Journal of Machine Learning Research 25 (2024) 1-38 Submitted 11/23; Revised 7/24; Published 7/24

Split Conformal Prediction and Non-Exchangeable Data

Roberto I. Oliveira RIMFO@IMPA.BR
IMPA, Rio de Janeiro, Brazil.

Paulo Orenstein PAULOO@IMPA.BR
IMPA, Rio de Janeiro, Brazil.

Thiago Ramos THIAGORR@IMPA.BR
IMPA, Rio de Janeiro, Brazil.

João Vitor Romano JOAO.VITOR@IMPA.BR
IMPA, Rio de Janeiro, Brazil.

Editor: Chris Oates

Abstract

Split conformal prediction (CP) is arguably the most popular CP method for uncertainty quantification, enjoying both academic interest and widespread deployment. However, the original theoretical analysis of split CP makes the crucial assumption of data exchangeability, which hinders many real-world applications. In this paper, we present a novel theoretical framework based on concentration inequalities and decoupling properties of the data, proving that split CP remains valid for many non-exchangeable processes by adding a small coverage penalty. Through experiments with both real and synthetic data, we show that our theoretical results translate to good empirical performance under non-exchangeability, e.g., for time series and spatiotemporal data. Compared to recent conformal algorithms designed to counter specific exchangeability violations, we show that split CP is competitive in terms of coverage and interval size, with the benefit of being extremely simple and orders of magnitude faster than alternatives.

Keywords: concentration inequalities, conformal prediction, non-exchangeable data, β-mixing, uncertainty quantification

1. Introduction

Conformal prediction (CP), introduced by Vovk et al. (2005), is a set of techniques for quantifying uncertainty in the predictions of any model, under very general assumptions on the data-generating distribution. CP yields finite-sample coverage guarantees of many kinds, and has generated much recent interest (Shafer and Vovk, 2008; Lei et al., 2018; Romano et al., 2019; Angelopoulos and Bates, 2023; Cauchois et al., 2021).

A concrete and popular formulation of CP is split conformal prediction (Papadopoulos et al., 2002; Lei et al., 2018). Consider a regression setting where the data is a random sample $(X_i, Y_i)_{i=1}^n$ of covariate and response pairs $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$. Split CP proceeds as follows: (a) partition the data indices in three parts: a training set I_{train}, a calibration set I_{cal} and a test set I_{test}, each with sizes n_{train}, n_{cal} and n_{test}; (b) train a nonconformity score $\widehat{s}_{\text{train}} \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, for example the residual $\widehat{s}_{\text{train}}(x, y) = |y - \widehat{\mu}(x)|$ of an arbitrary model $\widehat{\mu}$ trained on $(X_i, Y_i)_{i \in I_{\text{train}}}$; (c) compute the empirical $(1 - \alpha)$-quantile $\widehat{q}_{1-\alpha}$ of

$\{\widehat{s}_{\text{train}}(X_i, Y_i)\}_{i \in I_{\text{cal}}}$; and (d) for each $i \in I_{\text{test}}$, define a confidence set

$$C_{1-\alpha}(X_i) \coloneqq \{y \in \mathcal{Y} : \widehat{s}_{\text{train}}(X_i, y) \leq \widehat{q}_{1-\alpha}\}.$$

We note in passing that split CP itself does not depend on test data; I_{test} is included for notational ease and theoretical convenience, as the guarantees of the method pertain to unseen data. If the data $(X_i, Y_i)_{i=1}^n$ is exchangeable, then the usual theory of conformal prediction guarantees that the sets $C_{1-\alpha}(X_i)$ have good marginal coverage over the test set; that is, for any $i \in I_{\text{test}}$,

$$\mathbb{P}[Y_i \in C_{1-\alpha}(X_i)] \geq 1 - \alpha - \eta, \tag{1}$$

where $\eta = (1-\alpha)/(n_{\text{cal}} + 1)$. Equivalently, one can take the lower bound to be $1 - \alpha$ by employing $C_{1-\alpha+\eta}$ instead. Additionally, for independent and identically distributed (iid) data and $\eta \gg 1/\min\{n_{\text{cal}}, n_{\text{test}}\}$, Lei et al. (2018) prove empirical coverage over the test set; that is, for some positive constant $c > 0$ and $\delta = \exp(-c\,\eta^2 \min\{n_{\text{cal}}, n_{\text{test}}\})$,

$$\mathbb{P}\left[\frac{1}{n_{\text{test}}} \sum_{i \in I_{\text{test}}} \mathbb{1}\{Y_i \in C_{1-\alpha}(X_i)\} \geq 1 - \alpha - \eta\right] \geq 1 - \delta. \tag{2}$$

Unfortunately, the results above are strongly reliant on the data exchangeability. Similarly, most guarantees from the classical theory of CP do not apply to several important data processes. For instance, in a time series context it is typical to predict the current value Y_i via a time-lagged vector $X_i = (Y_{i-j})_{j=1}^p$, and the resulting process $(X_i, Y_i)_{i=1}^n$ is typically far from exchangeable. Spatial models and shifting data distributions will also break the exchangeability assumption.

Several recent papers have tried to address these issues (Chernozhukov et al., 2018; Xu and Xie, 2021, 2023; Jensen et al., 2022; Gibbs and Candès, 2021, 2024; Barber et al., 2023; Zaffran et al., 2022; Feldman et al., 2022), but they generally require the introduction of new CP algorithms specifically tailored to different types of non-exchangeability. Some of these are very computationally intensive, while others only possess asymptotic guarantees. A third group requires adaptativity—that is, recalibration of the prediction set $C_{1-\alpha}$ at each time step—, which may not be desirable in applications.

The main message of this work is that on many occasions there is no need to introduce specific CP methods for non-exchangeable data. We prove that in such cases split CP possesses the marginal and empirical guarantees above, up to the addition of a slightly larger penalty term η in (1) and (2). These guarantees hold in finite samples and make no underlying assumptions on model consistency. While the penalty depends on the nature of the non-exchangeability—more specifically, on decoupling and concentration properties of the process—, we show that in practice the effect is small even for moderately dependent data, and that increasing the calibration set size is a viable corrective. Importantly, split CP is computationally simple, avoiding intensive routines such as bootstrapping, ensembling or blocking. Finally, the method is exactly the same as the one used for the iid data and attests to its robustness, which is essential to ensure its validity in practical settings.

For example, Figure 1 shows how split CP's marginal coverage (7), calculated through 10 000 simulations, behaves for an AR(1) time series and three different underlying models. The data-generating mechanism is given by $Y_t = \lambda Y_{t-1} + \varepsilon_t$, $t \in \mathbb{N}$, $\lambda \in \mathbb{R}$ and $\varepsilon_t \sim N(0, 1)$

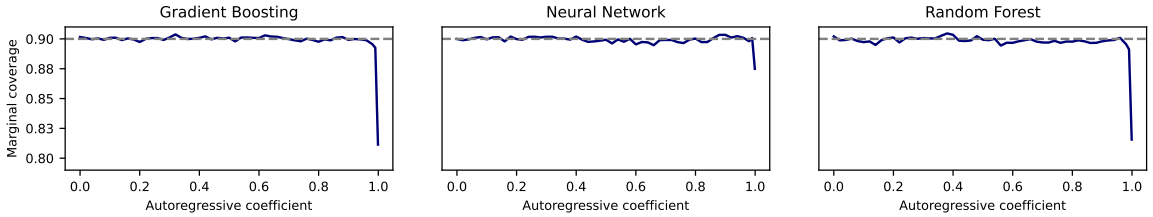

Figure 1: Marginal coverage for AR(1) process (solid) and nominally prescribed iid level of 90% (dashed) for different values of the autoregressive coefficient and three different models. Split CP holds well even under moderate dependence and undercoverage only happens at very high levels.

independently. Split conformal quantile regression (Romano et al., 2019) is employed to achieve a nominal level of 90%, with models trained on 11 lags (i.e. $X_t = (Y_{t-j})_{j=1}^{11}$) to predict the next element in the sequence. The x-axis is indexed by λ, which can be interpreted as a level of dependence in the data. Unless the dependence is very high, split CP still has adequate coverage: autoregressive coefficients up to $\lambda = 0.99$ achieve coverage higher than 89%. Significant losses of coverage only happen when $\lambda \geq 0.999$.

Our main contributions are as follows:

- We show coverage guarantees from split CP can be extended to large classes of dependent processes through the addition of a small coverage penalty;
- We do so by introducing a novel mathematical framework that is based on concentration inequalities and data decoupling properties rather than exchangeability;
- We present many generalizations that fit our framework, including both marginal, empirical and conditional coverage guarantees, as well as extensions to non-split CP methods and non-stationary data;
- We explicitly consider the broad class of stationary β-mixing processes (which includes, e.g., hidden Markov models, ARMA models and Markov chains), and show their empirical coverage bounds can match the order of the bounds under exchangeability;
- We conduct experiments that show split CP's success in generating prediction intervals in real and synthetic data, highlighting its advantages in terms of coverage, interval size, simplicity and speed compared to many recent algorithms, such as Gibbs and Candès (2024); Xu and Xie (2021, 2023); Barber et al. (2023).

Proofs are postponed to the appendices.

2. Related Works

There have been many new proposals in extending CP to non-exchangeable data. Barber et al. (2023) focus on distributional drift. They bound the *coverage gap*—i.e., the difference between nominal and actual coverage levels—by a measure of deviation from exchangeability,

which may be quite large for the time series or spatiotemporal data we deal with. Gibbs and Candès (2021, 2024) consider an online method where there is no calibration set and the quantile $\widehat{q}_{1-\alpha}$ is tuned online. They obtain very strong empirical coverage guarantees, but have no marginal coverage guarantees, and the online aspect of their method, which requires updates at every step, may be undesirable, e.g., in deployed applications. Similar comments apply to extensions and variants of this approach in Feldman et al. (2022) and Zaffran et al. (2022).

Chernozhukov et al. (2018) design prediction sets for time series data via a block-permutation method. Their Theorem 2 gives approximate guarantees in a setting including α-mixing processes, which is in principle weaker than our β-mixing assumption. However, this mixing condition interacts in a complicated way with the block structure (cf. Lemma 1 in Chernozhukov et al. 2018), which is a complicated hyperparameter that increases the computational cost. In other work, the same authors (Chernozhukov et al., 2021) achieve stronger conditional guarantees for a simpler method than in Chernozhukov et al. (2018), at the cost of much stronger assumptions. That is, they require that population objects be learned asymptotically from the data, which can be undesirable in real-world applications where one may want to be agnostic about the quality of the model. Similar comments apply to Xu and Xie (2021, 2023), which have additional hyperparameters and computationally demanding algorithms.

In contrast, our framework for split CP holds for a broad range of non-exchangeable data; retains finite-sample marginal, empirical and conditional guarantees; does not require online modifications; has no hyperparameters and is both simpler and orders of magnitude faster than these alternatives. These advantages can be seen in the experiments in Section 4.

3. Theoretical Results

We begin this section by introducing the notation for the theoretical results. Denote by $(X_i, Y_i)_{i=1}^n$ a sample of n random covariate/response pairs with stationary marginals. The pairs (X_i, Y_i) take values in $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are measurable spaces. An additional random pair (X_*, Y_*) with values in $\mathcal{X} \times \mathcal{Y}$, independent from the sample $(X_i, Y_i)_{i=1}^n$, will also be considered, and we assume $(X_i, Y_i) \sim (X_*, Y_*)$ for all $i \in [n]$, where $[n] \coloneqq \{1, \ldots, n\}$.

The data indices can be partitioned $[n] = I_{\text{train}} \sqcup I_{\text{cal}} \sqcup I_{\text{test}}$, where $n = n_{\text{train}} + n_{\text{cal}} + n_{\text{test}}$ and $I_{\text{train}} \coloneqq [n_{\text{train}}]$ corresponds to the training data, $I_{\text{cal}} \coloneqq [n_{\text{train}} + n_{\text{cal}}] \backslash [n_{\text{train}}]$ corresponds to calibration data, and $I_{\text{test}} \coloneqq [n] \backslash [n_{\text{train}} + n_{\text{cal}}]$ corresponds to test data.

For any function $s \colon (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}+1} \to \mathbb{R}$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$, denote a nonconformity score as

$$\widehat{s}_{\text{train}}(x, y) \coloneqq s((X_i, Y_i)_{i \in I_{\text{train}}}, (x, y)),$$

corresponding to the values of s when the first n_{train} pairs in the input are the training data. Intuitively, the role of $\widehat{s}_{\text{train}}$ is to measure how discrepant a prediction based on x_i is compared to the true y_i; e.g., $\widehat{s}_{\text{train}}(x, y) = |y - \widehat{\mu}(x)|$, where $\widehat{\mu}$ is some regression model trained on $(X_i, Y_i)_{i \in I_{\text{train}}}$. Several choices have been proposed in the literature (Lei et al., 2018; Hechtlinger et al., 2019; Romano et al., 2019; Angelopoulos et al., 2021).

Given $\phi \in [0,1)$, let $\widehat{q}_{\phi,\text{cal}}$ denote the empirical ϕ-quantile of $\widehat{s}_{\text{train}}(X_i, Y_i)$ over I_{cal}; i.e.,

$$\widehat{q}_{\phi,\text{cal}} \coloneqq \inf\left\{ t \in \mathbb{R} \,:\, \frac{1}{n_{\text{cal}}} \sum_{i \in I_{\text{cal}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \le t\} \ge \phi \right\}. \tag{3}$$

For $x \in \mathcal{X}$, the prediction sets are then defined via $C_\phi(x) \coloneqq \{y \in \mathcal{Y} \,:\, \widehat{s}_{\text{train}}(x,y) \le \widehat{q}_{\phi,\text{cal}}\}$. Also, given a measurable function $q\colon (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}} \to \mathbb{R}$ (which can be thought of as a quantile), define $q_{\text{train}} \coloneqq q((X_i, Y_i)_{i \in I_{\text{train}}})$ and the probability

$$P_{q,\text{train}} \coloneqq \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \le q_{\text{train}} \mid (X_i, Y_i)_{i \in I_{\text{train}}}]. \tag{4}$$

3.1 Marginal and Empirical Guarantees

This subsection details how marginal and empirical guarantees (1) and (2) can be extended when the data is not exchangeable. Some basic assumptions are needed, though they are satisfied by large classes of processes. Section 3.3 shows that is the case for stationary β-mixing data, but the general framework developed here generalizes to other processes.

