Training-Conditional Coverage Bounds under Covariate Shift

Anonymous authors
Paper under double-blind review

Abstract

Conformal prediction methodology has recently been extended to the covariate shift setting, where the distribution of covariates differs between training and test data. While existing results ensure that the prediction sets from these methods achieve marginal coverage above a nominal level, their coverage rate conditional on the training dataset—referred to as training-conditional coverage—remains unexplored. In this paper, we address this gap by deriving upper bounds on the tail of the training-conditional coverage distribution, offering probably approximately correct (PAC) guarantees for these methods. Our results characterize the reliability of the prediction sets in terms of the severity of distributional changes and the size of the training dataset.

1 Introduction

Conformal prediction is a framework for constructing *distribution-free* and *model-agnostic* predictive confidence regions under the exchangeability assumption for the training and test samples (Vovk et al., 2005) (also see Shafer and Vovk (2008); Vovk et al. (2009); Vovk (2012)). Let $((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n), (X_{\text{test}}, Y_{\text{test}}))$ denote a tuple of exchangeable data points, consisting of a training sequence of n samples $\mathbf{D}_n := ((X_1, Y_1), \ldots, (X_n, Y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ and one test sample $(X_{\text{test}}, Y_{\text{test}}) \in \mathcal{X} \times \mathcal{Y}$. For a fixed error rate level α, the conformal prediction framework, provides a prediction set $\hat{C}_\alpha(X_{\text{test}})$ for Y_{test} that satisfies

$$\mathbb{P}\left(Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right) \geq 1 - \alpha, \tag{1}$$

where $\hat{C}_\alpha : \mathcal{X} \to 2^{\mathcal{Y}}$ is a data-dependent map. This type of guarantee is referred to as *marginal* coverage, as it is averaged over both the training and test data.

One natural direction to stronger results is to provide a coverage guarantee *conditional* on the test data point X_{test}, i.e.,

$$\mathbb{P}\left(Y_{\text{test}} \in \hat{\mathcal{C}}_\alpha(X_{\text{test}}) | X_{\text{test}}\right) \geq 1 - \alpha.$$

However, when X_{test} has a continuous distribution, it has been shown in Vovk (2012); Foygel Barber et al. (2021); Lei and Wasserman (2014) that it is *impossible* to obtain (non-trivial) distribution-free prediction regions $\hat{\mathcal{C}}_\alpha(x)$ in the finite-sample regime; relaxed versions of this type of guarantee have been extensively studied (see Jung et al. (2022); Gibbs et al. (2023); Vovk et al. (2003) and references therein). As a different approach to stronger guarantees, several results (e.g., Vovk (2012); Bian and Barber (2023)) have been reported on the *training-conditional* guarantee by conditioning on $\mathbf{D}_n$, which is also more appealing than the marginal guarantee as can be seen below. Define the following miscoverage rate as a function of the training data,

$$P_e(\mathbf{D}_n) := \mathbb{P}\left(Y_{\text{test}} \notin \hat{\mathcal{C}}_\alpha(X_{\text{test}}) | \mathbf{D}_n\right).$$

Note that the marginal coverage in (1) is equivalent to $\mathbb{E}[P_e(\mathbf{D}_n)] \leq \alpha$. The training-conditional guarantees concern the concentration of the conditional error rate below the nominal level α and they have the following form, for some small $\delta > 0$,

$$\mathbb{P}\left(P_e(\mathbf{D}_n) \geq \alpha\right) \leq \delta$$

or its asymptotic variants. Roughly speaking, this guarantee means that the $(1-\alpha)$-level coverage lower bounds hold for a *generic* dataset.

Recently, Barber et al. (2021) proposed modified versions of jackknife and cross validation (CV), namely *jackknife+* and *CV+*, which can be used to compute conformal prediction sets. For the K-fold CV+ with ℓ samples in each fold, the training-conditional coverage bound

$$\mathbb{P}\left(P_e(\mathbf{D}_n) \geq 2\alpha + \sqrt{2\log(K/\delta)/\ell}\right) \leq \delta \tag{2}$$

is established in Bian and Barber (2023). Additionally, Liang and Barber (2023) proposed training-conditional coverage bounds for jackknife+ and full conformal prediction sets under the assumption that the training algorithm is symmetric and satisfies certain stability conditions (see Section 4 for more details). In this line of research, samples are assumed to be i.i.d., which is not only exchangeable but also ergodic and admits some nice concentration properties. However, this assumption can be violated in the application. In particular, the input data distribution during deployment can differ from the distribution observed during training. This phenomenon is called *distribution shift* and it is a crucial problem in trustworthy machine learning (see Section 2.3). In this regard, split and full conformal prediction methods as the central methods for distribution-free uncertainty quantification of black-box models have recently been extended to handle a popular type of distribution shift called *covariate shift* (Tibshirani et al., 2019). In the covariate shift setting, the distribution of covariates in the test data differs from the one observed in the training data, but, the conditional distribution of the response given the features remains the same across the training and test populations. A similar extension has been made for the jackknife+ method Prinster et al. (2022).

Despite these significant progress on handling distribution shift, they primarily focus on marginal coverage. Although the weighted conformal prediction methods (proposed in Tibshirani et al. (2019) and Prinster et al. (2022)) are guaranteed to keep the marginal coverage rate above the nominal level, they often reduce the concentration of the training-conditional coverage rate $\mathbb{P}(Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})|\mathbf{D}_n)$ (i.e., coverage rate conditional on the training data $\mathbf{D}_n$) around the nominal level $1-\alpha$. In particular, the weighting scheme leads to heavier tails for the training-conditional coverage rate. This phenomenon is illustrated in Figure 1, where ordinary and weighted conformal methods with 80% target coverage rate ($\alpha = 0.2$) are used for constructing prediction intervals under exchangeability and distribution shift settings, respectively. The distribution shift is introduced artificially via resampling according to an exponential tilt as in Section 2.3 from Tibshirani et al. (2019). It can be observed that the tails of the training conditional coverage get heavier when the weighting scheme is used to handle the distribution shift. In this paper, we explore the quality of the weighted conformal prediction sets by computing upper bounds on the tails of the training-conditional coverage distribution, quantifying the relationship between the tail behavior and the extent of distribution change. In other words, we examine the efficiency of weighted conformal prediction methods for a generic training dataset under distribution shift.

We present concentration bounds of the training-conditional coverage for weighted jackknife+ (JAW) (Prinster et al., 2022), full, and split conformal methods. Regarding the training algorithm, no assumption is made for the split conformal method. However, full conformal and jackknife+ methods are analyzed under the assumption of *uniform stability* as explained below. Let $\hat{\mu}_{\mathbf{D}_n}$ denote the predictor function estimated using the training data $\mathbf{D}_n$. A training algorithm is called uniformly stable if,

$$\|\hat{\mu}_{\mathbf{D}_n} - \hat{\mu}_{\mathbf{D}'_n}\|_\infty \leq \beta \tag{3}$$

with $\beta = O(1/n)$ for any two datasets $(\mathbf{D}_n, \mathbf{D}'_n)$ differing in one (training) data sample (Bousquet and Elisseeff, 2002). This is a stronger notion of algorithmic stability than the (m, n)-stability assumed in Liang and Barber (2023). Nevertheless, uniform stability is satisfied by a class of regression models known as RKHS regression (Bousquet and Elisseeff, 2002), i.e., regularized empirical risk minimization over an RKHS (Paulsen and Raghupathi, 2016; Schölkopf and Smola, 2002). Examples of reproducing kernel Hilbert spaces are certain Sobolev spaces of smooth functions (Wahba, 1990).

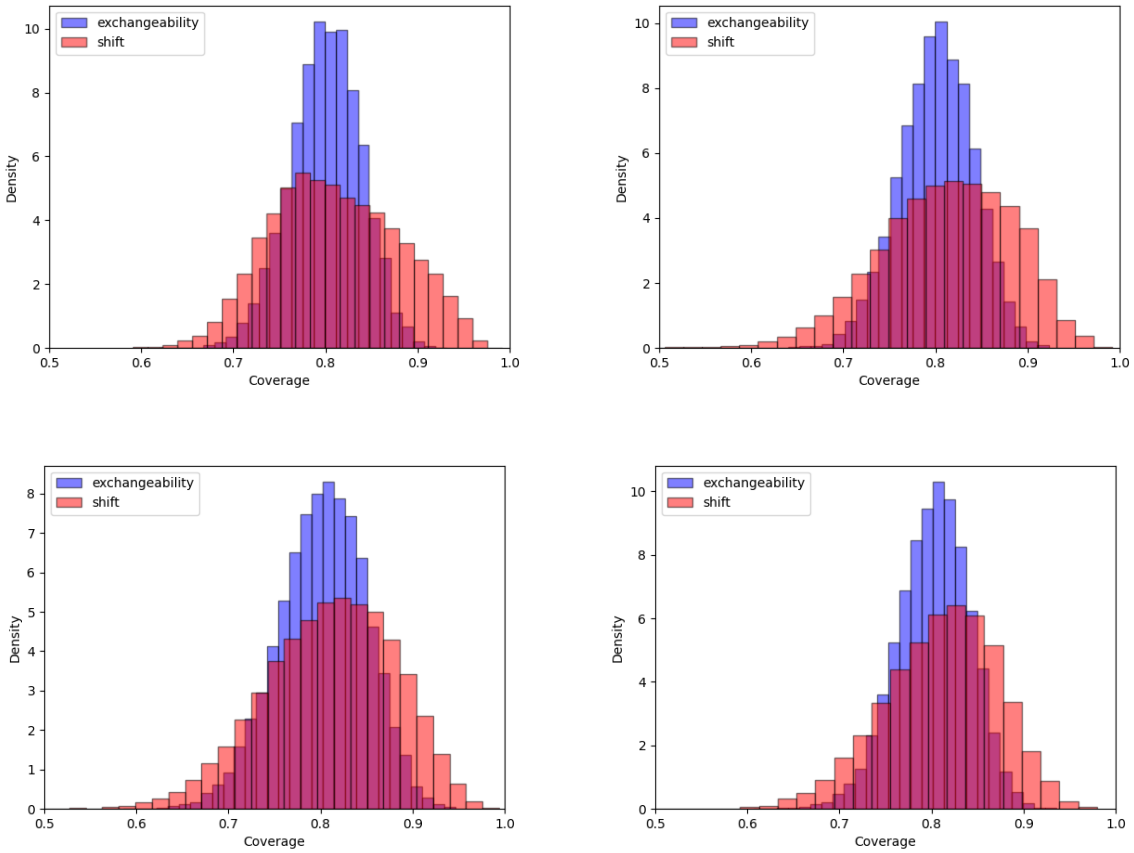

Figure 1: Histograms of the training-conditional coverage rates are presented for four datasets from the UCI Machine Learning Repository: Wine Quality (top left), Abalone (top right), Concrete Compressive Strength (bottom left), and Combined Cycle Power Plant (bottom right). See Appendix E for details of this simulation study.

2 Background and Related Work

2.1 Full conformal and split conformal

Define the shorthand $\mathbf{D}_n \cup (x, y) := ((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n), (x, y))$ and let $\hat{\mu}_{(x,y)} := T(\mathbf{D}_n \cup (x, y))$ denote the regression function obtained by running the training algorithm T on $\mathbf{D}_n \cup (x, y)$. Define the score function $s(x', y'; \mu) := f(\mu(x'), y')$ via some arbitrary (measurable) cost function $f : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and predictor function $\mu : \mathcal{X} \to \mathcal{Y}$. For instance, $s(x', y'; \mu) = |y' - \mu(x')|$ when $f(y, y') = |y - y'|$. For the data multiset $\mathcal{D}_n := \{(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} : i \in [n]\}$ where $[n] := \{1, 2, ..., n\}$, define

$$\mathcal{S}(\mathcal{D}_n; \mu) := \{s(x', y'; \mu) : (x', y') \in \mathcal{D}_n\}.$$

Observe that if T is symmetric, i.e., $T(\mathbf{D})$ is invariant under permutations of the elements of $\mathbf{D}$ for any data tuple $\mathbf{D}$, then the elements of $\mathcal{S}(\mathcal{D}_n \cup (X_{\text{test}}, Y_{\text{test}}); \hat{\mu}_{(X_{\text{test}}, Y_{\text{test}})})$ are exchangeable. Therefore,

$$\mathbb{P}\left(s(X_{\text{test}}, Y_{\text{test}}; \hat{\mu}_{(X_{\text{test}}, Y_{\text{test}})}) \leq \hat{F}^{-1}_{\mathcal{S}(\mathcal{D}_n \cup (X_{\text{test}}, Y_{\text{test}}); \hat{\mu}_{(X_{\text{test}}, Y_{\text{test}})})}(1 - \alpha)\right) \geq 1 - \alpha,$$

where $\hat{F}^{-1}_S(\cdot)$ denotes the *empirical* quantile function of S. Thus,

$$\mathbb{P}\left(Y_{\text{test}} \in \hat{C}^{\text{full}}_{\alpha}(X_{\text{test}})\right) \geq 1 - \alpha,$$

where the following confidence region is referred to as *full conformal* in the literature

$$\begin{aligned}\hat{C}_\alpha^{\text{full}}(x) &= \{y : s(x,y;\hat{\mu}_{(x,y)}) \le \hat{F}^{-1}_{\mathcal{S}(\mathcal{D}_n\cup(x,y);\hat{\mu}_{(x,y)})}(1-\alpha)\} \\ &\subseteq \{y : s(x,y;\hat{\mu}_{(x,y)}) \le \hat{F}^{-1}_{\mathcal{S}(\mathcal{D}_n;\hat{\mu}_{(x,y)})\cup\{\infty\}}(1-\alpha)\},\end{aligned} \tag{4}$$

