure that the estimators of the mean regression function derived from the transfer and debiasing steps closely approximate their respective population versions. These are just sup-norm consistency assumptions, which are typically met by lasso-type estimators under standard assumptions, fixed-dimension ordinary least squares with bounded predictors, and standard nonparametric regression estimators on compact domains. Similar conditions can be found in Lei et al. (2018). Notably, when $\mathcal{A}_{h_0}$ is not empty and $n_0 + N \gg n_0$, the convergence rates $\eta_{1,n}$ and $\eta_{2,n}$ are expected to be sharper than those obtained through conventional non-transfer regression methods. These assumptions are supported by various studies in transfer learning regression estimators (Li et al., 2022; Cai and Pu, 2022; Tian and Feng, 2023), demonstrating the benefits of leveraging information from informative source datasets to enhance the estimation of the mean regression function in target models.

Theorem 9 (Lower and upper bounds on coverage). *Assume Assumptions 1-4 hold. Let $\eta_n = \eta_{1,n} + \eta_{2,n}$, $h' = h_0 + h$ and $\rho_n = \rho_{1,n} + \rho_{2,n}$. Then the TCP defined in* (9) *satisfies*

$$\Pr\left\{Y_{new} \in \widehat{C}\left(X_{new}\right)\right\} \geq 1 - \alpha - O_p\left(\eta_n + \rho_n + h' + \Lambda + N^{-1/2}\right),$$

and

$$\Pr\left\{Y_{new} \in \widehat{C}\left(X_{new}\right)\right\} \leq 1 - \alpha + \widetilde{w}_{N+1}^{\mathcal{I}_1} + O_p\left(\eta_n + \rho_n + h' + \Lambda + N^{-1/2}\right),$$

where Λ is the bias correction term defined in (4) *and $\widetilde{w}_{N+1}^{\mathcal{I}_1}$ denotes a prespecified weight placed to data point (X_{new}, Y_{new}), which has the order of .*

Theorem 9 establishes the lower and upper bounds on the coverage of the proposed TCP method, which applies to general conditional mean functions. When $\mathcal{A}_{h_0}$ is not empty, these bounds are determined by the error terms in the estimates of the conditional means, i.e., η_n, ρ_n, h', Λ, and the sample sizes of both the target and source datasets. Notice that $\widetilde{w}_{N+1}^{\mathcal{I}_1} = 1/(w_1^{\mathcal{I}_1} + \cdots + w_N^{\mathcal{I}_1} + 1)$, if $w_1^{\mathcal{I}_1}, \cdots, w_N^{\mathcal{I}_1}$ are all fix values, the order of $\widetilde{w}_{N+1}^{\mathcal{I}_1}$ is typically $O(1/N)$. Specifically, if h' and Λ are much smaller than $N^{-1/2}$, and if N is significantly larger than n_0, with η_n and ρ_n observed to be smaller than $N^{-1/2}$ (or, in the case of nonparametric regression, smaller than $N^{-\beta_P/(2\beta_P+1)}$, where β_P denotes the smoothness degree of the unknown source function—see Definition 1 in Cai and Pu (2022) for more details), it follows that the upper and lower bounds of the TCP method approach the optimal order of $N^{-1/2}$ or $N^{-\beta_P/(2\beta_P+1)}$. As more relevant source data are incorporated (that is, as N increases), the convergence rate of $N^{-1/2}$ or $N^{-\beta_P/(2\beta_P+1)}$ improves, improving the prediction of the target, particularly when the target data set is small. Compared to Theorem 2 of Barber et al. (2023), as described in equation (2), incorporating additional source data (increasing N) when the target data set is small yields a better lower coverage bound, bringing it closer to $1 - \alpha$. Specific examples of models and

their corresponding coverage bounds are provided in Section 4.2. When $\mathcal{A}_{h_0}$ is empty, the lower and upper bounds of $\Pr\{Y_{new} \in \widehat{C}(X_{new})\}$ are $1-\alpha$ and $1-\alpha+O(1/n_0)$, respectively, which is the same as in split conformal prediction using only target data (Lei et al., 2018).

Remark 10 *As noted previously, the integration of additional informative source data (increasing N) enhances predictive accuracy for smaller target dataset by improving the convergence rate to $N^{-1/2}$ or $N^{-\beta_P/(2\beta_P+1)}$. As a result, the coverage gap, $O_p(N^{-1/2})$ or $O_p(N^{-\beta_P/(2\beta_P+1)})$, becomes negligible, bringing the proposed TCP coverage closer to $1-\alpha$, especially when the source sample size N is large. Importantly, this enhancement is achieved without imposing asymptotic assumptions on the target sample size n_0. Although a strict finite sample guarantee is not obtained, the method still provides highly reliable coverage in practical applications. The key strength of this approach lies in its ability to achieve coverage near $1-\alpha$ without requiring a large target sample size n_0.*

The robustness of our method stems from its ability to ensure valid prediction coverage not only under distributional shifts but also in the presence of model misspecification. On the one hand, our framework is designed for nonexchangeable data in a posterior drift setting, where the conditional distributions $P(Y|X)$ for the source domains may differ from those of the target. Our method restores coverage guarantees by detecting the informative sources whose conditional distributions are sufficiently close to that of the target and constructing the prediction sets using only these informative sources. On the other hand, the validity of our approach does not depend on $\widetilde{\mu}(X)$ or $\widehat{\mu}(X)$ being the true regression functions. Assumptions 3 and 4 require only that the estimation errors in the ℓ_∞ norm converge to zero with high probability. Thus, for our theoretical guarantees to hold, the fitted models need only approximate the truth uniformly well, rather than be perfectly specified.

Theorem 11 (Width of TCP). *Assume those conditions in Theorem 9 hold. The width of the TCP band is*

$$W_{trconf} = 2Q_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + 2\Lambda\left(\mathcal{I}_1\right) = 2q_2 + O_p\left(\eta_{2,n} + \rho_{2,n}\right),$$

where q_2 is the population $1-\alpha$ quantile of the distribution of $|Y - \mu_0(X)|$.

Theorem 11 demonstrates that the width of the proposed TCP interval is determined by $\max\{\eta_{2,n}, \rho_{2,n}\}$. This represents a notable departure from the results only based on target data as presented in Theorem 3.2 of Lei et al. (2018). Specifically, when $n_0 = o(N)$, Theorem 11 suggests a significantly narrower prediction interval compared to traditional conformal prediction intervals, as documented in works by (Lei et al., 2018; Tibshirani et al., 2019; Guan, 2023; Barber et al., 2023). This reduction in interval width is largely due to the fact that both $\eta_{2,n}$ and $\rho_{2,n}$ benefit from the increased effective sample size that encompasses the entire data set, not just the target data set.

Remark 12 *An intuitive alternative approach would involve leveraging transfer learning techniques during the training phase, followed by standard conformal calibration using the available target data. Although this method retains finite-sample guarantees, it typically*

produces overly wide prediction intervals when applied to limited target data, as discussed after Theorem 11. In particular, the simulation study and the analysis of real data in Sections 5-6 highlight the limitations of this intuitive approach, while also validating the theoretical advantages of our proposed method.

Remark 13 *It is essential to acknowledge the inherent trade-off between coverage guarantees and prediction interval width in conformal methods (Barber et al., 2021). Traditional approaches, such as split and full conformal prediction, which only depend on target data, typically provide strong coverage guarantees, with coverage errors bounded at a rate of n_0^{-1} (Barber et al., 2023). However, these methods often lack statistical efficiency, resulting in unnecessarily wide intervals. In contrast, our proposed TCP method capitalizes on valuable source data, effectively increasing the sample size to N, which is significantly larger than n_0 (Hu and Zhang, 2023). This enhanced sample size allows our method to produce substantially narrower prediction intervals compared to those generated by the split conformal method, which relies exclusively on target data.*

Remark 14 *Both our TCP framework and the work of Oliveira et al. (2024) aim to relax the classical exchangeability assumption, but they address fundamentally different regimes. Our TCP framework is designed for multi-source transfer learning under posterior drift, where coverage validity and efficiency are guaranteed by the aggregate source sample size N. In this setting, the bounds improve at rates determined by N, which is typically much larger than the target size n_0, yielding faster convergence and near-nominal coverage even when only limited target data are available. By contrast, Oliveira et al. (2024) demonstrated that split conformal prediction remains approximately valid for dependent data (e.g., β-mixing or spatiotemporal processes), but with a coverage penalty that scales directly with n_0. Hence, our problem setup extends to scenarios that fall beyond the scope of Oliveira et al. (2024).*

From a theoretical perspective, the coverage gap in our TCP framework arises because the usual finite-sample validity of conformal prediction critically depends on the exchangeability of the calibration and test residuals. In our transfer learning setting, the calibration residuals are drawn from informative but non-exchangeable source datasets, while the target residual distribution differs due to posterior drift. This lack of exchangeability breaks the finite-sample exactness of conformal prediction and necessarily introduces a coverage gap. Similar limitations have been noted in prior work on non-exchangeable settings, such as Gibbs and Candes (2021), Barber et al. (2023), and more recently Oliveira et al. (2024), where coverage can only be established approximately under dependence.

In our TCP framework, this gap is controlled through error terms that depend primarily on the aggregate source sample size N. As established in Theorems 9, and 11, the gap shrinks at rates $O_p(N^{-1/2})$ or $O_p(N^{-\beta_P/(2\beta_P+1)})$, depending on the regularity assumptions. This implies that with sufficiently large and informative source datasets, the coverage approaches the nominal level $1-\alpha$ asymptotically. Importantly, this convergence occurs without requiring asymptotic assumptions on the target sample size n_0, which distinguishes our results from those of Oliveira et al. (2024), whose guarantees depend directly on n_0. If the source and target distributions are exchangeable (i.e., no posterior drift), our method can achieve exact finite-sample validity $1-\alpha$. More generally, when posterior drift is mild

and the source sample size N is large, the asymptotic guarantees provide coverage very close to $1-\alpha$, as confirmed by our simulations.

We provide three specific statistical models and examine the corresponding coverage bounds and prediction interval widths for the proposed TCP method. These models include high-dimensional linear regression, high-dimensional generalized linear regression, and nonparametric regression. For a comprehensive discussion of the three models and detailed explanations of the parameters $\eta_{1,n}$, $\eta_{2,n}$, $\rho_{1,n}$, $\rho_{2,n}$, h_0, h, and Λ in the context of these regression models, please refer to Section A in the Appendix. Table 1 provides a summary of the coverage guarantees and prediction interval widths for each of the three regression models under the proposed TCP framework.

Table 1: Comparative summary of optimal order for coverage guarantees and prediction interval widths.s is the number of nonzero coefficients under the high-dimensional setting, c_1, c_2, c_3, c_4 are some positive constants relating to h, and β_P is the smoothness of the nonparametric function under the nonparametric setting.

	High-dimensional LM	High-dimensional GLM	Nonparametric
$\eta_{1,n}$	$s\left(\frac{\log(p)}{n_0+N}\right)^{1/2}$	$s\left(\frac{\log(p)}{n_0+N}\right)^{1/2}$	$(n_0+N)^{-\frac{\beta_P}{2\beta_P+1}}$
$\eta_{2,n}$	$s\left(\frac{\log(p)}{n_0+N}\right)^{1/2}+h$	$s\left(\frac{\log(p)}{n_0+N}\right)^{1/2}+h$	$(n_0+N)^{-\frac{\beta_P}{2\beta_P+1}}$
$\rho_{1,n}$	p^{-c_1}	p^{-c_3}	$(n_0+N)^{-\frac{\beta_P}{2\beta_P+1}}$
$\rho_{2,n}$	p^{-c_2}	p^{-c_4}	$(n_0+N)^{-\frac{\beta_P}{2\beta_P+1}}$
h_0	$N^{-1/2}$	$N^{-1/2}$	$N^{-1/2}$
h	$N^{-1/2}$	$N^{-1/2}$	$N^{-1/2}$
Λ	$N^{-1/2}$	$N^{-1/2}$	$N^{-1/2}$
Lower bound	$N^{-1/2}$	$N^{-1/2}$	$(n_0+N)^{-\frac{\beta_P}{2\beta_P+1}}$
Upper bound	$N^{-1/2}$	$N^{-1/2}$	$(n_0+N)^{-\frac{\beta_P}{2\beta_P+1}}$
Width	$\max\{\eta_{2,n},\rho_{2,n}\}$	$\max\{\eta_{2,n},\rho_{2,n}\}$	$(n_0+N)^{-\frac{\beta_P}{2\beta_P+1}}$

4.2 Conditional coverage guarantee

An important direction is to investigate the conditional coverage properties of our TCP method. This is particularly important in applications where uncertainty quantification must be reliable at specific covariate values X_{new}, rather than only on average across the covariate distribution. Examples include patient-specific risk prediction in medicine or localized demand forecasting in economics, where relying solely on marginal guarantees may obscure poor coverage in regions of low or high covariate density. Conditional coverage guarantee is strictly stronger than marginal coverage and, as noted in Lei et al. (2018) and Barber et al. (2021), is generally unattainable without additional structural assumptions or by relaxing finite-sample validity. To establish the conditional coverage properties of TCP, we strengthen the estimation assumptions so that they hold locally around X_{new}. In

particular, the global ℓ_∞ bounds in Assumptions 3 and 4 must be replaced by conditional ℓ_∞ error bounds that quantify the estimation accuracy of $\widetilde{\mu}$ and $\widehat{\mu}$ given X_{new}. This refinement parallels the approach taken in conditional conformal prediction (see, e.g. Guan (2023), Oliveira et al. (2024), and Hore and Barber (2025)), which extends marginal to conditional guarantees by imposing additional covariate-conditional assumptions. In the following, we present the conditional versions of these estimation assumptions.

Assumption 5 (ℓ_∞-estimation error bound, conditional on X_{new}) *For any fixed covariate value X_{new}, there exist sequences $\rho_{3,n} \to 0$ and $\eta_{3,n} \to 0$ such that*

$$\Pr\left(\|\widetilde{\mu} - \mu_{\mathcal{A}_{h_0}}\|_\infty \geq \eta_{3,n} | X_{new}\right) \leq \rho_{3,n}, \quad \text{as } N \to \infty.$$

Assumption 6 (ℓ_∞-estimation error bound, conditional on X_{new}) *For any fixed covariate value X_{new}, there exist sequences $\rho_{4,n} \to 0$ and $\eta_{4,n} \to 0$ such that*

$$\Pr\left(\|\widehat{\mu} - \mu_0\|_\infty \geq \eta_{4,n} | X_{new}\right) \leq \rho_{4,n}, \quad \text{as } N \to \infty.$$

Under these two assumptions, we establish the following result:

Theorem 15 (Lower and upper bounds on conditional coverage). *Assume Assumptions 1, 2, 5 and 6 hold. Let $\eta_n^c = \eta_{3,n} + \eta_{4,n}$, $h' = h_0 + h$ and $\rho_n^c = \rho_{3,n} + \rho_{4,n}$. Then the TCP satisfies*

$$\Pr\left\{Y_{new} \in \widehat{C}\left(X_{new}\right) | X_{new}\right\} \geq 1 - \alpha - O_p\left\{\eta_n^c + \rho_n^c + \eta_n + \rho_{2,n} + \Lambda + h' + N^{-1/2}\right\},$$

and

$$\Pr\left\{Y_{new} \in \widehat{C}\left(X_{new}\right) | X_{new}\right\} \leq 1 - \alpha + \widetilde{w}_{N+1}^{\mathcal{I}_1} + O_p\left\{\eta_n^c + \rho_n^c + \eta_n + \rho_{2,n} + \Lambda + h' + N^{-1/2}\right\}.$$

This theorem shows that, under the strengthened local assumptions, TCP achieves conditional coverage guarantees up to small residual error terms. The bounds depend on the local estimation accuracy of $\widehat{\mu}$ and $\widetilde{\mu}$ in a neighbourhood of X_{new}, rather than on global model correctness or exchangeability across domains. In other words, the procedure can remain valid at a fixed covariate value even when the overall model is misspecified, so long as the fitted regressions are sufficiently accurate locally. To support this theoretical claim, we conduct a simulation study that examines conditional coverage at a fixed covariate value in Appendix J. The results confirm that TCP remains robust in practice at specific regions of the covariate space.

In addition to conditional coverage at a fixed covariate value X_{new}, it is natural to study *group-conditional coverage*, as emphasized in Hore and Barber (2025). Exact test-conditional guarantees are known to be unattainable in a fully distribution-free setting (except for discrete X), whereas marginal guarantees are often too weak for practical applications. Group-conditional coverage offers a natural compromise: instead of requiring validity at a single point X_{new}, it requires validity averaged over subsets of the covariate space (Hore and Barber, 2025). Formally, let $\mathcal{B}$ denote a collection of measurable subsets of the feature space $\mathcal{X}$ (e.g., age intervals in biomedical applications, income brackets in

economics, or quantile bins of a covariate). The group-conditional requirement is that, for every $B \in \mathcal{B}$,

$$\Pr\{Y_{new} \in \widehat{C}(X_{new}) \mid X_{new} \in B\} \geq 1 - \alpha.$$

This guarantee is weaker than pointwise conditional coverage, because each B aggregates outcomes across many covariates rather than requiring validity at every individual X_{new}.

Our theoretical framework already establishes bounds for conditional coverage at a fixed X_{new} (Theorem 15). However, extending these results to group-conditional coverage is non-trivial. In particular, pointwise convergence at each fixed X_{new} does not automatically imply uniform validity over subsets $B \subseteq \mathcal{X}$, especially when B contains infinitely many covariate values. Obtaining group-conditional guarantees would require stronger assumptions to control uniform estimation error across subsets of the covariate space, and we leave the development of such conditions as an important direction for future work. To support this perspective empirically, we have conducted additional simulations where the covariate space is partitioned into meaningful subgroups $B \in \mathcal{B}$ (e.g., bins defined by the range of a selected covariate). We then evaluated TCP's empirical coverage within each subgroup. The results, presented in Appendix K, confirm that TCP maintains near-nominal coverage in every subgroup while continuing to produce narrower intervals than baseline methods.

4.3 Source detection consistency

In this section, we demonstrate that our transferable source data detection algorithm can accurately recover the level-h transferring set $\mathcal{A}_h$ for a specified h, with high probability, given certain conditions. To achieve this goal, we first define the population counterpart of $\widehat{cKL}$, which is given by:

$$cKL^{(k)} = \sum_{i=1}^{n_k} P^{(k)}\left(Y_i|X_i\right) \log\left\{\frac{P^{(k)}\left(Y_i|X_i\right)}{Q\left(Y_i|X_i\right)}\right\},$$

where $P^{(k)}\left(Y|X\right)$ denotes the population conditional probability associated with the k-th source dataset and $Q\left(Y|X\right)$ represents that of the target dataset. In addition, we need to introduce the following conditions on the identifiability of a specific $\mathcal{A}_h$.

Assumption 7 *There exist some sequences $h_p^{(k)} \to 0$ and $h_q \to 0$ such that*

$$\Pr\left(\sup_i \left|\widehat{P}^{(k)}\left(Y_i|X_i\right) - P^{(k)}\left(Y_i|X_i\right)\right| > \kappa h_p^{(k)}\right) \leq g_1^{(k)}(\kappa),$$
$$\Pr\left(\sup_i \left|\widehat{Q}\left(Y_i|X_i\right) - Q\left(Y_i|X_i\right)\right| > \kappa h_q\right) \leq g_2(\kappa),$$

where $g_1^{(k)}(\kappa), g_2(\kappa) \to 0$ as $\kappa \to \infty$.