First, it is necessary to have some form of concentration over the calibration data, as well as a degree of decoupling over the test data. We assume there exist $\varepsilon_{\text{cal}} \in (0,1)$ and $\delta_{\text{cal}} \in (0,1)$ such that

$$\mathbb{P}\left[\left| \frac{1}{n_{\text{cal}}} \sum_{i \in I_{\text{cal}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \le q_{\text{train}}\} - P_{q,\text{train}} \right| \le \varepsilon_{\text{cal}} \right] \ge 1 - \delta_{\text{cal}}, \tag{5}$$

where $P_{q,\text{train}}$ is defined as in (4). Intuitively, this condition requires that the empirical and population cumulative distribution functions of $\widehat{s}_{\text{train}}(X, Y)$ are close over calibration data. A key point here, however, is that this closeness should hold even when the cdf is computed over a point depending on training data.

Further, we assume that there exists $\varepsilon_{\text{train}}$ such that, for $i \in I_{\text{test}}$,

$$|\mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \le q_{\text{train}}] - \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \le q_{\text{train}}]| \le \varepsilon_{\text{train}}. \tag{6}$$

This means that (X_i, Y_i) for $i \in I_{\text{test}}$ essentially behaves like (X_*, Y_*), i.e., a data point that is independent of training data.

Conditions like (5) and (6) (with small error parameters) hold for many types of dependent data, such as strictly stationary processes whose "memory" is not too strong. Examples include finite Markov chains, hidden Markov chains and ARMA-type processes (both linear and nonlinear). A broader class of such processes can be described under so-called β-mixing conditions, which we discuss in detail in Section 3.3. Under these conditions, the usual marginal coverage guarantees can be recovered for split CP.

Theorem 1 (Marginal coverage over test data) *Given $\alpha \in (0,1)$ and $\delta_{\text{cal}} > 0$, if conditions (5) and (6) hold, then, for all $i \in I_{\text{test}}$:*

$$\mathbb{P}[Y_i \in C_{1-\alpha}(X_i)] \ge 1 - \alpha - \varepsilon_{\text{cal}} - \delta_{\text{cal}} - \varepsilon_{\text{train}}. \tag{7}$$

Additionally, if $\widehat{s}_{\text{train}}(X_*, Y_*)$ almost surely has a continuous distribution conditionally on the training data, then

$$|\mathbb{P}[Y_i \in C_{1-\alpha}(X_i)] - (1-\alpha)| \le \varepsilon_{\text{cal}} + \delta_{\text{cal}} + \varepsilon_{\text{train}}.$$

To also guarantee empirical coverage, suppose that instead of the decoupling assumption (6), there exists concentration of the empirical cumulative distribution function of the nonconformity score over the test data, that is, there exist $\varepsilon_{\text{test}}, \delta_{\text{test}} \in (0,1)$ such that

$$\mathbb{P}\left[\left|P_{q,\text{train}} - \frac{1}{n_{\text{test}}}\sum_{i\in I_{\text{test}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}\}\right| \leq \varepsilon_{\text{test}}\right] \geq 1 - \delta_{\text{test}}. \tag{8}$$

Theorem 2 (Empirical coverage over test data) *Given $\alpha \in (0,1)$, $\delta_{\text{cal}} > 0$ and $\delta_{\text{test}} > 0$, if (5) and (8) hold, then:*

$$\mathbb{P}\left[\frac{1}{n_{\text{test}}}\sum_{i\in I_{\text{test}}} \mathbb{1}\{Y_i \in C_{1-\alpha}(X_i)\} \geq 1 - \alpha - \eta\right] \geq 1 - \delta_{\text{cal}} - \delta_{\text{test}},$$

where $\eta = \varepsilon_{\text{cal}} + \varepsilon_{\text{test}}$. Additionally, if $\widehat{s}_{\text{train}}(X_*, Y_*)$ almost surely has a continuous distribution conditionally on the training data, then:

$$\mathbb{P}\left[\left|\frac{1}{n_{\text{test}}}\sum_{i\in I_{\text{test}}} \mathbb{1}\{Y_i \in C_{1-\alpha}(X_i)\} - (1-\alpha)\right| \leq \eta\right] \geq 1 - 2\delta_{\text{cal}} - 2\delta_{\text{test}}.$$

While the purpose of the above theorems is to extend split conformal guarantees to non-exchangeable data, they also readily apply to the iid case. Indeed, it is straightforward to show that, in such case, it suffices to take $\varepsilon_{\text{cal}} = \sqrt{(2n_{\text{cal}})^{-1}\log(2/\delta_{\text{cal}})}$ and $\varepsilon_{\text{test}} = \sqrt{(2n_{\text{test}})^{-1}\log(2/\delta_{\text{test}})}$.

3.2 Conditional Guarantees

Obtaining a conditional version of (1) and (2) is of interest in many cases. Barber et al. (2020) prove that coverage is not generally attainable, even for iid data. On the positive side, they show they can be achieved when conditioning on sets of finite VC dimension that are not too small. Our goal is to show similar guarantees for split conformal prediction under non-exchangeable data.

First, conditional versions of assumptions (5) and (6) are needed. For concentration over the calibration data, suppose there exist δ_{cal} and $\varepsilon_{\text{cal}} \in (0,1)$ such that, for $P_{q,\text{train}}(A) \coloneqq \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq q_{\text{train}} \mid (X_i, Y_i)_{i\in I_{\text{train}}}, X_* \in A]$ and $n_{\text{cal}}(A) \coloneqq \#\{i \in I_{\text{cal}} : X_i \in A\}$,

$$\mathbb{P}\left[\sup_{A\in\mathcal{A}}\left|\frac{1}{\max\{n_{\text{cal}}(A), 1\}}\sum_{i\in I_{\text{cal}}(A)} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}\} - P_{q,\text{train}}(A)\right| \leq \varepsilon_{\text{cal}}\right] \geq 1 - \delta_{\text{cal}}. \tag{9}$$

For a conditional version of marginal decoupling, assume there exists $\varepsilon_{\text{train}} \in (0,1)$ such that

$$|\mathbb{P}[\widehat{s}_{\text{train}}(X_k, Y_k) \leq q_{\text{train}} \mid X_k \in A] - \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq q_{\text{train}} \mid X_* \in A]| \leq \varepsilon_{\text{train}}. \tag{10}$$

These conditions suffice for conditional marginal coverage.

Theorem 3 (Conditional coverage over test data) *Given $\alpha \in (0,1)$ and $\delta_{\text{cal}} > 0$, if (9) and (10) hold, then, for each $A \in \mathcal{A} \subset \mathcal{X}$ and any $i \in I_{\text{test}}$:*

$$\mathbb{P}[Y_i \in C_{1-\alpha}(X_i; A) \mid X_i \in A] \geq 1 - \alpha - \varepsilon_{\text{cal}} - \delta_{\text{cal}} - \varepsilon_{\text{train}}. \tag{11}$$

Additionally, if $\widehat{s}_{\text{train}}(X_*, Y_*)$ almost surely has a continuous distribution conditionally on the training data, then:

$$|\mathbb{P}[Y_i \in C_{1-\alpha}(X_i; A) \mid X_i \in A] - (1-\alpha)| \leq \varepsilon_{\text{cal}} + \delta_{\text{cal}} + \varepsilon_{\text{train}}.$$

Appendix B includes a conditional version of the empirical coverage guarantee.

3.3 Application to Stationary β-Mixing Data

We now apply the framework from Section 3.1 to the class of stationary β-mixing processes. This class of non-exchangeable data is broad enough to cover many important applications, such as hidden Markov models and Markov chains (Doukhan, 2012) as well as ARMA and GARCH models (Carrasco and Chen, 2002; Mokkadem, 1988), while still providing explicit error terms in the bounds of Theorems 1 and 2.

Recall a sequence of random elements $\{Z_t\}_{t=-\infty}^{\infty}$ of a measurable space $\mathcal{Z}$ is stationary if its finite-dimensional distributions are time-invariant; that is, for any $t \in \mathbb{Z}$ and $m, k \in \mathbb{N}$,

$$Z_{t:(t+m)} = (Z_t, \ldots, Z_{t+m}) \overset{d}{=} (Z_{t+k}, \ldots, Z_{t+m+k}) = Z_{(t+k):(t+m+k)}.$$

Furthermore, for a stationary stochastic process $\{Z_t\}_{t=-\infty}^{\infty}$ and index $a \in \mathbb{N}$, the β-mixing coefficient of the process at a is defined as

$$\beta(a) = \|\mathbb{P}_{-\infty:0,a:\infty} - \mathbb{P}_{-\infty:0} \otimes \mathbb{P}_{a:\infty}\|_{\text{TV}},$$

where $\|\cdot\|_{\text{TV}}$ denotes the total variation norm, and $\mathbb{P}_{-\infty:0,a:\infty}$ is the joint distribution of the blocks $(Z_{-\infty:0}, Z_{a:\infty})$. The process is said to be β-mixing if $\beta(a) \to 0$ when $a \to \infty$.

The β-mixing condition allows us to replace independence with asymptotic independence and still retain some important concentration results. In particular, the so-called Blocking Technique (Yu, 1994; Mohri and Rostamizadeh, 2010; Kuznetsov and Mohri, 2017) allows one to compare a β-mixing process with another process made of independent blocks. The results below generally follow from combining the Blocking Technique with decoupling arguments and Bernstein's concentration inequality. Crucially, the guarantees might depend on block sizes, but split CP itself does not. This is an advantage over CP variants (Chernozhukov et al., 2018).

The sets of parameters we optimize over are defined as follows:

$$F_{\text{cal}} = \{(a, m, r) \in \mathbb{N}_{>0}^3 \ : \ 2ma = n_{\text{cal}} - r + 1, \ \delta_{\text{cal}} > 4(m-1)\beta(a) + \beta(r)\},$$
$$F_{\text{test}} = \{(a, m, s) \in \mathbb{N}_{>0}^2 \times \mathbb{N} \ : \ 2ma = n_{\text{test}} - s, \ \delta_{\text{test}} > 4(m-1)\beta(a) + \beta(n_{\text{cal}})\}.$$

These two sets correspond to block size choices in the calibration and test sets, respectively. For the calibration set, define the error term as follows:

$$\varepsilon_{\text{cal}} \coloneqq \inf_{(a,m,r) \in F_{\text{cal}}} \left\{ \widetilde{\sigma}(a) \sqrt{\frac{4}{n_{\text{cal}} - r + 1} \log\left(\frac{4}{\delta_{\text{cal}} - 4(m-1)\beta(a) - \beta(r)}\right)} \right. \tag{12}$$
$$\left. + \frac{1}{3m} \log\left(\frac{4}{\delta_{\text{cal}} - 4(m-1)\beta(a) - \beta(r)}\right) + \frac{r-1}{n_{\text{cal}}} \right\},$$

where $\widetilde{\sigma}(a) = \sqrt{\frac{1}{4} + \frac{2}{a}\sum_{j=1}^{a-1}(a-j)\beta(j)}$. Similarly, we define the test error correction factor for a stationary β-mixing process as

$$\varepsilon_{\text{test}} = \inf_{(a,m,s)\in F_{\text{test}}} \left\{ \widetilde{\sigma}(a)\sqrt{\frac{4}{n_{\text{test}}}\log\left(\frac{4}{\delta_{\text{test}} - 4(m-1)\beta(a) - \beta(n_{\text{cal}})}\right)} + \frac{1}{3m}\log\left(\frac{4}{\delta_{\text{test}} - 4(m-1)\beta(a) - \beta(n_{\text{cal}})}\right) + \frac{s}{n_{\text{test}}} \right\}. \tag{13}$$

We emphasize that this optimization of block sizes plays a role exclusively on the coverage guarantees below; the split CP algorithm itself remains unchanged.

With ε_{cal} as above, Theorem 1 yields the following result for stationary β-mixing processes:

Theorem 4 (Marginal coverage: stationary β-mixing processes) *Suppose the sample $(X_i, Y_i)_{i=1}^n$ is stationary β-mixing. Then given $\alpha \in (0,1)$ and $\delta_{\text{cal}} > 0$, for $i \in I_{\text{test}}$,*

$$\mathbb{P}\left[Y_i \in C_{1-\alpha}(X_i)\right] \geq 1 - \alpha - \eta,$$

with $\eta = \varepsilon_{\text{cal}} + \varepsilon_{\text{train}} + \delta_{\text{cal}}$, where ε_{cal} is as in (12) and $\varepsilon_{\text{train}} = \beta(i - n_{\text{train}})$. Additionally, if $\widehat{s}_{\text{train}}(X_*, Y_*)$ almost surely has a continuous distribution conditionally on the training data:

$$\left|\mathbb{P}[Y_i \in C_{1-\alpha}(X_i)] - (1-\alpha)\right| \leq \eta.$$

Under certain assumptions over the dependence of the processes, the stationary β-mixing bounds given by (12) are of the same asymptotic order as the corresponding iid bounds. Indeed, if $\beta(k) \leq k^{-b}$ and $\delta \geq n_{\text{cal}}^{-c}$ for $b > 1, c > 0$, with $1 + 2c < b$, as long as $m = o(n_{\text{cal}}^{(b-c)/(b+1)})$ and $\sqrt{n_{\text{cal}}\log(n_{\text{cal}})} = o(m)$, the bounds are of the same order. This is satisfied, for example, if $m = n_{\text{cal}}^{\lambda}$, $a = n_{\text{cal}}^{1-\lambda}/2$ with $1/2 < \lambda < (b-c)/(b+1)$.

Additionally, with $\varepsilon_{\text{test}}$ as above, Theorem 2 yields the following:

Theorem 5 (Empirical coverage: stationary β-mixing processes) *Suppose the sample $(X_i, Y_i)_{i=1}^n$ is stationary β-mixing. Then given $\alpha \in (0,1)$, $\delta_{\text{cal}} > 0$ and $\delta_{\text{test}} > 0$*

$$\mathbb{P}\left[\frac{1}{n_{\text{test}}}\sum_{i\in I_{\text{test}}} \mathbb{1}\{Y_i \in C_{1-\alpha}(X_i)\} \geq 1 - \alpha - \eta\right] \geq 1 - \delta_{\text{cal}} - \delta_{\text{test}},$$

with $\eta = \varepsilon_{\text{cal}} + \varepsilon_{\text{test}}$, and ε_{cal} and $\varepsilon_{\text{test}}$ defined in (12) and (13). Additionally, if $\widehat{s}_{\text{train}}(X_*, Y_*)$ almost surely has a continuous distribution conditionally on the training data, then:

$$\mathbb{P}\left[\left|\frac{1}{n_{\text{test}}}\sum_{i\in I_{\text{test}}} \mathbb{1}\{Y_i \in C_{1-\alpha}(X_i)\} - (1-\alpha)\right| \leq \eta\right] \geq 1 - 2\delta_{\text{cal}} - 2\delta_{\text{test}}.$$

We note in passing that the expression in (12) follows from a stationary β-mixing version of Bernstein's inequality which might be of independent interest. See Appendix C for the proof and for conditional extensions of the above results.