It is well-known that this approach can be computationally intensive when $\mathcal{Y} = \mathbb{R}$ since to find out whether $y \in \hat{C}_\alpha^{\text{full}}(x)$ one needs to train the model with the dataset including (x, y). One simple way to alleviate this issue is to split the data into training and calibration datasets, namely $\mathcal{D}_n = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{cal}}$. Without loss of generality (when splitting), we assume $\mathbf{D}^{\text{train}} = ((X_1, Y_1), \ldots, (X_{n-m}, Y_{n-m}))$ and $\mathbf{D}^{\text{cal}} = ((X_{n-m+1}, Y_{n-m+1}), \ldots, (X_n, Y_n))$. In the split conformal method, one first finds the regression predictor $\hat{\mu} = T(\mathbf{D}^{\text{train}})$ and treats $\hat{\mu}$ as fixed. Note that the elements of $\mathcal{S}(\mathcal{D}^{\text{cal}} \cup (X_{\text{test}}, Y_{\text{test}}); \hat{\mu})$ are exchangeable. Hence, we get

$$\mathbb{P}\left(s(X_{\text{test}}, Y_{\text{test}}; \hat{\mu}) \le \hat{F}^{-1}_{\mathcal{S}(\mathcal{D}^{\text{cal}}\cup(X_{\text{test}},Y_{\text{test}});\hat{\mu})}(1-\alpha)\right) \ge 1-\alpha.$$

Hence, we have $\mathbb{P}(Y_{\text{test}} \in \hat{C}_\alpha^{\text{split}}(X_{\text{test}})) \ge 1-\alpha$ for

$$\begin{aligned}\hat{C}_\alpha^{\text{split}}(x) &= \left\{y : s(x,y;\hat{\mu}) \le \hat{F}^{-1}_{\mathcal{S}(\mathcal{D}^{\text{cal}};\hat{\mu})\cup\{\infty\}}(1-\alpha)\right\} \\ &\supseteq \left\{y : s(x,y;\hat{\mu}) \le \hat{F}^{-1}_{\mathcal{S}(\mathcal{D}^{\text{cal}}\cup(x,y);\hat{\mu})}(1-\alpha)\right\}.\end{aligned} \tag{5}$$

2.2 Jackknife+

Although the split-conformal approach resolves the computational efficiency problem of the full conformal method, it is somewhat inefficient in using the data and may not be useful in situations where the number of samples is limited. A heuristic alternative has long been known in the literature, namely, jackknife or leave-one-out cross-validation that can provide a compromise between the full conformal and split conformal methods. In particular,

$$\hat{C}_\alpha^{\text{J}}(x) = \{y : s(x,y;\hat{\mu}) \le \hat{F}^{-1}_{\mathcal{S}^{\text{J}}}(1-\alpha)\},$$

where $\hat{\mu} = T(\mathbf{D}_n)$, $\mathcal{S}^{\text{J}} := \{s(X_i, Y_i; \hat{\mu}^{-i}) : 1 \le i \le n\}$ and $\hat{\mu}^{-i} := T(((X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)))$. Despite its effectiveness, no general finite-sample guarantees are known for jackknife. Recently, Barber et al. (2021) proposed jackknife+, a modified version of the jackknife for $\mathcal{Y} = \mathbb{R}$ and $f(y, y') = |y - y'|$, and established $(1 - 2\alpha)$ finite-sample coverage lower bound for it. Let

$$\begin{aligned}\mathcal{S}^-(x) &= \{\hat{\mu}^{-i}(x) - |Y_i - \hat{\mu}^{-i}(X_i)| : i \in [n]\} \cup \{-\infty\}, \\ \mathcal{S}^+(x) &= \{\hat{\mu}^{-i}(x) + |Y_i - \hat{\mu}^{-i}(X_i)| : i \in [n]\} \cup \{\infty\}.\end{aligned}$$

The jackknife+ prediction interval is defined as

$$\hat{C}_\alpha^{\text{J+}}(x) = [\hat{F}^{-1}_{\mathcal{S}^-(x)}(\alpha),\ \hat{F}^{-1}_{\mathcal{S}^+(x)}(1-\alpha)].$$

In the same paper, an ϵ-inflated version of the jackknife+

$$\hat{C}_\alpha^{\text{J+},\epsilon}(x) = [\hat{F}^{-1}_{\mathcal{S}^-(x)}(\alpha) - \epsilon,\ \hat{F}^{-1}_{\mathcal{S}^+(x)}(1-\alpha) + \epsilon]. \tag{6}$$

is proposed which has $1 - \alpha - 4\sqrt{\nu}$ coverage lower bound (instead of $1 - 2\alpha$), if the training procedure satisfies

$$\max_{i\in[n]} \mathbb{P}(|\hat{\mu}(X_{\text{test}}) - \hat{\mu}^{-i}(X_{\text{test}})| > \epsilon) < \nu.$$

Also, the jackknife+ has been generalized to CV+ for K-fold cross-validation, and $(1 - 2\alpha - \sqrt{2/n})$ coverage lower bound is established.

2.3 Conformal prediction under distribution shift

Covariate shift concerns the setting when the covariate distribution changes between training and test data, while the conditionals remain the same. Specifically, we have

$$Z_i := (X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P_X \times P_{Y|X} =: P \quad \text{training data;}$$
$$Z_{\text{test}} := (X_{\text{test}}, Y_{\text{test}}) \sim Q_X \times P_{Y|X} =: Q \quad \text{test data.}$$

This notion has been extensively studied in machine learning (e.g., see Sugiyama et al. (2007); Quinonero-Candela et al. (2008); Sugiyama and Kawanabe (2012); Wen et al. (2014); Reddi et al. (2015); Chen et al. (2016) and the references therein).

To handle the covariate shift, a weighted version of conformal prediction was first proposed in the seminal work by Tibshirani et al. (2019). The key assumption is that the likelihood ratio $dQ/dP = dQ_X/dP_X$, needs to be known; we follow the same assumption in this work. This reweighting scheme has been extended to various settings in conformal prediction Podkopaev and Ramdas (2021); Lei and Candès (2021); Fannjianga et al.; Guan (2023). The general domain adaption problem has also been an active area from a causal perspective (e.g., Zhang et al. (2013a); Peters et al. (2016); Gong et al. (2016); Chen and Bühlmann (2021); Du and Xiang (2023)).

The weighted procedures require computing a weight associated with each sample,

$$w_i = \frac{\frac{dQ}{dP}(Z_i)}{\frac{dQ}{dP}(Z_{\text{test}}) + \sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)}, \ i \in \mathcal{I} \quad \text{and} \quad w_{\text{test}} = \frac{\frac{dQ}{dP}(Z_{\text{test}})}{\frac{dQ}{dP}(Z_{\text{test}}) + \sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)}$$

where $\mathcal{I} = [n]$ for full conformal and jackknife+, and $\mathcal{I} \subset [n]$ for the split conformal method which corresponds to the calibration (or hold-out) dataset. In particular, to generalize the full and split conformal methods to distribution shift setting, according to Tibshirani et al. (2019) one needs to replace the empirical CDF used in construction of (4) and (5)

$$\hat{F}_{\mathcal{S}}(t) = \frac{1}{n+1} \sum_{s\in\mathcal{S}} \mathbf{1}\{s \le t\}, \quad \mathcal{S} = \{s(Z_i) : i \in \mathcal{I}\} \cup \{\infty\}$$

by the weighted version

$$F_{Q,|\mathcal{I}|}(t) = w_{\text{test}} \mathbf{1}\{t = \infty\} + \sum_{i\in\mathcal{I}} w_i \mathbf{1}\{s(Z_i) \le t\}.$$

Similarly, to extend the jackknife+ method, Prinster et al. (2022) proposes replacing $\hat{F}_{\mathcal{S}^-}$ and $\hat{F}_{\mathcal{S}^+}$ by their corresponding weighted versions where w_{test} is associated to $-\infty$ and ∞, respectively.

3 Conditional Coverage Guarantees

3.1 Split conformal

The training-conditional coverage guarantee for split conformal method under the i.i.d. setting has been studied in Vovk (2012); Bian and Barber (2023). The proof in this setting relies on the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality. The following theorem concerns training-conditional coverage guarantee under covariate shift, where the extension is made by developing a weighted version of the DKW inequality. Let $\mathbf{D}_n = (Z_1, \ldots, Z_n) \in (\mathcal{X} \times \mathcal{Y})^n$ denote an n-tuple of data points containing both the training and calibration data sets. Define the training conditional probability of error as

$$P_e^{\text{split}}(\mathbf{D}_n) := \mathbb{P}\left(Y_{\text{test}} \notin \hat{C}_\alpha^{\text{split}}(X_{\text{test}}) \middle| \mathbf{D}_n\right). \tag{7}$$

Theorem 1 *Let $m < n$ denote the size of the calibration data set. Assume that Q is absolutely continuous with respect to P ($Q \ll P$) and $dQ/dP \le B < \infty$. Then,*

$$\mathbb{P}\left(P_e^{split}(\mathbf{D}_n) > \alpha + \left(\sqrt{2B\log(4/\delta)} + 2C\right)\sqrt{\frac{B}{m}}\right) \le \delta.$$

for all $\delta > 0$, where $C > 0$ is a universal constant. The probability is taken with respect to P since each entry in $\mathbf{D}_n$ follows P.

Corollary 1 *Relaxing the assumption $dQ/dP \leq B$ to $\|dQ/dP\|_{P,2} := \int (dQ/dP)^2 \, dP \leq B$ via Lemma 2, we get*

$$\mathbb{P}\left(P_e^{split}(\mathbf{D}_n) > \alpha + \frac{2B(2C+1)}{\delta\sqrt{m}}\right) \leq \delta, \qquad 0 < \delta \leq 1. \tag{8}$$

To assess the robustness of weighted conformal prediction methods, we compute training-conditional error bounds with no assumptions on dQ/dP. The proof follows from Lemma 4 and the proof of Theorem 1.

Theorem 2 *Fix $B, \delta > 0$ and $\epsilon \in (0, 1/2)$. It holds that,*

$$P_e^{split}(\mathbf{D}_n) < \alpha + 3\delta + 4\,\mathbb{P}\left(\frac{dQ}{dP}(Z_{test}) > m\right) + 2C\sqrt{\frac{B}{m'}} + \mathbb{P}\left(\frac{dQ}{dP}(Z_{test}) > B\right),$$

with probability at least,

$$1 - 4e^{-m'\delta^2/(2B^2)} - 2e^{-2m^{2\epsilon}} - 2\,\mathbb{P}\left(\frac{dQ}{dP}(Z_{test}) > m\right) - \frac{8}{m\delta^2}\mathbb{E}\left(\left(\frac{dQ}{dP}(Z_{calib})\right)^2 \mathbf{1}\left\{\frac{dQ}{dP}(Z_{calib}) \leq m\right\}\right),$$

where $m' = m\,\mathbb{P}\left(\frac{dQ}{dP}(Z_{calib}) \leq B\right) - m^{1/2+\epsilon}$.

3.2 Full conformal and jackknife+

Let $\mu_\beta : \mathcal{X} \to \mathbb{R}$ denote a predictor function parameterized by $\beta \in \mathbb{R}^p$. By a slight abuse of notation, let the map $T : \cup_{n\geq 1}(\mathcal{X} \times \mathcal{Y})^n \to \mathbb{R}^p$ denote a training algorithm for estimating β, hence, $\hat{\beta}_n = T(\mathbf{D}_n)$. In this case, we have $\hat{\mu}_{\mathbf{D}_n} = \mu_{\hat{\beta}_n}$.

Assumption 1 (Uniform stability) *For all $i \in [n]$, we have*

$$\sup_{z_1,\ldots,z_n} \|\mu_{T(z_1,\ldots,z_{i-1},z_{i+1}\ldots,z_n)} - \mu_{T(z_1,\ldots,z_i,\ldots,z_n)}\|_\infty \leq \frac{c_n}{2}.$$

In the case of the ridge regression Hoerl and Kennard (1970) with $\mathcal{Y} = [-I, I]$ and $\mathcal{X} = \{x : \|x\|_2 \leq b\}$, this assumption holds with $c_n = 16\,b^2 I^2/(\lambda\, n)$, where λ denotes the regularization parameter. See Bousquet and Elisseeff (2002) for the general result on the uniform stability of RKHS regression.

Assumption 2 (Bi-Lipschitz continuity) *The map $\beta \mapsto \mu_\beta$ is bi-Lipschitz with respect to the ∞-norms, i.e.,*

$$\kappa_1 \|\beta - \beta'\|_\infty \;\leq\; \|\mu_\beta - \mu_{\beta'}\|_\infty \;\leq\; \kappa_2 \|\beta - \beta'\|_\infty,$$

for some constants $0 < \kappa_1 \leq \kappa_2 < \infty$.