Assumption 8 *Let $\mathcal{A}_h^c = \{1, \ldots, K\} \backslash \mathcal{A}_h$. Suppose that there exist constants l_3, u_3, l_4, u_4 such that $0 < l_3 \leq u_3 < 1$, $0 < l_4 \leq u_4 < 1$, $l_3 \leq P^{(k)}(Y|X) \leq u_3$ for all k, and $l_4 \leq$*

$Q(Y|X) \leq u_4$. *Furthermore, assume that* $h = o(1)$, $\underline{h} \geq \frac{(u_3+\kappa h_p^{(k)})(h+\kappa h_p^{(k)}+\kappa h_q)}{\min\{l_3,l_4\}(l_3-\kappa h_p^{(k)})} + \kappa h_p^{(k)} + \kappa h_q$, *and*

$$\sup_i \left| P^{(k)}\left(Y_i|X_i\right) - Q\left(Y_i|X_i\right) \right| \leq h, \forall k \in \mathcal{A}_h, \tag{12}$$

$$\inf_i \left| P^{(k)}\left(Y_i|X_i\right) - Q\left(Y_i|X_i\right) \right| \geq \underline{h}, \forall k \in \mathcal{A}_h^c. \tag{13}$$

Assumption 7 guarantees that as sample sizes increase, the maximum discrepancies represented by $\sup_i |\widehat{P}^{(k)}(Y_i|X_i) - P^{(k)}(Y_i|X_i)|$ and $\sup_i |\widehat{Q}(Y_i|X_i) - Q(Y_i|X_i)|$ consistently decrease. Assumptions 7 and 8 enhance the framework of $\mathcal{A}_h$ identifiability, introduced in Assumption 5 of Tian and Feng (2023), which primarily addresses the stability of coefficient estimates across source and target data within generalized linear models. Notice that, under the assumptions (12) and (13), a significant difference is observed between the population-level conditional distributions of Y given X from the source data and those associated with the target data, particularly when the source data falls outside the $\mathcal{A}_h$ transferring set. Our contribution extends their framework to the analysis of conditional distributions, specifically $P^{(k)}(Y|X)$ and $Q(Y|X)$. This extension is critical for conformal prediction, as it ensures compatibility with the model-free nature of conformal prediction and enhances its applicability across various statistical settings.

Theorem 16 (Detection consistency of $\mathcal{A}_h$). *Assume Assumptions 7 and 8 hold. For any* $\xi > 0$, *there exists* $N(\xi) > 0$ *such that when* $\min n_k > N(\xi)$,

$$\Pr\left(\widehat{\mathcal{A}} = \mathcal{A}_h\right) \geq 1 - \xi.$$

In addition, under the assumptions outlined in Theorem 9, the proposed detection algorithm exhibits the same high-probability upper and lower bounds of coverage probability as those specified in Theorem 9.

Theorem 16 establishes that, under certain conditions, $\widehat{\mathcal{A}}$ can be equated with $\mathcal{A}_h$ for a particular value of h. A crucial implication of this theorem is that our algorithm for detecting transferability works effectively without an explicit specification of h. When the source sample size is sufficiently large, employing $\widehat{\mathcal{A}}$ for transfer not only leads to a reduction in both the lower and upper bounds of the prediction interval but also narrows the interval width, in contrast to approaches that construct the prediction interval only based on the target data.

Remark 17 *In practice, the constants* $h_p^{(k)}$ *and* h_q *in Assumptions 7 and 8 depend on the estimation procedures used to compute the conditional densities* $\widehat{P}^{(k)}(Y|X)$ *and* $\widehat{Q}(Y|X)$. *Common density estimation techniques, such as kernel density estimation (KDE) or* $K-$*nearest neighbour methods (KNN), produce reliable estimates, especially when the underlying data distributions are sufficiently smooth. The choice of bandwidth (in KDE) or the number of neighbours (in KNN) plays a critical role in controlling the size of* $h_p^{(k)}$ *and* $h_{q'}$, *allowing them to decrease at rates that align with the assumptions in Theorem 16. For practical values of* h, *cross-validation methods can be employed to tune* h *in a way that balances detection consistency with estimation accuracy. Small values of* h *lead to good detection consistency, especially when large enough source datasets are used.*

Remark 18 *Our theoretical results (Theorems 9, 11, and 16) show that both coverage validity and interval width are primarily determined by the total number of informative source samples, $N = \sum_{k=1}^{K} n_k$. As long as N is sufficiently large, the coverage gap of our conformal prediction procedure decreases at the rate $O_p(N^{-1/2})$ or $O_p(N^{-\beta_P/(2\beta_P+1)})$, depending on the regularity conditions. At the same time, a larger N improves estimation accuracy, yielding narrower prediction intervals. From this perspective, the relative contributions of K and n_k matter only through their effect on N: both increasing the number of sources and enlarging the sample size within each source improve efficiency by expanding the overall pool of informative source data. In practice, however, increasing K can also diversify the environments represented in the data and is only beneficial if the additional sites are sufficiently informative. Thus, in theory, efficiency depends solely on N, while in applications, both K and n_k influence performance through their combined effect on N and the informativeness of the included sources.*

5. Simulation studies

In this section, we illustrate the empirical performance of the proposed TCP algorithms in several simulation settings. Specifically, we demonstrate the numerical performance of the proposed TCP method with known and unknown informative auxiliary set $\mathcal{A}_h$ in Section 5.1 and 5.3, respectively. The sensitivity of the discrepancy parameter δ for the proposed algorithms is discussed in Section 5.2. All results displayed average over 200 independent Monte Carlo trials.

5.1 Simulation with known $\mathcal{A}_h$

In this subsection, we investigate the empirical properties of the conformal prediction intervals under three simulated data settings when $\mathcal{A}_h$ is known. In each setting, the target samples $(X_i^{(0)}, Y_i^{(0)})$ along with auxiliary source samples $(X_i^{(k)}, Y_i^{(k)})$, $i = 1, \ldots, n_k$. $k = 1, \ldots, K$ are generated in an i.i.d. fashion, by first specifying $\mu_k(x^{(k)}) = \mathbb{E}\{Y_i^{(k)} | X_i^{(k)} = x^{(k)}\}$, then specifying a distribution for $X_i^{(k)}$, and lastly specifying a distribution for $\epsilon_i^{(k)} = Y_i^{(k)} - \mu(X_i^{(k)})$, $k = 0, \ldots, K$. These specifications are described below.

Setting 1 *(linear regression, high-dimensional):* Consider the simulation setting as follows: we take the target mean function $\mu_0(\boldsymbol{x}^{(0)}) = \boldsymbol{x}^{(0)\top}\boldsymbol{\beta}$ and source mean function $\mu_k(\boldsymbol{x}^{(k)}) = \boldsymbol{x}^{(k)\top}\boldsymbol{\theta}^{(k)}$, $k = 1, \ldots, K$. The predictors from target study $\boldsymbol{x}^{(0)} \overset{i.i.d.}{\sim} N(\mathbf{0}_p, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (0.5^{|i-j|})_{1 \le i,j \le p}$ and the predictors from source studies $\boldsymbol{x}^{(k)} \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma} + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top)$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_p, 0.3^2\boldsymbol{I}_p)$. For the target data, the coefficient is set to be $\boldsymbol{\beta} = (0.5 \cdot \mathbf{1}_s, \mathbf{0}_{p-s})^\top$, where $\mathbf{1}_s$ is has all s elements 1, $\mathbf{0}_{p-s}$ is a zero vector, and s is set to be 5. Let $\boldsymbol{r}_p^{(k)}$ be p independent Rademacher variables (taking values in {1, -1} with equal probability) for any k. Notice that $\boldsymbol{r}_p^{(k)}$ is independent with $\boldsymbol{r}_p^{(k')}$ for any $k \neq k'$. For any source data k in $\mathcal{A}_h$, we set $\boldsymbol{\theta}^{(k)} = \boldsymbol{\beta} + (h/p)\boldsymbol{r}_p^{(k)}$.

Setting 2 *(poisson regression, high-dimensional):* Same to Setting 1, but where the target mean function takes $\mu_0(\boldsymbol{x}^{(0)}) = \exp(\boldsymbol{x}^{(0)\top}\boldsymbol{\beta})$, the source mean function takes $\mu_k(\boldsymbol{x}^{(k)}) = \exp(\boldsymbol{x}^{(k)\top}\boldsymbol{\theta}^{(k)})$, $k = 1, \ldots, K$, and predictors are coordinate-wise truncation at ± 0.5. In this setting, we explore target sample sizes $n_0 = 75, 100, 150, 200$.

Setting 3 *(nonparametric regression):* We set the dimension to 1 and consider the target and source mean functions as follows:

$$\mu_0(x^{(0)}) = \sin(10\pi x^{(0)}) + (x^{(0)})^{3/2},$$
$$\mu_k(x^{(k)}) = \sin(10\pi x^{(k)}) + (x^{(k)})^{3/2} - \frac{h}{500}x^{(k)} + \zeta, \quad k = 1, \ldots, K,$$

where $\zeta \sim \mathcal{N}(0,1)$, and $x^{(k)} \sim \text{Unif}(0,10), k = 0, 1, \ldots, K$. The sample sizes of the target study are $n_0 = 25, 50, 100, 200$, and the sample size of source studies is fixed at $n_1, \ldots, n_K = 100$ with $K = 5$. In this setting, we utilize B-spline regression to investigate the performance of the proposed algorithms. The remaining setup is the same as in Setting 1.

We consider various target sample sizes, specifically $n_0 = 25, 50, 100, 200$ and $K = 5$ source studies with sample $n_1, \ldots, n_K = 100$. The dimension $p = 500$ for both target and source data. The transferring level h is considered at two levels, $h = 5, 30$, and the discrepancy parameter δ is set to 0.1. The number of bootstraps is fixed at 50, each with an equal sample size, denoted as $\left|\mathcal{I}_1^t\right| = |\mathcal{I}_1|$. For each setting, we evaluate the following methods, targeting a coverage level of $1 - \alpha = 0.9$. Specifically,

- **SCP**: We consider the original split conformal prediction, with $\widehat{\mu}$ the lasso regression fit using only data from the target study.
- **FS**: Applying the few-shot conformal prediction method (Fisch et al., 2021), we select a subset of the source data, $(X_i^{(k_1)}, Y_i^{(k_1)})$, $i = 1, \ldots, n_{k_1}$, $k_1 = 1, 2$, to train a transfer learning algorithm (Li et al., 2022) to fit $\widehat{\mu}$. Subsequently, we use the data from the last three sources, $(X_i^{(k_2)}, Y_i^{(k_2)})$, $i = 1, \ldots, n_{k_2}$, $k_2 = 3, 4, 5$ to estimate correction parameter $\Lambda(\mathcal{I}_1)$ without employing weights.
- **TL**: Analogous to the SCP method, we employ the transfer learning algorithm (Li et al., 2022) to fit $\widehat{\mu}$.
- **TL+UQ**: We run the proposed TCP algorithm without using weights, i.e., $\widetilde{w}_i^{\mathcal{I}_1} \equiv 1$ and $\widetilde{w}_i^t \equiv 1$, $t = 1, \ldots, T$ defined in Algorithm 1.
- **TL+WQ**: We also run the proposed TCP algorithm with unequal weights. The details for selecting weights $\widetilde{w}_i^{\mathcal{I}_1}$ and $\widetilde{w}_i^t$, for all three settings, can be found in Section B of the Appendix.

Our results are summarized in Tables 2-4 and Figure 3. In terms of coverage, we observe that the proposed methods, i.e., TL+UQ and TL+WQ, have coverage $\approx 90\%$ across all settings with $h = 5$, while for $h = 30$, the proposed TL+WQ performs better in maintaining the desired coverage level due to significantly non-exchangeable data setting. Meanwhile, as expected, both the SCP and TL methods exhibit poor performance in most cases – this is because of inaccuracies in estimating the regression function and errors in determining the $1 - \alpha$ quantile of the empirical distribution of the fitted residuals, especially when the target sample size n_0 is limited. Turning to the prediction interval width, we observe that for $h = 5$, TL+UQ and TL+WQ methods show similar mean widths. This is because a smaller h leads to a less distinct difference between source and target datasets. In addition,

variability is higher for TL+UQ than for TL+WQ in most cases. Conversely, when $h = 30$, we see that the TL+UQ method consistently yields wider prediction intervals than the TL+WQ method, which is to be expected due to the greater degree of nonexchangeability in the source data caused by the larger h value. Moreover, the SCP and TL methods perform poorly with a limited target sample size n_0 across all settings, although their performance improves as n_0 increases. This highlights the utility of the proposed algorithms for settings where the target data are limited. Furthermore, the FS method presupposes that source and target data are exchangeable. However, this assumption does not hold in our setting, resulting in poorer performance of the FS method in terms of coverage probability and interval width. Even in scenarios where the non-exchangeability between source and target data is minimal (e.g., when $h = 5$, the FS method still underperforms. This is primarily because it relies on source data to estimate the correction parameter $\Lambda(\mathcal{I}_1)$, and the number of source data is limited.

Remark 19 *While our theoretical guarantees are asymptotic, the simulation results show that the empirical coverage remains close to the nominal level $1-\alpha$, even when the target sample size is limited, and is comparable to that of classical conformal prediction. In addition, TCP consistently yields substantially narrower prediction intervals than conventional conformal prediction, thereby demonstrating clear efficiency gains while effectively preserving coverage at the nominal level.*

Additionaly, in Appendix F and G, we present simulation studies under Conformalized Quantile Regression (CQR) scores (see Romano et al. (2019)) and under binary outcomes, respectively. For the CQR setting, we compared five methods: CQR, Few Short learning under quantile regression conformity scores (FSQR), Transfer only under quantile regression conformity scores (TLQR), and our Transfer Learning under quantile regression conformity scores procedures with unweighted (TCQR+UQ) and weighted (TCQR+WQ) quantile calibration. As shown in Tables 12-13, TCQR+WQ achieves the smallest prediction set size while maintaining valid coverage. For example, at $n_0 = 50$, TCQR+WQ yields a set size of 4.973 with 89% coverage, outperforming CQR, which has a larger set size of 5.844. For the binary outcome setting, we compared five methods: SCP, FS, TL, TL+UQ and TL+WQ. As shown in Tables 14- 15, TL+WQ achieves the smallest prediction set size while maintaining valid coverage. For example, at $n_0 = 100$, TL+WQ yields a set size of 1.460 with 89.5% coverage, outperforming SCP, which has a larger set size of 1.970. These results demonstrate that our TCP method remains both effective and efficient under CQR scores and under binary outcomes, highlighting its adaptability to diverse prediction problems.

Instead of residual-based nonconformity scores, we consider an HPD-type score constructed from an estimated conditional density $\widehat{f}(,\cdot \mid x)$. Let (X, Y) denote a generic covariate–response pair from the target domain, and let (x, y) represent a covariate value and response value. We define the HPD-based nonconformity score as:

$$\widehat{s}_{\mathrm{HPD}}(x, y) := \widehat{\mathbb{P}}\left(\widehat{f}(Y \mid X) \geq \widehat{f}(y \mid x) \,\middle|\, X = x\right),$$

which corresponds to the estimated conditional probability mass assigned to outcomes whose estimated density is at least as large as that of y at the same covariate value x. Given target calibration data $\{(X_i, Y_i)\}_{i \in \mathcal{I}_{\mathrm{cal}}}$, we compute the calibration scores $\widehat{s}_{\mathrm{HPD}}(X_i, Y_i)$ and let $\widehat{q}_{1-\alpha}$

denote their $(1-\alpha)$ empirical quantile. The resulting split conformal prediction set is:

$$\widehat{C}_{\mathrm{HPD}}(x) = \{y : \widehat{s}_{\mathrm{HPD}}(x, y) \leq \widehat{q}_{1-\alpha}\}.$$

To empirically illustrate this extension, we have conducted additional simulation studies using the HPD-based nonconformity score and report the corresponding empirical coverage and interval widths alongside our existing baselines in Appendix L. Tables 24 and 25 show that our qualitative conclusions remain unchanged under this alternative choice of nonconformity measure. These additional experiments further demonstrate the flexibility of the proposed TCP framework, which naturally accommodates density-based nonconformity scores derived from HDP regions. This extension is motivated by recent work connecting conformal prediction to HPD regions and distribution-free uncertainty quantification (Izbicki et al., 2022; Dheur et al., 2024). Unlike residual- or quantile-based scores, HPD-based nonconformity measures exploit an estimated conditional density to assess the plausibility of a candidate response under the fitted predictive distribution.

We also conduct additional simulation studies under CATE settings. While the average treatment effect (ATE) is a well-established causal effect of interest, our framework, grounded in conformal prediction, is primarily designed to provide predictive inference at the individual level. Specifically, conformal methods construct valid prediction intervals for each test instance, thereby quantifying uncertainty for individual-level outcomes rather than population-level averages. Consequently, our framework aligns more naturally with the CATE (Lei and Candès, 2021). In this case, the target quantity is

$$\tau(x) = \mu^{(0)}_{treatment}(x) - \mu^{(0)}_{control}(x),$$

where $\mu^{(0)}_{treatment}(x) = \mathbb{E}[Y(1)|X = x]$ and $\mu^{(0)}_{control}(x) = \mathbb{E}[Y(0)|X = x]$ denote the conditional mean outcomes under treatment and control in the target population. To apply TCP, we proceed analogously to Algorithm 1 in our paper:

- Fit outcome regression models for treatment and control using both target and source data, obtaining $\widehat{\mu}_{treatment}(x)$ and $\widehat{\mu}_{control}(x)$, along with preliminary transfer-learned estimators $\widetilde{\mu}_{treatment}(x)$ and $\widetilde{\mu}_{control}(x)$ based on source data and target data.
- Estimate predicted treatment effects as $\widehat{\tau}(x) = \widehat{\mu}_{treatment}(x) - \widehat{\mu}_{control}(x)$, with a corresponding transfer-learned version $\widetilde{\tau}(x) = \widetilde{\mu}_{treatment}(x) - \widetilde{\mu}_{control}(x)$.
- Construct conformity scores using the residuals of $\tau(x)$ and apply the same debiasing and weighting steps as in TCP.

This yields prediction intervals for $\tau(x)$ that retain the validity and efficiency properties of TCP in the causal setting. As shown in Tables 18 and 19, our proposed TCP perform better than other methods. Details of the CATE data-generating process and results are provided in Appendix I.

5.2 Sensitivity of δ

In this subsection, we present additional simulation results to assess the sensitivity of δ that appeared in Algorithm 1, which affects the computation calibration conditional

Table 2: Comparison of prediction interval width in Setting 1. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

h	n_0	SCP	FS	TL	TL+UQ	TL+WQ
5	25	7.827(0.819)	4.944(0.562)	4.503(0.528)	4.071(0.381)	4.051(0.349)
	50	6.298(0.613)	4.529(0.411)	4.188(0.394)	3.471(0.261)	3.456(0.260)
	100	4.655(0.352)	3.909(0.319)	3.679(0.215)	3.392(0.182)	3.351(0.182)
	200	3.972(0.195)	3.758(0.248)	3.547(0.151)	3.215(0.129)	3.226(0.126)
30	25	8.094(0.881)	5.423(0.767)	4.774(0.579)	4.645(0.366)	4.482(0.427)
	50	6.379(0.655)	5.012(0.494)	4.299(0.371)	4.161(0.214)	3.928(0.269)
	100	4.603(0.343)	4.237(0.362)	3.823(0.226)	3.982(0.168)	3.619(0.176)
	200	3.908(0.193)	3.828(0.334)	3.595(0.160)	3.839(0.147)	3.407(0.130)

Table 3: Comparison of prediction interval width in Setting 2. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

h	n_0	SCP	FS	TL	TL+UQ	TL+WQ
5	75	4.999(0.675)	4.852(0.591)	4.604(0.542)	3.580(0.239)	3.462(0.196)
	100	4.174(0.495)	4.211(0.475)	3.913(0.361)	3.497(0.217)	3.396(0.189)
	150	4.096(0.432)	4.141(0.396)	3.851(0.321)	3.388(0.198)	3.312(0.173)
	200	3.846(0.315)	3.963(0.402)	3.699(0.229)	3.257(0.143)	3.218(0.141)
30	75	4.741(0.627)	5.010(0.690)	4.561(0.524)	3.865(0.241)	3.625(0.242)
	100	4.298(0.535)	4.506(0.518)	4.108(0.399)	3.730(0.238)	3.551(0.226)
	150	4.092(0.385)	4.202(0.548)	3.872(0.299)	3.649(0.194)	3.411(0.191)
	200	3.899(0.328)	4.152(0.536)	3.824(0.235)	3.565(0.187)	3.292(0.160)

conformal p-values. To achieve this, we specifically consider a sequence of δ values, i.e., $\delta \in \{0.01, 0.03, 0.05, 0.10, 0.15,\ 0.17, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$. For each δ value, we run Algorithm 1 with $h = 30$ in Setting 1 and repeat this procedure 200 times.