3.4 Extensions to Non-Split CP Methods and Non-Stationary Data

The results obtained in previous subsections also apply when the data is non-stationary. For brevity, we focus on marginal coverage. Our analysis is partly inspired by the recent work of Barber et al. (2023).

Replace the pair (X_*, Y_*) with an auxiliary process $(X_{*,i}, Y_{*,i})_{i\in[n]}$ that is an independent copy of the original data $(X_i, Y_i)_{i\in[n]}$. Let N_{cal} be a random number, uniformly distributed over I_{cal}, independently of the problem data and auxiliary process. For $j \in I_{\text{test}}$, the quantity $\delta^{(j)} := \|\text{Law}(X_j, Y_j) - \text{Law}(X_{N_{\text{cal}}}, Y_{N_{\text{cal}}})\|_{\text{TV}}$, for $j \in I_{\text{test}}$, measures how far the distribution of (X_j, Y_j) is to that of a randomly chosen point in the calibration data set (i.e., a measure of distributional drift).

For marginal coverage, we take the random variable q_{train} as before, but replace (4) by a time-inhomogeneous version, that is, $P^{(j)}_{q,\text{train}} := \mathbb{P}[\widehat{s}_{\text{train}}(X_{*,j}, Y_{*,j}) \le q_{\text{train}} \mid (X_i, Y_i)_{i\in I_{\text{train}}}]$. Furthermore, (5) and (6) are also replaced with time-inhomogeneous versions:

$$\mathbb{P}\left[\left|\frac{1}{n_{\text{cal}}}\sum_{i\in I_{\text{cal}}}(\mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \le q_{\text{train}}\} - P^{(i)}_{q,\text{train}})\right| \le \varepsilon_{\text{cal}}\right] \ge 1 - \delta_{\text{cal}}, \tag{14}$$

and, for $j \in I_{\text{test}}$,

$$|\mathbb{P}[\widehat{s}_{\text{train}}(X_j, Y_j) \le q_{\text{train}}] - \mathbb{P}[\widehat{s}_{\text{train}}(X_{*,j}, Y_{*,j}) \le q_{\text{train}}]| \le \varepsilon_{\text{train}}. \tag{15}$$

Theorem 6 (Marginal coverage for split CP on non-stationary data) *Given a level $\alpha \in (0, 1)$, if (14) and (15) hold, then for all $j \in I_{\text{test}}$:*

$$\mathbb{P}[\widehat{s}_{\text{train}}(X_j, Y_j) \le \widehat{q}_{1-\alpha,\text{cal}}] \ge 1 - \alpha - \eta - \delta^{(j)},$$

where $\eta = \varepsilon_{\text{cal}} + \delta_{\text{cal}} + \varepsilon_{\text{train}}$ is as in Theorem 1.

In particular, we recover Theorem 1 up to an error depending on how much distributional drift there is between j and the calibration set. This is similar to the main result in Barber et al. (2023), except that there the authors consider weighted calibration sets.

Finally, our framework also extends to other popular methods in the literature that are not based on split CP, such as rank-one-out (ROO) conformal prediction (Lei et al., 2018) and risk-controlling prediction sets (RCPS) (Bates et al., 2021). ROO calibrates the quantiles used for each test data point by using the remaining test points, so it is different from split CP. RCPS gives a general CP methodology that applies in a variety of settings, including regression, multiclass classification and image segmentation. Importantly, RCPS does not involve nonconformity scores but the construction of nested sets. Details for both of these extensions to non-exchangeable data are given in Appendix D.

4. Experiments

This section studies split CP's empirical performance in several numerical experiments. The first one uses real spatiotemporal climate data to compare split CP's coverage and average interval size to recent alternative conformal algorithms due to Barber et al. (2023) and Gibbs and Candès (2024). The second experiment benchmarks split CP on synthetic β-mixing

data against a popular conformal approach for time-series (Xu and Xie, 2021, 2023). In both cases, split conformal is used with absolute residuals as nonconformity score. The third example involves a hidden Markov model in which the bounds can be calculated explicitly, while the last shows that split CP's marginal and conditional guarantees work with real financial data, even when exchangeability is clearly violated. In these two final examples, as in the AR(1) experiment (cf. Figure 1), we employ split conformal quantile regression (Romano et al., 2019). The experiments were conducted on a server with 774 GB of RAM and two Intel Xeon Platinum 8354H processors, totalling 8 physical cores and 288 threads. All conformal implementations made use of multiprocessing for better performance. Further implementation details are discussed in Appendix E. Code to reproduce all the figures and tables is available at `https://github.com/jv-rv/split-conformal-nonexchangeable`.

Example 1 (Spatiotemporal climate data) Let $Y_{t,l} \in \mathbb{R}$ be the 14-day forward rolling average temperature in degrees Celsius, with t representing the first measurement day included in the average and $l \in L$ a location on Earth defined by latitudes and longitudes discretized in a $1.5^\circ \times 1.5^\circ$ grid totalling 7512 distinct locations. For all grid locations, measurements from the start of 1979 through the end of 2022 were retrieved via the National Oceanic and Atmospheric Administration (NOAA, 2023).

It will be notationally convenient to split the dates in year, month and day, so we consider instead the equivalent representation $Y_{y,m,d,l}$. A standard procedure in the meteorological literature to predict temperatures for a given date t and location l, called climatology, is to average the temperatures observed on the same day, month and location over the previous 30 years (Hwang et al., 2019; Arguez et al., 2012). More precisely, we take $X_{y,m,d,l} \coloneqq (Y_{y-j,m,d,l})_{j=1}^{30}$ as features and $\hat{\mu}(\cdot) = \text{Avg}(\cdot)$ as the prediction model. Since 30 years (1979–2008) of temperature data are needed for the features, our response series starts on 2009.

Note that the model $\hat{\mu}$ is fixed *a priori* and needs no training data, so $I_{\text{train}} = \emptyset$. For a new test point Y_{y_*,m_*,d_*,l_*}, the calibration set I_{cal} is taken as all the previous observations for that same day and month and the absolute residual is used as nonconformity score function. Thus, the set of nonconformity scores evaluated on calibration data is given by

$$\{|Y_{y,m,d,l} - \hat{\mu}(X_{y,m,d,l})| : y < y_*,\ m = m_*,\ d = d_*,\ l \in L\}. \tag{16}$$

The intuition behind this choice is that temperature measurements for a given location, day and month can be reasonably modeled as β-mixing and stationary over the years (Paura, 2006).

Split conformal prediction is efficient in generating prediction intervals in this scenario. At any given date, the calibration set for all 7512 geographic points is the same and taking the quantile of (16) allows us to build thousands of prediction intervals at once. We follow the online sequential procedure through time implied by (16), with no additional overhead due to the spatial dependence. Figure 2 shows the temperature measurements in the center, from which the spatial dependence becomes evident. The prediction intervals are represented by their minimum (left) and maximum (right). Coverage is generally attained and intervals are narrow enough to be useful.

We compare the standard and widespread split CP method (Papadopoulos et al., 2002; Vovk et al., 2005; Lei et al., 2018) against recent developments tailored to non-exchangeable

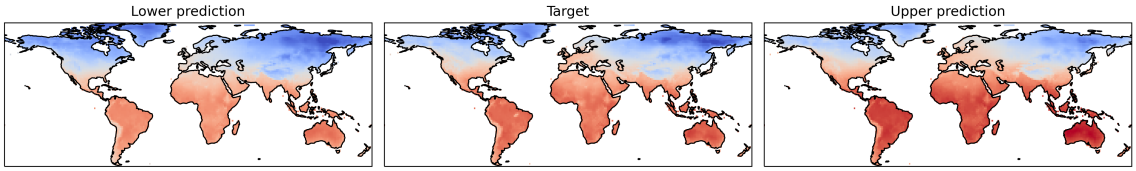

Figure 2: Average temperatures around the globe on 2022-12-31 (center), split CP's lower prediction (left) and upper prediction (right). Darker blues (reds) indicate lower (higher) temperatures. Approximately 92.6% of the 7512 grid points are inside the prediction interval, which is close to the 90% prescribed coverage. Averaging over all days, coverage is of 90.9%

data, namely NexCP due to Barber et al. (2023) and dynamically-tuned adaptive conformal inference (DtACI) due to Gibbs and Candès (2024), which we briefly describe next.

Barber et al. (2023) proposed variations of conformal prediction to deal with non-exchangeability. We focus on the split version of their NexCP algorithm, that is identical to split CP apart from the fact that quantiles are weighted to give more importance to more dependent residuals. In the original paper, this novel method is evaluated on time series data and shows improvement on coverage for data presenting distribution drift. For the non-exchangeable violations studied in our work, however, one would expect standard split CP to work well enough.

Using NexCP for spatiotemporal data requires a weighting scheme that takes into account both space and time. We deal with the time dimension as in the original paper via $\rho_{\text{time}}^{y_*-y}$ for some decay parameter $\rho_{\text{time}} \in (0, 1]$, giving less weight to observations from the distant past. In a similar vein, we employ the exponential decay $\rho_{\text{space}}^{d(l_*,l)}$ for $\rho_{\text{space}} \in (0, 1]$ and $d(l_*, l)$ a distance between the test point location l_* and the residual location l. A good approximation for the distance of two points on Earth defined in terms of latitudes and longitudes is achieved by assuming a spherical surface and making use of the haversine formula. We employ this procedure so that $d\colon \mathbb{R}^2 \times \mathbb{R}^2 \to [0, \pi]$ is the angular distance between l_* and l. The final weights are given by $\rho_{\text{time}}^{y_*-y} \cdot \rho_{\text{space}}^{d(l_*,l)}$. Calculating a weighted quantile has average time complexity $O(n_{\text{cal}} \log n_{\text{cal}})$ in contrast to $O(n_{\text{cal}})$ for the unweighted case. Moreover, NexCP must calculate a different quantile for each one of the 7512 geographic points, while Split CP makes one single calculation per time step. Therefore, one would expect NexCP to be much slower than Split CP for spatiotemporal data, even when parallelized (cf. Table 1).

Gibbs and Candès (2024) introduced dynamically-tuned adaptive conformal inference (DtACI) as an online learning method for constructing prediction intervals with a target coverage. As usual in the online setting, it is assumed that one new test point arrives at each time step. In the case of spatiotemporal data, with all grid points arriving at once, a reasonable strategy is to consider many individual time series. Indeed, this is how Covid-19 predictions are dealt with in the original paper and how we proceed.

Table 1 summarizes the comparison between split CP, NexCP and DtACI. The main takeaway is that split CP—a much simpler method, with no hyperparameters, and orders of

Method	Hyperparams	Coverage			Interval size (°C)	Avg time (min)
		Avg	SD (time)	SD (space)		
Split CP	None	90.9%	0.033	0.130	9.605	2.85
NexCP	$\rho_{\text{time}} = \rho_{\text{space}} = 0.99$	90.9%	0.033	0.130	9.601	1884.92
	$\rho_{\text{time}} = \rho_{\text{space}} = 0.9$	90.8%	0.033	0.131	9.554	
	$\rho_{\text{time}} = \rho_{\text{space}} = 0.7$	90.6%	0.033	0.132	9.446	
	$\rho_{\text{time}} = \rho_{\text{space}} = 0.5$	90.3%	0.035	0.133	9.360	
	$\rho_{\text{time}} = \rho_{\text{space}} = 0.3$	90.0%	0.038	0.134	9.289	
	$\rho_{\text{time}} = \rho_{\text{space}} = 0.1$	89.5%	0.042	0.136	9.205	
DtACI	$\lvert I\rvert = 1, \gamma \in (\frac{2^i}{1000})_{i=0}^8$	91.9%	0.034	0.096	9.658	75.15
	$\lvert I\rvert = 3, \gamma \in (\frac{2^i}{1000})_{i=0}^8$	92.1%	0.034	0.090	9.779	
	$\lvert I\rvert = 5, \gamma \in (\frac{2^i}{1000})_{i=0}^8$	92.1%	0.034	0.090	9.806	
	$\lvert I\rvert = 7, \gamma \in (\frac{2^i}{1000})_{i=0}^8$	92.1%	0.034	0.090	9.814	
	$\lvert I\rvert = 9, \gamma \in (\frac{2^i}{1000})_{i=0}^8$	92.2%	0.034	0.091	9.815	

Table 1: Comparison of conformal methods in terms of average running time, hyperparameters, average coverage, standard deviation of coverage over time and space, and average interval size. The prescribed level is 90%. While all methods are close to the nominal level, split CP is simpler, has no hyperparameter and is orders of magnitude faster for the spatiotemporal data considered.

magnitude faster—is comparable to the benchmarks both in terms of coverage and interval size, working as expected for the spatiotemporal data. NexCP's weighting scheme was useful in reducing the average interval size while maintaining a coverage above the prescribed 90% level, but the gains were small (less than 4% for the optimal choice ρ_{time} and ρ_{space} equal to 0.3). DtACI reduced the coverage standard deviation over space, but generated larger intervals.

Xu and Xie (2021, 2023) developed ensemble batch prediction intervals (EnbPI) specifically as a conformal method for time-series data. However, it crucially relies on fitting an underlying prediction algorithm on bootstrapped samples of the training data. Recall that the climatology model uses no training data and simply takes the average of the features, so $I_{\text{train}} = \emptyset$ and EnbPI cannot be applied. In order to compare split CP to this important recent development, we conducted the following experiment.

Example 2 (Hidden random walk on the cycle graph) Consider a β-mixing sequence generated from a random walk on the cycle graph. On any given state, there is a probability b of moving backward, f of moving forward and s of staying still. The dependence increases for larger values of s, which makes it more likely to have repeating values. Intuitively, the number of vertices also influence dependence, as it should take a larger sample to account for a larger number of states to be visited. Indeed, for $r \in \mathbb{N}_{>0}$ the β-mixing coefficient $\beta(r)$ for a cycle of v vertices decays at rate e^{-r/v^2}. A Gaussian noise with zero mean and variance of 10^{-6} is added to the resulting sequence to avoid draws.