Remark 1 *It is worth noting that if the parameter space $\Theta \subseteq \mathbb{R}^p$ is compact, $\Phi : U \to L^\infty(\mathcal{X})$ given by $\beta \mapsto \mu_\beta$ is continuously differentiable for some open $U \supseteq \Theta$, then $\kappa_2 < \infty$. Moreover, the inverse function theorem (for Banach spaces), gives the sufficient condition under which the inverse is continuously differentiable over $\Phi(U)$ and hence $\kappa_1 > 0$.*

In the case of linear regression with $\mathcal{X} = \{x : \|x\|_2 \leq b\}$, one can verify that Assumption 2 holds with $\kappa_1 = b$ and $\kappa_2 = \sqrt{p}b$. Let $\overline{\beta}_n = \mathbb{E}\,\hat{\beta}_n$, $\hat{\beta}_{-i} = T(Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$ with $Z_i = (X_i, Y_i)$ and $\overline{\beta}_{-i} = \mathbb{E}\,\hat{\beta}_{-i}$. Define

$$F^{(n-1)}(t) := \mathbb{P}_{Z_1\sim P}\left(\left|Y_1 - \mu_{\overline{\beta}_{-1}}(X_1)\right| \leq t\right),$$
$$F_Q^{(n-1)}(t) := \mathbb{P}_{Z_1\sim Q}\left(\left|Y_1 - \mu_{\overline{\beta}_{-1}}(X_1)\right| \leq t\right).$$

Assumption 3 (Bounded density) *$F'^{(n)} < L_n$ and $F_Q'^{(n)} < L_{Q,n}$ where $F'^{(n)}$ and $F_Q'^{(n)}$ denote the derivative of $F^{(n)}$ and $F_Q^{(n)}$, respectively.*

We introduce the following shorthand:

$$A(n,p,\epsilon) := 2\,\kappa_2\,c_{n-1}\left(\frac{1}{\kappa_1} + \sqrt{\frac{n}{2\kappa_1^2}\log\frac{2p}{\epsilon}}\right).$$

Theorem 3 (Jackknife+ under exchangeability) *Define $P_e^{J+}(\mathbf{D}_n) := \mathbb{P}\left(Y_{test} \notin \hat{C}_\alpha^{J+}(X_{test})\middle|\mathbf{D}_n\right)$. Under Assumptions 1—3, for all $\epsilon, \delta > 0$, it holds that*

$$\mathbb{P}\left(P_e^{J+}(\mathbf{D}_n) > \alpha + \sqrt{\frac{\log(2/\delta)}{2n}} + L_{n-1}\,A(n,p,\epsilon)\right) \leq \epsilon + \delta.$$

Now we present our result on jackknife+ under covariate shift when $dQ/dP \leq B$.

Theorem 4 (Jackknife+ under covariate shift) *Assume that Q is absolutely continuous with respect to P ($Q \ll P$) and $dQ/dP \leq B$. Under Assumptions 1—3, for all $\epsilon, \delta > 0$, it holds that*

$$\mathbb{P}\left(P_e^{J+}(\mathbf{D}_n) > \alpha + \left(\sqrt{2B\log 4/\delta} + 2C\right)\sqrt{\frac{B}{n}} + L_{Q,n-1}\,A(n,p,\epsilon)\right) \leq \epsilon + \delta,$$

where C is a universal constant and $A(n,p,\epsilon)$ is the same as in Theorem 3.

Using the same arguments as in the proof of this theorem, one can get a coverage bound for the CV+ as well. Unlike (2) which is meaningful only if the number of samples in each fold ℓ is large, the bound we present in the following corollary is suitable for cases where $\ell/n \to 0$.

Corollary 2 (CV+) *Define $P_e^{CV+}(\mathbf{D}_n) := \mathbb{P}\left(Y_{test} \notin \hat{C}_\alpha^{CV+}(X_{test})\middle|\mathbf{D}_n\right)$. Under Assumptions 1—3, for all $\epsilon, \delta > 0$, it holds that*

$$\mathbb{P}\left(P_e^{CV+}(\mathbf{D}_n) > \alpha + \sqrt{\frac{\log(2/\delta)}{2n}} + 2\,\ell\,L_{n-\ell}\,\kappa_2\,c_{n-\ell}\left(\frac{1}{\kappa_1} + \sqrt{\frac{n}{2\kappa_1^2}\log\frac{2p}{\epsilon}}\right)\right) \leq \epsilon + \delta.$$

The following theorem concerns the training-conditional guarantees for the full conformal prediction. We again introduce a shorthand:

$$E(n,p,\epsilon) := c_{n+1} + \sqrt{2n\log\frac{2p}{\epsilon}}\,\frac{\kappa_2\,c_n}{\kappa_1}.$$

Theorem 5 (Full conformal under exchangeability) *Define $P_e^{full}(\mathbf{D}_n) := \mathbb{P}\left(Y_{test} \notin \hat{C}_\alpha^{full}(X_{test})\middle|\mathbf{D}_n\right)$. Under Assumptions 1—3, for all $\epsilon, \delta > 0$, it holds that*

$$\mathbb{P}\left(P_e^{full}(\mathbf{D}_n) > \alpha + \sqrt{\frac{\log(2/\delta)}{2n}} + L_n\,E(n,p,\epsilon)\right) \leq \epsilon + \delta.$$

Theorem 6 (Full conformal under covariate shift) *Assume that Q is absolutely continuous with respect to P ($Q \ll P$) and $dQ/dP \leq B$. Under Assumptions 1—3, for all $\epsilon, \delta > 0$, it holds that*

$$\mathbb{P}\left(P_e^{full}(\mathbf{D}_n) > \alpha + \left(\sqrt{2B\log(4/\delta)} + 2C\right)\sqrt{\frac{B}{n}} + L_{Q,n}\,E(n,p,\epsilon)\right) \leq \epsilon + \delta.$$

where C is a universal constant and $E(n,p,\epsilon)$ is the same as in Theorem 5.

We note that similar to Remark 1, one can relax assumption $dQ/dP \leq B$ to $\|dQ/dP\|_{P,2} \leq B$ and in this case the bounds for the jackknife+ and full conformal methods hold with the slow rate $O(1/(\delta\sqrt{n}))$ instead of $O(\log(1/\delta)/\sqrt{n})$.

4 Discussion

We have presented training-conditional coverage guarantees for weighted conformal prediction methods under covariate shift. The split conformal method has been analyzed for the black-box training algorithm while the full conformal and jackknife+ have been analyzed under three assumptions. Although Assumptions 1 and 2 have been verified only for the ridge regression in this paper, we conjecture that they are satisfied by a truncated version of the general RKHS regressions which we leave for future research. Truncated and sketched versions of the RKHS models have been extensively studied in the previous literature from the computational efficiency perspective (see Amini (2021); Williams and Seeger (2000); Zhang et al. (2013b); Alaoui and Mahoney (2015); Cortes et al. (2010) and references therein). The results in this paper quantify the training sample size (or calibration sample size for split conformal) needed for the coverage under covariate shift.

Estimation of the likehood ratio. A natural question here is how to include the error arising from estimating the likelihood ratios in the analysis. Define

$$\hat{w}_i := \frac{V_i}{V} := \frac{\frac{dQ}{dP}(Z_i)}{\sum_{i\in[n]} \frac{dQ}{dP}(Z_i)}, \qquad \breve{w}_i := \frac{\hat{V}_i}{\hat{V}} := \frac{\widehat{\frac{dQ}{dP}}(Z_i)}{\sum_{i\in[n]} \widehat{\frac{dQ}{dP}}(Z_i)}$$

and let $s(\cdot)$ denote a fixed score function. In this case, one needs to include the following term to the weighted DKW inequalities derived in this paper,

$$\begin{aligned}
\sup_{t\in\mathbb{R}} \left| \sum_{i\in[n]} (\breve{w}_i - \hat{w}_i) \mathbf{1}\{s(Z_i) \leq t\} \right| &\leq \sum_{i\in[n]} |\breve{w}_i - \hat{w}_i| \\
&= \sum_{i\in[n]} \left| \frac{\hat{V}_i}{\hat{V}} - \frac{V_i}{V} \right| \\
&\leq \frac{\sum_{i\in[n]} \left(\left| \hat{V}_i V - V_i V \right| + \left| V_i V - V_i \hat{V} \right| \right)}{V\hat{V}} \\
&= \frac{V \sum_{i\in[n]} \left| \hat{V}_i - V_i \right| + \left| V - \hat{V} \right| \left(\sum_{i\in[n]} V_i \right)}{V\hat{V}} \\
&= \frac{\sum_{i\in[n]} \left| \hat{V}_i - V_i \right| + \left| V - \hat{V} \right|}{\hat{V}} \leq 2 \frac{\sum_{i\in[n]} \left| \hat{V}_i - V_i \right|}{\hat{V}},
\end{aligned}$$

where the concentration properties of this term around zero depends on the estimator of $dQ/dP(Z_i)$.

Comparison with the (m, n)-stability. The (m, n)-stability parameters were recently introduced in Liang and Barber (2023) and used to compute training-conditional coverage bounds for inflated full conformal and jackknife+ prediction intervals under exchangeability. Unlike uniform stability which is a distribution-free property of a training process, (m, n)-stability depends on both the training algorithm and the distributions of the data as follows

$$\psi^{\text{out}}_{m,n} := \mathbb{E}\, |\hat{\mu}_{\mathbf{D}_n}(X_{\text{test}}) - \hat{\mu}_{\mathbf{D}_{n+m}}(X_{\text{test}})|, \tag{9}$$

$$\psi^{\text{in}}_{m,n} := \mathbb{E}\, |\hat{\mu}_{\mathbf{D}_n}(X_1) - \hat{\mu}_{\mathbf{D}_{n+m}}(X_1)|, \tag{10}$$

with $X_{\text{test}} \perp\!\!\!\perp \mathbf{D}_{n+m}$, $\mathbf{D}_{n+m} = ((X_1, Y_1), ..., (X_{n+m}, Y_{n+m}))$, and $\hat{\mu}_{\mathbf{D}_n}$ denotes the predictor function obtained by training on $\mathbf{D}_n$. Although weaker than uniform stability, these parameters are yet not well-understood in

a practical sense. To elaborate on the difference between this approach and uniform stability, we evaluate their resulting training-conditional bounds for the ridge regression under exchangeability.

Assume $\mathcal{X} = \{x : \|x\| \leq b\}$ and $\mathcal{Y} = [-I, I]$. As stated in the previous section, this regression model satisfies $c_n = 16\, b^2 I^2/(\lambda\, n)$, $\kappa_1 = b$ and $\kappa_2 = \sqrt{p}\, b$. Hence, we get the following bound for both full conformal and jackknife+ methods,

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + O\big(n^{-1/2}\big(\sqrt{\log(1/\delta)} + \sqrt{p \log(2p/\epsilon)}\big)\big) \right) \leq \epsilon + \delta.$$

On the other hand, the following bound is proposed for the γ-inflated jackknife+ in Liang and Barber (2023),

$$\mathbb{P}\left(P_e^{\mathrm{J+},\gamma}(\mathbf{D}_n) > \alpha + 3\sqrt{\frac{\log(1/\delta)}{\min(m,n)}} + 2\sqrt[3]{\frac{\psi^{\mathrm{out}}_{m,n-1}}{\gamma}} \right) \leq 3\delta + \sqrt[3]{\frac{\psi^{\mathrm{out}}_{m,n-1}}{\gamma}}, \tag{11}$$

for all $m \geq 1$. We get $\psi^{\mathrm{out}}_{m,n} = O(mc_n)$ since $\psi^{\mathrm{out}}_{1,n} \leq c_{n+1}/2$ by definition (9) and Assumption 1, and $\psi^{\mathrm{out}}_{m,n} \leq \sum_{k=n}^{n+m-1} \psi^{\mathrm{out}}_{1,k}$ holds according to in Liang and Barber (2023, Lemma 5.2) . Substituting for $\psi^{\mathrm{out}}_{m,n-1}$ in bound (11), we obtain

$$\mathbb{P}\left(P_e^{\mathrm{J+},\gamma}(\mathbf{D}_n) > \alpha + O\left(\sqrt{\frac{\log(1/\delta)}{\min(m,n)}} + \sqrt[3]{\frac{m\, c_{n-1}}{\gamma}}\right) \right) \leq 3\delta + O\left(\sqrt[3]{\frac{mc_{n-1}}{\gamma}}\right). \tag{12}$$

Letting $m^{-1/2} = (m/n)^{1/3}$ to balance the two terms $\sqrt{\frac{\log(1/\delta)}{\min(m,n)}}$ and $\sqrt[3]{mc_{n-1}/\gamma}$, we get $m = n^{2/5}$. By plugging $m = n^{2/5}$ into (12), we have

$$\mathbb{P}\left(P_e^{\mathrm{J+},\gamma}(\mathbf{D}_n) > \alpha + O\big(n^{-1/5}\big(\sqrt{\log(1/\delta)} + \gamma^{-1/3}\big)\big) \right) \leq 3\delta + O\left(n^{-1/5}\gamma^{-1/3}\right). \tag{13}$$

This bound, although dimension-free (i.e., does not depend on p), is very slow with respect to the sample size. In Liang and Barber (2023), the same bound as (11) is established for γ-inflated full conformal method except with $\psi^{\mathrm{in}}_{m-1,n+1}$ instead of $\psi^{\mathrm{out}}_{m,n-1}$. Hence, the same bound as (13) can be obtained for the γ-inflated full conformal method via $\psi^{\mathrm{in}}_{m,n} = O(mc_n)$.

5 Conclusion

In this work, we have studied the training-conditional coverage bounds of full conformal, jackknife+, and CV+ prediction regions from a uniform stability perspective, which is well-understood for convexly regularized empirical risk minimization over reproducing kernel Hilbert spaces. We have derived new bounds via a concentration argument for the (estimated) predictor function. In the case of ridge regression, we have used the uniform stability parameter to derive a bound for (m, n)-stability and compare the resulting bounds from Liang and Barber (2023) to the bounds established in this paper. We have observed that our rates are faster in sample size but dependent on the dimension of the problem. Even though our work is theoretical in nature, it can potentially shed light on understanding a much broader downstream-task setups in reconstructive self-supervised learning (SSL), where the existing literature focuses on either linear regression or ridge regression (e.g., Lee et al. (2021); Teng et al. (2022); Du and Xiang (2024)). For split conformal, our result allows for flexible downstream setups in SSL well beyond simple regressions, while for jackknife+ it would be interesting to reveal the interplay between uniform stability and excess risk analysis in SSL. Another worthwhile direction is from the robustness perspective in conformal prediction and detection problems, including adversarial and label noise settings (e.g., Gendler et al. (2021); Einbinder et al. (2024); Zhang et al. (2025)). For instance, one can study the robustness of training conditional guarantees under adversarial attacks during inference time.