As depicted in Figure 4, both the coverage probability and interval width of predictive intervals exhibit uniform behavior across varying values of δ. Consequently, it should be noted that one has the flexibility to select any appropriate δ value tailored to their specific needs.

5.3 Simulation with unknown $\mathcal{A}_h$

In this subsection, we investigate the empirical properties of the conformal prediction intervals via the following simulated data settings when $\mathcal{A}_h$ is unknown. Specifically, we consider the same setting as Setting 1 in Section 5.1, but where the predictors from the target study $\boldsymbol{x}^{(0)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p)$ and two types of source studies belong to $\mathcal{A}_h$ or $\mathcal{A}_h^c$ are generated as follows: For any source data k in $\mathcal{A}_h$, the coefficient is set to be $\boldsymbol{\theta}^{(k)} = \boldsymbol{\beta} + (h/p)\boldsymbol{r}_p^{(k)}$;

Table 4: Comparison of prediction interval width in Setting 3. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

h	n_0	SCP	FS	TL	TL+UQ	TL+WQ
5	25	10.761(7.566)	2.491(0.287)	2.469(0.188)	2.332(0.122)	2.284(0.136)
	50	3.247(0.853)	2.160(0.206)	2.235(0.109)	2.163(0.074)	2.111(0.079)
	100	2.263(0.096)	2.206(0.204)	2.139(0.068)	2.148(0.063)	2.079(0.053)
	200	2.171(0.051)	2.096(0.198)	2.075(0.038)	2.111(0.060)	2.073(0.042)
30	25	10.734(7.566)	2.599(0.312)	2.642(0.210)	2.373(0.137)	2.333(0.150)
	50	2.985(0.579)	2.242(0.249)	2.435(0.126)	2.213(0.082)	2.167(0.089)
	100	2.315(0.097)	2.218(0.196)	2.245(0.073)	2.171(0.063)	2.129(0.065)
	200	2.181(0.059)	2.136(0.210)	2.180(0.053)	2.164(0.058)	2.137(0.046)

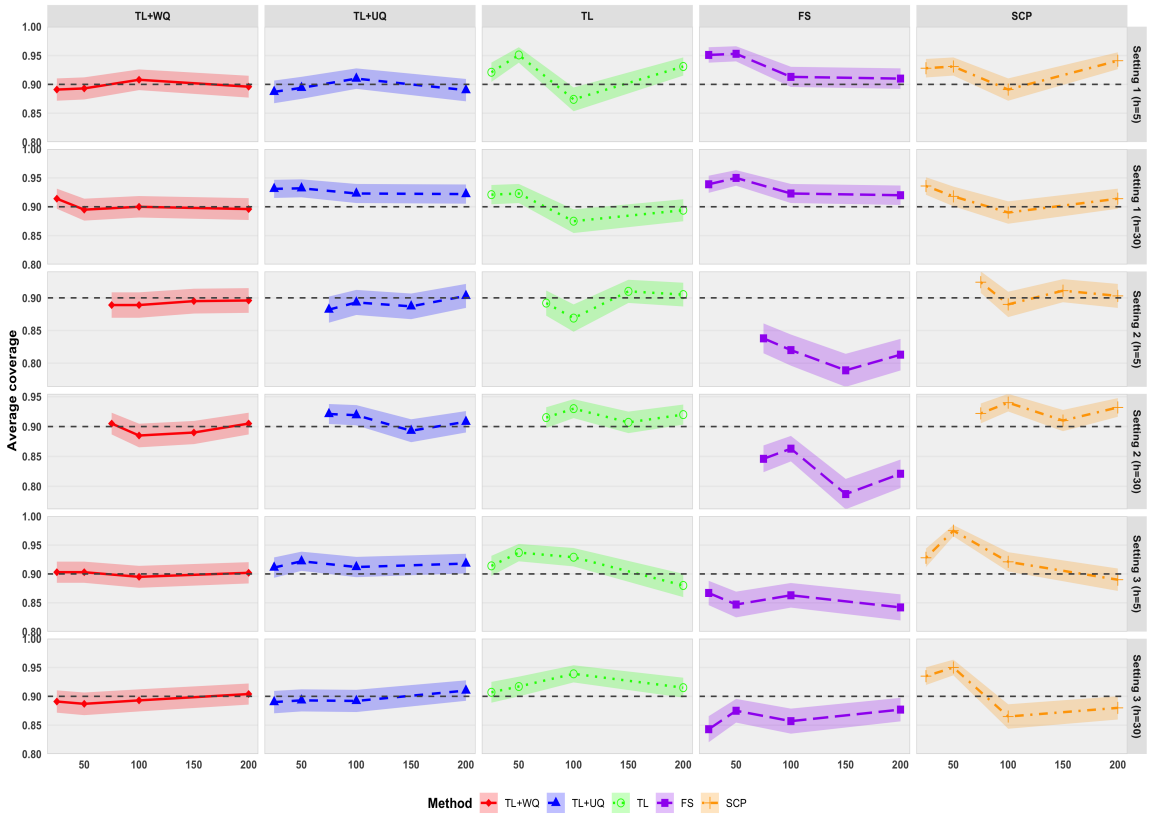

Figure 3: Simulation results showing mean prediction interval coverage averaged over 1000 independent trials across three settings. Shaded regions around each curve represent 95% confidence intervals for the estimated coverage rates.

For any source data k in $\mathcal{A}_h^c$, the jth element of the coefficient $\boldsymbol{\theta}^{(k)}$ is generated as:

$$\theta_j^{(k)} = \begin{cases} 0.5 + \dfrac{hr_j^{(k)}}{p}, & \text{if } j \in \{s+1, \ldots, 2s\} \cup S^{(k)}, \\ \dfrac{hr_j^{(k)}}{p}, & \text{otherwise,} \end{cases}$$

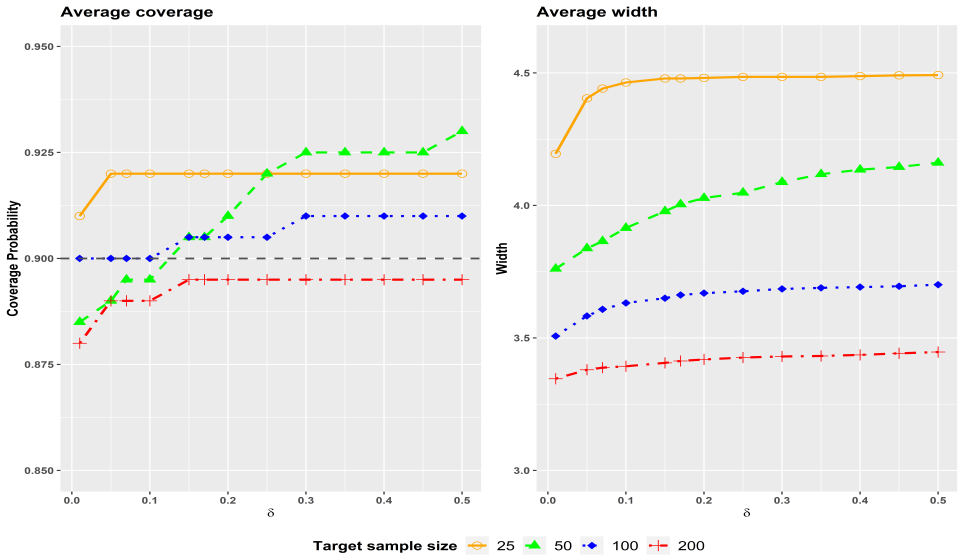

Figure 4: Coverage and width of predictive intervals across various δ values for $h = 30$ with $n_0 = 25, 50, 100, 200$ in Setting 1.

where we randomly select a subset, denoted as $S^{(k)}$, which has a dimensionality of s, from the set $\{2s+1, \ldots, p\}$, and $r_j^{(k)}$ is a Rademacher variable. We consider various target sample sizes $n_0 = 25, 50, 100, 200$ with dimensionality $p = 500$ and $s = 5$. The transferring level h is set to be 15 and the total number of source studies $K = 10$ with $|\mathcal{A}_h| = 5$ and $|\mathcal{A}_h^c| = 5$.

In this setup, we know that the kth source data and target data are both generated from the multivariate Gaussian distribution. Hence, we use the following cKL divergence between two multivariate Gaussian formulas to detect whether the kth source is informative or not,

$$cKL^{(k)} = \log\left\{\frac{\sigma_k(x)}{\sigma_0(x)}\right\} + \frac{\sigma_0^2(x) + \{\mu_0(x) - \mu_k(x)\}^2}{2\sigma_k^2(x)} - \frac{1}{2}, \quad k = 1, \ldots, K,$$

where $\mu_0(x)$ and $\mu_k(x)$ are the mean functions from the target data and the kth source data, respectively, and $\sigma_0(x)$ and $\sigma_k(x)$ are their respective covariance functions. Notice that the estimated probability densities may become too small due to the curse of dimensionality when considering the high-dimensional setting. To address this issue, we first conduct principal component analysis (PCA), selecting the top ten principal components to compute the cKL divergence. We then apply the proposed transferable source detection algorithm to select the informative transferring set $\widehat{\mathcal{A}}$, with $\widehat{\mathcal{A}}$ limited to five sources.

For comparison, we also run the proposed TCP algorithms with and without weights based on source data in $\mathcal{A}_h$, denoted as TLUQ($\mathcal{A}_h$) and TLWQ($\mathcal{A}_h$), respectively. Meanwhile, we consider the transferable source data detection to identify the informative transferring set $\widehat{\mathcal{A}}$ and then proceed to run the proposed TCP algorithms with and without weights based on source data in $\widehat{\mathcal{A}}$, denoted as TLWQ($\widehat{\mathcal{A}}$) and TLUQ($\widehat{\mathcal{A}}$), respectively. Our coverage results are shown in Figure 5, and the prediction interval width results are summarized in Table 5. Obviously, TLWQ($\mathcal{A}_h$) consistently outperforms its counterparts

in performance across all settings, which aligns with expectations, as TLWQ($\mathcal{A}_h$) benefits from incorporating information transferred from the sources within $\mathcal{A}_h$. Notably, the performance of TLWQ($\widehat{\mathcal{A}}$) shows a remarkable similarity to that of TLWQ($\mathcal{A}_h$), suggesting that the algorithm for detecting transferable sources effectively identifies $\mathcal{A}_h$.

Table 5: Comparison of average prediction interval width. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	TLUQ($\mathcal{A}_h$)	TLWQ($\mathcal{A}_h$)	TLUQ($\widehat{\mathcal{A}}$)	TLWQ($\widehat{\mathcal{A}}$)
25	3.972(0.390)	3.942(0.375)	4.388(0.432)	4.303(0.435)
50	3.455(0.245)	3.465(0.247)	3.637(0.267)	3.555(0.287)
100	3.270(0.174)	3.282(0.178)	3.524(0.240)	3.418(0.216)
200	3.155(0.121)	3.128(0.120)	3.408(0.212)	3.221(0.148)

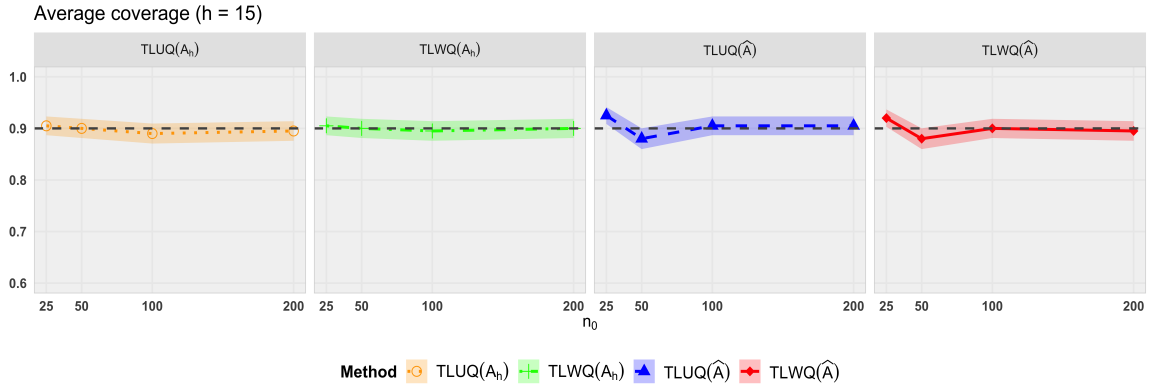

Figure 5: Comparison of average prediction interval coverage over 1000 independent trials. Shaded regions around each curve represent 95% confidence intervals for the estimated coverage rates.

We clarify that the cKL divergence can be computed using nonparametric methods without requiring prior knowledge of the posterior drift. Specifically, we estimate the cKL divergence between the target and each source using a nonparametric kernel density estimation (KDE) approach. To address the potential high dimensionality of covariates, we first conduct principal component analysis and retain the leading principal component. For each observation, we then apply a univariate Gaussian KDE to estimate the conditional density of Y given this component. For each source–target pair, we compute the estimated conditional densities $\widehat{P}^{(k)}(y \mid x)$ and $\widehat{Q}(y \mid x)$ over a discretized grid, and approximate the cKL divergence numerically via

$$\widehat{cKL}^{(k)} = \sum_{i=1}^{n} \int \widehat{Q}\left(y \mid x_i\right) \log \left\{ \frac{\widehat{Q}\left(y \mid x_i\right)}{\widehat{P}^{(k)}\left(y \mid x_i\right)} \right\} dy.$$

We then apply the same transferable source detection procedure as before, selecting the sources with the lowest estimated $\widehat{cKL}^{(k)}$ to form the set $\widehat{\mathcal{A}}$. Specifically, we adopt the

feature distribution from Setting 1 in Section 5.1, where target predictors are i.i.d. samples from $\mathcal{N}(0,1)$ with $p=1$. Unlike the linear model in earlier settings, the target response is generated via the nonlinear model $y^{(0)} = \sin(x_1^{(0)}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 0.5^2)$. The source studies are divided into two types: for informative sources ($k \in \mathcal{A}_h$), the response is $y^{(k)} = \sin(x^{(k)} + 0.2) + \varepsilon$, introducing a mild posterior drift relative to the target; for uninformative sources ($k \in \mathcal{A}_h^c$), the response is $y^{(k)} = \sin(x^{(k)} + 1.2) + \varepsilon$, introducing a stronger shift. We consider target sample sizes $n_0 = 25, 50, 100, 200$ with $K = 10$ total sources and $|\mathcal{A}_h| = 5$.

Table 6 reports the average prediction interval widths, and Figure 6 reports the corresponding coverages. The results are consistent with those from the earlier multivariate Gaussian simulations: TLWQ($\mathcal{A}_h$) delivers the most efficient prediction intervals while maintaining reliable coverage, and TLWQ($\widehat{\mathcal{A}}$) performs comparably to TLWQ($\mathcal{A}_h$). This robustness further confirms the effectiveness of our KDE-based cKL divergence approach for detecting informative sources, even when conditional distributions are complex and unknown. In this KDE-based implementation, we do not rely on any parametric assumptions about the data-generating process; instead, the conditional densities $\widehat{P}^{(k)}(y \mid x)$ and $\widehat{Q}(y \mid x)$ are approximated directly using Gaussian kernel density estimation.

Table 6: Comparison of average prediction interval width. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	TLUQ($\mathcal{A}_h$)	TLWQ($\mathcal{A}_h$)	TLUQ($\widehat{\mathcal{A}}$)	TLWQ($\widehat{\mathcal{A}}$)
25	2.261(0.279)	2.252(0.271)	2.417(0.385)	2.270(0.282)
50	1.994(0.244)	1.971(0.231)	2.010(0.256)	2.005(0.252)
100	1.879(0.251)	1.857(0.235)	1.953(0.244)	1.903(0.236)
200	1.714(0.160)	1.712(0.158)	1.794(0.166)	1.747(0.163)

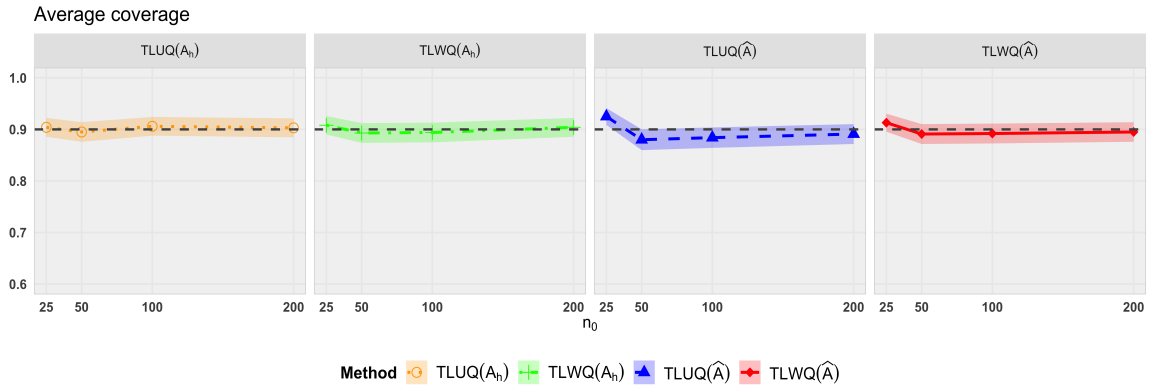

Figure 6: Comparison of average prediction interval coverage over 1000 independent trials. Shaded regions around each curve represent 95% confidence intervals for the estimated coverage rates.

6. Election data set

In this section, we illustrate the usefulness of the proposed TCP method by predicting how Americans voted in the 2020 U.S. presidential election (Cherian and Bronner, 2020; Gibbs and Candes, 2021). Our experiments are designed to investigate the county-wise representation of the relative changes in the number of votes for the Democratic Candidate in 2016 versus 2020, defined as: $Y = \frac{\text{Dem}_{2020} - \text{Dem}_{2016}}{\text{Dem}_{2016}}$, where Dem_{2020} is the number of Democratic Party votes in a given county in 2020, with analogous definitions for 2016. In our study, the covariate will include 27 potential factors for each county in our analysis such as ethnicity, age, gender, median income, and educational metrics based on the 2020 dataset. The detailed information for this dataset can be found in the Appendix in Barber et al. (2023).

In this experiment, our primary focus is on swing states that are characterized by a large number of counties. To make this experiment precise, we have excluded Alaska and Washington, D.C. from our analysis of the 49 states, selecting Texas as our target state, which presents a diverse political landscape, with the percentage of counties voting Democratic between 10% and 90%. Texas, with its substantial tally of 254 counties, serves as a key state for extensive analysis. Beyond Texas, we incorporate 12 additional states as our primary source datasets: North Carolina (100 counties), Virginia (133 counties), Missouri (115 counties), Kentucky (120 counties), Pennsylvania (67 counties), Nevada (17 counties), Colorado (64 counties), Georgia (159 counties), Illinois (102 counties), Maryland (24 counties), California (58 counties), and New Jersey (21 counties). Specifically for Texas, we divide the counties into two separate datasets for our analysis: the first dataset comprises the first 25 counties for training data and the remaining 229 counties for testing data. Similarly, the second dataset consists of the first 100 counties for training data and the remaining 154 counties for testing data.

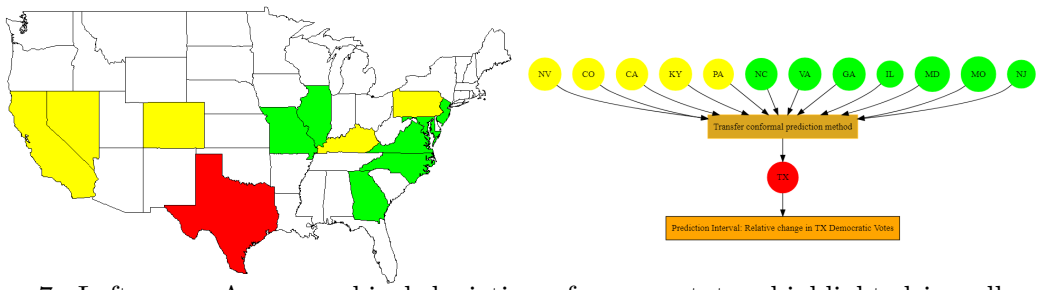

Figure 7: Left map: A geographical depiction of source states, highlighted in yellow and green, with the target state, Texas, marked in red. Right map: A diagram provides a visual explanation of how the TCP method is applied to forecast changes in Democratic votes in Texas.