Figure 3 compares the rolling coverage $\frac{1}{500}\sum_{i=j}^{499+j} \mathbb{1}\{Y_i \in C_{1-\alpha}(X_i)\}$ for $j \in \{1, \ldots, 2501\}$ of split CP and EnbPI, both using a random forest trained on 11 lags to predict the next element in the sequence as model and α set to 0.1. The implementation of EnbPI used

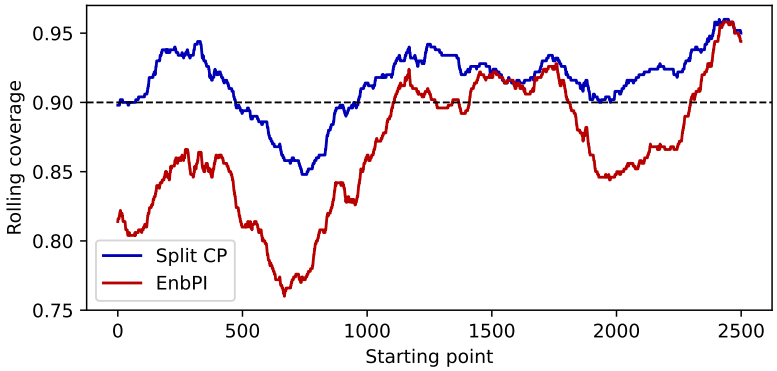

Figure 3: Split CP's rolling coverage for β-mixing data generated from a random walk on the cycle graph is close to the prescribed 90% level, whereas EnbPI significantly undercovers at some points. Considering all test points, Split CP's coverage was of 91.4% with an average interval size of 0.34; EnbPi's coverage was of 86.6% with an average interval size of 0.15.

was from `MAPIE` (Taquet et al., 2022). In this experiment, split CP was more than 8 times faster and had a rolling coverage much closer to the prescribed level. Fifteen other random sequences besides the one presented here were generated from the random walk on the cycle and the conclusion was invariably the same.

Example 3 (Two-state hidden Markov model) Let $(W_0, W_1, W_2, \ldots)$ be a Markov chain with state space $\mathcal{W} = \{0, 1\}$, probabilities $\mathbb{P}[W_t = 1 | W_{t-1} = 0] = p$ and $\mathbb{P}[W_t = 0 | W_{t-1} = 1] = q$, following the stationary distribution with $\pi = \begin{bmatrix} \frac{q}{p+q} & \frac{p}{p+q} \end{bmatrix}$, with $p, q \in (0, 1)$ and $p+q > 0$. This data is stationary β-mixing, and the mixing coefficients can be found explicitly (McDonald et al., 2015). When $p = q = 0.5$, $\beta(r) \equiv 0$ for all $r \in \mathbb{N}_{>0}$, so the Markov chain reduces to a sequence of iid Bernoulli trials. On the other hand, as p and q tend towards zero, $\beta(r)$ becomes large for every r and dependence increases. We construct a hidden Markov model by adding a Gaussian noise with zero mean and variance of 10^{-6} to the Markov chain

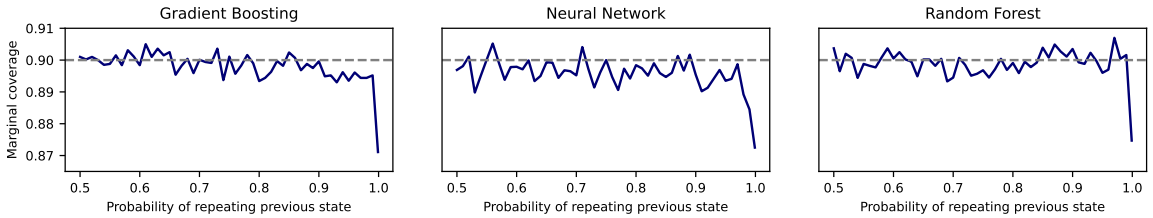

Figure 4: Marginal coverage for two-state hidden Markov model (solid) and nominally prescribed 90% target (dashed) for different levels of dependence and three different models. Coverage is above 89% for all but very extreme levels.

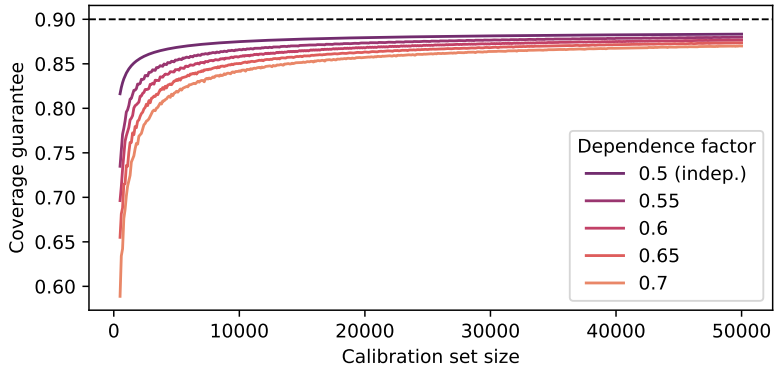

Figure 5: Marginal coverage for different calibration set sizes and dependence levels; the guarantees under dependence converge to the iid case with larger calibration sizes.

with $p = q$, and consider predicting a single data point by using the 11 preceding elements in the series as features. Figure 4 shows how marginal coverage (7), calculated through 10 000 simulations, is affected by increasing levels of dependence $1 - p$ for three different models (boosting, neural network and random forest) and $n_{\text{train}} = 1000, n_{\text{cal}} = 500$ and $n_{\text{test}} = 1$. We note in passing that the AR(1) experiment (cf. Figure 1) used the same amount of data for training, calibration and testing. Marginal coverage observed is close to nominal iid value of 90% for the independent case ($p = 0.5$) and weak to medium dependence, measured by the probabilities $1 - p$ of repeating the previous state. Coverage remains above 89% even for large values of dependence, and falls below 88% only after $1 - p = 0.999$. Also, Figure 5 shows how the correction $\varepsilon_{\text{cal}} + \delta_{\text{cal}} + \varepsilon_{\text{train}}$ in Theorem 1 depends on the calibration set sizes, quickly converging to the iid limit even for moderately dependent data.

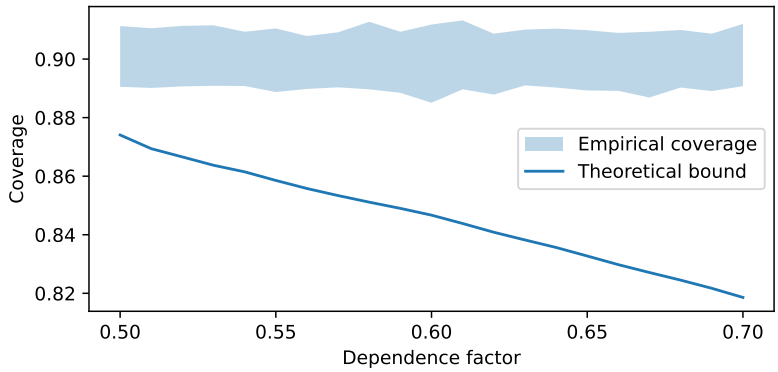

Figure 6: Empirical coverage bound; while the worst-case bound decreases with dependence, empirical coverage remains close to the iid level.

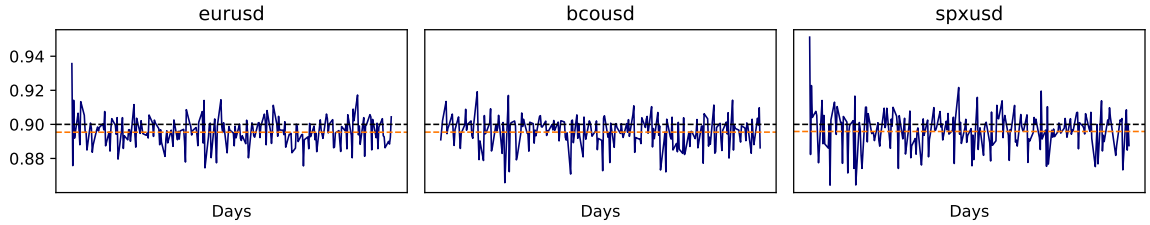

Figure 7: Daily marginal coverages of minute-by-minute online prediction for financial time series `eurusd`, `bcousd` and `spxusd` (solid blue), prescribed iid levels of $1-\alpha = 0.9$ (dashed black) and observed marginal coverages over the entire year (dashed orange), above 0.895 in all cases.

Further, Figure 6 shows the empirical coverage for a gradient boosting model over a thousand simulations with $n_{\text{train}} = 1000$, $n_{\text{cal}} = n_{\text{test}} = 15\,000$ and $\delta_{\text{cal}} = \delta_{\text{test}} = 0.005$. Note the empirical coverage revolves around the prescribed iid level 0.9, and it remains above the worst-case theoretical bound, which decreases with the dependence level $1-p$.

Example 4 (Financial time series) We study split CP's performance on three real-world time series: the euro spot exchange rate (`eurusd`), Brent crude oil future (`bcousd`) and S&P 500 stock index future (`spxusd`). Minute-by-minute data is retrieved directly from HistData (2022) via an open-source API (Rémy, 2022). We compute linear returns by dividing a price at minute t by the price at minute $t-1$ and subtracting 1. Due to market closures, Fridays and Sundays were discarded. We use gradient boosting to predict the price at time t using the prices at times $t-11, \ldots, t-1$. Then, we apply online conformal prediction over a sliding window of 1000 training points, 500 calibration points and 1 single test point for the entire year of 2021. Figure 7 shows the daily coverage of split CP. The dashed black line represents the iid nominal coverage of 90% and the dashed orange one the marginal coverage over the entire year. Marginal coverage is slightly below 90%, but never drastically so.

Data set	Cal. set size	Conditional coverage			
		Uptrend	Downtrend	High vol.	Low vol.
`eurusd`	500	88.76%	88.82%	87.64%	90.07%
	1000	89.19%	89.17%	88.38%	90.19%
	5000	90.03%	89.98%	89.85%	90.08%
`bcousd`	500	88.94%	88.72%	87.10%	89.43%
	1000	89.35%	89.04%	87.65%	89.95%
	5000	89.78%	89.77%	89.33%	89.98%
`spxusd`	500	89.12%	89.01%	88.87%	89.68%
	1000	89.53%	89.48%	88.84%	90.03%
	5000	90.04%	89.73%	89.53%	90.30%

Table 2: Conditional coverage for distinct trend and volatility events and varying calibration set size (before conditioning). Note that conditional coverage is generally close to nominal iid level $1-\alpha = 0.9$ and results improve given more calibration points.

Table 2 presents the conditional coverage (11) on four events of interest for all three financial data sets. Uptrend (respectively, downtrend) stands for two consecutive observations of positive (negative) returns. High (low) volatility events are taken to be those in which the standard deviation of the previous 10 returns observed is above (below) a given threshold. Note that conditioning on all such events still yields coverage close to the nominal iid level, on all three data sets. As previously noted, larger calibration sets have an important effect in improving coverage.

All experiments in this section were also conducted for nominally prescribed iid levels of 85% and 95% and the conclusions remained the same, as expected.

5. Conclusion

This paper shows that many of the appealing properties of split conformal prediction hold well even when the data is not exchangeable. Theorems 1 and 2 give marginal and empirical coverage guarantees for broad classes of non-exchangeable data, and Theorem 3 gives marginal coverage guarantees. We also consider the concrete case of stationary β-mixing sequences in Theorems 4 and 5, and show that our error bounds for such non-exchangeable data may match the order of the iid bounds. To prove these results, we introduce a novel mathematical framework that frees split CP from its exchangeability hypotheses, replacing it with concentration inequalities and decoupling properties.

We further show that these guarantees translate to robust empirical performance by split CP beyond independent data, with examples included in Section 4 ranging from synthetic to real data. In the same section, in spite of its generality, split CP is shown to fare well when compared to recent conformal algorithms developed for specific kinds of non-exchangeability. Except for extreme levels of dependence or very small calibration sets—small "effective sample size" regimes—, performance holds exceedingly well for standard split CP without any modifications.

Finally, we note that the general framework introduced in Section 3.1 can be extended in many directions. Section 3.4 looks at how results generalize for non-stationary data, as well as for other conformal prediction methods that are not based on split CP. One particular open direction is the application of concentration inequalities for weighted sums of exchangeable random variables (e.g., Barber 2024) to analyze weighted conformal methods such as NexCP (Barber et al., 2023). More generally, we expect that several results and methods (e.g., Chernozhukov et al. 2021, 2018; Zaffran et al. 2022; Feldman et al. 2022) can be analyzed and extended via our unified framework, which remains an avenue for future work.

Acknowledgments

RIO is supported by CNPq grants 432310/2018-5 (Universal) and 304475/2019-0 (Produtividade em Pesquisa), and FAPERJ grants 202.668/2019 (Cientista do Nosso Estado) and 290.024/2021 (Edital Inteligência Artificial). PO is supported by FAPERJ grant SEI-260003/001545/2022.

Appendix A. Proofs of Section 3.1

For the proofs below, we need to introduce certain population quantiles for $\widehat{s}_{\text{train}}(X_*, Y_*)$ conditionally on the training data.