References

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

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

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

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. *arXiv preprint arXiv:2209.15145*, 2022.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. *Technical Report*, 2003.

Michael Bian and Rina Foygel Barber. Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044–2066, 2023.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. 2021.

Ruiting Liang and Rina Foygel Barber. Algorithmic stability implies training-conditional coverage for distribution-free prediction methods. *arXiv preprint arXiv:2311.04295*, 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.

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

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Vern I Paulsen and Mrinal Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*, volume 152. Cambridge university press, 2016.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Grace Wahba. *Spline models for observational data*. SIAM, 1990.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. MIT Press, 2008.

Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. 2012.

Junfeng Wen, Chun-Nam Yu, and Russell Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning*, pages 631–639. PMLR, 2014.

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

Xiangli Chen, Mathew Monfort, Anqi Liu, and Brian D Ziebart. Robust covariate shift regression. In *Artificial Intelligence and Statistics*, pages 1270–1279. PMLR, 2016.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in artificial intelligence*, pages 844–853. PMLR, 2021.

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.

Clara Fannjianga, Stephen Batesa, Anastasios Angelopoulosa, Jennifer Listgartena, and Michael I Jordana. Conformal prediction for the design problem.

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

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013a.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International Conference on Machine Learning*, pages 2839–2848. PMLR, 2016.

Yuansi Chen and Peter Bühlmann. Domain adaptation under structural causal models. *Journal of Machine Learning Research*, 22:1–80, 2021.

Kang Du and Yu Xiang. Learning invariant representations under general interventions on the response. *IEEE Journal on Selected Areas in Information Theory*, 2023.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Arash A Amini. Spectrally-truncated kernel ridge regression and its free lunch. *Electronic Journal of Statistics*, 15(2):3743–3761, 2021.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression. In *Conference on learning theory*, pages 592–617. PMLR, 2013b.

Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. *Advances in neural information processing systems*, 28, 2015.

Corinna Cortes, Mehryar Mohri, and Ameet Talwalkar. On the impact of kernel approximation on learning accuracy. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 113–120. JMLR Workshop and Conference Proceedings, 2010.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

Jiaye Teng, Weiran Huang, and Haowei He. Can pretext-based self-supervised learning be boosted by downstream data? a theoretical analysis. In *International conference on artificial intelligence and statistics*, pages 4198–4216. PMLR, 2022.

Kang Du and Yu Xiang. Low-rank approximation of structural redundancy for self-supervised learning. In *Causal Learning and Reasoning*, pages 1008–1032. PMLR, 2024.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2021.

Bat-Sheva Einbinder, Shai Feldman, Stephen Bates, Anastasios N Angelopoulos, Asaf Gendler, and Yaniv Romano. Label noise robustness of conformal prediction. *Journal of Machine Learning Research*, 25(328): 1–66, 2024.

Daofu Zhang, Mehrdad Pournaderi, Hanne Clifford, Yu Xiang, and Pramod Varshney. On the adversarial robustness of learning-based conformal novelty detection. *arXiv preprint arXiv:2510.00463*, 2025.

Aad W Van Der Vaart and Jon A Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer, 1996.

John Lafferty, Han Liu, and Larry Wasserman. Concentration of measure. *On-line. Available: https://www.stat.cmu.edu/ larry/=sml/Concentration.pdf*, 2008.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in Combinatorics*, 141(1):148–188, 1989.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

A Technical Lemmas: Weighted DKW Inequalities

A.1 Background

In this section, we first briefly introduce the notion of *bracketing number* and four foundamental results, which we will be applying in our proofs.

Bracketing Numbers. A measure of complexity for a class of functions is the bracketing number. For a pair of functions f and g, the bracket $[f, g]$ is defined as follows,

$$[f, g] = \{h : f \leq h \leq g\}.$$

A set of brackets $[f_1, g_1], \ldots, [f_n, g_n]$ is a ϵ-$L^q(P)$ bracketing of function class $\mathcal{F}$ if $\mathcal{F} \subseteq \bigcup_{1 \leq i \leq n} [f_i, g_i]$ and $(\int |f_i - g_i|^q dP)^{1/q} \leq \epsilon$ for all $i \in [n]$. The bracketing number $N_{[]}(\epsilon, \mathcal{F}, L^q(P))$ is the size of the smallest ϵ-$L^q(P)$ bracketing of $\mathcal{F}$.

Theorem 7 (Van Der Vaart and Wellner (1996)(Theorem 2.14.2)) *Let $\mathcal{F}$ be a class of measurable functions with a measurable envelope F. Then, for some universal constant $C > 0$,*

$$\mathbb{E}\left(\sup_{f \in \mathcal{F}} \left|\sqrt{n} \int f \; d(P_n - P)\right|\right) \leq C \|F\|_{P,2} \int_0^1 \sqrt{1 + \log N_{[]}\left(\epsilon \|F\|_{P,2}, \mathcal{F}, L^2(P)\right)} \; d\epsilon. \tag{14}$$

Theorem 8 (Lafferty et al. (2008)(Theorem 7.86)) *Let $\mathcal{F}$ be a class of measurable functions with $A = \sup_{f\in\mathcal{F}} \|f\|_{P,1}$ and $B = \sup_{f\in\mathcal{F}} \|f\|_\infty$. Then, for $\epsilon < 2A/3$,*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\int f\ d(P_n - P)\right| > \epsilon\right) \leq 4N_{[]}\left(\epsilon/8, \mathcal{F}, L^1(P)\right)\exp\left(-\frac{96n\epsilon^2}{76AB}\right). \tag{15}$$

Theorem 9 (Durrett (2019) (Theorem 2.2.11)) *For each n let $X_{n,k}$, $1 \leq k \leq n$ be independent. Let $b_n > 0$ with $b_n \to \infty$, and let $\tilde{X}_{n,k} = X_{n,k}\mathbf{1}\{|X_{n,k}| \leq b_n\}$. Suppose that as $n \to \infty$ (i) $\sum_{k=1}^n \mathbb{P}(|X_{n,k}| > b_n) \to 0$, and (ii) $b_n^{-2}\sum_{k=1}^n \mathbb{E}\,\tilde{X}_{n,k}^2 \to 0$. If we let $S_n = X_{1,n} + \ldots + X_{n,n}$ and put $a_n = \sum_{k=1}^n \mathbb{E}\,\tilde{X}_{n,k}$, then $(S_n - a_n)/b_n \to 0$ in probability.*

Theorem 10 (Durrett (2019) (Theorem 2.2.12)) *Let $X_1, X_2, \ldots$ be i.i.d. with $x\,\mathbb{P}(|X_i| > x) \to 0$ as $x \to \infty$. Let $S_n = X_1 + \ldots + X_n$ and let $\mu_n = \mathbb{E}\,X_1\mathbf{1}\{|X_1| \leq n\}$. Then $S_n/n - \mu_n \to 0$ in probability.*

A.2 Weighted DKW Inequalities

With these results in place, we are ready to proceed with our proofs. Let $F_Q(x) = \mathbb{P}_{Z\sim Q}(s(Z) \leq x)$ where $s(\cdot)$ denotes some fixed score function. For $Z_i \overset{\text{i.i.d.}}{\sim} P$, we define

$$\hat{F}_{Q,n}(x) := \sum_{i\in[n]} \hat{w}_i\mathbf{1}\{s(Z_i) \leq x\}, \qquad \hat{w}_i = \frac{\frac{dQ}{dP}(Z_i)}{\sum_{i\in[n]}\frac{dQ}{dP}(Z_i)}. \tag{16}$$

Lemma 1 (Bounded likelihood ratio) *Assume that Q is absolutely continuous with respect to P ($Q \ll P$) and $dQ/dP \leq B$. Then, for all $\delta > 0$*

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}\left|\hat{F}_{Q,n}(x) - F_Q(x)\right| > \delta + 2C\sqrt{\frac{B}{n}}\right) \leq 4e^{-n\delta^2/(2B^2)},$$

where $C > 0$ is some universal constant.

Proof: By substituting the formula for $\hat{w}_i$ in the definition of $\hat{F}_{Q,n}$, we get

$$\hat{F}_{Q,n} = \frac{\frac{1}{n}\sum_{i\in[n]}\frac{dQ}{dP}(Z_i)\mathbf{1}\{s(Z_i) \leq x\}}{\frac{1}{n}\sum_{i\in[n]}\frac{dQ}{dP}(Z_i)}\ .$$

Hence we have

$$\begin{aligned}\sup_{x\in\mathbb{R}}|\hat{F}_{Q,n}(x) - F_Q(x)| \leq \sup_{x\in\mathbb{R}}&\left|\frac{1}{n}\sum_{i\in[n]}\frac{dQ}{dP}(Z_i)\mathbf{1}\{s(Z_i) \leq x\} - F_Q(x)\right| + \\ &\left|\frac{1}{\frac{1}{n}\sum_{i\in[n]}\frac{dQ}{dP}(Z_i)} - 1\right| \sup_{x\in\mathbb{R}}\left|\frac{1}{n}\sum_{i\in[n]}\frac{dQ}{dP}(Z_i)\mathbf{1}\{s(Z_i) \leq x\}\right| \\ = \sup_{x\in\mathbb{R}}&\left|\frac{1}{n}\sum_{i\in[n]}\frac{dQ}{dP}(Z_i)\mathbf{1}\{s(Z_i) \leq x\} - F_Q(x)\right| + \left|1 - \frac{1}{n}\sum_{i\in[n]}\frac{dQ}{dP}(Z_i)\right|.\end{aligned} \tag{17}$$

Regarding the first term, we have,

$$\begin{aligned}\sup_{x\in\mathbb{R}}\left|\frac{1}{n}\sum_{i\in[n]}\frac{dQ}{dP}(Z_i)\mathbf{1}\{s(Z_i) \leq x\} - F_Q(x)\right| &= \sup_{x\in\mathbb{R}}\left|\int \frac{dQ}{dP}(z)\mathbf{1}\{s(z) \leq x\}\ d(P_n - P)\right| \\ &= \sup_{x\in\mathbb{R}}\left|\int f_x\ d(P_n - P)\right|,\end{aligned}$$

where $f_x \in L^1(\mathcal{X} \times \mathcal{Y}, P)$ is defined as $f_x(z) := \frac{dQ}{dP}(z)\mathbf{1}\{s(z) \leq x\}$, and $P_n(A) = \frac{1}{n}\sum_{i\in[n]} \mathbf{1}\{Z_i \in A\}$ is the empirical measure.

We note,

$$1 \leq \|dQ/dP\|_{P,2}^2 = \int (dQ/dP)^2 \; dP \leq \|dQ/dP\|_\infty \int dQ/dP \; dP \leq B.$$

according to the Holder's inequality. Hence, $\|dQ/dP\|_{P,2} \leq \sqrt{B}$ and by Theorem 7 we have

$$\mathbb{E}\left(\sup_{x\in\mathbb{R}} \left| \int f_x \; d(P_n - P) \right|\right) \leq C\sqrt{\frac{B}{n}} \int_0^1 \sqrt{1 + \log N_{[]}\left(\epsilon \|dQ/dP\|_{P,2}, \mathcal{F}, L^2(P)\right)} \; d\epsilon, \tag{18}$$

with some universal constant $C > 0$, and $N_{[]}\left(\epsilon\|dQ/dP\|_{P,2}, \mathcal{F}, L^2(P)\right)$ denotes the bracketting number for function class $\mathcal{F} = \{f_x : x \in \mathbb{R}\}$. It remains to compute an upper bound for $N_{[]}\left(\epsilon, \mathcal{F}, L^2(P)\right)$. Let $t_0 = -\infty$ and define,

$$t_{i+1} := \sup\left\{ t \in \mathbb{R} : \int f^2(z)\mathbf{1}\{s(z) \in (t_i, t]\} \; dP < \epsilon^2 \right\}, \tag{19}$$

where $f = dQ/dP$. We note that there exists $\ell < \infty$ such that $t_\ell = \infty$. This is true since $t_{j+1} < \infty$ implies

$$\int f^2(z)\mathbf{1}\{s(z) \in (t_i, t_{i+1}]\} \; dP \geq \epsilon^2, \quad 0 \leq i \leq j,$$

according to (19) and dominated convergence theorem. Hence,

$$(j+1)\epsilon^2 \leq \sum_{i=0}^{j} \int f^2(z)\mathbf{1}\{s(z) \in (t_i, t_{i+1}]\} \; dP \leq \int f^2 \; dP = \|dQ/dP\|_{P,2}^2, \tag{20}$$

and therefore $j + 1 \leq (\|dQ/dP\|_{P,2}/\epsilon)^2$. Let $k = \min\{\ell \in \mathbb{Z} : t_\ell = \infty\}$. Now define ϵ-brackets as follows

$$b_i = \{g(z) : f(z)\mathbf{1}\{s(z) \leq t_i\} \leq g(z) \leq f(z)\mathbf{1}\{s(z) < t_{i+1}\}\}$$

where $0 \leq i \leq k-1$. Clearly, $\mathcal{F} \subset \cup_{i=0}^k b_i$ with $b_k = \{f\}$. Brackets $b_0, \ldots, b_{k-1}$ have size

$$\left(\int f^2(z)\mathbf{1}\{s(z) \in (t_i, t_{i+1})\} \; dP\right)^{1/2} \leq \epsilon, \quad 0 \leq i \leq k-1,$$

according to (19) and bracket b_k has size 0. Hence, we have $k+1$ brackets in total with size smaller than ϵ. By (20), we have $k - 1 \leq (\|dQ/dP\|_{P,2}/\epsilon)^2$ which implies that $N_{[]}\left(\epsilon, \mathcal{F}, L^2(P)\right) \leq k + 1 \leq 2 + (\|dQ/dP\|_{P,2}/\epsilon)^2$. Plugging this into (20), we get

$$N_{[]}\left(\epsilon\|dQ/dP\|_{P,2}, \mathcal{F}, L^2(P)\right) \leq 2 + \frac{1}{\epsilon^2}.$$