Figure 7 displays the chosen U.S. states, highlighting both the source states and the target state (on the left map), along with a schematic representation of the TCP method (on the right map), illustrating its use in predicting relative changes in Democratic votes in Texas. To run the experiment, we first apply our transferable source detection algorithm to identify states that share notable similarities with Texas. The relative importance of these states is illustrated in Figure 8, where lower values correspond to higher relevance. Based

on a significant gap in the cKL divergence between KY and IL, we select the top five states: CO, CA, PA, NV, and KY, as the estimated transferring set $\widehat{\mathcal{A}}$. In Figure 7, the source states included in $\widehat{\mathcal{A}}$ are highlighted in yellow, while those excluded are shown in green. We then proceed to evaluate the four conformal methods: SCP, TL, TL+UQ, and TL+WQ, as in previous analyses. To implement the proposed method, the experimental parameters used are consistent with those in Setting 1 in Section 5.1, including the discrepancy parameter δ, the number of bootstrap iterations, and the configuration of the weight vectors w_i. For this experiment, we evaluate the above methods using the target coverage level of 0.9.

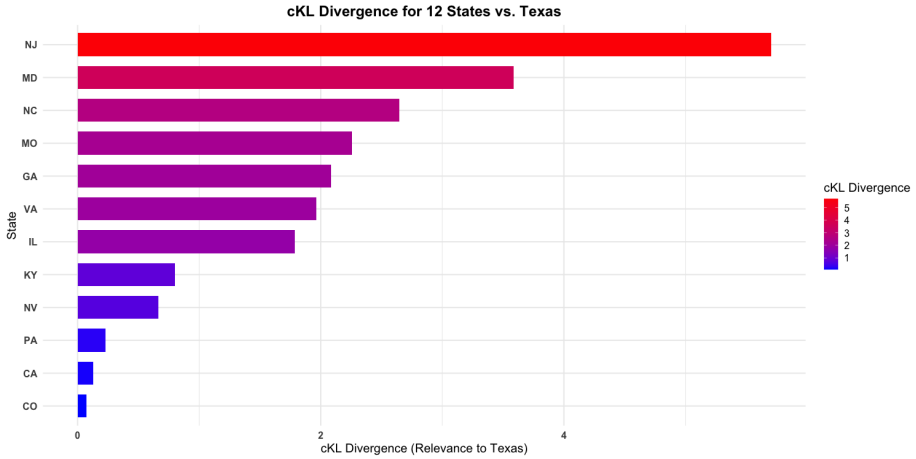

Figure 8: The cKL divergence between 12 states and Texas.

Table 7: Election data results showing coverage, and interval width averaged over all test counties.

	n_0=25		n_0=100	
	Width	Coverage	Width	Coverage
TL+WQ	0.692	0.905	0.497	0.896
TL+UQ	1.021	0.925	0.583	0.916
TL	8.377	0.985	0.626	0.948
SCP	8.478	0.985	0.685	0.916

Table 7 shows the predictive coverage and interval width, averaged over all the test data points for each of the four methods. It can be seen that the SCP and TL methods often exceed the necessary coverage, leading to considerably wider intervals, especially when the target data size, n_0 is small. On the contrary, both the TL+WQ and TL+UQ methods closely achieve the desired 90% coverage while maintaining narrower intervals. Furthermore, the TL+WQ method consistently outperforms TL+UQ, which is attributable to the specific structure of the non-exchangeable source data.

By leveraging information from related source counties, our TCP method produces substantially tighter prediction intervals while maintaining valid coverage guarantees. This enables more accurate estimation of shifts in Democratic support at the county level, par-

ticularly within swing states. Such precision is especially valuable in politically competitive regions, where even small changes in voter preferences can decisively affect election outcomes. Accurate uncertainty quantification in these contexts not only informs interpretations of election results but also guides strategic decision-making in campaign planning and resource allocation.

7. Conclusion and Future Work

This paper presents a novel TCP algorithm that enhances the accuracy of estimated residual distribution quantiles. The proposed TCP method not only maintains the desired coverage probability but also ensures narrower prediction intervals, regardless of the sample size or the underlying distribution of the target data. Additionally, we introduce a transferable source detection algorithm that effectively distinguishes between informative and non-informative sources under certain conditions. We provide some theoretical insights into the proposed TCP method, illustrating its advantages over traditional conformal methods. Our findings are validated through extensive simulation studies and an analysis of real-world data.

Several directions for future research are worth exploring. One direction is to concern adaptive selection of the coverage level. In line with standard practice in CP (e.g., Vovk et al. (2005), Lei et al. (2018), Romano et al. (2019), Barber et al. (2023), Candès et al. (2023)), we fix the coverage level in advance (e.g., $1 - \alpha = 0.90$) to guarantee validity and ensure comparability across methods. However, in practice, practitioners may be tempted to tune α post hoc to obtain narrower sets, thereby undermining theoretical guarantees. Recent work by Gauthier et al. (2025) addressed this by inverting the problem: adaptively selecting prediction sets (e.g., minimizing size) and then estimating the achieved coverage. it remains limited to classification with exchangeable data and does not address challenges such as distributional shift or regression outcomes, which our TCP framework explicitly tackles. Extending adaptive CP strategies to transfer learning therefore constitutes an exciting future direction. A second avenue involves leveraging surrogate or auxiliary variables to enhance efficiency. Recent work by Gao et al. (2024) demonstrated that surrogate outcomes can yield tighter conformal intervals for causal effects while preserving validity under appropriate assumptions. Building on this insight, our TCP framework could be extended to exploit surrogate information, either to refine estimation of $\mu(x)$ or to guide weighted calibration. Such an extension would be particularly valuable in domains such as healthcare or manufacturing, where surrogate measurements (e.g., biomarkers, sensor data) are abundant, whereas primary outcomes are scarce. Investigating how surrogate-assisted conformal methods interact with our posterior-drift framework under distributional shift or non-exchangeability provides another fruitful line of research.

Acknowledgments

We are grateful to three anonymous reviewers for their insightful comments and suggestions, which have helped improve the presentation of this paper. Jinhan Xie was supported by the National Key R&D Program of China (102022YFA1003701) and the National Natural Science Foundation of China (No. 12501388). Ting Li's research was supported by

the National Natural Science Foundation of China (Grant No. 12571304), the Shanghai Pujiang Program (Grant No. 24PJIC030), and the Program for Innovative Research Team of Shanghai University of Finance and Economics. Bei Jiang and Linglong Kong were partially supported by grants from the Canada CIFAR AI Chairs program, the Alberta Machine Intelligence Institute (AMII), and Natural Sciences and Engineering Council of Canada (NSERC), and Linglong Kong was also partially supported by grants from the Canada Research Chair program from NSERC.

Appendix A. Verification of the assumptions for three statistical models

In this section, we verify why Assumptions 3 and 4 are applicable to three specific statistical models. Additionally, we provide the explicit forms of $\eta_{1,n}$, $\eta_{2,n}$, $\rho_{1,n}$, and $\rho_{2,n}$ for these models in Section 4.2.

A.1 High-dimensional linear regression

In this subsection, we consider a high-dimensional linear regression setting to validate Assumptions 3-4. For the linear regression models, we assume that the fitted regression function is denoted by $\widetilde{\mu} = X^\top \widehat{\boldsymbol{\theta}}$, and the debiased fitted regression function by $\widehat{\mu} = X^\top \widehat{\boldsymbol{\beta}}$. The true parameters for the target model are represented by $\boldsymbol{\beta}$, and for the kth source model by $\boldsymbol{\theta}^{(k)}$. Here $\boldsymbol{\theta}^{(k)} \in \mathbb{R}^p$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ are both assumed to be a sparse vector with $s \ll \min\{n, p\}$ nonzero entries. Following the definition of transferring set in (Li et al., 2022), i.e., $\|\boldsymbol{\beta} - \boldsymbol{\theta}^{(k)}\|_1 \leq h_1$, we assume that $\|X^{(k)}\|_\infty$ is bounded by a constant C_0. Analogous to the level-h_0 transferring set $\mathcal{A}_{h_0}$ defined in equation (11), we introduce the level-h_0 transferring set for high-dimensional linear regression, denoted as $\mathcal{A}_{h_0}^{\mathrm{LM}}$:

$$\mathcal{A}_{h_0}^{\mathrm{LM}} = \left\{ k \in \{1, \ldots, K\} : \left\| X^{(k)\top} \left(\boldsymbol{\theta}^{(k)} - \boldsymbol{\beta} \right) \right\|_\infty \leq h_0 \right\},$$

where $h_0 = C_0 h_1$, defining the boundary within which the source domains are considered transferable.

In high-dimensional linear regression, usually, we have the following oracle inequalities (Van de Geer, 2008; Van de Geer et al., 2014) for lasso problems, i.e., we have

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1 \leq K_1 s \sqrt{\frac{\log(p)}{n_0 + N}},$$

with a probability at least $1 - p^{-c_1}$, where c_1 and K_1 are some positive constants. Consequently, Assumption 3 is satisfied for $\widetilde{\mu} = X^\top \widehat{\boldsymbol{\theta}}$, $\mu_{\mathcal{A}_{h_0}} = X^\top \boldsymbol{\theta}$, $\eta_{1,n} = O(s\sqrt{\log(p)/(n_0 + N)})$ and $\rho_{1,n} = p^{-c_1}$. This leads to the bound:

$$\Pr\left\{ \|X^\top (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|_\infty \geq s \sqrt{\frac{\log(p)}{n_0 + N}} \right\} \leq p^{-c_1}.$$

Furthermore, according to Theorem 1 of Li et al. (2022), if $s \log p / N + h_1 \left(\log p / n_0\right)^{1/2} = o(1)$ and $h_1 \ll s\sqrt{\log p / n_0}$ with $n_0 + N \gg n_0$, they first establish that, for any $B \in \Theta_1(s, h_1)$,

$$\begin{aligned} &\inf_{B \in \Theta_1(s,h_1)} \mathbb{P}\left(\frac{1}{n_0} \left\| X^{(0)} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \right\|_2^2 \vee \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \lesssim \frac{s \log p}{n_0 + N} + \frac{s \log p}{n_0} \wedge \eta_{h_1} \right) \\ &\quad \geq 1 - \exp\left(-c_2 \log p\right), \end{aligned}$$

where $\eta_{h_1} = h_1 \sqrt{\log p / n_0} \wedge h_1^2$. From this, it follows that, with probability at least $1 - p^{-c_2}$,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \leq K_2 s \sqrt{\frac{\log p}{n_0 + N}} + K_2 h_1,$$

where c_2 and K_2 are positive constants. Consequently, Assumption 4 is verified with $\widehat{\mu} = X^\top\widehat{\boldsymbol{\beta}}$, $\mu_0 = X^\top\boldsymbol{\beta}$, $\eta_{2,n} = C_2 s\sqrt{\log(p)/(n_0+N)} + C_2 h_1$ and $\rho_{2,n} = p^{-c_2}$, where C_2 is any constant. Thus, we have

$$\Pr\left\{\left\|X^\top\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\|_\infty \geq C_2 s\sqrt{\frac{\log(p)}{n_0+N}} + C_2 h_1\right\} \leq p^{-c_2}.$$

A.2 High-dimensional generalized linear regression

In this subsection, we validate Assumptions 3-4 using high-dimensional generalized linear models. We assume that the fitted regression functions in the transferring step, $\widetilde{\mu} = G(X^\top\widehat{\boldsymbol{\theta}})$, and the debiased fitted regression function, $\widehat{\mu} = G(X^\top\widehat{\boldsymbol{\beta}})$, are nonlinear in X, where $G(\cdot)$ is the link function with bounded first derivation as discussed in Tian and Feng (2023). Similarly to Section A.1, the true parameters for the target model and the kth source model are denoted as $\boldsymbol{\beta}$ and $\boldsymbol{\theta}^{(k)}$, respectively. We maintain that $\|\boldsymbol{\beta}-\boldsymbol{\theta}^{(k)}\|_1 \leq h_1$ and that $\|X^{(k)}\|_\infty$ is bounded by some positive constant C_0. Building on the definition of the level-h_0 transferring set $\mathcal{A}_{h_0}$ in equation (11), we introduce the corresponding transferring set for high-dimensional generalized linear regression, denoted as $\mathcal{A}_{h_0}^{\mathrm{GLM}}$:

$$\mathcal{A}_{h_0}^{\mathrm{GLM}} = \left\{k \in \{1,\ldots,K\} : \left\|G\left(X^{(k)\top}\boldsymbol{\theta}^{(k)}\right) - G\left(X^{(k)\top}\boldsymbol{\beta}\right)\right\|_\infty \leq h_0\right\},$$

where h_0 defines the threshold beyond which source datasets are considered non-transferable.

Similar to high-dimensional linear regression with lasso problems in Section A.1, oracle inequalities are also typically established for generalized linear models; see Van de Geer (2008); Van de Geer et al. (2014), i.e.,

$$\Pr\left\{\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_1 \geq K_3 s\sqrt{\frac{\log(p)}{n_0+N}}\right\} \leq p^{-c_3},$$

where K_3 and c_3 are some positive constants. Consequently, Assumption 3 is satisfied with $\widetilde{\mu} = G(X^\top\widehat{\boldsymbol{\theta}})$, $\mu_{\mathcal{A}_{h_0}} = G(X^\top\boldsymbol{\theta})$, $\eta_{1,n} = C_3 s\sqrt{\log(p)/n_0+N}$, $\rho_{1,n} = p^{-c_3}$, ensuring that

$$\Pr\left\{\left\|G\left(X^\top\widehat{\boldsymbol{\theta}}\right) - G\left(X^\top\boldsymbol{\theta}\right)\right\|_\infty \geq C_3 s\sqrt{\frac{\log(p)}{n_0+N}}\right\} \leq p^{-c_3},$$

where C_3 is some positive constant. Furthermore, according to Theorem 1 in Tian and Feng (2023), when $h_1 \ll s\sqrt{\log(p)/n_0}$, $n_0+N \gg n_0$, we have

$$\Pr\left\{\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \geq K_4 s\sqrt{\frac{\log(p)}{n_0+N}} + K_4 h_1\right\} \leq p^{-c_4},$$

where K_4 and c_4 are some positive constants. Consequently, Assumption 4 is verified with $\widehat{\mu} = G(X^\top\widehat{\boldsymbol{\beta}})$, $\mu_0 = G(X^\top\boldsymbol{\beta})$, $\eta_{2,n} = C_4 s\sqrt{\log(p)/(n_0+N)} + C_4 h_1$ and $\rho_{2,n} = p^{-c_4}$, that is

$$\Pr\left\{\left\|G\left(X^\top\widehat{\boldsymbol{\beta}}\right) - G\left(X^\top\boldsymbol{\beta}\right)\right\|_\infty \geq C_4 s\sqrt{\frac{\log(p)}{n_0+N}} + C_4 h_1\right\} \leq p^{-c_4},$$

where C_4 is some positive constant.

A.3 Nonparametric regression

In this subsection, we verify Assumptions 3-4 in the framework of nonparametric regressions. The nonparametric regression model for the target data is defined as follows:

$$Y_i^{(0)} = f(X_i^{(0)}) + \zeta_i^{(0)},$$

where $f(\cdot)$ is an unknown function of interest, assumed to be β_Q smooth, and $\zeta_i^{(0)}$ are i.i.d. random noise variables with a mean zero. The model for the kth source data is given by:

$$Y_i^{(k)} = g_k(X_i^{(k)}) + \zeta_i^{(k)}, k = 1, \ldots, K,$$

where $g_k, k = 1, \ldots, K$ are unknown functions that are $\beta_P^{(k)}$ smooth, and $\zeta_i^{(k)}$ are i.i.d. with a mean zero. For detailed definitions of β_Q and $\beta_P^{(k)}$, please refer to Definition 1 in Cai and Pu (2022). Building upon the discussions in Sections A.1 and A.2, we define the level-h_0 transferring set for nonparametric regression, denoted as

$$\mathcal{A}_{h_0}^{\mathrm{NM}} = \left\{ k \in \{1, \ldots, K\} : \left\| g_k\left(X^{(k)}\right) - f\left(X^{(k)}\right) \right\|_\infty \leq h_0 \right\},$$

indicating that the kth source data are close enough to $f(\cdot)$ to be potentially useful to transfer learning.

Recall that $N = |\mathcal{S}_1| = |\mathcal{S}_2|$, and p denotes the dimensionality of X. Let g represent the combined function of g_k for $k = 1, \ldots, K$, along with the target function f. According to the minimax optimal rate of estimation for local polynomial regression (Fan, 1993), we have

$$\mathbb{E}\left(\widehat{g} - g\right)^2 = O\left((N + n_0)^{-\frac{2\beta_P}{2\beta_P + p}} \right).$$

Given that $\mathbb{E}(\widehat{g} - g) = 0$, applying Chebyshev's inequality yields

$$\Pr\left(\|\widehat{g} - g\|_\infty \geq (N + n_0)^{-\frac{\beta_P}{2\beta_P + p}} \right) \leq C_5 \left(N + n_0\right)^{-\frac{\beta_P}{2\beta_P + p}},$$

where C_5 is some positive constant. Therefore, let $\widetilde{\mu} = \widehat{g}$, $\mu_{\mathcal{A}_{h_0}} = g$, $\eta_{1,n} = (N + n_0)^{-\frac{\beta_P}{2\beta_P + p}}$ and $\rho_{1,n} = C_5 \left(N + n_0\right)^{-\frac{\beta_P}{2\beta_P + p}}$, Assumption 3 is verified as follows:

$$\Pr\left(\|\widehat{g} - g\|_\infty \geq \eta_{1,n} \right) \leq \rho_{1,n}.$$

According to the Theorems 1 and 2 in Cai and Pu (2022), the estimation risk for transfer learning nonparametric regression, $\mathbb{E}(\widehat{f} - f)^2$, is proportional to:

$$n_{\max}^{-\frac{2\beta_{\max}}{2\beta_{\max} + p}} + \min\left\{ h_0, n_0^{-\frac{\beta_Q}{2\beta_Q + p}} \right\} \cdot n_0^{-\frac{\beta_Q}{2\beta_Q + p}} + \frac{1}{n_0},$$

where $\beta_{\max} = \max\{\beta_Q, \beta_P\}$ and $n_{\max} = \max\{n_0, N\}$. Assuming that $h_0 \ll n_0^{-\frac{\beta_Q}{2\beta_Q + p}}$, $\beta_P > \beta_Q$ and $N \gg n_0$, the minimax risk for transfer learning is smaller than the minimax risk for

estimating f using data from target domain alone. Let $\widehat{\mu} = \widehat{f}$, $\mu_0 = f$, $\eta_{2,n} = (N+n_0)^{-\frac{\beta_P}{2\beta_P+p}}$ and $\rho_{2,n} = C_6\{(N+n_0)^{-\frac{\beta_P}{2\beta_P+p}}\}$, where C_6 is some positive constant. Therefore, Assumption 4 is verified by

$$\Pr(\|\widehat{f} - f\|_\infty \geq \eta_{2,n}) \leq \rho_{2,n}.$$

Appendix B. Details of settings 2 and 3 in simulation study

This section presents the process of selecting weights for three different settings in the simulation study. For Settings 1 and 3 with continuous responses, we begin by establishing the distances between each data point $(X_i, Y_i) \in \mathcal{S}_1$ and test point (X_{n_0+1}, Y_{n_0+1}):

$$\begin{aligned} D_i^t &= \{|\widetilde{\mu}_t(X_i) - \widehat{\mu}_t(X_{n_0+1})| + \|X_i - X_{n_0+1}\|_2\}\,|Y_i - \widetilde{\mu}_t(X_i)|, \\ D_i^{\mathcal{I}_1} &= \{|\widetilde{\mu}_{\mathcal{I}_1}(X_i) - \widehat{\mu}_{\mathcal{I}_1}(X_{n_0+1})| + \|X_i - X_{n_0+1}\|_2\}\,|Y_i - \widetilde{\mu}_{\mathcal{I}_1}(X_i)|, \end{aligned}$$

where $i = 1, \ldots, n_{\text{cal}}$ and $n_{\text{cal}} = |\mathcal{S}_1|$. We then arrange the residuals $\widetilde{R}_i^t = |Y_i - \widetilde{\mu}_t(X_i)|$ for each data point $(X_i, Y_i) \in \mathcal{S}_1$ in ascending order based on the respective distances D_i^t. In a similar manner, the residuals $\widetilde{R}_i^{\mathcal{I}_1} = |Y_i - \widetilde{\mu}_{\mathcal{I}_1}(X_i)|$ are sorted in ascending order for all data points $(X_i, Y_i) \in \mathcal{S}_1$ according to $D_i^{\mathcal{I}_1}$. For the algorithm implementation, we apply weights $w_i^t = e^{-i/(n_{\text{cal}}+1)}$ to the ascending ordered residuals $\widetilde{R}_i^t$ with $w_{n_{\text{cal}}+1}^t = 1$. Similarly, weights $w_i^{\mathcal{I}_1} = e^{-i/(n_{\text{cal}}+1)}$ for the ascending ordered residuals $\widetilde{R}_i^{\mathcal{I}_1}$, setting $w_{n_{\text{cal}}+1}^{\mathcal{I}_1} = 1$. The corresponding normalized weights $\widetilde{w}_i^{\mathcal{I}_1}$ and $\widetilde{w}_i^t$ are defined as follows:

$$\begin{aligned} \widetilde{w}_i^t &= \frac{w_i^t}{w_1^t + \cdots + w_{n_{\text{cal}}}^t + 1},\ i = 1, \ldots, n_{\text{cal}}, \\ \widetilde{w}_i^{\mathcal{I}_1} &= \frac{w_i^{\mathcal{I}_1}}{w_1^{\mathcal{I}_1} + \cdots + w_{n_{\text{cal}}}^{\mathcal{I}_1} + 1},\ i = 1, \ldots, n_{\text{cal}}. \end{aligned}$$

For Setting 2 poisson regression with discrete responses, the weights are constructed the same as Settings 1 and 3, but the distances D_i^t and $D_i^{\mathcal{I}_1}$ between each data point $(X_i, Y_i) \in \mathcal{S}_1$ and test point (X_{n_0+1}, Y_{n_0+1}) are different,

$$\begin{aligned} D_i^t &= \{|\widetilde{\mu}_t(X_i) - \widehat{\mu}_t(X_{n_0+1})| + \|\exp(X_i) - \exp(X_{n_0+1})\|_2\}\,|Y_i - \widetilde{\mu}_t(X_i)|, \\ D_i^{\mathcal{I}_1} &= \{|\widetilde{\mu}_{\mathcal{I}_1}(X_i) - \widehat{\mu}_{\mathcal{I}_1}(X_{n_0+1})| + \|\exp(X_i) - \exp(X_{n_0+1})\|_2\}\,|Y_i - \widetilde{\mu}_{\mathcal{I}_1}(X_i)|, \end{aligned}$$

where $i = 1, \ldots, n_{\text{cal}}$.