Definition 7 (Conditional ϕ-quantile of the conformity score) *Given $\phi \in [0,1)$, let $q_{\phi,\text{train}}$ denote the ϕ-quantile of $\widehat{s}_{\text{train}}(X_*, Y_*)$ conditioned on the training data; that is:*

$$q_{\phi,\text{train}} \coloneqq \inf\{t \in \mathbb{R} : \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq t \mid \{(X_i, Y_i)\}_{i \in I_{\text{train}}}] \geq \phi\}.$$

Alternatively, define, for a deterministic $(x_i, y_i)_{i=1}^{n_{\text{train}}} \in (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}}$, the ϕ-quantile:

$$q_\phi((x_i, y_i)_{i=1}^{n_{\text{train}}}) \coloneqq \inf\{t \in \mathbb{R} : \mathbb{P}[s((x_i, y_i)_{i=1}^{n_{\text{train}}}, (X_*, Y_*)) \leq t] \geq \phi\},$$

and set $q_{\phi,\text{train}} \coloneqq q_\phi((X_i, Y_i)_{i \in I_{\text{train}}})$. We also define:

$$p_{\phi,\text{train}} \coloneqq \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq t \mid (X_i, Y_i)_{i \in I_{\text{train}}}]$$

Remark 8 *When the conditional law of $\widehat{s}_{\text{train}}(X_*, Y_*)$ given the training data is continuous, we have $p_{\phi,\text{train}} = \phi$. Otherwise, it only holds that $p_{\phi,\text{train}} \geq \phi$.*

Proof [Theorem 1] First we show that the event $F = \{q_{1-\alpha-\varepsilon_{\text{cal}},\text{train}} \leq \widehat{q}_{1-\alpha,\text{cal}}\}$ satisfies $\mathbb{P}[F] \geq 1 - \delta_{\text{cal}}$. Indeed, by (3) and Definition 7, if condition (5) holds, for any $\ell \in \mathbb{N}_{>0}$, with probability at least $1 - \delta_{\text{cal}}$,

$$\begin{aligned}\frac{1}{n_{\text{cal}}} \sum_{i \in I_{\text{cal}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{1-\alpha-\varepsilon_{\text{cal}},\text{train}} - 1/\ell\} &\leq \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq q_{1-\alpha-\varepsilon_{\text{cal}},\text{train}} - 1/\ell] + \varepsilon_{\text{cal}} \\ &< 1 - \alpha - \varepsilon_{\text{cal}} + \varepsilon_{\text{cal}} = 1 - \alpha \\ &\leq \frac{1}{n_{\text{cal}}} \sum_{i \in I_{\text{cal}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq \widehat{q}_{1-\alpha,\text{cal}}\}.\end{aligned}$$

This implies that the event

$$E_\ell = \left\{ \frac{1}{n_{\text{cal}}} \sum_{i \in I_{\text{cal}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{1-\alpha-\varepsilon_{\text{cal}},\text{train}} - 1/\ell\} < \frac{1}{n_{\text{cal}}} \sum_{i \in I_{\text{cal}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq \widehat{q}_{1-\alpha,\text{cal}}\} \right\}$$

satisfies $\mathbb{P}[E_\ell] \geq 1 - \delta_{\text{cal}}$ for all $\ell \in \mathbb{N}_{>0}$, and since $E_{\ell+1} \subset E_\ell$, we have $1 - \delta_{\text{cal}} \leq \lim_{\ell \to \infty} \mathbb{P}[E_\ell] = \mathbb{P}[E_\infty]$, where,

$$E_\infty = \left\{ \frac{1}{n_{\text{cal}}} \sum_{i \in I_{\text{cal}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{1-\alpha-\varepsilon_{\text{cal}},\text{train}}\} \leq \frac{1}{n_{\text{cal}}} \sum_{i \in I_{\text{cal}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq \widehat{q}_{1-\alpha,\text{cal}}\} \right\},$$

proving that $\mathbb{P}[F] \geq 1 - \delta_{\text{cal}}$. Therefore, given $i \in I_{\text{test}}$, using the fact that the function $t \mapsto \mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \leq t]$ is increasing,

$$\begin{aligned}\mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \leq \widehat{q}_{1-\alpha,\text{cal}}] &\geq \mathbb{P}[\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq \widehat{q}_{1-\alpha,\text{cal}}\} \cap F] \\ &\geq \mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{1-\alpha-\varepsilon_{\text{cal}},\text{train}}] - \delta_{\text{cal}}.\end{aligned}$$

Hence, by condition (6) and a conditioning argument,

$$\begin{aligned}\mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \leq \widehat{q}_{1-\alpha,\text{cal}}] &\geq \mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{1-\alpha-\varepsilon_{\text{cal}},\text{train}}] - \delta_{\text{cal}} \\ &\geq \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq q_{1-\alpha-\varepsilon_{\text{cal}},\text{train}}] - \varepsilon_{\text{test}} - \delta_{\text{cal}} \\ &\geq 1 - \alpha - \varepsilon_{\text{cal}} - \varepsilon_{\text{test}} - \delta_{\text{cal}},\end{aligned} \tag{17}$$

proving the first part of the theorem.

For the second part, note that by Definition 7 and condition (5) we have with probability at least $1 - \delta_{\text{cal}}$,

$$\frac{1}{n_{\text{cal}}} \sum_{i \in I_{\text{cal}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{1-\alpha+\varepsilon_{\text{cal}},\text{train}}\} \geq \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq q_{1-\alpha+\varepsilon_{\text{cal}},\text{train}}] - \varepsilon_{\text{cal}} \geq 1 - \alpha,$$

and since $\widehat{q}_{1-\alpha-\varepsilon_{\text{cal}},\text{cal}}$ is the smallest possible value satisfying the expression above, the event $G = \{\widehat{q}_{1-\alpha,\text{cal}} \leq q_{1-\alpha+\varepsilon_{\text{cal}},\text{train}}\}$ satisfies $\mathbb{P}[G] \geq 1 - \delta_{\text{cal}}$. Then,

$$\begin{aligned}\mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \leq \widehat{q}_{1-\alpha,\text{cal}}] &\leq \mathbb{P}[\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq \widehat{q}_{1-\alpha,\text{cal}}\} \cap G] + \delta_{\text{cal}} \\ &\leq \mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{1-\alpha+\varepsilon_{\text{cal}},\text{train}}] + \delta_{\text{cal}}.\end{aligned}$$

Hence, if $\widehat{s}_{\text{train}}(X_*, Y_*)$ almost surely has a continuous distribution conditionally on the training data, by condition (6)

$$\begin{aligned}\mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \leq \widehat{q}_{1-\alpha,\text{cal}}] &\leq \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq q_{1-\alpha+\varepsilon_{\text{cal}},\text{train}}] + \varepsilon_{\text{test}} + \delta_{\text{cal}} \\ &= 1 - \alpha + \varepsilon_{\text{cal}} + \varepsilon_{\text{test}} + \delta_{\text{cal}}.\end{aligned}$$

Putting this together with (17) concludes the second part. ■

Proof [Theorem 2] Assuming that condition (5) and condition (8) hold and using a similar argument as we did in the proof of Theorem 1, it is straightforward to show that the event $F = \{\widehat{q}_{1-\alpha-\varepsilon_{\text{cal}}-\varepsilon_{\text{test}},\text{test}} \leq \widehat{q}_{1-\alpha,\text{cal}}\}$ satisfies $\mathbb{P}[F] \geq 1 - \delta_{\text{cal}} - \delta_{\text{test}}$. But then,

$$\begin{aligned}&\mathbb{P}\left[\frac{1}{n_{\text{test}}} \sum_{i \in I_{\text{test}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq \widehat{q}_{1-\alpha,\text{cal}}\} \geq 1 - \alpha - \varepsilon_{\text{cal}} - \varepsilon_{\text{test}}\right] \\ &\geq \mathbb{P}\left[\frac{1}{n_{\text{test}}} \sum_{i \in I_{\text{test}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq \widehat{q}_{1-\alpha-\varepsilon_{\text{cal}}-\varepsilon_{\text{test}},\text{test}}\} \geq 1 - \alpha - \varepsilon_{\text{cal}} - \varepsilon_{\text{test}}\right] - \mathbb{P}[F^c] \\ &\geq 1 - \delta_{\text{cal}} - \delta_{\text{test}},\end{aligned}$$

proving the first part. For the second part, note that $G = \{\widehat{q}_{1-\alpha+\varepsilon_{\text{cal}}+\varepsilon_{\text{test}},\text{test}} \geq \widehat{q}_{1-\alpha,\text{cal}}\}$ also has probability at least $1 - \delta_{\text{cal}} - \delta_{\text{test}}$, therefore the event $F \cap G = \{\widehat{q}_{1-\alpha+\varepsilon_{\text{cal}}+\varepsilon_{\text{test}},\text{test}} \geq \widehat{q}_{1-\alpha-\varepsilon_{\text{cal}}-\varepsilon_{\text{test}},\text{test}}\}$ has probability at least $1 - 2\delta_{\text{cal}} - 2\delta_{\text{test}}$ Hence, if $\widehat{s}_{\text{train}}(X_*, Y_*)$ almost surely has a continuous distribution conditionally on the training data, using the same argument as we did in the proof of Theorem 1 concludes the theorem. ■

Proof [Application to the iid case] First, note that, in the iid case, when $i \in I_{\text{test}}$,

$$\mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}] = \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq q_{\text{train}}],$$

showing that condition (6) holds with $\varepsilon_{\text{test}} = 0$.

Moreover, using the fact that $(\mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}\})_{i=1}^n$ is an iid sample of bounded random variables, by Hoeffding's inequality, with probability at least $1 - \delta$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}\} - P_{q,\text{train}} \right| \leq \sqrt{\frac{1}{2n} \log\left(\frac{2}{\delta}\right)}.$$

Therefore, conditions (5) and (8) are proved by taking

$$\varepsilon_{\text{cal}} = \sqrt{\frac{1}{2n_{\text{cal}}} \log\left(\frac{2}{\delta_{\text{cal}}}\right)} \quad \text{and} \quad \varepsilon_{\text{test}} = \sqrt{\frac{1}{2n_{\text{test}}} \log\left(\frac{2}{\delta_{\text{test}}}\right)}.$$ ■

Appendix B. Proofs of Section 3.2 and Further Results

Let $\mathcal{A}$ be a family of measurable subsets of $\mathcal{X}$. Given $\phi \in [0, 1)$ and $A \in \mathcal{A}$, let $I_{\text{cal}}(A) \coloneqq \{i \in I_{\text{cal}} : X_i \in A\}$ and $n_{\text{cal}}(A) \coloneqq \#I_{\text{cal}}(A)$. Denote the empirical ϕ-quantile of $\widehat{s}_{\text{train}}(X_i, Y_i)$ over $i \in I_{\text{cal}}$ as:

$$\widehat{q}_{\phi,\text{cal}}(A) \coloneqq \inf\left\{ t \in \mathbb{R} \,:\, \frac{1}{n_{\text{cal}}(A)} \sum_{i \in I_{\text{cal}}(A)} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq t\} \geq \phi \right\},$$

and, for $x \in A$, define the prediction set:

$$C_\phi(x; A) \coloneqq \{y \in \mathcal{Y} \,:\, \widehat{s}_{\text{train}}(x, y) \leq \widehat{q}_{\phi,\text{cal}}(A)\}.$$

For $A \in \mathcal{A}$ with $\mathbb{P}[X \in A] > 0$, let

$$P_{q,\text{train}}(A) \coloneqq \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq q_{\text{train}} \mid (X_i, Y_i)_{i \in I_{\text{train}}}, X_* \in A]. \tag{18}$$

Proof [Theorem 3] Following the same strategy as in Theorem 1, we have that, with probability at least $1 - \delta_{\text{cal}}$, the event $F_{\text{cal}} = \{q_{1-\alpha-\varepsilon_{\text{cal}}}(A) \leq \widehat{q}_{1-\alpha,\text{cal}}(A),\ \forall A \in \mathcal{A}\}$ satisfies $\mathbb{P}[F_{\text{cal}}] \geq 1 - \delta_{\text{cal}}$. Now, using the fact that the function $t \mapsto \mathbb{P}[\widehat{s}_{\text{train}}(X_k, Y_k) \leq t \mid X_k \in A]$ is increasing, for any $k \in I_{\text{test}}$ and all $A \in \mathcal{A}$,

$$\begin{aligned}
\mathbb{P}[\widehat{s}_{\text{train}}(X_k, Y_k) \leq \widehat{q}_{1-\alpha,\text{cal}}(A) \mid X_k \in A] &\geq \mathbb{P}[\{\widehat{s}_{\text{train}}(X_k, Y_k) \leq \widehat{q}_{1-\alpha,\text{cal}}(A)\} \cap F_{\text{cal}} \mid X_k \in A] \\
&\geq \mathbb{P}[\{\widehat{s}_{\text{train}}(X_k, Y_k) \leq q_{1-\alpha-\varepsilon_{\text{cal}}}(A)\} \cap F_{\text{cal}} \mid X_k \in A] \\
&\geq \mathbb{P}[\widehat{s}_{\text{train}}(X_k, Y_k) \leq q_{1-\alpha-\varepsilon_{\text{cal}}}(A) \mid X_k \in A] - \delta_{\text{cal}}.
\end{aligned}$$

Then, to conclude,

$$\begin{aligned}
&\mathbb{P}[\widehat{s}_{\text{train}}(X_k, Y_k) \leq \widehat{q}_{1-\alpha,\text{cal}}(A) \mid X_k \in A] \\
&\geq \mathbb{P}[\widehat{s}_{\text{train}}(X_k, Y_k) \leq q_{1-\alpha-\varepsilon_{\text{cal}}}(A) \mid X_k \in A] - \delta_{\text{cal}} \\
&\geq \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq q_{1-\alpha-\varepsilon_{\text{cal}}}(A) \mid X_* \in A] - \varepsilon_{\text{train}} - \delta_{\text{cal}} \\
&= \mathbb{E}[\mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq q_{1-\alpha-\varepsilon_{\text{cal}}}(A) \mid (X_i, Y_i)_{i \in I_{\text{train}}}, X_* \in A]] - \varepsilon_{\text{train}} - \delta_{\text{cal}} \\
&\geq 1 - \alpha - \varepsilon_{\text{cal}} - \varepsilon_{\text{train}} - \delta_{\text{cal}}.
\end{aligned}$$ ■

We now proceed to give an empirical conditional coverage guarantee. For a conditional version of (8), suppose there exist δ_{test} and $\varepsilon_{\text{test}} \in (0,1)$ such that

$$\mathbb{P}\left[\sup_{A\in\mathcal{A}}\left|P_{q,\text{train}}(A) - \frac{1}{n_{\text{test}}(A)}\sum_{i\in I_{\text{test}}(A)}\mathbb{1}\{\widehat{s}_{\text{train}}(X_i,Y_i)\leq q_{\text{train}}\}\right| \leq \varepsilon_{\text{test}}\right] \geq 1-\delta_{\text{test}}, \quad (19)$$

where $P_{q,\text{train}}(A)$ is defined as in (18). This suffices for empirical conditional coverage.

Theorem 9 (Empirical conditional coverage over test data) *Given $\alpha \in (0,1)$, $\delta_{\text{cal}} > 0$ and $\delta_{\text{test}} > 0$, if (9) and (19) hold, then for each $A \in \mathcal{A}$:*

$$\mathbb{P}\left[\inf_{A\in\mathcal{A}}\frac{1}{n_{\text{test}}(A)}\sum_{i\in I_{\text{test}}(A)}\mathbb{1}\{Y_i \in C_{1-\alpha}(X_i;A)\} \geq 1-\alpha-\eta\right] \geq 1-\delta_{\text{cal}}-\delta_{\text{test}},$$

where $\eta = \varepsilon_{\text{cal}}+\varepsilon_{\text{test}}$. Additionally, if $\widehat{s}_{\text{train}}(X_*,Y_*)$ almost surely has a continuous distribution conditionally on the training data, then:

$$\mathbb{P}\left[\sup_{A\in\mathcal{A}}\left|\frac{1}{n_{\text{test}}(A)}\sum_{i\in I_{\text{test}}(A)}\mathbb{1}\{Y_i \in C_{1-\alpha}(X_i;A)\} - (1-\alpha)\right| \leq \eta\right] \geq 1-2\delta_{\text{cal}}-2\delta_{\text{test}}.$$

Proof As in the proof of empirical coverage over test data (Theorem 2), the event $F = \{\widehat{q}_{1-\alpha-\varepsilon_{\text{cal}}-\varepsilon_{\text{test}},\text{test}}(A) \leq \widehat{q}_{1-\alpha,\text{cal}}(A) \forall A \in \mathcal{A}\}$ satisfies $\mathbb{P}[F] \geq 1-\delta_{\text{cal}}-\delta_{\text{test}}$ and the remaining of the proof is just a direct application of the definition of conditional empirical quantile calibrated over the test data. ■

The results above are directly applicable in the iid case. It is straightforward to show that (10) holds with $\varepsilon_{\text{test}} = 0$ and, if the family $\mathcal{A}$ has finite VC dimension $\text{VC}(\mathcal{A}) = d$ and $\mathbb{P}[A] > \gamma$ for some $\gamma > 0$ and all $A \in \mathcal{A}$, it suffices to take $\varepsilon_{\text{cal}} = \gamma^{-1}(4\sqrt{\log(2(n+1)^d)/n} + 2\sqrt{\log(4/\delta)/(2n)})$.