Therefore we have,

$$\mathbb{E}\left(\sup_{x\in\mathbb{R}} \left| \int f_x \; d(P_n - P) \right|\right) \leq 2C\sqrt{\frac{B}{n}} = C'\sqrt{\frac{B}{n}}, \tag{21}$$

which follows from the observation that

$$\int_0^1 \sqrt{1 + \log(2 + 1/\epsilon^2)} \leq \left(1 + \int_0^1 \log(2 + 1/\epsilon^2)\right)^{1/2} \leq \left(1 + 2\int_0^1 (\epsilon^2 - \log\epsilon)\right)^{1/2} \leq 2.$$

Now let $P_n^{(i)}$ denote the empirical measure obtained after replacing Z_i by some arbitrary Z_i'. In this case, we observe that

$$\left| \sup_{x\in\mathbb{R}} \left| \int f_x \; d(P_n - P) \right| - \sup_{x\in\mathbb{R}} \left| \int f_x \; d(P_n^{(i)} - P) \right| \right| \leq \sup_{x\in\mathbb{R}} \left| \int f_x \; d(P_n - P_n^{(i)}) \right|$$

$$\leq \frac{1}{n} \sup_{x \in \mathbb{R}} |f_x(Z_i) - f_x(Z_i')| \leq \frac{B}{n}.$$

Therefore, by McDiarmid's inequality McDiarmid et al. (1989) we get

$$\mathbb{P}\left(\left| \sup_{x \in \mathbb{R}} \left| \int f_x \; d(P_n - P) \right| - \mathbb{E}\left(\sup_{x \in \mathbb{R}} \left| \int f_x \; d(P_n - P) \right| \right) \right| > \delta \right) \leq 2e^{-2n\delta^2/B^2}.$$

Combining with (21), we obtain,

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} \left| \int f_x \; d(P_n - P) \right| > \delta + C' \sqrt{\frac{B}{n}} \right) \leq 2e^{-2n\delta^2/B^2}.$$

Back to (17), for the second term, Hoeffding's inequality implies

$$\mathbb{P}\left(\left| 1 - \frac{1}{n} \sum_{i \in [n]} \frac{dQ}{dP}(Z_i) \right| > \delta \right) \leq 2e^{-2n\delta^2/B^2}. \tag{22}$$

Combining the bounds for the two terms, we obtain,

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} \left| \hat{F}_{Q,n}(x) - F_Q(x) \right| > \delta + C' \sqrt{\frac{B}{n}} \right) \leq 4e^{-n\delta^2/(2B^2)}. \qquad \blacksquare$$

The following version relaxes the assumption $dQ/dP \leq B$ to $\|dQ/dP\|_{P,2} \leq K$ at the cost of slower rates.

Lemma 2 (Alternative version: bounded second moment) *Assume $Q \ll P$ and $\|dQ/dP\|_{P,2} \leq K$. Then, for all $\delta > 0$*

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} \left| \hat{F}_{Q,n}(x) - F_Q(x) \right| > \delta \right) \leq \frac{4\,CK}{\delta\sqrt{n}} + \frac{4(K^2 - 1)}{n\delta^2}.$$

where $C > 0$ is some universal constant.

Proof: By the same argument that led to (21) in the proof of Lemma 1, we get

$$\mathbb{E}\left(\sup_{x \in \mathbb{R}} \left| \int f_x \; d(P_n - P) \right| \right) \leq \frac{2CK}{\sqrt{n}} = \frac{C'K}{\sqrt{n}}.$$

Combining this with Markov's inequality, we obtain

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} \left| \int f_x \; d(P_n - P) \right| > \delta \right) \leq \frac{C'K}{\delta\sqrt{n}}. \tag{23}$$

This bounds the first term of (17). Regarding the second term, Chebyshev's inequality gives

$$\mathbb{P}\left(\left| 1 - \frac{1}{n} \sum_{i \in [n]} \frac{dQ}{dP}(Z_i) \right| > \delta \right) \leq \frac{\mathrm{Var}\left(\frac{1}{n} \sum_{i \in [n]} \frac{dQ}{dP}(Z_i) \right)}{\delta^2} \leq \frac{K^2 - 1}{n\delta^2}. \tag{24}$$

The result follows from combining (23) and (24). $\blacksquare$

In the following, we present yet another version by dropping the dependency on the constant C.

Lemma 3 (Alternative version: bounded likelihood ratio) *Assume $Q \ll P$ and $dQ/dP \leq B$. Then,*

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} |\hat{F}_{Q,n}(x) - F_Q(x)| > \delta \right) \leq (72/\delta)e^{-n\delta^2/(4B)} + 2e^{-n\delta^2/(2B^2)}$$

for all $\delta > 0$.

Proof: To prove this lemma, we compute the upper bound for the first term of (17) differently. According to Theorem 8, we get

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}\left|\int f_x \; d(P_n - P)\right| > \epsilon\right) \leq 4N_{[]}(\epsilon/8, \mathcal{F}, L^1(P)) \; e^{-n\epsilon^2/B} \tag{25}$$

for $\epsilon \leq 2/3$, where $N_{[]}(\epsilon/8, \mathcal{F}, L^1(P))$ denotes the bracketting number for function class $\mathcal{F} = \{f_x : x \in \mathbb{R}\}$. Similar to the proof of Lemma 1, it can be shown that $N_{[]}(\epsilon, \mathcal{F}, L^1(P)) \leq 2 + 1/\epsilon$. Therefore,

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}\left|\int f_x \; d(P_n - P)\right| > \epsilon\right) \leq 8(1 + 4/\epsilon) \; e^{-n\epsilon^2/B} \leq (36/\epsilon)e^{-n\epsilon^2/B}.$$

Combining with (22) for the second term of (17), we obtain

$$\begin{aligned}\mathbb{P}\left(\sup_{x\in\mathbb{R}}|\hat{F}_{Q,n}(x) - F_Q(x)| > \delta\right) &\leq (72/\delta)e^{-n\delta^2/(4B)} + 2e^{-n\delta^2/(2B^2)} \\ &\leq (74/\delta)\exp\left(-\frac{n\delta^2}{2(B+1)^2}\right). \qquad \blacksquare\end{aligned}$$

Lemma 4 (Alternative version: unbounded likelihood ratio) *Fix $B, \delta > 0$ and $\epsilon \in (0, 1/2)$. It holds that,*

$$\sup_{x\in\mathbb{R}}\left|F_Q(x) - \hat{F}_{Q,n}(x)\right| \leq 3\delta + 4\,\mathbb{P}\left(\frac{dQ}{dP}(Z_{test}) > n\right) + C'\sqrt{\frac{B}{n'}} + \mathbb{P}\left(\frac{dQ}{dP}(Z_{test}) > B\right),$$

with probability at least,

$$\begin{aligned}1 - 4e^{-n'\delta^2/(2B^2)} - 2e^{-2n^{2\epsilon}} - 2\,\mathbb{P}\left(\frac{dQ}{dP}(Z_{test}) > n\right) \\ - \frac{8}{n\delta^2}\mathbb{E}\left(\left(\frac{dQ}{dP}(Z_{calib})\right)^2 \mathbf{1}\left\{\frac{dQ}{dP}(Z_{calib}) \leq n\right\}\right),\end{aligned}$$

where $n' = n\,\mathbb{P}\left(\frac{dQ}{dP}(Z_{calib}) \leq B\right) - n^{1/2+\epsilon}$.

Proof: To drop the assumption on dQ/dP, we start by writing $\mathcal{X} = \mathcal{X}_B \cup \mathcal{X}_B^c$, where $\mathcal{X}_B = \{x : dQ_X/dP_X(x) \leq B\}$. Define

$$\begin{aligned}F_{Q|\mathcal{X}_B} &:= \mathbb{P}\left(s(Z_{\text{test}}) \leq x \middle| \frac{dQ}{dP}(Z_{\text{test}}) \leq B\right), \\ \hat{F}_{Q,n|\mathcal{X}_B} &:= \frac{\sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)\mathbf{1}\{s(Z_i) \leq x\}\mathbf{1}\left\{\frac{dQ}{dP}(Z_i) \leq B\right\}}{\sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)\mathbf{1}\left\{\frac{dQ}{dP}(Z_i) \leq B\right\}},\end{aligned}$$

where the summations are taken over calibration data points generated from P. We note,

$$\begin{aligned}\sup_{x\in\mathbb{R}}\left|F_Q(x) - \hat{F}_{Q,n}(x)\right| \leq \sup_{x\in\mathbb{R}}\left|F_Q(x) - F_{Q|\mathcal{X}_B}(x)\right| + \sup_{x\in\mathbb{R}}\left|F_{Q|\mathcal{X}_B}(x) - \hat{F}_{Q,n|\mathcal{X}_B}(x)\right| \\ + \sup_{x\in\mathbb{R}}\left|\hat{F}_{Q,n|\mathcal{X}_B}(x) - \hat{F}_{Q,n}(x)\right|.\end{aligned}$$

To upper-bound the first term, we observe that

$$\begin{aligned}\left|F_Q(x) - F_{Q|\mathcal{X}_B}(x)\right| &= \left|\mathbb{P}\left(s(Z_{\text{test}}) \leq x\right) - \mathbb{P}\left(s(Z_{\text{test}}) \leq x \middle| \frac{dQ}{dP}(Z_{\text{test}}) \leq B\right)\right| \\ &= \left|\mathbb{P}\left(s(Z_{\text{test}}) \leq x \middle| \frac{dQ}{dP}(Z_{\text{test}}) \leq B\right)\left(\mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) \leq B\right) - 1\right)\right.\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{P}\left(s(Z_{\text{test}}) \le x \middle| \frac{dQ}{dP}(Z_{\text{test}}) > B\right) \mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) > B\right) \Bigg| \\
&\le \mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) > B\right) \Bigg| \mathbb{P}\left(s(Z_{\text{test}}) \le x \middle| \frac{dQ}{dP}(Z_{\text{test}}) > B\right) \\
&\qquad\qquad - \mathbb{P}\left(s(Z_{\text{test}}) \le x \middle| \frac{dQ}{dP}(Z_{\text{test}}) \le B\right) \Bigg| \\
&\le \mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) > B\right).
\end{aligned}
$$

Regarding the second term, using the same argument as in the proof of Lemma 1, we get

$$
\mathbb{P}\left(\sup_{x\in\mathbb{R}} \left|F_{Q|\mathcal{X}_B}(x) - \hat{F}_{Q,n|\mathcal{X}_B}(x)\right| \le \delta + C'\sqrt{B/n_B} \,\middle|\, n_B\right) \ge 1 - 4e^{-n_B\delta^2/(2B^2)},
$$

where $n_B = \sum_{i\in\mathcal{I}} \mathbf{1}\left\{\frac{dQ}{dP}(Z_i) \le B\right\}$. Now let $n' = n\,\mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{calib}}) \le B\right) - n^{1/2+\epsilon}$ and observe that

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{x\in\mathbb{R}} \left|F_{Q|\mathcal{X}_B}(x) - \hat{F}_{Q,n|\mathcal{X}_B}(x)\right| \le \delta + C'\sqrt{B/n'}\right) \\
&\ge \mathbb{P}\left(\sup_{x\in\mathbb{R}} \left|F_{Q|\mathcal{X}_B}(x) - \hat{F}_{Q,n|\mathcal{X}_B}(x)\right| \le \delta + C'\sqrt{B/n_B},\, n_B \ge n'\right) \\
&= \mathbb{E}\left(\mathbb{P}\left(\sup_{x\in\mathbb{R}} \left|F_{Q|\mathcal{X}_B}(x) - \hat{F}_{Q,n|\mathcal{X}_B}(x)\right| \le \delta + C'\sqrt{B/n_B} \middle| n_B\right)\mathbf{1}\{n_B \ge n'\}\right) \\
&\ge \mathbb{E}\left(\left(1 - 4e^{-n_B\delta^2/(2B^2)}\right)\mathbf{1}\{n_B \ge n'\}\right) \\
&\ge \left(1 - 4e^{-n'\delta^2/(2B^2)}\right)\mathbb{P}\left(n_B \ge n'\right) \\
&\overset{(*)}{\ge} \left(1 - 4e^{-n'\delta^2/(2B^2)}\right)\left(1 - 2e^{-2n^{2\epsilon}}\right) \\
&\ge 1 - 4e^{-n'\delta^2/(2B^2)} - 2e^{-2n^{2\epsilon}},
\end{aligned}
$$

where $(*)$ holds according to Hoeffding's inequality.