Appendix C. All Technique Proofs

In this section, we provide proof of the theorems and lemmas in the manuscript. To facilitate our proof, let's introduce some notations.

- Let $N = |\mathcal{S}_1| = |\mathcal{S}_2|$ and $n_0/2 = |\mathcal{I}_1| = |\mathcal{I}_1^t| = |\mathcal{I}_2|$, for ease of illustration.

- Population regression functions. We denote by $\mu_0(X^{(0)}) = \mathbb{E}[Y^{(0)} \mid X = X^{(0)}]$ the target regression function, and by $\mu_k(X^{(k)}) = \mathbb{E}[Y^{(k)} \mid X = X^{(k)}]$ the regression function for the k-th source dataset. The transferring set $\mathcal{A}_{h_0}$ is associated with the combined regression function $\mu_{\mathcal{A}_{h_0}}(x) := \mathbb{E}\left[Y^{(k)} \mid X = X^{(k)}, \text{ data from } k \in \mathcal{A}_{h_0}\right]$. Preliminary regression estimators, such as $\widetilde{\mu}_{\mathcal{I}_1}, \widetilde{\mu}_{\mathcal{I}_2}$, or $\widetilde{\mu}_t$, are understood as approximations to $\mu_{\mathcal{A}_{h_0}}$, while debiased transfer–learning estimators, such as $\widehat{\mu}_{\mathcal{I}_1}, \widehat{\mu}_{\mathcal{I}_2}$, or $\widehat{\mu}_t$, approximate μ_0. For brevity, we use $\widetilde{\mu}$ and $\widehat{\mu}$ to represent generic elements from these two classes of estimators.

- Population residuals. We denote $R_{1,i} = |Y_i - \mu_{\mathcal{A}_{h_0}}(X_i)|$ and $R_{2,i} = |Y_i - \mu_0(X_i)|$, with distributions F_1, F_2 and quantiles q_1, q_2.

- Fitted residuals. We denote $\widetilde{R}_i = |Y_i - \widetilde{\mu}(X_i)|$ and $\widehat{R}_i = |Y_i - \widehat{\mu}(X_i)|$, with corresponding distributions $F_{\widetilde{R}}, F_{\widehat{R}}$ and quantiles $\widetilde{q}, \widehat{q}$.

- Test residuals. For a fresh point $(X_{\text{new}}, Y_{\text{new}})$, we write $\widetilde{R}_{\text{new}} = |Y_{\text{new}} - \widetilde{\mu}(X_{\text{new}})|$ and $\widehat{R}_{\text{new}} = |Y_{\text{new}} - \widehat{\mu}(X_{\text{new}})|$.

- Correction parameter. We clarified that the population correction parameter is $\Lambda = |q_2 - q_1|$, with the empirical analogue $\Lambda(\mathcal{I}_1)$ defined from calibration residuals.

- Recall that $\eta_n = \eta_{1,n} + \eta_{2,n}$, $h' = h_0 + h$, $\rho_n = \rho_{1,n} + \rho_{2,n}$, $\eta_n^c = \eta_{3,n} + \eta_{4,n}$, and $\rho_n^c = \rho_{3,n} + \rho_{4,n}$ for simplicity.

C.1 Some lemmas

Lemma 20 *Assume Assumptions 1-4 hold. For a new data point (X_{new}, Y_{new}), we have*

$$\Pr\left(\left|\widetilde{R}_{new} - \widehat{R}_{new}\right| \geq \eta_n + h_0\right) \leq \rho_n,$$

where h_0 is defined in equation (11).

Lemma 21 *Assume Assumptions 1-4 hold, if f_1 and f_2 are both upper bounded by some positive constants u_1, u_2 and lower bounded by some positive constants l_1, and l_2 respectively, we then have*

$$|\Lambda(\mathcal{I}_1) - \Lambda| = O_p\left\{\eta_n + \rho_{2,n} + N^{-1/2}\right\}.$$

Lemma 22 *Assume Assumptions 1, 2, 5, and 6 hold. Conditional on a fixed X_{new}, we have*

$$\Pr\left(\left|\widetilde{R}_{new} - \widehat{R}_{new}\right| \geq \eta_n^c + h_0 \,\middle|\, X_{new}\right) \leq \rho_n^c.$$

C.2 Proofs of lemmas

Proof of Lemma 20. Using the definition of $\mathcal{A}_{h_0}$ and Assumption 2, we then have

$$\begin{aligned}
&\Pr\left(\left|\widetilde{R}_{new}-\widehat{R}_{new}\right|\geq\eta_n+h_0\right)\\
=&\Pr\left\{\left|\left|Y_{new}-\widetilde{\mu}_{\mathcal{I}_2}\left(X_{new}\right)\right|-\left|Y_{new}-\widehat{\mu}_{\mathcal{I}_2}\left(X_{new}\right)\right|\right|\geq\eta_n+h_0\right\}\\
\leq&\Pr\left\{\left|\widetilde{\mu}_{\mathcal{I}_2}\left(X_{new}\right)-\widehat{\mu}_{\mathcal{I}_2}\left(X_{new}\right)\right|\geq\eta_n+h_0\right\}\\
\leq&\Pr\left\{\left|\widetilde{\mu}_{\mathcal{I}_2}\left(X_{new}\right)-\widehat{\mu}_{\mathcal{I}_2}\left(X_{new}\right)\right|\geq\eta_n+h_0,\ \left|\widetilde{\mu}_{\mathcal{I}_2}\left(X_{new}\right)-\mu_{\mathcal{A}_{h_0}}\left(X_{new}\right)\right|\leq\eta_{1,n},\right.\\
&\left.\left|\widehat{\mu}_{\mathcal{I}_2}\left(X_{new}\right)-\mu_0\left(X_{new}\right)\right|\leq\eta_{2,n}\right\}+\Pr\left\{\left|\widetilde{\mu}_{\mathcal{I}_2}\left(X_{new}\right)-\mu_{\mathcal{A}_{h_0}}\left(X_{new}\right)\right|\geq\eta_{1,n}\right\}\\
&+\Pr\left\{\left|\widehat{\mu}_{\mathcal{I}_2}\left(X_{new}\right)-\mu_0\left(X_{new}\right)\right|\geq\eta_{2,n}\right\}\\
\leq&\Pr\left\{\left|\mu_{\mathcal{A}_{h_0}}\left(X_{new}\right)-\mu_0\left(X_{new}\right)\right|\geq h_0\right\}+\rho_{1,n}+\rho_{2,n}\\
=&\rho_n.
\end{aligned}$$

where the first inequality is derived from the Triangle's inequality and the last inequality holds due to Assumptions 3-4.

Proof of Lemma 21. Notice that f_2 is bounded by $u_2>0$, using the Theorem 3.1 in Lei et al. (2018), for any $m>0$, we then have

$$\begin{aligned}
F_{\widehat{R}}\left(m\right)=&\Pr\left(\widehat{R}<m\right)\\
\leq&\Pr\left(\widehat{R}<m,\ \left|\widehat{R}-R_2\right|\leq\eta_{2,n}\right)+\Pr\left(\left|\widehat{R}-R_2\right|\geq\eta_{2,n}\right)\\
\leq&\Pr\left(R_2<m+\eta_{2,n}\right)+\rho_{2,n}\\
\leq&F_2\left(m\right)+u_2\eta_{2,n}+\rho_{2,n}.
\end{aligned}$$

Additionally,

$$\begin{aligned}
F_2\left(m\right)=&\Pr\left(R<m\right)\\
\leq&\Pr\left(R<m,\ \left|\widehat{R}-R_2\right|\leq\eta_{2,n}\right)+\Pr\left(\left|\widehat{R}-R_2\right|\geq\eta_{2,n}\right)\\
\leq&\Pr\left(\widehat{R}_2<m+\eta_{2,n}\right)+\rho_{2,n}\\
\leq&F_{\widehat{R}}\left(m\right)+u_2\eta_{2,n}+\rho_{2,n}.
\end{aligned}$$

Thus, we have

$$\sup_{m>0}\left|F_{\widehat{R}}\left(m\right)-F_2\left(m\right)\right|\leq u_2\eta_{2,n}+\rho_{2,n}. \tag{C.1}$$

Moreover since f_2 is lower bounded by l_2, it then follows that

$$\left|\widehat{q}-q_2\right|\leq\left(1/l_2\right)\left(u_2\eta_{2,n}+\rho_{2,n}\right). \tag{C.2}$$

Let $\mathrm{Q}_{1-\alpha}(\sum_{i=1}^{N+1}\widetilde{w}_i\cdot\delta_{\widetilde{R}_i})$ denote the weighted $1-\alpha$ quantile of the empirical CDF of $\widetilde{R}_i, i=1,\ldots,N$, and $\mathrm{Q}_{1-\alpha}(\sum_{i=1}^{N+1}\widetilde{w}_i\cdot\delta_{R_{1,i}})$ denote the weighted $1-\alpha$ quantile of the

empirical CDF of $R_{1,i}, i = 1, \ldots, N$, for all $(X_i, Y_i) \in \mathcal{S}_1$. Notice that by Triangle's inequality and Assumption 3, and on the event $\{\|\widetilde{\mu} - \mu_{\mathcal{A}_{h_0}}\|_\infty\} \leq \eta_{1,n}$, we have $|\widetilde{R}_i - R_{1,i}| = \left||Y_i - \widetilde{\mu}(X_i)| - |Y_i - \mu_{\mathcal{A}_{h_0}}(X_i)|\right| \leq \eta_{1,n}$ for $i = 1, \ldots, N$. Therefore, we have

$$\Pr\left(\left|\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i \cdot \delta_{\widetilde{R}_i}\right) - \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i \cdot \delta_{R_{1,i}}\right)\right| \geq \eta_{1,n}\right) \leq \rho_{1,n}. \tag{C.3}$$

Using Corollary 21.5 in Van der Vaart (2000) and the assumption that f_1 is lower bounded by l_1, we obtain

$$\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i \cdot \delta_{R_{1,i}}\right) = q_1 + O_p\left(N^{-1/2}\right). \tag{C.4}$$

Combining (C.3) and (C.4), we have

$$\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i \cdot \delta_{\widetilde{R}_i}\right) = q_1 + O_p\left(\eta_{1,n} + N^{-1/2}\right). \tag{C.5}$$

By the definition of correction parameter $\Lambda(\mathcal{I}_1)$, we can treat $\mathrm{Q}_{1-\alpha}(\sum_{i=1}^{N+1} \widetilde{w}_i \cdot \delta_{\widetilde{R}_i}) + \Lambda(\mathcal{I}_1)$ as the $1-\alpha$ quantile of empirical CDF of $\widehat{R}_j,\ j = 1, \cdots, n_0(T-1)/2$ for all $(X_j, Y_j) \in \mathcal{B}_{-t}$, where T is the number of random bootstrapping samples of $\mathcal{I}_1$. Furthermore, using the same arguments as (C.4) along with (C.2), we get

$$\begin{aligned}\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i \cdot \delta_{\widetilde{R}_i}\right) + \Lambda\left(\mathcal{I}_1\right) =& \widehat{q} + O_p\left\{(T-1)^{-1/2} n_0^{-1/2}\right\} \\ =& q_2 + O_p\left\{\eta_{2,n} + \rho_{2,n} + (T-1)^{-1/2} n_0^{-1/2}\right\}. \end{aligned} \tag{C.6}$$

Therefore, when T is sufficient large and combining (C.5) and (C.6), we can conclude that

$$\left|\Lambda\left(\mathcal{I}_1\right) - \Lambda\right| = O_p\left\{\eta_n + \rho_{2,n} + N^{-1/2}\right\}.$$

Proof of Lemma 22. Conditional on a fixed X_{new}, we proceed as follows. By the definition of $\mathcal{A}_{h_0}$ and Assumption 2, we have

$$
\begin{aligned}
&\Pr\left(\left|\widetilde{R}_{new}-\widehat{R}_{new}\right|\geq \eta_n^c+h_0\,\middle|\,X_{new}\right)\\
=&\Pr\left(\left|\,|Y_{new}-\widetilde{\mu}_{\mathcal{I}_2}(X_{new})|-|Y_{new}-\widehat{\mu}_{\mathcal{I}_2}(X_{new})|\,\right|\geq \eta_n^c+h_0\,\middle|\,X_{new}\right)\\
\leq&\Pr\left(|\widetilde{\mu}_{\mathcal{I}_2}(X_{new})-\widehat{\mu}_{\mathcal{I}_2}(X_{new})|\geq \eta_n^c+h_0|X_{new}\right)\\
\leq&\Pr\left(\begin{array}{l}|\widetilde{\mu}_{\mathcal{I}_2}(X_{new})-\widehat{\mu}_{\mathcal{I}_2}(X_{new})|\geq \eta_n^c+h_0,\\ \left|\widetilde{\mu}_{\mathcal{I}_2}(X_{new})-\mu_{\mathcal{A}_{h_0}}(X_{new})\right|\leq \eta_{3,n},\\ |\widehat{\mu}_{\mathcal{I}_2}(X_{new})-\mu_0(X_{new})|\leq \eta_{4,n}\end{array}\,\middle|\,X_{new}\right)\\
&+\Pr\left(\left|\widetilde{\mu}_{\mathcal{I}_2}(X_{new})-\mu_{\mathcal{A}_{h_0}}(X_{new})\right|>\eta_{3,n}|X_{new}\right)\\
&+\Pr\left(|\widehat{\mu}_{\mathcal{I}_2}(X_{new})-\mu_0(X_{new})|>\eta_{4,n}|X_{new}\right)\\
\leq&\Pr\left(\left|\mu_{\mathcal{A}_{h_0}}(X_{new})-\mu_0(X_{new})\right|\geq h_0|X_{new}\right)+\rho_{3,n}+\rho_{4,n}\\
\leq&\rho_n^c.
\end{aligned}
$$

C.3 Proofs of theorems

Proof of Theorem 9. For simplification, we hereafter denote $\widetilde{R}_{N+1}^{\mathcal{I}_1}$ as $\widetilde{R}_{new}$. We first establish the lower bound for coverage probability. Notice that the definition of the TCP interval, as stated in equation (9), reveals

$$
Y_{new}\notin\widehat{C}\left(X_{new}\right)\Longleftrightarrow\widehat{R}_{new}>\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1}\widetilde{w}_i^{\mathcal{I}_1}\cdot\delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right)+\Lambda\left(\mathcal{I}_1\right),
$$

Then, using the same arguments as (C.1) in Lemma 21, we obtain

$$
\sup_{m>0}\left|F_{\widetilde{R}}\left(m\right)-F_1\left(m\right)\right|\leq u_1\eta_{1,n}+\rho_{1,n}. \tag{C.7}
$$

According to Lemma 20 and (C.7), we have

$$
\begin{aligned}
&\Pr\left\{Y_{new} \notin \widehat{C}\left(X_{new}\right)\right\} \\
=&\Pr\left\{\widehat{R}_{new} > \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right)\right\} \\
\leq&\Pr\left\{\widehat{R}_{new} > \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right), \left|\widehat{R}_{new} - \widetilde{R}_{new}\right| \leq \eta_n + h_0\right\} \\
&+ \Pr\left\{\left|\widehat{R}_{new} - \widetilde{R}_{new}\right| \geq \eta_n + h_0\right\} \\
\leq&\Pr\left\{\widetilde{R}_{new} > \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right) - \eta_n - h_0\right\} + \rho_n \\
\leq&\Pr\left\{\widetilde{R}_{new} > \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right)\right\} + F_{\widetilde{R}}\left\{\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right)\right\} \\
&- F_{\widetilde{R}}\left\{\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right) - \eta_n - h_0\right\} + \rho_n \\
\leq&\Pr\left\{\widetilde{R}_{new} > \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right)\right\} + F_1\left\{\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right)\right\} \\
&- F_1\left\{\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right) - \eta_n - h_0\right\} + 2u_1\eta_{1,n} + 2\rho_{1,n} + \rho_n \\
\leq&\Pr\left\{\widetilde{R}_{new} > \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right)\right\} + u_1\left|\min\left(0, \Lambda\left(\mathcal{I}_1\right) - \eta_n - h_0\right)\right| \\
&+ 2u_1\eta_{1,n} + 2\rho_{1,n} + \rho_n.
\end{aligned}
$$

Using the same argument as Theorem 2 in Barber et al. (2023), we know

$$
\Pr\left\{\widetilde{R}_{new} > \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right)\right\} \leq \alpha + \sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \mathrm{d_{TV}}\left(\widetilde{R}^{\mathcal{I}_1}, \widetilde{R}^{\mathcal{I}_1,i}\right),
$$

where $\widetilde{R}^{\mathcal{I}_1}$ denotes the fitted residual sequence, i.e., $\widetilde{R}^{\mathcal{I}_1} = (\widetilde{R}_1^{\mathcal{I}_1}, \ldots, \widetilde{R}_{N+1}^{\mathcal{I}_1})$ and $\widetilde{R}^{\mathcal{I}_1,i}$ is a new sequence after swapping the test residual $\widetilde{R}_{N+1}^{\mathcal{I}_1}$, i.e. $\widetilde{R}_{new}$, with the ith fitted residual $\widetilde{R}_i^{\mathcal{I}_1}$ which has the elements

$$
\left(\widetilde{R}^{\mathcal{I}_1,i}\right)_j = \begin{cases} \widetilde{R}_j^{\mathcal{I}_1}, & \text{if } j \neq i \text{ and } j \neq N+1, \\ \widetilde{R}_{N+1}^{\mathcal{I}_1}, & \text{if } j = i, \\ \widetilde{R}_i^{\mathcal{I}_1}, & \text{if } j = N+1. \end{cases}
$$