Appendix C. Proofs of Section 3.3 and Further Results

C.1 Standard Coverage Guarantees

Our goal is to check conditions (5), (6) and (8) when $(X_i,Y_i)_{i=1}^n$ is a stationary β-mixing process. As stated in the main text, the main tool we will use is the so-called Blocking Technique (Yu, 1994; Mohri and Rostamizadeh, 2010; Kuznetsov and Mohri, 2017). It allows one to measure the difference in expectation between a function of a β-mixing process and the same function over an independent process, thereby transforming the original dependent problem into an independent one with the addition of a penalty factor.

Proposition 10 (Blocking Technique) *Let $\{Z_t\}_{t=1}^T$ be a sample of a stationary β-mixing process. Split the sample into $2m$ interleaved blocks, with even blocks of size a and odd blocks of size b, such that $T = m(a+b)$. Denote each block by $B_j = \{Z_i\}_{i=l(j)}^{u(j)}$, where*

$l(j) = 1+\lceil (j-2)/2 \rceil a+\lfloor j/2 \rfloor b$ *and* $u(j) = \lfloor j/2 \rfloor a+\lceil j/2 \rceil b$*, so the set of odd blocks, each of size* b*, is given by* $B_{\text{odd}} = (B_1, B_3, \ldots, B_{2m-1})$*. Consider also the set* $B^*_{\text{odd}} = (B^*_1, B^*_3, \ldots, B^*_{2m-1})$ *where* B^*_j *are independent for* $j = 1, 3, \ldots, 2m-1$*, and* $B^*_j \overset{d}{=} B_j$*. If* $h : \mathbb{R}^{mb} \to \mathbb{R}$ *is a Borel-measurable function with* $|h| \leq M$ *for some* $M > 0$*, then*

$$|\mathbb{E}[h(B_{\text{odd}})] - \mathbb{E}[h(B^*_{\text{odd}})]| \leq 2M(m-1)\beta(a),$$

where $\beta(a)$ *is the* β*-mixing coefficient of* $\{Z_t\}_{t=1}^T$*.*

Using the Blocking Technique, we can prove that up to a error correction factor, we can transform our stationary β-mixing problem into an iid one:

Lemma 11 (Mohri and Rostamizadeh 2009) *Let* $Z_1, \ldots, Z_n$ *be a sample drawn from a stationary* β*-mixing distribution and* $\mathcal{F}$ *be a class of functions from* $\mathcal{X}$ *to* $[0,1]$*. Split the sample into* $2m$ *blocks, with blocks of size* a *with* $n = 2ma$*. Denote the blocks by* $B_j = \{Z_i\}_{i=l(j)}^{u(j)}$ *where* $l(j) = 1 + (j-1)a$ *and* $u(j) = ja$*, with* $B_{\text{odd}} = (B_1, B_3, \ldots, B_{2m-1})$*. Call the independent version of* B_{odd} *by* $B^*_{\text{odd}} = (B^*_1, B^*_3, \ldots, B^*_{2m-1})$*, where* B^*_j *are independent with* $B^*_j \overset{d}{=} B_j$*, and let* $\mathbb{P}_*$ *be their law. Then,*

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}} \left|\mathbb{E}[f(Z_1)] - \frac{1}{n}\sum_{i=1}^n f(Z_i)\right| > \varepsilon\right] \leq 2\mathbb{P}_*\left[\sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[f(Z_1)\right] - \frac{1}{ma}\sum_{Z_j \in B^*_{\text{odd}}} f(Z_j)\right| > \varepsilon\right] + 4(m-1)\beta(a).$$

Finally, using Lemma 11 and Bernstein's Inequality, we are ready to prove a concentration inequality for stationary β-mixing processes.

Lemma 12 *Let* $Z_1, \ldots, Z_n$ *be a sample drawn from a stationary* β*-mixing distribution with* $Z_1 \in [0,1]$ *and* $\text{Var}[Z_1] = v < \infty$*. Then, for any* $m, a, s \in \mathbb{N}_+$ *with* $m > 1$*,* $n = 2ma + s$ *and* $\delta > 4(m-1)\beta(a))$*, with probability at least* $1-\delta$ *it holds that*

$$\left|\mathbb{E}\left[Z_1\right] - \frac{1}{n}\sum_{i=1}^n Z_i\right| \leq \varepsilon,$$

where

$$\varepsilon = \widetilde{\sigma}(a)\sqrt{\frac{4}{n}\log\left(\frac{4}{\delta - 4(m-1)\beta(a)}\right)} + \frac{1}{3m}\log\left(\frac{4}{\delta - 4(m-1)\beta(a)}\right) + \frac{s}{n},$$

and $\widetilde{\sigma}(a) = \sqrt{v + \frac{2}{a}\sum_{k=1}^{a-1}(a-k)\beta(k)}$*.*

Proof By an application of Lemma 11 taking $\mathcal{F} = \{x \mapsto x, x \in [0,1]\}$ and Bernstein's inequality over the m independent blocks, with probability at least $1-\delta$

$$\begin{aligned}\left|\mathbb{E}\left[Z_1\right] - \frac{1}{n}\sum_{i=1}^n Z_i\right| &\leq \frac{2ma}{n}\left|\mathbb{E}\left[Z_1\right] - \frac{1}{2ma}\sum_{i=1}^{2ma} Z_i\right| + \frac{s}{n} \qquad (20)\\ &\leq \sigma\sqrt{\frac{2}{m}\log\left(\frac{4}{\delta - 4(m-1)\beta(a)}\right)} + \frac{1}{3m}\log\left(\frac{4}{\delta - 4(m-1)\beta(a)}\right) + \frac{s}{n},\end{aligned}$$

where $\sigma^2 = \text{Var}\left[\frac{1}{a}\sum_{i:Z_i \in B_j} Z_i\right]$. To estimate σ^2, note that by stationarity,

$$\text{Var}\left[\frac{1}{a}\sum_{i=1}^{a} Z_i\right] = \frac{1}{a}\mathbb{E}\left[Z_1^2\right] - \mathbb{E}[Z_1]^2 + \frac{1}{a^2}\sum_{k=1}^{a-1}\sum_{|i-j|=k}^{a} \mathbb{E}\left[Z_i Z_j\right].$$

Now, using the fact that $\{Z_i\}_{i \in B_j}$ is β-mixing, we have

$$\begin{aligned}
\text{Var}\left[\frac{1}{a}\sum_{i=1}^{a} Z_i\right] &\leq \frac{1}{a}\mathbb{E}\left[Z_1^2\right] - \mathbb{E}[Z_1]^2 + \frac{1}{a^2}\sum_{k=1}^{a-1}\sum_{|i-j|=k}^{a}\Big(\mathbb{E}\left[Z\right]^2 + \beta(k)\Big) \\
&= \frac{1}{a}\mathbb{E}\left[Z_1^2\right] - \mathbb{E}[Z_1]^2 + \frac{1}{a^2}\sum_{k=1}^{a-1}\sum_{|i-j|=k}^{a}\mathbb{E}\left[Z\right]^2 + \frac{1}{a^2}\sum_{k=1}^{a-1}\sum_{|i-j|=k}^{a}\beta(k) \\
&= \frac{1}{a}\text{Var}[Z_1] + \frac{1}{a^2}\sum_{k=1}^{a-1}\sum_{|i-j|=k}\beta(k) = \frac{1}{a}\left(\text{Var}[Z_1] + \frac{2}{a}\sum_{k=1}^{a-1}(a-k)\beta(k)\right).
\end{aligned}$$

That is,

$$\sigma \leq \sqrt{\frac{1}{a}\left(v + \frac{2}{a}\sum_{k=1}^{a-1}(a-k)\beta(k)\right)}.$$

Plugging the above expression in (20) and using the fact that $2ma = n$, yields the result. ■

Now we are ready to prove conditions (5), (6) and (8) for the stationary β-mixing case.

Proposition 13 *If $(X_i, Y_i)_{i=1}^n$ is stationary β-mixing, then condition (5) holds with*

$$\begin{aligned}
\varepsilon_{\text{cal}} = \inf_{(a,m,r)\in F_{\text{cal}}} \Bigg\{ &\widetilde{\sigma}(a)\sqrt{\frac{4}{n_{\text{cal}} - r + 1}\log\left(\frac{4}{\delta_{\text{cal}} - 4(m-1)\beta(a) - \beta(r)}\right)} \\
&+ \frac{1}{3m}\log\left(\frac{4}{\delta_{\text{cal}} - 4(m-1)\beta(a) - \beta(r)}\right) + \frac{r-1}{n_{\text{cal}}}\Bigg\}
\end{aligned}$$

for

$$F_{\text{cal}} = \{(a, m, r) \in \mathbb{N}_{>0}^3 \ : \ 2ma = n_{\text{cal}} - r + 1, \ \delta_{\text{cal}} > 4(m-1)\beta(a) + \beta(r)\},$$

where $\widetilde{\sigma}(a) = \sqrt{\frac{1}{4} + \frac{2}{a}\sum_{k=1}^{a-1}(a-k)\beta(k)}$.

Proof We want to use Lemma 12 for the random variables $\{\mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}\}\}_{i \in I_{\text{cal}}}$, however, since the random variables $(X_i, Y_i)_{i \in I_{\text{cal}}}$ are dependent to $\widehat{s}_{\text{train}}$ and the quantile q_{train}, we cannot simply apply the result. To fix this problem, it will be necessary to create a gap between our training and calibration data and use the Blocking Technique, Proposition 10, to transpose our problem to an independent setting.

For $\varepsilon > 0$ and $r \in \{1, \ldots, n_{\text{cal}}\}$, let $I_{\text{cal},r} = \{n_{\text{train}} + r, \ldots, n_{\text{train}} + n_{\text{cal}}\}$ and define the event

$$E(r, \varepsilon) = \left\{\left|\frac{1}{n_{\text{cal}} - r + 1}\sum_{i \in I_{\text{cal},r}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}\} - P_{q,\text{train}}\right| > \varepsilon\right\},$$

we want to show that there exists $\varepsilon > 0$ such that $\mathbb{P}[E(1,\varepsilon)] \leq \delta$. Note that if $E(1,\varepsilon)$ holds, then

$$\left| \frac{1}{n_{\text{cal}}} \sum_{i \in I_{\text{cal},r}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}\} - P_{q,\text{train}} \right| > \varepsilon - \frac{r-1}{n_{\text{cal}}},$$

and since $n_{\text{cal}} \geq n_{\text{cal}} - r + 1$,

$$\left| \frac{1}{n_{\text{cal}} - r + 1} \sum_{i \in I_{\text{cal},r}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}\} - P_{q,\text{train}} \right| > \varepsilon - \frac{r-1}{n_{\text{cal}}},$$

that is, $E(1,\varepsilon) \subset E(r, \varepsilon - (r-1)/n_{\text{cal}})$. Now, define $\mathbb{P}_* = \mathbb{P}_1^{n_{\text{train}}} \otimes \mathbb{P}_{n_{\text{train}}+r}^{n_{\text{train}}+n_{\text{cal}}}$, so under $\mathbb{P}_*$, we have that $(X_i, Y_i)_{i \in I_{\text{train}}}$ and $(X_i, Y_i)_{i \in I_{\text{cal},r}}$ are independent. Then, by Proposition 10,

$$\begin{aligned}
\mathbb{P}[E(1,\varepsilon)] &\leq \mathbb{P}[E(r, \varepsilon - (r-1)/n_{\text{cal}})] \\
&\leq \mathbb{P}_*[E(r, \varepsilon - (r-1)/n_{\text{cal}})] + \beta(r) \\
&= \mathbb{E}_*[\mathbb{P}_*[E(r, \varepsilon - (r-1)/n_{\text{cal}}) \mid (X_i, Y_i)_{i \in I_{\text{train}}}]] + \beta(r).
\end{aligned}$$

Note that by Lemma 12, for any $m, a \in \mathbb{N}_+$ with $n_{\text{cal}} - (r+s) + 1 = 2ma$ and $\delta_{\text{cal}} > 4(m-1)\beta(a))$, using the fact that $\text{Var}\left[\mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}\}\right] \leq 1/4$, taking

$$\varepsilon = \widetilde{\sigma}(a) \sqrt{\frac{4}{n_{\text{cal}} - r + 1} \log\left(\frac{4}{\delta - 4(m-1)\beta(a)}\right)} + \frac{1}{3m} \log\left(\frac{4}{\delta - 4(m-1)\beta(a)}\right) + \frac{r-1}{n_{\text{cal}}},$$

implies that $\mathbb{P}_*[E(r, \varepsilon - (r-1)/n_{\text{cal}}) \mid (X_i, Y_i)_{i \in I_{\text{train}}}] \leq \delta_{\text{cal}}$. Therefore, $\mathbb{P}[E(1,\varepsilon)] \leq \delta_{\text{cal}} + \beta(r)$, which is equivalent to $\mathbb{P}[E(1,\varepsilon')] \leq \delta_{\text{cal}}$ taking

$$\begin{aligned}
\varepsilon' = \widetilde{\sigma}(a) &\sqrt{\frac{4}{n_{\text{cal}} - r + 1} \log\left(\frac{4}{\delta - 4(m-1)\beta(a) - \beta(r)}\right)} \\
&+ \frac{1}{3m} \log\left(\frac{4}{\delta - 4(m-1)\beta(a) - \beta(r)}\right) + \frac{s + r - 1}{n_{\text{cal}}}.
\end{aligned}$$