For the third term, we note

$$
\begin{aligned}
|\hat{F}_{Q,n|\mathcal{X}_B}(x) - \hat{F}_{Q,n}(x)| &= \left|\frac{W^<(x)}{W^<} - \frac{W^<(x) + W^>(x)}{W^< + W^>}\right| \\
&= \left|\frac{W^<(x)W^> - W^<W^>(x)}{W^<(W^< + W^>)}\right| \\
&\le 2\,\frac{W^<W^>}{W^<(W^< + W^>)} = 2\,\frac{W^>}{W^< + W^>},
\end{aligned} \tag{26}
$$

where

$$
\begin{aligned}
W^<(x) &= \sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)\mathbf{1}\{s(Z_i) \le x\}\mathbf{1}\left\{\frac{dQ}{dP}(Z_i) \le B\right\}, \\
W^>(x) &= \sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)\mathbf{1}\{s(Z_i) \le x\}\mathbf{1}\left\{\frac{dQ}{dP}(Z_i) > B\right\}, \\
W^< &= \sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)\mathbf{1}\left\{\frac{dQ}{dP}(Z_i) \le B\right\}, \\
W^> &= \sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)\mathbf{1}\left\{\frac{dQ}{dP}(Z_i) > B\right\},
\end{aligned}
$$

and the inequalty 26 follows from the fact that $W^{<}(x) \leq W^{<}$ and $W^{>}(x) \leq W^{>}$ for all $x \in \mathbb{R}$. Hence,

$$\sup_{x\in\mathbb{R}} |\hat{F}_{Q,n|\mathcal{X}_B}(x) - \hat{F}_{Q,n}(x)| \leq 2\ \frac{\sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)\mathbf{1}\left\{\frac{dQ}{dP}(Z_i) > B\right\}}{\sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)} \overset{a.s.}{\to} 2\,\mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) > B\right).$$

We note,

$$\begin{aligned}
&\left|\frac{\frac{1}{n}\sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)\mathbf{1}\left\{\frac{dQ}{dP}(Z_i) > B\right\}}{\frac{1}{n}\sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)} - \mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) > B\right)\right| \leq \\
&\left|\frac{1}{n}\sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)\mathbf{1}\left\{\frac{dQ}{dP}(Z_i) > B\right\} - \mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) > B\right)\right| + \\
&\qquad \left|\frac{1}{\frac{1}{n}\sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)} - 1\right| \left|\frac{1}{n}\sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)\mathbf{1}\left\{\frac{dQ}{dP}(Z_i) > B\right\}\right| \leq \\
&\left|\frac{1}{n}\sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)\mathbf{1}\left\{\frac{dQ}{dP}(Z_i) > B\right\} - \mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) > B\right)\right| + \left|\frac{1}{n}\sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i) - 1\right| \leq \\
&\left|\frac{1}{n}\sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i)\mathbf{1}\left\{\frac{dQ}{dP}(Z_i) > B\right\} - \mathbb{P}\left(B < \frac{dQ}{dP}(Z_{\text{test}}) \leq n\right)\right| + \\
&\mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) > B, \frac{dQ}{dP}(Z_{\text{test}}) > n\right) + \left|\frac{1}{n}\sum_{i\in\mathcal{I}} \frac{dQ}{dP}(Z_i) - \mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) \leq n\right)\right| + \mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) > n\right) \\
&\leq \delta + 2\,\mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) > n\right),
\end{aligned}$$

with probability at least

$$1 - 2\,\mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) > n\right) - \frac{8}{n\delta^2}\mathbb{E}\left(\left(\frac{dQ}{dP}(Z_{\text{calib}})\right)^2 \mathbf{1}\left\{\frac{dQ}{dP}(Z_{\text{calib}}) \leq n\right\}\right),$$

where in the last step we have used (the proof of) Theorem 9 with $\epsilon = \delta/2$ and $b_n = n$ twice; once for the first term with $X_{n,k} = \frac{dQ}{dP}(Z_k)\mathbf{1}\left\{\frac{dQ}{dP}(Z_k) > B\right\}$, $k \in \mathcal{I}$, and the other time for the third term with $X_{n,k} = dQ/dP(Z_k)$, $k \in \mathcal{I}$. The requirements of the theorem are satisfied by

$$x\,\mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{calib}}) > x\right) \leq \mathbb{E}\left(\left|\frac{dQ}{dP}(Z_{\text{calib}})\right| \mathbf{1}\left\{\left|\frac{dQ}{dP}(Z_{\text{calib}})\right| > x\right\}\right) \to 0$$

as $x \to \infty$ and then according to (the proof of) Theorem 10 ; also recall that $\mathcal{I} \subset [n]$ for the split conformal method corresponds to the calibration dataset. Putting everything together, we have

$$\sup_{x\in\mathbb{R}} \left|F_Q(x) - \hat{F}_{Q,n}(x)\right| \quad \leq \quad 3\delta \ + \ 4\,\mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) > n\right) \ + \ C'\sqrt{\frac{B}{n'}} \ + \ \mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) > B\right),$$

with probability at least,

$$\begin{aligned}
1 - 4e^{-n'\delta^2/(2B^2)} - 2e^{-2n^{2\epsilon}} &- 2\,\mathbb{P}\left(\frac{dQ}{dP}(Z_{\text{test}}) > n\right) \\
&- \frac{8}{n\delta^2}\mathbb{E}\left(\left(\frac{dQ}{dP}(Z_{\text{calib}})\right)^2 \mathbf{1}\left\{\frac{dQ}{dP}(Z_{\text{calib}}) \leq n\right\}\right).
\end{aligned}$$

B Proof of Theorem 1

Recall from Section 2.3, we have

$$F_{Q,|\mathcal{I}|}(x) = w_{\text{test}}\mathbf{1}\{t=\infty\} + \sum_{i\in\mathcal{I}} w_i \mathbf{1}\{s(Z_i)\le x\}, \qquad w_i = \frac{\frac{dQ}{dP}(Z_i)}{\frac{dQ}{dP}(Z_{\text{test}}) + \sum_{i\in\mathcal{I}}\frac{dQ}{dP}(Z_i)}, \tag{27}$$

where $Z_i \overset{\text{i.i.d.}}{\sim} P$, $Z_{\text{test}} \sim Q$, and $\mathcal{I}$ denotes the set of indices corresponding to the calibration dataset. We note

$$\begin{aligned} P_e(\mathbf{D}_n) = \mathbb{P}\left(Y_{\text{test}} \notin \hat{C}_\alpha^{\text{split}}(X_{\text{test}})\middle|\mathbf{D}_n\right) &= \mathbb{P}\left(s(X_{\text{test}}, Y_{\text{test}}; \hat{\mu}) > F_{Q,m}^{-1}(1-\alpha)\middle|\mathbf{D}_n\right) \\ &\le \mathbb{P}\left(s(X_{\text{test}}, Y_{\text{test}}; \hat{\mu}) > \hat{F}_{Q,m}^{-1}(1-\alpha)\middle|\mathbf{D}_n\right) \\ &= 1 - F_Q(\hat{F}_{Q,m}^{-1}(1-\alpha)), \end{aligned}$$

where $F_Q(x) = \mathbb{P}_{Z\sim Q}(s(Z)\le x)$ and $\hat{F}_{Q,m}^{-1}$ is defined according to (16) with index set $\mathcal{I}$ instead of $[n]$. Using the weighted DKW inequality from Lemma 1 we get

$$F_Q(\hat{F}_{Q,m}^{-1}(1-\alpha)) \ge \hat{F}_{Q,m}(\hat{F}_{Q,m}^{-1}(1-\alpha)) - \delta - C'\sqrt{\frac{B}{m}} \ge 1-\alpha-\delta-C'\sqrt{\frac{B}{m}}, \quad \text{whp.}$$

Hence, we have

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + \delta + C'\sqrt{\frac{B}{m}}\right) \le 4e^{-m\delta^2/(2B^2)}$$

for all $\delta > 0$, or equivalently,

$$\begin{aligned} \mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + \sqrt{\frac{2B^2}{m}\log\frac{4}{\delta}} + C'\sqrt{\frac{B}{m}}\right) &= \mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + \left(\sqrt{2B\log 4/\delta} + C'\right)\sqrt{\frac{B}{m}}\right) \\ &\le \delta. \end{aligned}$$

This completes the proof of Theorem 1.

B.1 Proof of Remark 1

To see (8), we note that by using Lemma 2 instead of Lemma 1 we get

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + \delta\right) \le \frac{6\,CK}{\delta\sqrt{m}} + \frac{4(K^2-1)}{m\delta^2}$$

for all $\delta > 0$. Hence,

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + 2\left(\frac{2CK + \sqrt{\delta(K^2-1)}}{\delta\sqrt{m}}\right)\right) \le \delta$$

for all $\delta > 0$, which implies

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + \frac{2K(2C+1)}{\delta\sqrt{m}}\right) \le \delta, \qquad 0 < \delta \le 1.$$

■.

C Proof for Jackknife+

Lemma 5 *If Assumption 1 and 2 hold, then*

$$\mathbb{P}\left(\left\|\hat{\beta}_n - \mathbb{E}\,\hat{\beta}_n\right\|_\infty \geq \epsilon\right) \leq 2p\exp\left(-\frac{2\kappa_1^2\epsilon^2}{nc_n^2}\right).$$

Proof: Assumption 1 and 2 imply that

$$\sup_{z_1,\ldots,z_n,z_i'} \|T(z_1,\ldots,z_i\ldots,z_n) - T(z_1,\ldots,z_i',\ldots,z_n)\|_\infty \leq \frac{c_n}{\kappa_1}.$$

By McDiarmid's inequality McDiarmid et al. (1989) we get

$$\mathbb{P}\left(\|\hat{\beta}_n - \mathbb{E}\,\hat{\beta}_n\|_\infty \geq \epsilon\right) = \mathbb{P}\left(\left\|T(Z_1,\ldots,Z_n) - \mathbb{E}\;T(Z_1,\ldots,Z_n)\right\|_\infty \geq \epsilon\right) \leq 2p\exp\left(-\frac{2\kappa_1^2\epsilon^2}{nc_n^2}\right)$$

for independent Z_i and all $\epsilon > 0$. ■

Lemma 6 *Under Assumptions 1 and 2 we have*

$$\mathbb{P}\left(\max_i \left\|\mu_{\hat{\beta}_{-i}} - \mu_{\overline{\beta}_{-1}}\right\|_\infty \geq \epsilon\right) \leq 2p\exp\left(-\frac{2\kappa_1^2}{n}\left(\frac{\epsilon}{\kappa_2 c_{n-1}} - \frac{1}{\kappa_1}\right)^2\right).$$

Proof: From Assumption 1 and 2, it follows that

$$\max_{i,j} \|\hat{\beta}_{-i} - \hat{\beta}_{-j}\|_\infty \leq \frac{c_{n-1}}{\kappa_1}. \tag{28}$$

Also, according to (5), we have $\|\hat{\beta}_{-1} - \overline{\beta}_{-1}\|_\infty < \epsilon$ with probability at least $1 - 2p\exp(-2\kappa_1^2\epsilon^2/(nc_{n-1}^2))$. We note that,

$$\begin{aligned}
\mathbb{P}\left(\max_i \left\|\mu_{\hat{\beta}_{-i}} - \mu_{\overline{\beta}_{-1}}\right\|_\infty \geq \epsilon\right) &\overset{(*)}{\leq} \mathbb{P}\left(\kappa_2 \max_i \left\|\hat{\beta}_{-i} - \overline{\beta}_{-1}\right\|_\infty \geq \epsilon\right) \\
&\overset{(**)}{\leq} \mathbb{P}\left(\kappa_2\left(\frac{c_{n-1}}{\kappa_1} + \left\|\hat{\beta}_{-1} - \overline{\beta}_{-1}\right\|_\infty\right) \geq \epsilon\right) \\
&\leq 2p\exp\left(-\frac{2\kappa_1^2}{n}\left(\frac{\epsilon}{\kappa_2 c_{n-1}} - \frac{1}{\kappa_1}\right)^2\right),
\end{aligned}$$

where (*) and (**) hold according to Assumption 2 and (28), respectively. ■

C.1 Proof of Theorem 3

We note,

$$\begin{aligned}
\hat{\mathcal{C}}_\alpha^{\mathrm{J+}}(X_{\mathrm{test}}) &\supseteq \left\{y \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\left|Y_i - \mu_{\hat{\beta}_{-i}}(X_i)\right| \geq \left|y - \mu_{\hat{\beta}_{-i}}(X_{\mathrm{test}})\right|\right\} > \alpha\right\} \\
&\supseteq \left\{y \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\left|Y_i - \mu_{\overline{\beta}_{-1}}(X_i)\right| - \left|\mu_{\hat{\beta}_{-i}}(X_i) - \mu_{\overline{\beta}_{-1}}(X_i)\right| \geq \right.\right. \\
&\qquad\qquad \left.\left.\left|y - \mu_{\overline{\beta}_{-1}}(X_{\mathrm{test}})\right| + \left|\mu_{\hat{\beta}_{-i}}(X_{\mathrm{test}}) - \mu_{\overline{\beta}_{-1}}(X_{\mathrm{test}})\right|\right\} > \alpha\right\},
\end{aligned}$$

where the first relation holds according to Bian and Barber (2023). This first step can also be recovered by letting $w_i = 1/(n+1)$ and $\hat{w}_i = 1/n$ in the first four steps of C.2. Assuming $\max_i \|\mu_{\hat{\beta}_{-i}} - \mu_{\overline{\beta}_{-1}}\|_\infty < \epsilon$, we obtain

$$\begin{aligned}\hat{\mathcal{C}}_\alpha^{\mathrm{J+}}(X_{\mathrm{test}}) &\supseteq \left\{ y \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{ \left|Y_i - \mu_{\overline{\beta}_{-1}}(X_i)\right| \geq \left|y - \mu_{\overline{\beta}_{-1}}(X_{\mathrm{test}})\right| + 2\epsilon \right\} > \alpha \right\} \\ &\supseteq \left\{ y \in \mathbb{R} : 1 - \hat{F}^{(n-1)}\left(\left|y - \mu_{\overline{\beta}_{-1}}(X_{\mathrm{test}})\right| + 2\epsilon \right) > \alpha \right\}.\end{aligned}$$