Consider a data sequence $Z = (Z_1, \cdots, Z_{N+1})$ with $Z_i = \{X_i, Y_i\}$, we then define Z^i as a new sequence obtained by swapping the ith data Z_i with Z_{N+1} in Z. As all the data

are independent, and the total variation distance between any function applied to each of Z and Z^i cannot exceed $\mathrm{d_{TV}}\left(Z, Z^i\right)$ itself, we apply Lemma 1 in Barber et al. (2023) to obtain

$$\mathrm{d_{TV}}\left(\widetilde{R}^{\mathcal{I}_1}, \widetilde{R}^{\mathcal{I}_1,i}\right) \leq \mathrm{d_{TV}}\left(Z, Z^i\right) \leq 2\ \mathrm{d_{TV}}\left(Z_i, Z_{N+1}\right).$$

Let A_i be the domain of probability distributions $P_{(i)}$ for each data point $Z_i = \{X_i, Y_i\}$, where $P_{(i)} = P^{(k)}$ if Z_i belongs to the k-th source data. Using equation (4.1) in Levin and Peres (2017) and posterior drift assumption, it then follows that

$$\begin{aligned}
\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \mathrm{d_{TV}}\left(\widetilde{R}^{\mathcal{I}_1}, \widetilde{R}^{\mathcal{I}_1,i}\right) \leq & 2 \sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \mathrm{d_{TV}}\left(Z_i, Z_{N+1}\right) \\
\leq & 2 \sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \max_{(X,Y)\in A_i} \left|P_{(i)}\left(X,Y\right) - Q\left(X,Y\right)\right| \\
= & 2 \sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \max_{(X,Y)\in A_i} \left|P_{(i)}\left(Y|X\right) P_{(i)}\left(X\right) - Q\left(Y|X\right) Q\left(X\right)\right| \\
\leq & 2h,
\end{aligned}$$

where the third inequality is due to (3). Therefore, combining the above results and Lemma 21, we obtain

$$\Pr\left\{Y_{new} \notin \widehat{C}\left(X_{new}\right)\right\} \leq \alpha + O_p\left\{\eta_n + \rho_n + h' + \Lambda + N^{-1/2}\right\}.$$

Next, we show the upper bound for coverage. Using the same arguments as the lower bound on coverage, the coverage event for new data point $Z_{new} = (X_{new}, Y_{new})$ can be described as follows

$$Y_{new} \in \widehat{C}\left(X_{new}\right) \Longleftrightarrow \widehat{R}_{new} \leq \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right),$$

Thus,

$$
\begin{aligned}
&\Pr\left\{Y_{new} \in \widehat{C}\left(X_{new}\right)\right\} \\
=&\Pr\left\{\widehat{R}_{new} \leq \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right)\right\} \\
\leq&\Pr\left\{\widehat{R}_{new} \leq \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right),\ \left|\widehat{R}_{new} - \widetilde{R}_{new}\right| \leq \eta_n + h_0\right\} \\
&+ \Pr\left\{\left|\widehat{R}_{new} - \widetilde{R}_{new}\right| \geq \eta_n + h_0\right\} \\
\leq&\Pr\left\{\widetilde{R}_{new} \leq \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right) + \eta_n + h_0\right\} + \rho_n \\
=&\Pr\left\{\widetilde{R}_{new} \leq \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right)\right\} + F_{\widetilde{R}}\left\{\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right) + \eta_n + h_0\right\} \\
&- F_{\widetilde{R}}\left\{\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right)\right\} + \rho_n \\
\leq&\Pr\left\{\widetilde{R}_{new} \leq \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right)\right\} + F_1\left\{\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right) + \eta_n + h_0\right\} \\
&- F_1\left\{\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right)\right\} + 2u_1\eta_{1,n} + 2\rho_{1,n} + \rho_n \\
\leq&\Pr\left\{\widetilde{R}_{new} \leq \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right)\right\} + u_1\left|\max\left(0, \Lambda\left(\mathcal{I}_1\right) + \eta_n + h_0\right)\right| \\
&+ 2u_1\eta_{1,n} + 2\rho_{1,n} + \rho_n.
\end{aligned}
$$

A direct argument as Theorem 3 in Barber et al. (2023), we have

$$
\begin{aligned}
\Pr\left\{\widetilde{R}_{new} \leq \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right)\right\} \leq& 1 - \alpha + \widetilde{w}_{N+1}^{\mathcal{I}_1} + \sum_{i=1}^{N} \widetilde{w}_i \cdot \mathrm{d_{TV}}\left(\widetilde{R}^{\mathcal{I}_1}, \widetilde{R}^{\mathcal{I}_1,i}\right) \\
\leq& 1 - \alpha + \widetilde{w}_{N+1}^{\mathcal{I}_1} + a_1 h,
\end{aligned}
$$

for some positive constant a_1. Combining the above results and Lemma 21, we have

$$
\Pr\left\{Y_{new} \in \widehat{C}\left(X_{new}\right)\right\} \leq 1 - \alpha + \widetilde{w}_{N+1}^{\mathcal{I}_1} + O_p\left\{\eta_n + \rho_n + h' + \Lambda + N^{-1/2}\right\}.
$$

We have completed the proof of this theorem.

Proof of Theorem 11. According to the construction of TCP (9) and (C.6), the width of prediction interval W_{trconf} satisfies

$$
\begin{aligned}
W_{trconf} =& 2Q_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + 2\Lambda\left(\mathcal{I}_1\right) \\
=& 2q_2 + O_p\left\{\eta_{2,n} + \rho_{2,n}\right\}.
\end{aligned}
$$

Proof of Theorem 15. Following the same proof strategy as the lower bound on coverage in Theorem 9, and according to Lemma 22 we have that,

$$
\begin{aligned}
&\Pr\left\{Y_{new} \notin \widehat{C}\left(X_{new}\right)|X_{new}\right\} \\
=&\Pr\left\{\widehat{R}_{new} > \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right) \middle| X_{new}\right\} \\
\leq&\Pr\left\{\widehat{R}_{new} > \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right),\ \left|\widehat{R}_{new} - \widetilde{R}_{new}\right| \leq \eta_n^c + h_0 \middle| X_{new}\right\} \\
&+ \Pr\left\{\left|\widehat{R}_{new} - \widetilde{R}_{new}\right| \geq \eta_n^c + h_0 \middle| X_{new}\right\} \\
\leq&\Pr\left\{\widetilde{R}_{new} > \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right) - \eta_n^c - h_0 \middle| X_{new}\right\} + \rho_n^c \\
\leq&\Pr\left\{\widetilde{R}_{new} > \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) \middle| X_{new}\right\} + F_{\widetilde{R}|X_{new}}\left\{\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right)\right\} \\
&- F_{\widetilde{R}|X_{new}}\left\{\mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) + \Lambda\left(\mathcal{I}_1\right) - \eta_n^c\right\} + \rho_n^c \\
\leq&\Pr\left\{\widetilde{R}_{new} > \mathrm{Q}_{1-\alpha}\left(\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \delta_{\widetilde{R}_i^{\mathcal{I}_1}}\right) \middle| X_{new}\right\} + O_p\left\{\eta_n^c + \rho_n^c + \eta_n + \rho_{2,n} + \Lambda + h_0 + N^{-1/2}\right\} \\
\leq&\alpha + \sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \mathrm{d_{TV}}\left(\widetilde{R}^{\mathcal{I}_1}|X_{new}, \widetilde{R}^{\mathcal{I}_1,i}|X_{new}\right) + O_p\left\{\eta_n^c + \rho_n^c + \eta_n + \rho_{2,n} + \Lambda + h_0 + N^{-1/2}\right\},
\end{aligned}
$$

where $\mathrm{d_{TV}}\left(\widetilde{R}^{\mathcal{I}_1}|X_{new}, \widetilde{R}^{\mathcal{I}_1,i}|X_{new}\right)$ denotes the total variation distance between conditional distribution, $\widetilde{R}^{\mathcal{I}_1}$ and $\widetilde{R}^{\mathcal{I}_1,i}$ conditional on X_{new}. Let B_i be the domain of conditional probability distributions $P_{(i)}(\cdot|X_{new})$ for each data point $Z_i = \{X_i, Y_i\}$, where $P_{(i)}(\cdot|X_{new}) = P^{(k)}(\cdot|X_{new})$ if Z_i belongs to the k-th source data. Using equation (4.1) in Levin and Peres (2017) and posterior drift assumption, it then follows that

$$
\begin{aligned}
\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \mathrm{d_{TV}}\left(\widetilde{R}^{\mathcal{I}_1}|X_{new}, \widetilde{R}^{\mathcal{I}_1,i}|X_{new}\right) \leq& 2\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \mathrm{d_{TV}}\left(Z_i|X_{new}, Z_{N+1}|X_{new}\right) \\
=&2\sum_{i=1}^{N+1} \widetilde{w}_i^{\mathcal{I}_1} \cdot \max_{(X,Y)\in B_i} \left|P_{(i)}\left(Y|X_{new}\right) - Q\left(Y|X_{new}\right)\right| \\
\leq&2h,
\end{aligned}
$$

where the third inequality is due to (3). Therefore, combining the above results, we obtain

$$
\Pr\left\{Y_{new} \notin \widehat{C}\left(X_{new}\right)|X_{new}\right\} \leq \alpha + O_p\left\{\eta_n^c + \rho_n^c + \eta_n + \rho_{2,n} + \Lambda + h' + N^{-1/2}\right\}.
$$

Next, we establish an upper bound on the coverage. Using the same arguments as for the lower bound on conditional coverage and the same proof strategy as for the upper bound in Theorem 9, we have

$$\Pr\left\{Y_{new} \in \widehat{C}\left(X_{new}\right)|X_{new}\right\} \leq 1-\alpha+\widetilde{w}_{N+1}^{\mathcal{I}_1}+O_p\left\{\eta_n^c+\rho_n^c+\eta_n+\rho_{2,n}+\Lambda+h'+N^{-1/2}\right\}.$$

Proof of Theorem 16. Using the Mean Value Theorem (MVT), we obtain

$$\left|\log P^{(k)}\left(Y|X\right)-\log Q\left(Y|X\right)\right| \leq \frac{1}{\min\left\{l_3,l_4\right\}}\left|P^{(k)}\left(Y|X\right)-Q\left(Y|X\right)\right|, \forall k \in \mathcal{A}_h.$$

Then, according to Assumptions 7-8 and Triangle's inequality, for $k \in \mathcal{A}_h$, we have

$$\begin{aligned}
\left|\widehat{cKL}^{(k)}\right| \leq & \sum_{i=1}^{n_k} \widehat{P}^{(k)}\left(Y_i|X_i\right)\left|\log\left\{\frac{\widehat{P}^{(k)}\left(Y_i|X_i\right)}{\widehat{Q}\left(Y_i|X_i\right)}\right\}\right| \\
\leq & \frac{1}{\min\left\{l_3,l_4\right\}}\sum_{i=1}^{n_k} \widehat{P}^{(k)}\left(Y_i|X_i\right)\left|\widehat{P}^{(k)}\left(Y_i|X_i\right)-\widehat{Q}\left(Y_i|X_i\right)\right| \\
\leq & \frac{1}{\min\left\{l_3,l_4\right\}}\sum_{i=1}^{n_k} \widehat{P}^{(k)}\left(Y_i|X_i\right)\left(\left|P^{(k)}\left(Y_i|X_i\right)-Q\left(Y_i|X_i\right)\right|\right. \\
& \left.+\left|\widehat{P}^{(k)}\left(Y_i|X_i\right)-P^{(k)}\left(Y_i|X_i\right)\right|+\left|\widehat{Q}\left(Y_i|X_i\right)-Q\left(Y_i|X_i\right)\right|\right) \\
\leq & \frac{1}{\min\left\{l_3,l_4\right\}}\sum_{i=1}^{n_k}\left\{P^{(k)}\left(Y_i|X_i\right)+\kappa h_p^{(k)}\right\}\left(h+\kappa h_p^{(k)}+\kappa h_q\right) \\
\leq & \frac{n_k\left(u_3+\kappa h_p^{(k)}\right)\left(h+\kappa h_p^{(k)}+\kappa h_q\right)}{\min\left\{l_3,l_4\right\}} \\
\leq & D_0,
\end{aligned}$$

simultaneously with probability at least $1-\max_{k\in\mathcal{A}_h}\{g_1^{(k)}(\kappa)+g_2(\kappa)\}$. On the other hand, using Assumption 8 and Triangle's inequality, for $k \in \mathcal{A}_h^c$, we know

$$\begin{aligned}
\left|\widehat{cKL}^{(k)}\right| \geq & \sum_{i=1}^{n_k} \widehat{P}^{(k)}\left(Y_i|X_i\right)\left|\widehat{P}^{(k)}\left(Y_i|X_i\right)-\widehat{Q}\left(Y_i|X_i\right)\right| \\
\geq & \sum_{i=1}^{n_k} \widehat{P}^{(k)}\left(Y_i|X_i\right)\left\{\left|P^{(k)}\left(Y_i|X_i\right)-Q\left(Y_i|X_i\right)\right|\right. \\
& \left.-\left|\widehat{P}^{(k)}\left(Y_i|X_i\right)-P^{(k)}\left(Y_i|X_i\right)\right|-\left|\widehat{Q}\left(Y_i|X_i\right)-Q\left(Y_i|X_i\right)\right|\right\} \\
\geq & \sum_{i=1}^{n_k}\left\{P^{(k)}\left(Y_i|X_i\right)-\kappa h_p^{(k)}\right\}\left(\underline{h}-\kappa h_p^{(k)}-\kappa h_q\right) \\
\geq & n_k\left(l_3-\kappa h_p^{(k)}\right)\left(\underline{h}-\kappa h_p^{(k)}-\kappa h_q\right) \\
\geq & D_0,
\end{aligned}$$

simultaneously with probability at least $1-\max_{k\in\mathcal{A}_h^c}\{g_1^{(k)}(\kappa)+g_2(\kappa)\}$, where D_0 is a constant. Combining the above results, we have

$$\begin{aligned}\Pr\left(\widehat{\mathcal{A}}\neq\mathcal{A}_h\right)\leq&\Pr\left\{\bigcup_{k\in\mathcal{A}_h}\left(\left|\widehat{cKL}^{(k)}\right|>D_0\right)\bigcup\bigcup_{k\in\mathcal{A}_h^c}\left(\left|\widehat{cKL}^{(k)}\right|\leq D_0\right)\right\}\\ \leq&\sum_{k\in\mathcal{A}_h}\Pr\left(\left|\widehat{cKL}^{(k)}\right|>D_0\right)+\sum_{k\in\mathcal{A}_h^c}\Pr\left(\left|\widehat{cKL}^{(k)}\right|\leq D_0\right)\\ \leq&|\mathcal{A}_h|\max_{k\in\mathcal{A}_h}\left\{g_1^{(k)}(\kappa)+g_2(\kappa)\right\}+|\mathcal{A}_h^c|\max_{k\in\mathcal{A}_h^c}\left\{g_1^{(k)}(\kappa)+g_2(\kappa)\right\}.\end{aligned}$$

For any given $\xi>0$, we can find constants $D_0'(\xi)$ and $\kappa'>0$ such that when $D_0=D_0'(\xi)$, $K\max_{k\in\mathcal{A}_h}\{g_1^{(k)}(\kappa')+g_2(\kappa')\}\leq\xi/2$, $K\max_{k\in\mathcal{A}_h^c}\{g_1^{(k)}(\kappa')+g_2(\kappa')\}<\xi/2$. Correspondingly, there exists $N'=N'(\xi)>0$, such that when $\min n_k>N'(\xi)$, $l_3>\kappa' h_p^{(k)}$ and $n_k(u_3+\kappa' h_p^{(k)})(h+\kappa' h_p^{(k)}+\kappa' h_q)/\min\{l_3,l_4\}$ is sufficiently small to satisfy

$$\frac{n_k\left(u_3+\kappa' h_p^{(k)}\right)\left(h+\kappa' h_p^{(k)}+\kappa' h_q\right)}{\min\{l_3,l_4\}}\leq D_0'(\xi)\leq n_k\left(l_3-\kappa' h_p^{(k)}\right)\left(\underline{h}-\kappa' h_p^{(k)}-\kappa' h_q\right).$$

Combining the above results, given any $\xi>0$, there are constants $D_0'(\xi)$ and $N'(\xi)>0$ such that if $D_0=D_0'(\xi)$ and $\min n_k>N'(\xi)$, the probability that $\widehat{\mathcal{A}}$ is not equal to $\mathcal{A}_h$ is less than or equal to ξ. We have completed the proof of this theorem.

Appendix D. Additional Simulation Studies: cKL vs. KL Divergence

In this section, we conducted some simulations studies to provide further insight into scenarios, where the marginal distribution of X differs substantially across source and target domains, while the regression function $\mu(X)$ remains largely similar across domains (achieved by setting $h=5$). Specifically, we modified the baseline data generation process described in Setting 1 in Section 5.1 with $h=5$ by altering the marginal distribution of covariates X across source and target datasets, while keeping the response generation mechanism unchanged. In the target domain, covariates are generated as:

$$\boldsymbol{x}^{(0)}\overset{\text{i.i.d.}}{\sim}\mathcal{N}\left(\mathbf{0},\boldsymbol{I}_p\right),$$

whereas in each source domain $k\in 1,\ldots,K$, covariates are generated as

$$\boldsymbol{x}^{(k)}\overset{\text{i.i.d.}}{\sim}\mathcal{N}\left(5\cdot\mathbf{1}_p,\boldsymbol{I}_p+\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right)$$

with $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0}_p,0.3^2\boldsymbol{I}_p)$. This construction induces strong covariate shifts in the form of mean differences across all dimensions.

For comparison, we evaluate two strategies for identifying informative source datasets. The first strategy employs cKL divergence to select the transfer set and applies our proposed TCP algorithms with and without weighting, denoted as TLUQ(cKL) and TLWQ(cKL), respectively. The second strategy uses full KL divergence for source selection, followed

by the same TCP algorithms with and without weighting, denoted as TLUQ(KL) and TLWQ(KL). As expected, cKL divergence successfully identifies all sources as transferable (since $P(Y|X)$ is aligned), whereas KL divergence fails due to differences in the marginal covariate distributions. The results, reported in Tables 8 and 9, corroborate our theoretical motivation: selecting sources via cKL divergence, which explicitly targets similarity in the conditional distribution $Y|X$, leads to more accurate estimation of the conditional mean function and consequently yields tighter conformal intervals with valid coverage.

Table 8: Comparison of average prediction interval width. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	TL+UQ(KL)	TL+WQ(KL)	TL+UQ(cKL)	TL+WQ(cKL)
25	7.806(1.827)	7.806(1.827)	4.082(0.988)	4.044(0.963)
50	6.280(1.373)	6.280(1.373)	3.491(0.599)	3.476(0.567)
100	4.637(0.789)	4.637(0.789)	3.378(0.388)	3.377(0.367)
200	3.957(0.437)	3.957(0.437)	3.220(0.325)	3.217(0.284)

Table 9: Comparison of average prediction interval coverage. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	TL+UQ(KL)	TL+WQ(KL)	TL+UQ(cKL)	TL+WQ(cKL)
25	0.932(0.008)	0.933(0.008)	0.912(0.009)	0.898(0.009)
50	0.917(0.009)	0.915(0.009)	0.894(0.010)	0.893(0.010)
100	0.891(0.010)	0.899(0.009)	0.897(0.009)	0.897(0.009)
200	0.918(0.009)	0.910(0.009)	0.903(0.009)	0.905(0.009)

Appendix E. Additional Simulation Studies: Robustness Properties of the Method

We build upon Setting 1 in Section 5.1, with the same covariate generation structure. That is, for target and source datasets, we simulate the covariates $\boldsymbol{x}^{(0)}$ and $\boldsymbol{x}^{(k)}$ $(k = 1, \ldots, K)$ independently from multivariate normal distributions: $\boldsymbol{x}^{(0)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma})$, $\boldsymbol{x}^{(k)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}_p, \boldsymbol{\Sigma} + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\right)$, where $\boldsymbol{\Sigma} = (0.5^{|i-j|})_{1\le i,j\le p}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_p, 0.3^2\boldsymbol{I}_p)$. However, unlike Setting 1, where both target and source response variables were generated from linear models, here we generate the target mean function according to a nonlinear mean function:

$$\mu_0(\boldsymbol{x}) = \sin\left(\boldsymbol{x}^\top \boldsymbol{\beta}_0\right),$$

where $\boldsymbol{\beta}_0 = (0.5, \ldots, 0.5)^\top \in \mathbb{R}^p$ with the first s entries being 0.5 and the remaining entries being 0. The response variable for the target data is then simulated as

$$Y_i^{(0)} = \mu_0\left(\boldsymbol{x}_i^{(0)}\right) + \varepsilon_i^{(0)}, \quad \varepsilon_i^{(0)} \sim \mathcal{N}(0, 1).$$

Source mean function are generated similarly as those in Setting 1:

$$\mu_k(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\theta}^{(k)}, \quad Y_i^{(k)} = \mu_k\left(\boldsymbol{x}_i^{(k)}\right) + \varepsilon_i^{(k)},$$

where $\boldsymbol{\theta}^{(k)} = \boldsymbol{\beta}_0 + (h/p) \cdot \boldsymbol{r}_p^{(k)}$ and $\boldsymbol{r}_p^{(k)}$ consists of i.i.d. Rademacher entries (taking values ± 1 with equal probability). Thus, the fitted regression model (linear) deviates from the true target mean function (nonlinear sin transformation), leading to model misspecification.