Finally, since this is true for any choice of $a, m, r \in \mathbb{N}_{>0}$ and $s \in \mathbb{N}$ with $s + 2ma = n_{\text{cal}} - r + 1$ and $\delta > 4(m-1)\beta(a) + \beta(r)$, we can choose a, m, r optimally such that the value of ε' is minimized and there is no need to optimize in s in this case. ■

Proposition 14 *If $(X_i, Y_i)_{i=1}^n$ is stationary β-mixing, then condition (6) holds with $\varepsilon_{\text{train}} = \beta(k - n_{\text{train}})$. Moreover, since $\beta(k - n_{\text{train}}) \leq \beta(1 - n_{\text{train}})$, it is possible to find $\varepsilon_{\text{train}}$ not depending on k.*

Proof Given $k \in I_{\text{test}}$, define $\mathbb{P}_* = \mathbb{P}_1^{n_{\text{train}}} \otimes \mathbb{P}_k^k$, so under $\mathbb{P}_*$ the random variable (X_k, Y_K) is independent of the training data $(X_i, Y_i)_{i \in I_{\text{train}}}$. Then, if $\beta_k = \beta(k - n_{\text{train}})$ we have,

$$\begin{aligned}
\beta_k &\geq \left|\mathbb{P}[\widehat{s}_{\text{train}}(X_k, Y_k) \leq q_{\text{train}}] - \mathbb{P}_*[\widehat{s}_{\text{train}}(X_k, Y_k) \leq q_{\text{train}}]\right| \\
&= \left|\mathbb{P}[\widehat{s}_{\text{train}}(X_k, Y_k) \leq q_{\text{train}}] - \mathbb{E}_*[\mathbb{P}_*[\widehat{s}_{\text{train}}(X_k, Y_k) \leq q_{\text{train}} \mid (X_i, Y_i)_{i \in I_{\text{train}}}]]\right| \\
&= \left|\mathbb{P}[\widehat{s}_{\text{train}}(X_k, Y_k) \leq q_{\text{train}}] - \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq q_{\text{train}} \mid (X_i, Y_i)_{i \in I_{\text{train}}}]\right|,
\end{aligned}$$

where the β_k penalty follows from Proposition 10. Note that the larger the k the smaller the penalty incurred by the dependence in the β-mixing process. Moreover, since $\beta_k \leq \beta_1$, it is possible to define $\varepsilon_{\text{train}} = \beta_1$ not depending on $k \in I_{\text{test}}$. ■

Proposition 15 *If $(X_i, Y_i)_{i=1}^n$ is stationary β-mixing, then condition (8) holds with*

$$\varepsilon_{\text{test}} = \inf_{(a,m)\in F_{\text{test}}} \left\{ \widetilde{\sigma}(a)\sqrt{\frac{4}{n_{\text{test}}}\log\left(\frac{4}{\delta_{\text{test}} - 4(m-1)\beta(a) - \beta(n_{\text{cal}})}\right)} + \frac{1}{3m}\log\left(\frac{4}{\delta_{\text{test}} - 4(m-1)\beta(a) - \beta(n_{\text{cal}})}\right) + \frac{s}{n_{\text{test}}} \right\}$$

for

$$F_{\text{test}} = \{(a, m, s) \in \mathbb{N}_{>0}^2 \times \mathbb{N} \ : \ s + 2ma = n_{\text{test}},\ \delta > 4(m-1)\beta(a) + \beta(n_{\text{cal}})\},$$

and $\widetilde{\sigma}(a) = \sqrt{\frac{1}{4} + \frac{2}{a}\sum_{k=1}^{a-1}(a-k)\beta(k)}$.

Proof The proof is similar to the proof of Proposition 13. Let the event $E(\varepsilon)$ be

$$E(\varepsilon) = \left\{ \left| P_{q,\text{train}} - \frac{1}{n_{\text{test}}} \sum_{i\in I_{\text{test}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}\} \right| > \varepsilon \right\}.$$

Define, $\mathbb{P}_* = \mathbb{P}_1^{n_{\text{train}}} \otimes \mathbb{P}_{n_{\text{train}}+n_{\text{cal}}}^{n_{\text{train}}+n_{\text{cal}}+n_{\text{test}}}$, so under $\mathbb{P}_*$ we have that $(X_i, Y_i)_{i\in I_{\text{train}}}$ and $(X_i, Y_i)_{i\in I_{\text{test}}}$ are independent. By Proposition 10 we have

$$\mathbb{P}[E(\varepsilon)] \leq \mathbb{P}_*[E(\varepsilon)] + \beta(n_{\text{cal}}) = \mathbb{E}_*[\mathbb{P}_*[E(\varepsilon) \mid (X_i, Y_i)_{i\in I_{\text{train}}}]] + \beta(n_{\text{cal}}).$$

Now we can apply Lemma 12 and conclude that, just as we did in Proposition 13, that for any $m, a \in \mathbb{N}_+$, $s \in \mathbb{N}$ with $n_{\text{test}} = 2ma - s$ and $\delta_{\text{test}} > 4(m-1)\beta(a)) + \beta(n_{\text{cal}})$, it is true that $\mathbb{P}[E(\varepsilon)] \leq \delta$, where

$$\varepsilon = \widetilde{\sigma}(a)\sqrt{\frac{4}{n_{\text{test}}}\log\left(\frac{4}{\delta_{\text{test}} - 4(m-1)\beta(a) - \beta(n_{\text{cal}})}\right)} + \frac{1}{3m}\log\left(\frac{4}{\delta_{\text{test}} - 4(m-1)\beta(a) - \beta(n_{\text{cal}})}\right) + \frac{s}{n_{\text{test}}},$$

and $\widetilde{\sigma}(a) = \sqrt{\frac{1}{4} + \frac{2}{a}\sum_{k=1}^{a-1}(a-k)\beta(k)}$. Finally, since this is true for any choice of $a, m \in \mathbb{N}_{>0}$ and $s \in \mathbb{N}$, with $s + 2ma = n_{\text{test}}$ and $\delta_{\text{test}} > 4(m-1)\beta(a) + \beta(n_{\text{cal}})$, we can choose a, m, s optimally to minimize ε. ■

Proof [Theorem 4] The result follows from applying Propositions 13 and 14 in Theorem 1. ■

Proof [Theorem 5] The result follows from applying Propositions 13 and 15 in Theorem 2. ■

C.2 Conditional Guarantees

To obtain conditional guarantees for stationary β-mixing processes, we need to specify a family $\mathcal{A}$ of measurable sets in $\mathcal{X}$ satisfying certain conditions. In particular, we will assume that, for a fixed value $\gamma > 0$, the family $\mathcal{A}$ has finite VC dimension $\mathrm{VC}(\mathcal{A}) = d$ and $\mathbb{P}[X_* \in A] > \gamma$ for all $A \in \mathcal{A}$.

Then, given $\delta_{\mathrm{cal}} > 0$ and $\alpha \in (0,1)$, define the calibration error correction factor for a stationary β-mixing process conditioned to the family $\mathcal{A}$ as

$$\varepsilon_{\mathrm{cal}} = \inf_{(a,m,r)\in G_{\mathrm{cal}}} \left\{ \frac{1}{\gamma} \left(\frac{\kappa(m,r)}{n_{\mathrm{cal}}} + \sqrt{\frac{2}{m} \log\left(\frac{16}{\delta_{\mathrm{cal}} - 16(m-1)\beta(a) - \beta(r)} \right)} \right) \right\}, \tag{21}$$

where $\kappa(m,r) = 4n_{\mathrm{cal}}\sqrt{\log(2(m+1)^d)/m} + 2(r-2)$ and

$$G_{\mathrm{cal}} = \{(a,m,r) \in \mathbb{N}^3_{>0} \ : \ 2ma = n_{\mathrm{cal}} - r + 1, \delta_{\mathrm{cal}} > 16(m-1)\beta(a) + \beta(r)\}.$$

Note the factor $1/\gamma$ in $\varepsilon_{\mathrm{cal}}$: for η to be small, we need $\varepsilon_{\mathrm{cal}}$ to be small and consequently m has to be large. This is quite natural, since if γ is too small, the probability $\mathbb{P}[X_* \in A]$ can be close to zero, and thus a larger sample is necessary to estimate the empirical quantile well. Therefore, Theorem 3 yields the following.

Theorem 16 (Conditional coverage: stationary β-mixing processes) *Suppose that $(X_i, Y_i)_{i=1}^n$ is stationary β-mixing. Then given $\alpha \in (0,1)$, $\gamma > 0$ and $\delta_{\mathrm{cal}} > 0$, for each $A \in \mathcal{A}$ and any $i \in I_{\mathrm{test}}$*

$$\mathbb{P}[Y_i \in C_{1-\alpha}(X_i; A) \mid X_i \in A] \geq 1 - \alpha - \eta,$$

with $\eta = \varepsilon_{\mathrm{cal}} + \varepsilon_{\mathrm{test}}$, where $\varepsilon_{\mathrm{cal}}$ is as in (21) and $\varepsilon_{\mathrm{test}} = \beta(i - n_{\mathrm{train}})$.

Additionally, if $\widehat{s}_{\mathrm{train}}(X_*, Y_*)$ almost surely has a continuous distribution conditionally on the training data, then:

$$|\mathbb{P}[Y_i \in C_{1-\alpha}(X_i; A) \mid X_i \in A] - (1-\alpha)| \leq \varepsilon_{\mathrm{cal}} + \delta_{\mathrm{cal}} + \varepsilon_{\mathrm{test}}.$$

Now, denote the test error correction factor for a stationary β-mixing process conditioned to the family $\mathcal{A}$ as

$$\varepsilon_{\mathrm{test}} = \inf_{(a,m,s)\in G_{\mathrm{test}}} \left\{ \frac{1}{\gamma} \left(\frac{\tilde{\kappa}(m,r)}{n_{\mathrm{test}}} + \sqrt{\frac{2}{m} \log\left(\frac{8}{\delta_{\mathrm{test}} - 8(m-1)\beta(a) - \beta(n_{\mathrm{cal}})} \right)} \right) \right\}, \tag{22}$$

where $\tilde{\kappa}(m,r) = 4n_{\mathrm{test}}\sqrt{\log(2(m+1)^d)/m} + 2s$ and

$$G_{\mathrm{test}} = \{(a,m,s) \in \mathbb{N}^2_{>0} \times \mathbb{N} \ : \ 2ma = n_{\mathrm{test}} - s, \delta_{\mathrm{test}} > 8(m-1)\beta(a) + \beta(n_{\mathrm{cal}})\}.$$

The following result then follows from Theorem 9.

Theorem 17 (Empirical conditional coverage: stationary β-mixing processes) *Suppose that $(X_i, Y_i)_{i=1}^n$ is stationary β-mixing, then given $\alpha \in (0,1)$, $\gamma > 0$, $\delta_{\mathrm{cal}} > 0$ and $\delta_{\mathrm{test}} > 0$, for each $A \in \mathcal{A}$:*

$$\mathbb{P}\left[\inf_{A\in\mathcal{A}} \frac{1}{n_{\mathrm{test}}(A)} \sum_{i \in I_{\mathrm{test}}(A)} \mathbb{1}\{Y_i \in C_{1-\alpha}(X_i; A)\} \geq 1 - \alpha - \eta \right] \geq 1 - \delta_{\mathrm{cal}} - \delta_{\mathrm{test}},$$

where $\eta = \varepsilon_{\text{cal}} + \varepsilon_{\text{test}}$, *for* ε_{cal} *as in (21) and* $\varepsilon_{\text{test}}$ *as in (22).*

Additionally, if $\widehat{s}_{\text{train}}(X_*, Y_*)$ *almost surely has a continuous distribution conditionally on the training data, then:*

$$\mathbb{P}\left[\sup_{A\in\mathcal{A}}\left|\frac{1}{n_{\text{test}}(A)}\sum_{i\in I_{\text{test}}(A)}\mathbb{1}\{Y_i \in C_{1-\alpha}(X_i; A)\} - (1-\alpha)\right| \leq \eta\right] \geq 1 - 2\delta_{\text{cal}} - 2\delta_{\text{test}}.$$

The proofs in this section are very similar to the proofs in Section C.1, however, since we are dealing with a family of Borel measurable sets $\mathcal{A}$, we will need concentration results that allow us to uniformly control certain quantities over the family $\mathcal{A}$. First, we state such classical results for iid sequences.

Theorem 18 *Let* $Z_1, \ldots, Z_n$ *be iid random variables taking values on* $\mathcal{X}$ *and* $\mathcal{F}$ *be a class of functions from* $\mathcal{X}$ *to* $\{0, 1\}$. *Then*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}f(Z) - \frac{1}{n}\sum_{i=1}^{n} f(Z_i)\right|\right] \leq 2\sqrt{\frac{\log(2\mathcal{S}_{\mathcal{F}}(n))}{n}}.$$

Using the Blocking Technique, we can prove that up to a error correction factor, we can transform our stationary β-mixing problem into a iid one. For example,

Corollary 19 *Let* $Z_1, \ldots, Z_n$ *be a sample drawn from a stationary* β*-mixing distribution and* $\mathcal{F}$ *be a class of functions from* $\mathcal{X}$ *to* $\{0, 1\}$. *Then, for any* $a, m, s \in \mathbb{N}_+$, *with* $m > 1$, $n = 2ma + s$ *and* $\delta > 4(m-1)\beta(a)$, *it holds that*

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(Z_1)\right] - \frac{1}{n}\sum_{i=1}^{n} f(Z_i)\right| \leq \varepsilon_0(a, m, \delta)\right] \geq 1 - \delta,$$

where

$$\varepsilon_0(a, m, s, \delta) = 2\sqrt{\frac{\log(2\mathcal{S}_{\mathcal{F}}(m))}{m}} + \sqrt{\frac{1}{2m}\log\left(\frac{4}{\delta - 4(m-1)\beta(a)}\right)} + \frac{s}{n}. \tag{23}$$

Proof By an application of Lemma 11 and McDiarmids's inequality over the m independent blocks, it follows that

$$\mathbb{P}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}[f(Z_1)] - \frac{1}{n}\sum_{i=1}^{n} f(Z_i)\right| > \varepsilon\right] \leq 4(e^{-2m\varepsilon'^2} + (m-1)\beta(a)).$$

where

$$\varepsilon' = \varepsilon - \mathbb{E}_*\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}[f(Z_1)] - \frac{1}{m}\sum_{j:B_j\in B^*_{\text{odd}}}\left(\frac{1}{a}\sum_{i:Z_i\in B_j} f(Z_i)\right)\right|\right] - \frac{s}{n}. \tag{24}$$

Denote by $Z_j^{(i)}$ the ith random variable of the jth block $B_j \in B_{\text{odd}}^*$, therefore the expectation in (24) can be written as

$$\mathbb{E}_* \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}[f(Z_1)] - \frac{1}{a} \sum_{j=1}^{a} \left(\frac{1}{m} \sum_{i=1}^{m} f(Z_j^{(i)}) \right) \right| \right] \leq \frac{1}{a} \sum_{j=1}^{a} \mathbb{E}_* \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}[f(Z_1)] - \frac{1}{m} \sum_{i=1}^{m} f(Z_j^{(i)}) \right| \right],$$

where the inequality comes from the triangular inequality and the monotonicity of the supremum.