Assuming $\left\|\hat{F}^{(n-1)} - F^{(n-1)}\right\|_\infty < \delta$, we obtain

$$\begin{aligned}\hat{\mathcal{C}}_\alpha^{\mathrm{J+}}(X_{\mathrm{test}}) &\supseteq \left\{ y \in \mathbb{R} : 1 - F^{(n-1)}\left(\left|y - \mu_{\overline{\beta}_{-1}}(X_{\mathrm{test}})\right| + 2\epsilon \right) > \alpha + \delta \right\} \\ &\supseteq \left\{ y \in \mathbb{R} : 1 - F^{(n-1)}\left(\left|y - \mu_{\overline{\beta}_{-1}}(X_{\mathrm{test}})\right| \right) > \alpha + \delta + 2\epsilon L_{n-1} \right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}P_e(\mathbf{D}_n) = \mathbb{P}(Y_{\mathrm{test}} \notin \hat{\mathcal{C}}_\alpha^{\mathrm{J+}}(X_{\mathrm{test}})|\mathbf{D}_n) &\leq \mathbb{P}\left(1 - F^{(n-1)}\left(\left|Y_{\mathrm{test}} - \mu_{\overline{\beta}_{-1}}(X_{\mathrm{test}})\right|\right) \leq \alpha + \delta + 2\epsilon L_{n-1}\right) \\ &= \alpha + \delta + 2\epsilon L_{n-1}\end{aligned}$$

for $\mathbf{D}_n \in \mathcal{A} \cap \mathcal{B}$ where $\mathcal{A} := \left\{ D : \max_i \|\mu_{\hat{\beta}_{-i}} - \mu_{\overline{\beta}_{-1}}\|_\infty < \epsilon \right\}$ and $\mathcal{B} := \left\{ D : \left\|\hat{F}^{(n-1)} - F^{(n-1)}\right\|_\infty < \delta \right\}$. From Lemma 6, we know

$$\mathbb{P}(\mathbf{D}_n \notin \mathcal{A}) \leq 2p \exp\left(-\frac{2\kappa_1^2}{n}\left(\frac{\epsilon}{\kappa_2 c_{n-1}} - \frac{1}{\kappa_1} \right)^2 \right).$$

Also, according to Dvoretzky–Kiefer–Wolfowitz inequality Dvoretzky et al. (1956), we have $\mathbb{P}(\mathbf{D}_n \notin \mathcal{B}) \leq 2e^{-2n\delta^2}$. Thus,

$$\mathbb{P}(P_e(\mathbf{D}_n) > \alpha + \delta + \epsilon) \leq \mathbb{P}\left((\mathcal{A} \cap \mathcal{B})^c\right) \leq 2e^{-2n\delta^2} + 2p \exp\left(-\frac{2\kappa_1^2}{n}\left(\frac{\epsilon}{2L_{n-1}\kappa_2 c_{n-1}} - \frac{1}{\kappa_1} \right)^2 \right),$$

or equivalently,

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + \sqrt{\frac{\log(2/\delta)}{2n}} + 2\, L_{n-1}\, \kappa_2\, c_{n-1} \left(\frac{1}{\kappa_1} + \sqrt{\frac{n}{2\kappa_1^2} \log \frac{2p}{\epsilon}} \right) \right) \leq \epsilon + \delta. \quad \blacksquare$$

C.2 Proof of Theorem 4

We note,

$$\begin{aligned}\hat{\mathcal{C}}_\alpha^{\mathrm{J+}}(X_{\mathrm{test}}) = &\left\{ y \in \mathbb{R} : \sum_{i=1}^n w_i \mathbf{1}\left\{ \mu_{\hat{\beta}_{-i}}(X_{\mathrm{test}}) + \left|Y_i - \mu_{\hat{\beta}_{-i}}(X_i)\right| < y \right\} < 1 - \alpha \right\} \bigcap \\ &\left\{ y \in \mathbb{R} : \sum_{i=1}^n w_i \mathbf{1}\left\{ \mu_{\hat{\beta}_{-i}}(X_{\mathrm{test}}) - \left|Y_i - \mu_{\hat{\beta}_{-i}}(X_i)\right| > y \right\} \leq 1 - \alpha \right\} \\ \supseteq &\left\{ y \in \mathbb{R} : \sum_{i=1}^n w_i \mathbf{1}\left\{ \left|Y_i - \mu_{\hat{\beta}_{-i}}(X_i)\right| < \left|y - \mu_{\hat{\beta}_{-i}}(X_{\mathrm{test}})\right| \right\} < 1 - \alpha \right\} \bigcap\end{aligned}$$

$$
\begin{aligned}
&\left\{y \in \mathbb{R} : \sum_{i=1}^{n} w_i \mathbf{1}\left\{\left|\mu_{\hat{\beta}_{-i}}(X_{\text{test}}) - y\right| > \left|Y_i - \mu_{\hat{\beta}_{-i}}(X_i)\right|\right\} < 1-\alpha\right\} \\
&= \left\{y \in \mathbb{R} : \sum_{i=1}^{n} w_i \mathbf{1}\left\{\left|Y_i - \mu_{\hat{\beta}_{-i}}(X_i)\right| < \left|y - \mu_{\hat{\beta}_{-i}}(X_{\text{test}})\right|\right\} < 1-\alpha\right\} \\
&\supseteq \left\{y \in \mathbb{R} : \sum_{i=1}^{n} \hat{w}_i \mathbf{1}\left\{\left|Y_i - \mu_{\hat{\beta}_{-i}}(X_i)\right| < \left|y - \mu_{\hat{\beta}_{-i}}(X_{\text{test}})\right|\right\} < 1-\alpha\right\} \\
&= \left\{y \in \mathbb{R} : \sum_{i=1}^{n} \hat{w}_i \mathbf{1}\left\{\left|Y_i - \mu_{\hat{\beta}_{-i}}(X_i)\right| \geq \left|y - \mu_{\hat{\beta}_{-i}}(X_{\text{test}})\right|\right\} > \alpha\right\} \\
&\supseteq \left\{y \in \mathbb{R} : \sum_{i=1}^{n} \hat{w}_i \mathbf{1}\left\{\left|Y_i - \mu_{\overline{\beta}_{-1}}(X_i)\right| - \left|\mu_{\hat{\beta}_{-i}}(X_i) - \mu_{\overline{\beta}_{-1}}(X_i)\right| \geq \right. \right. \\
&\qquad\qquad \left.\left. \left|y - \mu_{\overline{\beta}_{-1}}(X_{\text{test}})\right| + \left|\mu_{\hat{\beta}_{-i}}(X_{\text{test}}) - \mu_{\overline{\beta}_{-1}}(X_{\text{test}})\right|\right\} > \alpha\right\},
\end{aligned}
$$

where the first and last relations hold by the definition of $\hat{\mathcal{C}}_\alpha^{\text{J+}}(X_{\text{test}})$ and triangle inequality, respectively. Assuming $\max_i \|\mu_{\hat{\beta}_{-i}} - \mu_{\overline{\beta}_{-1}}\|_\infty < \epsilon$, we obtain

$$
\begin{aligned}
\hat{\mathcal{C}}_\alpha^{\text{J+}}(X_{\text{test}}) &\supseteq \left\{y \in \mathbb{R} : \sum_{i=1}^{n} \hat{w}_i \mathbf{1}\left\{\left|Y_i - \mu_{\overline{\beta}_{-1}}(X_i)\right| \geq \left|y - \mu_{\overline{\beta}_{-1}}(X_{\text{test}})\right| + 2\epsilon\right\} > \alpha\right\} \\
&\supseteq \left\{y \in \mathbb{R} : \sum_{i=1}^{n} \hat{w}_i \mathbf{1}\left\{\left|Y_i - \mu_{\overline{\beta}_{-1}}(X_i)\right| \geq \left|y - \mu_{\overline{\beta}_{-1}}(X_{\text{test}})\right| + 2\epsilon\right\} > \alpha\right\} \\
&\supseteq \left\{y \in \mathbb{R} : 1 - \hat{F}_Q^{(n-1)}\left(\left|y - \mu_{\overline{\beta}_{-1}}(X_{\text{test}})\right| + 2\epsilon\right) > \alpha\right\},
\end{aligned}
$$

where $\hat{F}_Q^{(n-1)}(t) := \sum_{i=1}^{n} \hat{w}_i \mathbf{1}\left\{\left|Y_i - \mu_{\overline{\beta}_{-1}}(X_i)\right| \leq t\right\}$ and, $\hat{w}_i$ and w_i are defined in (16) and (27), respectively. Define,

$$
F_Q^{(n-1)}(t) := \mathbb{P}_{Z_1 \sim Q}\left(\left|Y_1 - \mu_{\overline{\beta}_{-1}}(X_1)\right| \leq t\right).
$$

Assuming $\left\|\hat{F}_Q^{(n-1)} - F_Q^{(n-1)}\right\|_\infty < \delta$, we obtain

$$
\begin{aligned}
\hat{\mathcal{C}}_\alpha^{\text{J+}}(X_{\text{test}}) &\supseteq \left\{y \in \mathbb{R} : 1 - F_Q^{(n-1)}\left(\left|y - \mu_{\overline{\beta}_{-1}}(X_{\text{test}})\right| + 2\epsilon\right) > \alpha + \delta\right\} \\
&\supseteq \left\{y \in \mathbb{R} : 1 - F_Q^{(n-1)}\left(\left|y - \mu_{\overline{\beta}_{-1}}(X_{\text{test}})\right|\right) > \alpha + \delta + 2\epsilon L_{Q,n-1}\right\}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
P_e(\mathbf{D}_n) = \mathbb{P}(Y_{\text{test}} \notin \hat{\mathcal{C}}_\alpha^{\text{J+}}(X_{\text{test}})|\mathbf{D}_n) &\leq \mathbb{P}\left(1 - F_Q^{(n-1)}\left(\left|Y_{\text{test}} - \mu_{\overline{\beta}_{-1}}(X_{\text{test}})\right|\right) \leq \alpha + \delta + 2\epsilon L_{Q,n-1}\right) \\
&= \alpha + \delta + 2\epsilon L_{Q,n-1}
\end{aligned}
$$

for $\mathbf{D}_n \in \mathcal{A} \cap \mathcal{B}(\delta)$ where $\mathcal{A} := \left\{D : \max_i \|\mu_{\hat{\beta}_{-i}} - \mu_{\overline{\beta}_{-1}}\|_\infty < \epsilon\right\}$ and $\mathcal{B}(\delta) := \left\{D : \left\|\hat{F}_Q^{(n-1)} - F_Q^{(n-1)}\right\|_\infty \leq \delta\right\}$. From Lemma 6, we know

$$
\mathbb{P}(\mathbf{D}_n \notin \mathcal{A}) \leq 2p \exp\left(-\frac{2\kappa_1^2}{n}\left(\frac{\epsilon}{\kappa_2 c_{n-1}} - \frac{1}{\kappa_1}\right)^2\right).
$$

Also, according to weighted DKW inequality from Lemma 1, we have

$$\mathbb{P}\left(\mathbf{D}_n \notin \mathcal{B}\left(\delta + 2C\sqrt{\frac{B}{n}}\right)\right) \leq 4e^{-n\delta^2/(2B^2)}.$$

Thus,

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + \delta + \epsilon + 2C\sqrt{\frac{B}{n}}\right) \leq 4e^{-n\delta^2/(2B^2)} + 2p\exp\left(-\frac{2\kappa_1^2}{n}\left(\frac{\epsilon}{2L_{Q,n-1}\kappa_2 c_{n-1}} - \frac{1}{\kappa_1}\right)^2\right),$$

or equivalently,

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + \left(\sqrt{2B\log\frac{4}{\delta}} + 2C\right)\sqrt{\frac{B}{n}} + 2\,L_{Q,n-1}\,\kappa_2\,c_{n-1}\left(\frac{1}{\kappa_1} + \sqrt{\frac{n}{2\kappa_1^2}\log\frac{2p}{\epsilon}}\right)\right) \leq \epsilon + \delta. \qquad \blacksquare$$

D Proof for full conformal

Lemma 7 *Under Assumptions 1 and 2, we have*

$$\mathbb{P}\left(\left\|\mu_{\hat{\beta}_n} - \mu_{\overline{\beta}_n}\right\|_\infty \geq \epsilon\right) \leq 2p\exp\left(-\frac{2\kappa_1^2\epsilon^2}{n\kappa_2^2 c_n^2}\right).$$

Proof: According to Lemma 5, we have $\|\hat{\beta}_n - \overline{\beta}_n\|_\infty < \epsilon$ with probability at least $1 - 2p\exp\left(-\frac{2\kappa_1^2\epsilon^2}{nc_n^2}\right)$. It follows from Assumption 2 that,

$$\mathbb{P}\left(\left\|\mu_{\hat{\beta}_n} - \mu_{\overline{\beta}_n}\right\|_\infty \geq \epsilon\right) \leq \mathbb{P}\left(\kappa_2\left\|\hat{\beta}_n - \overline{\beta}_n\right\|_\infty \geq \epsilon\right) \leq 2p\exp\left(-\frac{2\kappa_1^2\epsilon^2}{n\kappa_2^2 c_n^2}\right). \qquad \blacksquare$$

We need the following notation for the proof of Theorem 5:

$$\hat{\beta}_{X_{\text{test}},y} := T(((X_1,Y_1),\ldots,(X_n,Y_n),(X_{\text{test}},y))).$$