We apply the same conformal prediction procedures as in Section 5, including SCP, FS, TL, TL+UQ, and TL+WQ. As shown in Tables 10 and 11, TL+WQ achieves the smallest prediction set size while maintaining valid coverage. For example, at $n_0 = 25$, TL+WQ yields a set size of 4.117 with 89.2% coverage, outperforming SCP, which has a larger set size of 8.352. When compared with the results from the correctly specified Setting 1, the coverage and set sizes here show only a modest degradation, indicating that the proposed method is robust to model misspecification.

Table 10: Comparison of average prediction interval width. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	SCP	FS	TL	TL+UQ	TL+WQ
25	8.352(1.950)	5.187(0.800)	4.588(1.000)	4.164(0.880)	4.117(0.850)
50	6.718(1.450)	4.741(0.980)	4.286(0.900)	3.544(0.600)	3.535(0.590)
100	4.869(0.820)	4.084(0.750)	3.736(0.500)	3.451(0.420)	3.436(0.420)
200	4.155(0.470)	3.894(0.580)	3.630(0.360)	3.294(0.300)	3.277(0.290)

Table 11: Comparison of average prediction interval coverage. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	SCP	FS	TL	TL+UQ	TL+WQ
25	0.943(0.007)	0.928(0.008)	0.932(0.008)	0.881(0.010)	0.892(0.010)
50	0.926(0.008)	0.928(0.008)	0.925(0.008)	0.882(0.010)	0.889(0.010)
100	0.913(0.009)	0.912(0.009)	0.919(0.009)	0.893(0.009)	0.896(0.009)
200	0.910(0.009)	0.916(0.009)	0.911(0.009)	0.896(0.009)	0.897(0.009)

Appendix F. Additional Simulation Studies: TCP under Quantile Regression

Our method is not restricted to absolute residual—it naturally extends to Conformalized Quantile Regression (CQR) scores (see Romano et al. (2019)). In this case, the target population's α_τ-th quantile is defined as:

$$q_\tau^0(X^{(0)}) = \inf\left\{t \in \mathbb{R} : P^{(0)}\left(Y^{(0)} \leq t | X^{(0)}\right) \geq \alpha_\tau\right\}. \tag{14}$$

We determine the lower and upper quantiles by setting $\alpha_l = \alpha/2$ and $\alpha_h = 1 - \alpha/2$, allowing the calculation of the corresponding conditional quantile functions $q_l^0(X^{(0)})$ and $q_h^0(X^{(0)})$.

Furthermore, in the context of transfer learning, we include additional observations from the auxiliary source domains K, denoted $(X_i^{(k)}, Y_i^{(k)}) \sim P^{(k)}$, for $i = 1, \ldots, n_k$ and $k = 1, \ldots, K$. The τ-th conditional quantile function of the k-th source is defined as:

$$q_\tau^k(X^{(k)}) = \inf \left\{ t \in \mathbb{R} : P^{(k)} \left(Y^{(k)} \leq t | X^{(k)} \right) \geq \alpha_\tau \right\}. \tag{15}$$

Similarly, we compute the conditional quantile functions $q_l^k(X^{(k)})$ and $q_h^k(X^{(k)})$ for each source data at the levels of α_l and α_h, respectively. Building on these definitions, we propose a transfer learning CQR (TCQR) procedure, which is organized into three steps:

- **Step 1.** Estimate preliminary quantile regression functions $\widetilde{Q}_{l,\mathcal{I}_1}$ and $\widetilde{Q}_{h,\mathcal{I}_1}$ at levels $\alpha_l = \alpha/2$ and $\alpha_h = 1 - \alpha/2$ using both $\mathcal{S}_1$ and $\mathcal{I}_1$. Compute the empirical weighted $(1-\alpha)$ quantile of

  $$\widetilde{E}_i^1 = \max\{\widetilde{Q}_{l,\mathcal{I}_1}(X_i) - Y_i,\ Y_i - \widetilde{Q}_{h,\mathcal{I}_1}(X_i)\},$$

  denoted as $\widetilde{q}_{1-\alpha}^1$, based on source data $\mathcal{S}_1$.

- **Step 2.** Apply transfer learning quantile regression to obtain debiased estimators $\widehat{Q}_{l,\mathcal{I}_2}$ and $\widehat{Q}_{h,\mathcal{I}_2}$ at the same quantile levels, using both $\mathcal{S}_2$ and $\mathcal{I}_2$.

- **Step 3.** For a new covariate X_{new}, construct the TCQR prediction interval for Y_{new} as

  $$\widehat{C}\left(X_{new}\right) = \left[\widehat{Q}_{l,\mathcal{I}_2}(X_{new}) - \{\widetilde{q}_{1-\alpha}^1 + \Lambda(\mathcal{I}_1)\},\ \widehat{Q}_{h,\mathcal{I}_2}(X_{new}) + \{\widetilde{q}_{1-\alpha}^1 + \Lambda(\mathcal{I}_1)\} \right],$$

  where $\Lambda(\mathcal{I}_1)$ is the bias correction term. Its computation follows the same procedure as in Algorithm 1 in our manuscript, except that the residuals are replaced by CQR conformity scores derived from the estimated quantile regression functions.

To illustrate this extension, we consider additional simulation studies under CQR scores. Same design as in Setting 1 of our manuscript, but instead of estimating the conditional mean function, we compared five methods: Conformalized Quantile Regression (CQR), Few Short learning under quantile regression conformity scores (FSQR), Transfer Only under quantile regression conformity scores (TLQR), and our Transfer Learning under quantile regression conformity scores procedures with unweighted (TCQR+UQ) and weighted (TCQR+WQ) quantile calibration. As shown in Tables 12 and 13, TCQR+WQ achieves the smallest prediction set size while maintaining valid coverage. For example, at $n_0 = 50$, TCQR+WQ yields a set size of 4.973 with 89.3% coverage, outperforming CQR, which has a larger set size of 5,844. These results demonstrate that our TCP method remains effective and efficient under quantile regression conformity scores, offering a promising extension to discrete-response prediction problems.

Appendix G. Additional Simulation Studies: TCP with the Binary Outcomes

When the response is binary, the TCP method adapts by using transfer learning logistic regression (or another classifier) to estimate probabilities, and defining conformity scores in

Table 12: Comparison of average prediction interval width. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	CQR	FSQR	TLQR	TCQR+UQ	TCQR+WQ
50	5.844 (1.467)	5.777 (1.329)	5.123 (1.219)	4.982 (1.005)	4.973 (0.994)
100	5.303 (1.037)	5.219 (1.003)	4.753 (0.813)	4.713 (0.831)	4.705 (0.830)
150	5.102 (0.967)	4.921 (0.931)	4.721 (0.865)	4.656 (0.743)	4.651 (0.731)
200	5.041 (0.876)	4.734 (0.841)	4.632 (0.729)	4.571 (0.712)	4.564 (0.696)

Table 13: Comparison of average prediction interval coverage. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	CQR	FSQR	TLQR	TCQR+UQ	TCQR+WQ
50	0.946(0.007)	0.933(0.008)	0.878(0.010)	0.891(0.009)	0.893(0.009)
100	0.923(0.008)	0.929(0.008)	0.880(0.009)	0.893(0.009)	0.896(0.009)
150	0.887(0.010)	0.881(0.010)	0.883(0.010)	0.902(0.009)	0.903(0.009)
200	0.891(0.009)	0.893(0.009)	0.885(0.009)	0.896(0.009)	0.901(0.009)

terms of the misfit between predicted probabilities and each candidate label. In this case, same as the mean regression function, but with binary outcomes where the target label is generated from a Bernoulli distribution:

$$Y_i^{(0)} \sim \text{Bernoulli}\left(p^{(0)}\left(\boldsymbol{x}_i^{(0)}\right)\right), \quad p^{(0)}\left(\boldsymbol{x}_i^{(0)}\right) = \Pr\left(Y = 1 \mid X = \boldsymbol{x}_i^{(0)}\right).$$

Each source $k = 1, \ldots, K$ follows its own logistic mean function:

$$Y_i^{(k)} \sim \text{Bernoulli}\left(p^{(k)}\left(\boldsymbol{x}_i^{(k)}\right)\right), \quad p^{(k)}\left(\boldsymbol{x}_i^{(k)}\right) = \Pr\left(Y = 1 \mid X = \boldsymbol{x}_i^{(k)}\right).$$

When the response is binary, our TCP framework extends naturally by replacing the regression step with transfer learning generalized linear models (GLMs) (Tian and Feng, 2023), such as logistic regression.

- **Step 1.** Using $(\mathcal{S}_2, \mathcal{I}_2)$, we fit a transfer learning GLM to obtain the estimated success probability

$$\widehat{p}^{(0)}(\boldsymbol{x}) = \Pr(Y = 1 \mid X = \boldsymbol{x}).$$

- **Step 2.** Using $(\mathcal{S}_1, \mathcal{I}_1)$, we estimate preliminary success probilities $\widetilde{p}(\boldsymbol{x})$ and define the corresponding conformity scores for each candidate label $y \in 0, 1$ as

$$\widetilde{s}_y = |1 - \widetilde{p}(\boldsymbol{x}) \cdot y - (1 - \widetilde{p}(\boldsymbol{x})) \cdot (1 - y)|.$$

  This reduces to $\widetilde{s}_0 = |\widetilde{p}(\boldsymbol{x})|$ when $y = 0$ and $\widetilde{s}_1 = |1 - \widetilde{p}(\boldsymbol{x})|$ when $y = 1$. Then we find the empirical $(1 - \alpha)$ quantile of conformity score $\widetilde{s}_y$, defined as $\widetilde{q}_{1-\alpha}^{\mathcal{I}_1}$.

- **Step 3.** To construct a prediction set $\widehat{C}(\boldsymbol{x}_{new}) \subseteq \{0,1\}$, we first evaluate the conformity score for both candidate labels $y \in \{0,1\}$ using:

$$\widehat{s}_y = \left|1 - \widehat{p}^{(0)}(\boldsymbol{x}_{new}) \cdot y - (1 - \widehat{p}^{(0)}(\boldsymbol{x}_{new})) \cdot (1-y)\right|.$$

  The prediction set $\widehat{C}(\boldsymbol{x}_{new})$ contains all labels y such that $\widehat{s}_y \leq \widetilde{q}_{1-\alpha}^{\mathcal{I}_1} + \Lambda(\mathcal{I}_1)$, where $\Lambda(\mathcal{I}_1)$ is the bias correction term. Its computation follows the same procedure as in Algorithm 1 in our manuscript, except that the residuals are replaced by conformity scores derived from the estimated probabilities.

Same design as in Setting 1 of our manuscript, but instead of estimating the conditional mean function, we estimate the conditional class probabilities. To illustrate this, we have carried out additional simulations in the case of TCP under the binary classification setting. We compared five methods: Standard Split Conformal Prediction (SCP), Few Short (FS), Transfer Only (TL), and our Transfer Learning procedures with unweighted (TL+UQ) and weighted (TL+WQ) quantile calibration. As shown in Tables 14 and 15, TL+WQ achieves the smallest prediction set size while maintaining valid coverage. For example, at $n_0 = 100$, TL+WQ yields a set size of 1.460 with 89.3% coverage, outperforming SCP, which has a larger set size of 1.970. These results demonstrate that our TCP method remains effective and efficient for binary outcomes, offering a promising extension to discrete-response prediction problems.

Table 14: Comparison of average prediction interval width. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	SCP	FS	TL	TL+UQ	TL+WQ
100	1.970(0.733)	1.830(0.714)	1.670(0.609)	1.575(0.507)	1.460(0.500)
150	1.840(0.633)	1.660(0.604)	1.540(0.503)	1.525(0.511)	1.421(0.501)
200	1.640(0.624)	1.625(0.511)	1.505(0.513)	1.300(0.463)	1.280(0.454)
250	1.542(0.601)	1.503(0.512)	1.412(0.493)	1.200(0.462)	1.190(0.450)

Table 15: Comparison of average prediction interval coverage. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	SCP	FS	TL	TL+UQ	TL+WQ
100	0.953(0.007)	0.922(0.008)	0.918(0.008)	0.887(0.010)	0.893(0.009)
150	0.928(0.008)	0.915(0.009)	0.881(0.010)	0.890(0.009)	0.895(0.009)
200	0.881(0.010)	0.888(0.009)	0.879(0.010)	0.892(0.009)	0.905(0.009)
250	0.887(0.009)	0.893(0.009)	0.883(0.009)	0.901(0.009)	0.903(0.009)

Appendix H. Additional Simulation Studies: Comparison of TCP with Lee et al. (2023) and Duchi et al. (2024)

In this section, we further empirically validate our TCP method. We conducted some simulations comparing our TCP approach with representative baselines inspired by Duchi

et al. (2024) and hierarchical conformal prediction (Lee et al., 2023). Specifically, we denote the hierarchical conformal prediction method as HCP, implemented following Section 2.2 of Lee et al. (2023), and the resized multi-environment split conformal method as MHCP, corresponding to Algorithm 7 in Duchi et al. (2024). The data generation process follows Setting 1 in Section 5.1 in our manuscript. The results, summarized in Tables 16 and 17, demonstrate that in our setting, especially when the target sample size n_0 is small, TCP consistently achieves better coverage while producing tighter prediction intervals. Our method is thus both more flexible and more efficient in the transfer learning context. In contrast, when applied to non-hierarchical data, hierarchical methods such as HCP and MHCP typically yield intervals that are much wider than those from TCP, even though the coverage levels are comparable (92% vs. our 90% target). This inefficiency arises because hierarchical methods must account for potential groupwise dependencies, even in cases where no such structure exists, whereas our cKL-based weighting adapts directly to the empirical similarity between the source and target domains.

Table 16: Comparison of average prediction interval width. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	HCP(Lee)	MHCP(Duchi)	TL+UQ	TL+WQ
25	10.277(4.043)	6.865(1.904)	4.682(0.862)	4.459(0.952)
50	8.533(2.602)	4.417(1.454)	4.251(0.573)	4.042(0.513)
100	7.762(1.772)	4.179(0.892)	3.990(0.357)	3.783(0.370)
200	6.944(1.389)	3.931(0.426)	3.866(0.331)	3.401(0.214)

Table 17: Comparison of average prediction interval coverage. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	HCP(Lee)	MHCP(Duchi)	TL+UQ	TL+WQ
25	0.927(0.008)	0.923(0.008)	0.915(0.009)	0.897(0.010)
50	0.921(0.008)	0.928(0.008)	0.893(0.010)	0.893(0.010)
100	0.919(0.009)	0.917(0.009)	0.899(0.009)	0.896(0.009)
200	0.911(0.009)	0.913(0.009)	0.901(0.009)	0.899(0.009)

Our approach is both more flexible and more efficient in our target setting: Tables 16 and 17 show that when applied to non-hierarchical data, hierarchical methods like HCP and MHCP typically produce intervals much wider than ours while achieving comparable coverage (92% vs. our 90% target). The inefficiency arises because hierarchical methods must account for potential groupwise dependencies even when none exist, whereas our cKL-based weighting adapts directly to the empirical similarity between sources and target.

Appendix I. Details of CATE in Simulation Study

We simulate a linear regression setting to evaluate the performance of transfer conformal inference for estimating confidence intervals for the CATE. Similar to Setting 1, but we set

the number of the coefficients is 20 and the potential outcomes $Y^{(0)}(0), Y^{(0)}(1)$ for target data are defined as:

$$\mu^{(0)}_{control}(\boldsymbol{x}^{(0)}) = \boldsymbol{x}^{(0)}\boldsymbol{\beta} + \varepsilon, \quad \mu^{(0)}_{treatment}(\boldsymbol{x}^{(0)}) = \mu^{(0)}_{control}(\boldsymbol{x}^{(0)}) + 1 + 0.5\sin\left(x_3^{(0)}\right) + \varepsilon,$$

The binary treatment assignment $T \in \{0, 1\}$ is generated according to a logistic model

$$\Pr(T = 1 \mid X) = \frac{1}{1 + \exp\left(-x_1^{(0)} + 0.5x_2^{(0)}\right)},$$

where only the first two covariates contribute to the propensity score. Then the observed outcome is defined as

$$Y^{(0)} = T \cdot \mu^{(0)}_{treatment}(\boldsymbol{x}^{(0)}) + (1 - T) \cdot \mu^{(0)}_{control}(\boldsymbol{x}^{(0)}).$$

For the k-th source dataset, the potential outcomes are defined analogously as

$$\mu^{(k)}_{control}(\boldsymbol{x}^{(k)}) = \boldsymbol{x}^{(0)}\boldsymbol{\theta}^{(k)} + \varepsilon, \quad \mu^{(k)}_{treatment}(\boldsymbol{x}^{(k)}) = \mu^{(k)}_{control}(\boldsymbol{x}^{(k)}) + 1 + 0.5\sin\left(x_3^{(k)}\right) + \varepsilon,$$

and the binary treatment assignment $T \in \{0, 1\}$ is generated according to a logistic model

$$\Pr(T = 1 \mid X) = \frac{1}{1 + \exp\left(-x_1^{(k)} + 0.5x_2^{(k)}\right)}.$$

Then the observed outcomes of the k-th source dataset are similarly obtained as

$$Y^{(k)} = T \cdot \mu^{(k)}_{treatment}(\boldsymbol{x}^{(k)}) + (1 - T) \cdot \mu^{(k)}_{control}(\boldsymbol{x}^{(k)}).$$

We aim to construct valid and efficient prediction intervals for the individual treatment effect $\tau(x) = Y^{(0)}(1) - Y^{(0)}(0) = \mu^{(0)}_{treatment}(x) - \mu^{(0)}_{control}(x)$ using our method. To evaluate performance, we compare five methods: SCP, FS, TL, TL+UQ, and TL+WQ. The results, summarized in Tables 18 and 19, show that transfer-based methods (TL, TL+UQ, TL+WQ) consistently yield narrower intervals than SCP and FS, while maintaining coverage close to the nominal level. Notably, TL+WQ consistently yields the narrowest intervals while preserving coverage, showing that weighted quantile adjustment leads to more informative and reliable inference.

Table 18: Comparison of average prediction interval width. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	SCP	FS	TL	TL+UQ	TL+WQ
25	13.234(4.034)	9.721(3.211)	8.422(2.899)	7.334(2.445)	6.928(2.157)
50	10.313(3.102)	8.567(2.731)	7.753(2.463)	7.063(2.118)	6.733(1.965)
100	9.778(2.881)	8.184(2.510)	7.216(2.207)	6.822(2.034)	6.519(1.932)
200	7.122(2.218)	7.131(2.233)	6.937(2.104)	6.637(1.921)	6.448(1.899)

Table 19: Comparison of average prediction interval coverage. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	SCP	FS	TL	TL+UQ	TL+WQ
25	0.953(0.007)	0.931(0.008)	0.924(0.008)	0.889(0.010)	0.912(0.009)
50	0.931(0.008)	0.924(0.008)	0.921(0.008)	0.891(0.010)	0.907(0.009)
100	0.927(0.008)	0.915(0.009)	0.918(0.009)	0.893(0.009)	0.895(0.009)
200	0.912(0.009)	0.912(0.009)	0.909(0.009)	0.896(0.009)	0.903(0.009)

Appendix J. Additional Simulation Studies: Validity of Conditional Coverage

To complement the theory, we have conducted a simulation study under Setting 1 (Section 5.1), focusing on the conditional coverage at a fixed test point $X_{new} = \mathbf{1}_p$ and comparing five procedures: SCP, FS, TL, TL+UQ, and TL+WQ. We compute the conditional coverage rate and average prediction interval width at this specific covariate value. The results in Tables 20 and 21 confirm that our method maintains near-nominal conditional coverage while yielding narrower intervals compared to baselines, validating its practical robustness at specific covariate levels.