D.1 Proof of Theorem 5

We note,

$$\begin{aligned}
\hat{\mathcal{C}}_\alpha^{\text{full}}(X_{\text{test}}) &\supseteq \left\{y \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\left|Y_i - \mu_{\hat{\beta}_{X_{\text{test}},y}}(X_i)\right| \geq \left|y - \mu_{\hat{\beta}_{X_{\text{test}},y}}(X_{\text{test}})\right|\right\} > \alpha\right\} \\
&\supseteq \left\{y \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\left|Y_i - \mu_{\hat{\beta}_n}(X_i)\right| - \left|\mu_{\hat{\beta}_n}(X_i) - \mu_{\hat{\beta}_{X_{\text{test}},y}}(X_i)\right| \geq \right.\right. \\
&\qquad\qquad \left.\left.\left|y - \mu_{\hat{\beta}_n}(X_{\text{test}})\right| + \left|\mu_{\hat{\beta}_n}(X_{\text{test}}) - \mu_{\hat{\beta}_{X_{\text{test}},y}}(X_{\text{test}})\right|\right\} > \alpha\right\} \\
&\supseteq \left\{y \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\left|Y_i - \mu_{\hat{\beta}_n}(X_i)\right| \geq \left|y - \mu_{\hat{\beta}_n}(X_{\text{test}})\right| + c_{n+1}\right\} > \alpha\right\},
\end{aligned}$$

where the first and last relations hold according to the definition of $\hat{\mathcal{C}}_\alpha^{\text{full}}(X_{\text{test}})$ and Assumption 1. Assuming $\|\mu_{\hat{\beta}_n} - \mu_{\overline{\beta}_n}\|_\infty < \epsilon$, we obtain

$$\begin{aligned}\hat{\mathcal{C}}_\alpha^{\text{full}}(X_{\text{test}}) &\supseteq \left\{ y \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{ \left|Y_i - \mu_{\overline{\beta}_n}(X_i)\right| \geq \left|y - \mu_{\overline{\beta}_n}(X_{\text{test}})\right| + c_{n+1} + 2\epsilon \right\} > \alpha \right\} \\ &\supseteq \left\{ y \in \mathbb{R} : 1 - \hat{F}^{(n)}\left(\left|y - \mu_{\overline{\beta}_n}(X_{\text{test}})\right| + c_{n+1} + 2\epsilon \right) > \alpha \right\}.\end{aligned}$$

Assuming $\left\|\hat{F}^{(n)} - F^{(n)}\right\|_\infty < \delta$, we obtain

$$\begin{aligned}\hat{\mathcal{C}}_\alpha^{\text{full}}(X_{\text{test}}) &\supseteq \left\{ y \in \mathbb{R} : 1 - F^{(n)}\left(\left|y - \mu_{\overline{\beta}_n}(X_{\text{test}})\right| + c_{n+1} + 2\epsilon \right) > \alpha + \delta \right\} \\ &\supseteq \left\{ y \in \mathbb{R} : 1 - F^{(n)}\left(\left|y - \mu_{\overline{\beta}_n}(X_{\text{test}})\right| \right) > \alpha + \delta + (2\epsilon + c_{n+1})L_n \right\}.\end{aligned}$$

Therefore,

$$\begin{aligned}P_e(\mathbf{D}_n) &= \mathbb{P}(Y_{\text{test}} \notin \hat{\mathcal{C}}_\alpha^{\text{full}}(X_{\text{test}})|\mathbf{D}_n) \\ &\leq \mathbb{P}\left(1 - F^{(n)}\left(\left|Y_{\text{test}} - \mu_{\overline{\beta}_n}(X_{\text{test}})\right|\right) \leq \alpha + \delta + (2\epsilon + c_{n+1})L_n\right) \\ &= \alpha + \delta + (2\epsilon + c_{n+1})L_n\end{aligned}$$

for $\mathbf{D}_n \in \mathcal{A} \cap \mathcal{B}$ where $\mathcal{A} := \left\{ D : \|\mu_{\hat{\beta}_n} - \mu_{\overline{\beta}_n}\|_\infty < \epsilon \right\}$ and $\mathcal{B} := \left\{ D : \left\|\hat{F}^{(n)} - F^{(n)}\right\|_\infty < \delta \right\}$. From Lemma 6, we know $\mathbb{P}(\mathbf{D}_n \notin \mathcal{A}) \leq 2p\exp\left(-\frac{2\kappa_1^2\epsilon^2}{n\kappa_2^2 c_n^2}\right)$. Also, according to Dvoretzky–Kiefer–Wolfowitz inequality, we have $\mathbb{P}(\mathbf{D}_n \notin \mathcal{B}) \leq 2e^{-2n\delta^2}$. Thus,

$$\mathbb{P}(P_e(\mathbf{D}_n) > \alpha + \delta + \epsilon) \leq \mathbb{P}\left((\mathcal{A} \cap \mathcal{B})^c\right) \leq 2e^{-2n\delta^2} + 2p\exp\left(-\left(\frac{\kappa_1(\epsilon/L_n - c_{n+1})}{\sqrt{2n}\kappa_2 c_n}\right)^2\right),$$

or equivalently,

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + \sqrt{\frac{\log(2/\delta)}{2n}} + L_n\left(c_{n+1} + \sqrt{2n\log\frac{2p}{\epsilon}}\,\frac{\kappa_2\, c_n}{\kappa_1}\right)\right) \leq \epsilon + \delta. \quad \blacksquare$$

D.2 Proof of Theorem 6

We note,

$$\begin{aligned}\hat{\mathcal{C}}_\alpha^{\text{full}}(X_{\text{test}}) &= \left\{ y \in \mathbb{R} : \sum_{i=1}^n w_i \mathbf{1}\left\{ \left|Y_i - \mu_{\hat{\beta}_{X_{\text{test}},y}}(X_i)\right| < \left|y - \mu_{\hat{\beta}_{X_{\text{test}},y}}(X_{\text{test}})\right| \right\} < 1 - \alpha \right\} \\ &\supseteq \left\{ y \in \mathbb{R} : \sum_{i=1}^n \hat{w}_i \mathbf{1}\left\{ \left|Y_i - \mu_{\hat{\beta}_{X_{\text{test}},y}}(X_i)\right| < \left|y - \mu_{\hat{\beta}_{X_{\text{test}},y}}(X_{\text{test}})\right| \right\} < 1 - \alpha \right\} \\ &= \left\{ y \in \mathbb{R} : \sum_{i=1}^n \hat{w}_i \mathbf{1}\left\{ \left|Y_i - \mu_{\hat{\beta}_{X_{\text{test}},y}}(X_i)\right| \geq \left|y - \mu_{\hat{\beta}_{X_{\text{test}},y}}(X_{\text{test}})\right| \right\} > \alpha \right\} \\ &\supseteq \left\{ y \in \mathbb{R} : \sum_{i=1}^n \hat{w}_i \mathbf{1}\left\{ \left|Y_i - \mu_{\hat{\beta}_n}(X_i)\right| - \left|\mu_{\hat{\beta}_n}(X_i) - \mu_{\hat{\beta}_{X_{\text{test}},y}}(X_i)\right| \geq \right.\right. \\ &\qquad\qquad \left.\left. \left|y - \mu_{\hat{\beta}_n}(X_{\text{test}})\right| + \left|\mu_{\hat{\beta}_n}(X_{\text{test}}) - \mu_{\hat{\beta}_{X_{\text{test}},y}}(X_{\text{test}})\right| \right\} > \alpha \right\}\end{aligned}$$

$$\supseteq \left\{ y \in \mathbb{R} : \sum_{i=1}^{n} \hat{w}_i \mathbf{1}\left\{ \left| Y_i - \mu_{\hat{\beta}_n}(X_i) \right| \geq \left| y - \mu_{\hat{\beta}_n}(X_{\text{test}}) \right| + c_{n+1} \right\} > \alpha \right\},$$

where the first and last relations hold according to the definition of $\hat{\mathcal{C}}_\alpha^{\text{full}}(X_{\text{test}})$ (under covariate shift) and Assumption 1. Assuming $\|\mu_{\hat{\beta}_n} - \mu_{\overline{\beta}_n}\|_\infty < \epsilon$, we obtain

$$\begin{aligned}
\hat{\mathcal{C}}_\alpha^{\text{full}}(X_{\text{test}}) &\supseteq \left\{ y \in \mathbb{R} : \sum_{i=1}^{n} \hat{w}_i \mathbf{1}\left\{ \left| Y_i - \mu_{\overline{\beta}_n}(X_i) \right| \geq \left| y - \mu_{\overline{\beta}_n}(X_{\text{test}}) \right| + c_{n+1} + 2\epsilon \right\} > \alpha \right\} \\
&\supseteq \left\{ y \in \mathbb{R} : 1 - \hat{F}_Q^{(n)} \left(\left| y - \mu_{\overline{\beta}_n}(X_{\text{test}}) \right| + c_{n+1} + 2\epsilon \right) > \alpha \right\}.
\end{aligned}$$

Assuming $\left\| \hat{F}_Q^{(n)} - F_Q^{(n)} \right\|_\infty < \delta$, we obtain

$$\begin{aligned}
\hat{\mathcal{C}}_\alpha^{\text{full}}(X_{\text{test}}) &\supseteq \left\{ y \in \mathbb{R} : 1 - F_Q^{(n)} \left(\left| y - \mu_{\overline{\beta}_n}(X_{\text{test}}) \right| + c_{n+1} + 2\epsilon \right) > \alpha + \delta \right\} \\
&\supseteq \left\{ y \in \mathbb{R} : 1 - F_Q^{(n)} \left(\left| y - \mu_{\overline{\beta}_n}(X_{\text{test}}) \right| \right) > \alpha + \delta + (2\epsilon + c_{n+1}) L_{Q,n} \right\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
P_e(\mathbf{D}_n) &= \mathbb{P}(Y_{\text{test}} \notin \hat{\mathcal{C}}_\alpha^{\text{full}}(X_{\text{test}}) | \mathbf{D}_n) \\
&\leq \mathbb{P}\left(1 - F_Q^{(n)} \left(\left| Y_{\text{test}} - \mu_{\overline{\beta}_n}(X_{\text{test}}) \right| \right) \leq \alpha + \delta + (2\epsilon + c_{n+1}) L_{Q,n} \right) \\
&= \alpha + \delta + (2\epsilon + c_{n+1}) L_{Q,n}
\end{aligned}$$

for $\mathbf{D}_n \in \mathcal{A} \cap \mathcal{B}(\delta)$ where $\mathcal{A} := \left\{ D : \|\mu_{\hat{\beta}_n} - \mu_{\overline{\beta}_n}\|_\infty < \epsilon \right\}$ and $\mathcal{B}(\delta) := \left\{ D : \left\| \hat{F}_Q^{(n)} - F_Q^{(n)} \right\|_\infty \leq \delta \right\}$. From Lemma 6, we know $\mathbb{P}(\mathbf{D}_n \notin \mathcal{A}) \leq 2p \exp\left(-\frac{2\kappa_1^2 \epsilon^2}{n \kappa_2^2 c_n^2} \right)$. Also, according to weighted DKW inequality from Lemma 1, we have

$$\mathbb{P}\left(\mathbf{D}_n \notin \mathcal{B}\left(\delta + 2C\sqrt{\frac{B}{n}} \right) \right) \leq 4e^{-n\delta^2/(2B^2)}.$$

Thus,

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + \delta + 2C\sqrt{\frac{B}{n}} + \epsilon \right) \leq 4e^{-n\delta^2/(2B^2)} + 2p \exp\left(-\left(\frac{\kappa_1(\epsilon/L_{Q,n} - c_{n+1})}{\sqrt{2n}\kappa_2 c_n} \right)^2 \right),$$

or equivalently,

$$\mathbb{P}\left(P_e(\mathbf{D}_n) > \alpha + \left(\sqrt{2B \log 4/\delta} + 2C \right) \sqrt{\frac{B}{n}} + L_{Q,n} \left(c_{n+1} + \sqrt{2n \log \frac{2p}{\epsilon}} \frac{\kappa_2 c_n}{\kappa_1} \right) \right) \leq \epsilon + \delta.$$

■

E Experiment Details for Figure 1

The sizes of the training, calibration, and test datasets are 525, 225, and 247, respectively, for the Wine Quality, Abalone, and Combined Cycle Power Plant datasets. For the Concrete Compressive Strength dataset, the corresponding sizes are 360, 155, and 169. Distribution shift is introduced following the approach of Tibshirani et al. (2019), by resampling data points with probabilities proportional to $\exp(x^\top \beta)$, where

$$\beta_{\text{wine quality}} = [0.5, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0],$$

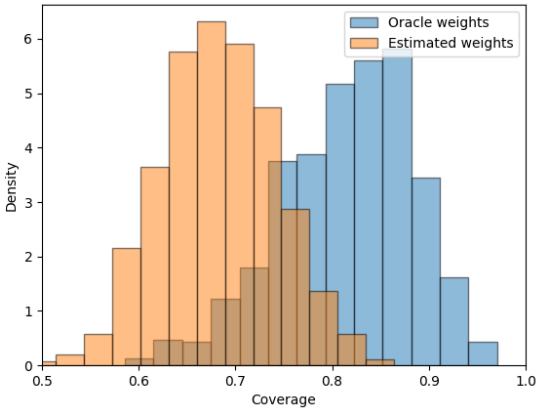

Figure 2: The likelihood ratio is estimated using kernel density estimator with Gaussian kernel.

$$\beta_{\text{abalone}} = [0, 10, 10, 0, 0, 0, 0, 0],$$
$$\beta_{\text{CCS}} = [0.01, 0.01, 0, 0, 0, 0, 0, 0],$$
$$\beta_{\text{CCPP}} = [0.2, 0, 0, 0].$$

As a simple illustration of the role of likelihood ratio estimation, as discussed in Section 4, Figure 2 shows the impact of likelihood ratio estimation on the distribution of the training-conditional error for the Concrete Compressive Strength dataset.