Table 20: Comparison of average prediction interval width under group-conditional coverage. Results are averaged over 1000 replications, with subgroup membership defined by bins of X_1.

n_0	SCP	FS	TL	TL+UQ	TL+WQ
25	8.731(1.944)	5.289(0.877)	4.583(1.012)	4.321(0.901)	4.217(0.864)
50	6.882(1.521)	4.812(0.933)	4.303(0.897)	3.912(0.652)	3.633(0.581)
100	5.244(0.899)	4.313(0.762)	3.951(0.533)	3.602(0.482)	3.492(0.427)
200	4.557(0.641)	3.981(0.629)	3.572(0.374)	3.412(0.362)	3.280(0.318)

Table 21: Comparison of average prediction interval coverage under group-conditional coverage. Results are averaged over 1000 replications, with subgroup membership defined by bins of X_1.

n_0	SCP	FS	TL	TL+UQ	TL+WQ
25	0.939(0.008)	0.927(0.008)	0.922(0.009)	0.888(0.010)	0.896(0.009)
50	0.932(0.008)	0.926(0.008)	0.925(0.008)	0.890(0.010)	0.899(0.010)
100	0.924(0.008)	0.913(0.009)	0.920(0.009)	0.888(0.010)	0.902(0.009)
200	0.945(0.007)	0.910(0.009)	0.918(0.009)	0.895(0.010)	0.906(0.009)

Appendix K. Additional Simulation Studies: Validity of Group-Conditional Coverage

To complement the theoretical results on group-conditional coverage, we have carried out a simulation study in which the test covariates were divided into predefined groups and coverage was assessed within each group. This setup reflects the guarantee

$$\Pr\{Y_{new} \in \widehat{C}(X_{new}) \mid X_{new} \in B\} \geq 1 - \alpha,$$

for all subsets $B \in \mathcal{B}$ of the covariate space, as formalized in Hore and Barber (2025). The simulations use the same baseline data-generating mechanism as in Setting 1 of Section 5.1. To evaluate group-conditional validity, we stratify the test feature space by the first coordinate X_1, which captures a meaningful direction of variation across datasets. Four disjoint groups are defined:

$$B_1 = \{X : X_1 < -1\}, \quad B_2 = \{-1 \leq X_1 < 0\},$$
$$B_3 = \{0 \leq X_1 < 1\}, \quad B_4 = \{X : X_1 \geq 1\}.$$

Each test sample is placed into one of these bins, and coverage is evaluated separately within each B_j.

In each replication, we generate 1000 test points from the target distribution. For a given method, the empirical coverage within group B_j is computed as

$$\widehat{\mathrm{Cov}}_r(B_j) = \frac{1}{|T_r(B_j)|} \sum_{i \in T_r(B_j)} \mathbf{1}\{Y_i \in \widehat{C}(X_i)\},$$

where $T_r(B_j)$ is the set of test indices assigned to group B_j in replication r. This yields four coverage estimates per replication. Averaging over $R = 1000$ replications gives the final estimate for each group:

$$\widehat{\mathrm{Cov}}(B_j) = \frac{1}{R} \sum_{r=1}^{R} \widehat{\mathrm{Cov}}_r(B_j).$$

Prediction interval widths are computed in the same way, averaged within each group and across replications. We compared five procedures: SCP, FS, TL, TL+UQ, and TL+WQ. Tables 22 and 23 present the results. Across all target sample sizes and groups, TCP achieves coverage close to the nominal 90% level while producing substantially narrower intervals than SCP, FS, and other baselines. These findings reinforce the theoretical guarantees: even when conditioning on broad subgroups of the covariate space, TCP delivers both validity and efficiency.

Appendix L. Additional Simulation Studies: Extension to HPD-Split Nonconformity Measures

In this section, we further illustrate the flexibility of our TCP framework by extending it to a density-based nonconformity measure based on highest predictive density (HPD)

Table 22: Comparison of average prediction interval width. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	SCP	FS	TL	TL+UQ	TL+WQ
25	8.042(1.933)	5.022(0.821)	4.367(1.078)	4.264(0.931)	4.125(0.862)
50	6.624(1.431)	4.735(0.971)	4.231(0.889)	3.747(0.631)	3.503(0.581)
100	4.773(0.810)	4.077(0.731)	3.722(0.531)	3.483(0.478)	3.471(0.422)
200	4.266(0.577)	3.822(0.598)	3.392(0.3720)	3.282(0.359)	3.211(0.319)

Table 23: Comparison of average prediction interval coverage. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	SCP	FS	TL	TL+UQ	TL+WQ
25	0.937(0.008)	0.928(0.008)	0.921(0.009)	0.883(0.010)	0.891(0.010)
50	0.931(0.008)	0.927(0.008)	0.924(0.008)	0.887(0.010)	0.897(0.010)
100	0.923(0.008)	0.911(0.009)	0.919(0.009)	0.885(0.010)	0.904(0.009)
200	0.944(0.007)	0.908(0.009)	0.916(0.009)	0.893(0.010)	0.903(0.009)

regions. This extension is motivated by recent developments that connect conformal prediction to the highest density regions and distribution-free uncertainty quantification (Izbicki et al., 2022; Dheur et al., 2024). Unlike residual- or quantile-based scores, HPD-based nonconformity measures leverage an estimated conditional density to assess how plausible a candidate response value is under the fitted predictive distribution.

Specifically, we adopt the same data-generating mechanism as in Setting 1 (linear, high-dimensional regression) in Section 5.1 to isolate the effect of the nonconformity score while keeping all other aspects of the simulation unchanged. Target covariates $\boldsymbol{x}^{(0)}$ are generated independently from $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}_{ij} = 0.5^{|i-j|}$, while source covariates $\boldsymbol{x}^{(k)}$ are drawn from $\mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma} + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top)$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_p, 0.3^2 \boldsymbol{I}_p)$. The target response satisfies:

$$Y_i^{(0)} = \boldsymbol{x}_i^{(0)\top}\boldsymbol{\beta} + \varepsilon_i^{(0)}, \qquad \varepsilon_i^{(0)} \sim \mathcal{N}(0,1),$$

with a sparse coefficient vector $\boldsymbol{\beta}$, while source responses are generated analogously with perturbed coefficients, as described in Setting 1. Given a fitted regression function $\widehat{\mu}(x)$, we construct a plug-in conditional density estimator

$$\widehat{f}(y \mid x) = \phi\left(\frac{y - \widehat{\mu}(x)}{\widehat{\sigma}}\right) \Big/ \widehat{\sigma},$$

where $\widehat{\sigma}$ is estimated from training residuals and $\phi(\cdot)$ denotes the standard normal density. Under this Gaussian plug-in model, the HPD nonconformity score admits the following closed-form:

$$\widehat{s}_{\mathrm{HPD}}(x, y) = 2\Phi\left(\frac{|y - \widehat{\mu}(x)|}{\widehat{\sigma}}\right) - 1,$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Subsequently, we compute HPD scores on the target calibration set and obtain the $(1-\alpha)$ empirical quantile

$\widehat{q}_{1-\alpha}$. The resulting HPD-split conformal prediction set for a new covariate value X_{new} is

$$\widehat{C}_{\text{HPD}}(X_{\text{new}}) = \{y : \widehat{s}_{\text{HPD}}(X_{\text{new}}, y) \leq \widehat{q}_{1-\alpha}\}.$$

We consider two implementations: a target-only HPD-split baseline and an HPD-based TCP method, and compare five procedures: SCP, FS, TL, TL+UQ, and TL+WQ. The results are reported in Tables 24 and 25. Across all target sample sizes, the HPD-based TCP method achieves empirical coverage close to the nominal level $1 - \alpha = 0.9$. Compared to the target-only HPD-split baseline, the HPD-based TCP intervals are consistently shorter, particularly when the target sample size n_0 is small. This behavior mirrors what we observe for residual- and quantile-based nonconformity scores, confirming that the performance gains of TCP are not tied to a specific choice of nonconformity measure. Overall, these results demonstrate that the proposed framework naturally extends to HPD-based scores, providing a density-based alternative aligned with recent developments in conformal prediction while preserving the benefits of transfer learning.

Table 24: Comparison of average prediction interval width. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	SCP	FS	TL	TL+UQ	TL+WQ
25	5.867(1.466)	4.166(1.033)	3.923(0.889)	4.002(0.352)	3.391(0.337)
50	4.991(1.471)	4.003(1.021)	3.777(0.661)	3.661(0.344)	3.170(0.320)
100	4.054(1.470)	3.713(1.013)	3.482(0.663)	3.332(0.351)	3.129(0.327)
200	3.894(1.435)	3.657(1.008)	3.482(0.663)	3.306(0.353)	3.022(0.314)

Table 25: Comparison of average prediction interval coverage. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	SCP	FS	TL	TL+UQ	TL+WQ
25	0.932(0.012)	0.927(0.009)	0.922(0.009)	0.885(0.008)	0.891(0.011)
50	0.924(0.010)	0.924(0.009)	0.918(0.009)	0.891(0.009)	0.894(0.009)
100	0.919(0.009)	0.915(0.009)	0.912(0.010)	0.894(0.010)	0.899(0.009)
200	0.914(0.007)	0.914(0.008)	0.910(0.009)	0.896(0.009)	0.902(0.007)

Appendix M. Additional Simulation Studies: When KL Divergence Outperforms cKL

In this section, we clarify that cKL is well suited to settings where conditional discrepancies dominate and sufficient covariate overlap holds, whereas KL divergence on the joint distribution (X, Y) may be preferable when transferability is constrained by overlap failure, substantial variance inflation, or heavy-tail behavior. To illustrate, we include a dedicated simulation study comparing cKL- and KL-based source criteria. Specifically, we fix the dimension p and sample sizes n_0 (target) and $\{n_k\}_{k=1}^K$ (sources) as in Setting 1. Target

covariates are generated as:

$$X^{(0)} \in \mathbb{R}^p, \qquad X^{(0)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_p),$$

with responses

$$Y^{(0)} = \sin\big(X_1^{(0)}\big) + \xi^{(0)}, \qquad \xi^{(0)} \sim \mathcal{N}(0, \sigma^2_{\text{noise}}).$$

Among the K sources, the first K_a are informative and share the same conditional mean, differing only through mild covariance perturbations. For $k \le K_a$, we generate

$$X^{(k)} = Z^{(k)} + u^{(k)}\varepsilon^\top, \qquad Z^{(k)} \sim \mathcal{N}(0, I_p), \quad u^{(k)} \sim \mathcal{N}(0,1), \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_p),$$

with

$$Y^{(k)} = \sin\big(X_1^{(k)}\big) + \xi^{(k)}, \qquad \xi^{(k)} \sim \mathcal{N}(0, \sigma^2_{\text{noise}}).$$

The remaining sources are non-informative and exhibit weak overlap through both large mean shifts and variance inflation:

$$X^{(k)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_{\text{shift}}\mathbf{1}_p,\ \sigma_X^2 I_p), \qquad k > K_a,$$

with large σ_X^2. To induce tail-driven uncertainty, we contaminate responses via

$$\begin{aligned} Y^{(k)} &= \sin\big(X_1^{(k)}\big) + T^{(k)} + \xi^{(k)}, \qquad \xi^{(k)} \sim \mathcal{N}(0, \sigma^2_{\text{noise}}), \\ T^{(k)} &= B^{(k)} Z^{(k)}, \qquad B^{(k)} \sim \text{Bernoulli}(\pi_{\text{tail}}), \quad Z^{(k)} \sim \mathcal{N}(0, \tau^2_{\text{tail}}), \end{aligned}$$

Unless otherwise stated, we set $K = 10$, $K_a = 5$, $\sigma_{\text{noise}} = 0.5$, $\sigma_\varepsilon = 0.5$, $\mu_{\text{shift}} = 6$, $\sigma_X^2 = 6$, $\pi_{\text{tail}} = 0.25$, and $\tau_{\text{tail}} = 4$. This design yields a regime in which (i) the high-probability regions of $X^{(k)}$ for $k > K_a$ are far from that of $X^{(0)}$ in both location and scale, and (ii) rare but large response outliers can dominate calibration quantiles when such sources are included.

We compare two source criteria strategies: cKL-based criterion followed by our TCP methods (TLUQ(cKL), TLWQ(cKL)), and KL-based criterion followed by the same procedures (TLUQ(KL), TLWQ(KL)). Results in Tables 26 and 27 show that, across all target sample sizes, KL-based criterion yields consistently shorter prediction intervals while maintaining empirical coverage near the nominal level. Intuitively, cKL may admit shifted or heavy-tailed sources because conditional similarity is assessed locally and can underrepresent regions with weak covariate overlap; once such sources are included, contamination inflates the calibration quantiles and leads to wider prediction intervals. In contrast, KL-based criterion penalizes both marginal shifts in X and tail behavior in Y, enabling more effective exclusion of non-informative sources in this regime. These findings clarify that cKL is advantageous when conditional mismatch dominates under sufficient overlap, whereas KL on (X, Y) can outperform cKL when variance shifts, overlap failure, and heavy-tailed responses are the primary barriers to transfer.

Table 26: Comparison of average prediction interval width. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	TL+UQ(KL)	TL+WQ(KL)	TL+UQ(cKL)	TL+WQ(cKL)
25	2.549(0.519)	2.385(0.465)	6.012(0.831)	3.564(0.450)
50	2.410(0.448)	2.209(0.408)	5.116(0.788)	2.977(0.431)
100	2.334(0.442)	2.155(0.412)	4.213(0.637)	2.691(0.433)
200	2.301(0.401)	2.111(0.407)	4.005(0.616)	2.487(0.429)

Table 27: Comparison of average prediction interval coverage. All quantities have been averaged over 1000 independent trials, and the standard errors are in parentheses.

n_0	TL+UQ(KL)	TL+WQ(KL)	TL+UQ(cKL)	TL+WQ(cKL)
25	0.933(0.008)	0.910(0.007)	0.941(0.012)	0.932(0.011)
50	0.927(0.011)	0.913(0.008)	0.936(0.010)	0.934(0.010)
100	0.911(0.009)	0.908(0.008)	0.927(0.009)	0.917(0.009)
200	0.915(0.009)	0.906(0.008)	0.923(0.009)	0.918(0.009)

References

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

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

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

Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

T Tony Cai and Hongming Pu. Transfer learning for nonparametric regression: Non-asymptotic minimax analysis and adaptive procedure. *arXiv preprint arXiv:0000.0000*, 2022.

T TONY Cai and HONGJI Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128, 2021.

Emmanuel Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45, 2023.

John Cherian and Lenny Bronner. How the washington post estimates outstanding votes for the 2020 presidential election, 2020.

Victor Chernozhukov, Kaspar Wüthrich, and Zhu Yinchu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On learning theory*, pages 732–749. PMLR, 2018.

Victor Dheur, Tanguy Bosser, Rafael Izbicki, and Souhaib Ben Taieb. Distribution-free conformal joint prediction regions for neural marked temporal point processes. *Machine Learning*, 113(9):7055–7102, 2024.

John C Duchi, Suyash Gupta, Kuanhao Jiang, and Pragya Sur. Predictive inference in multi-environment scenarios. *arXiv preprint arXiv:2403.16336*, 2024.

Robin Dunn, Larry Wasserman, and Aaditya Ramdas. Distribution-free prediction sets for two-layer hierarchical models. *Journal of the American Statistical Association*, 118(544): 2491–2502, 2023.

Jianqing Fan. Local linear regression smoothers and their minimax efficiencies. *The annals of Statistics*, pages 196–216, 1993.

Clara Fannjiang, Stephen Bates, Anastasios N Angelopoulos, Jennifer Listgarten, and Michael I Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Few-shot conformal prediction with auxiliary tasks. In *International Conference on Machine Learning*, pages 3329–3339. PMLR, 2021.

Chenyin Gao, Peter B Gilbert, and Larry Han. On the role of surrogates in conformal inference of individual causal effects. *arXiv preprint arXiv:2412.12365*, 2024.

Etienne Gauthier, Francis Bach, and Michael I Jordan. Backward conformal prediction. *arXiv preprint arXiv:2505.13732*, 2025.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Peter Hoff. Bayes-optimal prediction with frequentist coverage control. *Bernoulli*, 29(2): 901–928, 2023.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):549–578, 2025.

Xiaonan Hu and Xinyu Zhang. Optimal parameter-transfer learning by semiparametric model averaging. *Journal of Machine Learning Research*, 24(358):1–53, 2023.

Kexin Huang, Ying Jin, Emmanuel Candes, and Jure Leskovec. Uncertainty quantification over graph with conformalized graph neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Rafael Izbicki, Gilson T Shimizu, and Rafael B Stern. Flexible distribution-free conditional predictive bands using density estimators. *arXiv preprint arXiv:1910.05575*, 2019.

Rafael Izbicki, Gilson Shimizu, and Rafael B Stern. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87):1–32, 2022.

Danijel Kivaranovic, Kory D Johnson, and Hannes Leeb. Adaptive, distribution-free prediction intervals for deep networks. In *International Conference on Artificial Intelligence and Statistics*, pages 4346–4356. PMLR, 2020.

Yonghoon Lee, Rina Foygel Barber, and Rebecca Willett. Distribution-free inference with hierarchical data. *arXiv preprint arXiv:2306.06342*, 2023.

Jing Lei, Alessandro Rinaldo, and Larry Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74:29–43, 2015.

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

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

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022.

Henrik Linusson, Ulf Johansson, Henrik Boström, and Tuve Löfström. Efficiency comparison of unstable transductive and inductive conformal classifiers. In *Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings 10*, pages 261–270. Springer, 2014.

Ruiqi Liu, Kexuan Li, and Zuofeng Shang. A computationally efficient classification algorithm in posterior drift model: Phase transition and minimax adaptivity. *arXiv e-prints*, pages arXiv–2011, 2020.

Yi Liu, Alexander W Levis, Sharon-Lise Normand, and Larry Han. Multi-source conformal inference under distribution shift. *Proceedings of machine learning research*, 235:31344, 2024.

Subha Maity, Diptavo Dutta, Jonathan Terhorst, Yuekai Sun, and Moulinath Banerjee. A linear adjustment-based approach to posterior drift in transfer learning. *Biometrika*, 111 (1):31–50, 2024.

Roberto I Oliveira, Paulo Orenstein, Thiago Ramos, and João Vitor Romano. Split conformal prediction for dependent data. *arXiv preprint arXiv:2203.15885*, 2022.

Roberto I Oliveira, Paulo Orenstein, Thiago Ramos, and Joao Vitor Romano. Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, 25 (225):1–38, 2024.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pages 345–356. Springer, 2002.

Henry WJ Reeve, Timothy I Cannings, and Richard J Samworth. Adaptive transfer learning. *The Annals of Statistics*, 49(6):3618–3649, 2021.

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

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551, 2020.

Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34:6304–6315, 2021.

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

Simon J Sheather and Michael C Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):683–690, 1991.

Aldo Solari and Vera Djordjilović. Multi split conformal prediction. *Statistics & Probability Letters*, 184:109395, 2022.

David Stutz, Ali Taylan Cemgil, Arnaud Doucet, et al. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*, 2021.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697, 2023.

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

Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Sara A Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

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

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

Ran Xie, Rina Foygel Barber, and Emmanuel J Candès. Boosted conformal prediction intervals. *arXiv preprint arXiv:2406.07449*, 2024.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866. PMLR, 2022.

Yao Zhang and Emmanuel J Candès. Posterior conformal prediction. *arXiv preprint arXiv:2409.19712*, 2024.