Note that in $\frac{1}{m} \sum_{i=1}^{m} f(Z_j^{(i)})$ we are considering only one element of each independent block $B_j \in B_{\text{odd}}^*$, therefore this is a sum over iid random variables. Hence, by Theorem 18

$$\mathbb{E}_* \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}[f(Z_1)] - \frac{1}{m} \sum_{j : B_j \in B_{\text{odd}}^*} \left(\frac{1}{a} \sum_{i : Z_i \in B_j} f(Z_i) \right) \right| \right] \leq 2 \sqrt{\frac{\log(2\mathcal{S}_{\mathcal{F}}(m))}{m}}.$$

That is, $\varepsilon' > \varepsilon - 2\sqrt{\frac{\log(2\mathcal{S}_{\mathcal{F}}(m))}{m}} - \frac{s}{n}$. So taking $\delta > 4(e^{-2m\varepsilon'^2} + (m-1)\beta(a))$ and $\varepsilon = \varepsilon_0(a, m, \delta)$ yields

$$\mathbb{P} \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}[f(Z_1)] - \frac{1}{n} \sum_{i=1}^{n} f(Z_i) \right| > \varepsilon_0(a, m, \delta) \right] \leq \delta. \qquad \blacksquare$$

Corollary 20 *Let $(X_*, Y_*), \ldots, (X_n, Y_n)$ be a sample drawn from a stationary β-mixing distribution, $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be any deterministic function and $\mathcal{A}$ be a family of Borel measurable sets in $\mathcal{X}$ with finite* VC *dimension* $\text{VC}(\mathcal{A}) = d$.

Then, for any $m, a \in \mathbb{N}_+$ with $m > 1$, $n = 2ma$ and $\delta > 4(m-1)\beta(a)$, it holds that

$$\mathbb{P} \left[\sup_{A \in \mathcal{A}} \left| \mathbb{P}\left[X_* \in A \right] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \in A\} \right| \leq \varepsilon_0(a, m, \delta) \right] \geq 1 - \delta,$$

where $\varepsilon_0(a, m, \delta)$ is as defined in (23).

Proof Taking $\mathcal{F} = \{x \mapsto \mathbb{1}\{X_* \in A\} \ : \ A \in \mathcal{A}\}$ in Corollary 19 and using Sauer-Shelah lemma (Sauer, 1972; Mohri et al., 2018) yields the result. $\blacksquare$

Lemma 21 *Let $X_1, \ldots, X_n$ be a sample drawn from a stationary β-mixing distribution, $\gamma \in (0, 1)$ and $\mathcal{A}$ be a family of Borel measurable sets in $\mathcal{X}$ with finite* VC *dimension* $\text{VC}(\mathcal{A}) = d$ *such that $\mathbb{P}[X_* \in A] > \gamma$ for all $A \in \mathcal{A}$. For $m, a \in \mathbb{N}_+$ with $m > 1$, $n = 2ma$ and $\delta > 4(m-1)\beta(a)$ suppose that $\frac{2}{\gamma}\varepsilon_0(a, m, \delta) < 1$, with $\varepsilon_0(a, m, \delta)$ as in (23). Then,*

$$\mathbb{P} \left[\inf_{A \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \in A\} > \frac{\gamma}{2} \right] \geq 1 - \delta.$$

Proof By Corollary 20, for any $m, a \in \mathbb{N}_+$ with $m > 1$, $n = 2ma$ and $\delta > 4(m-1)\beta(a)$, using the fact that $\varepsilon_0(a, m, \delta) < \gamma/2$,

$$\begin{aligned}
\mathbb{P}\left[\inf_{A\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \in A\} \leq \frac{\gamma}{2}\right] &= \mathbb{P}\left[\sup_{A\in\mathcal{A}} \gamma - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \in A\} \geq \frac{\gamma}{2}\right] \\
&\leq \mathbb{P}\left[\sup_{A\in\mathcal{A}} \left|\mathbb{P}[X_* \in A] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \in A\}\right| \geq \frac{\gamma}{2}\right] \\
&\leq \mathbb{P}\left[\sup_{A\in\mathcal{A}} \left|\mathbb{P}[X_* \in A] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \in A\}\right| \geq \varepsilon_0(a, m, \delta)\right] \\
&\leq \delta.
\end{aligned}$$

■

Lemma 22 *Let $(X_*, Y_*), \ldots, (X_n, Y_n)$ be a sample drawn from a stationary β-mixing distribution, $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a deterministic function, $\gamma \in (0, 1)$ and $\mathcal{A}$ be a family of Borel measurable sets in $\mathcal{X}$ with finite* VC *dimension $\mathrm{VC}(\mathcal{A}) = d$ such that $\mathbb{P}[X_* \in A] > \gamma$ for all $A \in \mathcal{A}$. For $m, a \in \mathbb{N}_+$ with $m > 1$, $n = 2ma$ and $\delta > 8(m-1)\beta(a)$, if $\varepsilon \coloneqq \frac{2}{\gamma}\varepsilon_0(a, m, \delta/2) < 1$, for $\varepsilon_0(a, m, \delta/2)$ as in (23), then with probability at least $1 - \delta$*

$$\sup_{\substack{A\in\mathcal{A} \\ t\in\mathbb{R}}} \left| \frac{\mathbb{P}[s(X_*, Y_*) \leq t, X_* \in A]}{\mathbb{P}[X_* \in A]} - \frac{\sum_{i=1}^{n} \mathbb{1}\{s(X_i, Y_i) \leq t\}\, \mathbb{1}\{X_i \in A\}}{\sum_{i=1}^{n} \mathbb{1}\{X_i \in A\}} \right| \leq \varepsilon.$$

Proof Define ε as in the lemma statement. We want to show that:

$$C = \left\{ \sup_{\substack{A\in\mathcal{A} \\ t\in\mathbb{R}}} \left| \frac{\mathbb{P}[s(X_*, Y_*) \leq t, X_* \in A]}{\mathbb{P}[X_* \in A]} - \frac{\sum_{i=1}^{n} \mathbb{1}\{s(X_i, Y_i) \leq t\}\, \mathbb{1}\{X_i \in A\}}{\sum_{i=1}^{n} \mathbb{1}\{X_i \in A\}} \right| > \varepsilon \right\}$$

has probability at most δ. To this end, we define the following auxiliary event, which controls the random denominator term in C:

$$B = \left\{ \inf_{A\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \in A\} > \frac{\gamma}{2} \right\}.$$

By Lemma 21, $\mathbb{P}[B^c] < \delta/2$, so it suffices to show that $\mathbb{P}[E] \leq \frac{\delta}{2}$ where $E \coloneqq C \cap B$. Note that, if E holds, then the quotient $\frac{\sum_{i=1}^{n} \mathbb{1}\{s(X_i, Y_i) \leq t\}\, \mathbb{1}\{X_i \in A\}}{\sum_{i=1}^{n} \mathbb{1}\{X_i \in A\}}$ is well defined and

$$\begin{aligned}
\varepsilon &< \sup_{\substack{A\in\mathcal{A} \\ t\in\mathbb{R}}} \left| \frac{\mathbb{P}[s(X_*, Y_*) \leq t, X_* \in A]}{\mathbb{P}[X_* \in A]} - \frac{\sum_{i=1}^{n} \mathbb{1}\{s(X_i, Y_i) \leq t\}\, \mathbb{1}\{X_i \in A\}}{\sum_{i=1}^{n} \mathbb{1}\{X_i \in A\}} \right| \\
&\leq \sup_{\substack{A\in\mathcal{A} \\ t\in\mathbb{R}}} \left| \frac{\mathbb{P}[s(X_*, Y_*) \leq t, X_* \in A] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{s(X_i, Y_i) \leq t\}\, \mathbb{1}\{X_i \in A\}}{\mathbb{P}[X_* \in A]} \right| \\
&+ \sup_{\substack{A\in\mathcal{A} \\ t\in\mathbb{R}}} \left| \frac{\sum_{i=1}^{n} \mathbb{1}\{s(X_i, Y_i) \leq t\}\, \mathbb{1}\{X_i \in A\} \left(\mathbb{P}[X_* \in A] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \in A\}\right)}{\mathbb{P}[X_* \in A] \sum_{i=1}^{n} \mathbb{1}\{X_i \in A\}} \right|
\end{aligned}$$

$$\leq \sup_{\substack{A\in\mathcal{A}\\ t\in\mathbb{R}}} \left| \frac{\mathbb{P}[s(X_*,Y_*)\leq t, X_*\in A] - \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{s(X_i,Y_i)\leq t\}\,\mathbb{1}\{X_i\in A\}}{\gamma} \right|$$
$$+ \sup_{A\in\mathcal{A}} \left| \frac{\mathbb{P}[X_*\in A] - \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i\in A\}}{\gamma} \right|,$$

Moreover, for any $A\in\mathcal{A}$:

$$\left| \frac{\mathbb{P}[X_*\in A] - \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i\in A\}}{\gamma} \right|$$
$$= \lim_{t\to+\infty} \left| \frac{\mathbb{P}[s(X_*,Y_*)\leq t, X_*\in A] - \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{s(X_i,Y_i)\leq t\}\,\mathbb{1}\{X_i\in A\}}{\gamma} \right|$$

We deduce that:

$$E \text{ holds } \Rightarrow \varepsilon < 2 \sup_{A\in\mathcal{A}} \left| \frac{\mathbb{P}[X_*\in A] - \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i\in A\}}{\gamma} \right|. \tag{25}$$

By Corollary 20, we know that, with our choice of ε:

$$\mathbb{P}\left\{ \sup_{A\in\mathcal{A}} \left| \frac{\mathbb{P}[X_*\in A] - \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i\in A\}}{\gamma} \right| > \frac{\varepsilon}{2} \right\} \leq \frac{\delta}{2}.$$

By (25), we also have $\mathbb{P}[E]\leq\delta/2$. This finishes the proof. ■

Proposition 23 *Let*

$$\varepsilon = \inf_{(a,m,r)\in G_{\text{cal}}} \left\{ \frac{2}{\gamma}\left(\varepsilon_0\left(a,m,\frac{\delta-\beta(r)}{4}\right) + \frac{2(r-1)}{n_{\text{cal}}} \right) \right\}$$

where $\varepsilon_0(a,m,\delta/2)$ as in (23) and

$$G_{\text{cal}} = \{(a,m,r)\in\mathbb{N}^3_{>0} \ : \ 2ma = n_{\text{cal}} - r + 1, \delta > 16(m-1)\beta(a) + \beta(r)\}.$$

If $\varepsilon<1$, then condition (9) holds with $\varepsilon_{\text{cal}}=\varepsilon$.

Proof The proof is similar to the proof of Proposition 13. For $\varepsilon>0$ and $r\in\{1,\ldots,n_{\text{cal}}\}$, let $I_{\text{cal},r} = \{n_{\text{train}}+r,\ldots,n_{\text{train}}+n_{\text{cal}}\}$ and $I_{\text{cal},r}(A) = \{i\in I_{\text{cal},r} \ : \ X_i\in A\}$. Define the events

$$E(r,\varepsilon') = \left\{ \inf_{A\in\mathcal{A}} \left| \frac{\sum_{i\in I_{\text{cal},r}(A)} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i,Y_i)\leq q_{\text{train}}\}}{\#I_{\text{cal},r}(A)} - P_{q,\text{train}}(A) \right| > \varepsilon' \right\},$$
$$C = \left\{ \inf_{A\in\mathcal{A}} \frac{1}{n_{\text{cal}}} \sum_{i\in I_{\text{cal}}} \mathbb{1}\{X_i\in A\} > \frac{\gamma}{2} \right\} \quad \text{and} \quad B(r,\varepsilon') = E(r,\varepsilon')\cap C.$$

We want to show that there exists $\varepsilon' > 0$ such that if $\varepsilon' < 1$ then $\mathbb{P}[E(1,\varepsilon')] \leq \delta$. Note that if $B(1,\varepsilon')$ holds, then for all $A \in \mathcal{A}$,

$$\left| \frac{\sum_{i \in I_{\text{cal},r}(A)} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}\}}{\#I_{\text{cal},r}(A)} - P_{q,\text{train}}(A) \right| > \varepsilon' - \frac{2(r-1)}{\gamma n_{\text{cal}}}.$$

That is, $B(1,\varepsilon') \subset B\left(r, \varepsilon' - \frac{2(r-1)}{\gamma n_{\text{cal}}}\right)$. Now, define $\mathbb{P}_* = \mathbb{P}_1^{n_{\text{train}}} \otimes \mathbb{P}_{n_{\text{train}}+r}^{n_{\text{train}}+n_{\text{cal}}}$, so under $\mathbb{P}_*$ we have that $(X_i, Y_i)_{i \in I_{\text{train}}}$ and $(X_i, Y_i)_{i \in I_{\text{cal},r}}$ are independent. By Proposition 10 we have

$$\mathbb{P}[B(1,\varepsilon')] \leq \mathbb{P}\left[B\left(r, \varepsilon' - \frac{2(r-1)}{\gamma n_{\text{cal}}}\right)\right] \leq \mathbb{P}_*\left[B\left(r, \varepsilon' - \frac{2(r-1)}{\gamma n_{\text{cal}}}\right)\right] + \beta(r).$$

But this implies that $\mathbb{P}[E(1,\varepsilon')] \leq \mathbb{P}_*\left[B\left(r, \varepsilon' - \frac{2(r-1)}{\gamma n_{\text{cal}}}\right)\right] + \beta(r) + 1 - \mathbb{P}[C]$. For any $m, a \in \mathbb{N}_+$ with $n_{\text{cal}} - r + 1 = 2ma$ and $\delta_{\text{cal}} > 8(m-1)\beta(a)$, if we take

$$\varepsilon' = \frac{1}{\gamma}\left(4\sqrt{\frac{\log(2(m+1)^d)}{m}} + 2\sqrt{\frac{1}{2m}\log\left(\frac{8}{\delta - 8(m-1)\beta(a)}\right)} + \frac{2(r-1)}{n_{\text{cal}}}\right),$$

and assume that $\varepsilon' < 1$ then

$$\frac{1}{\gamma}\left(4\sqrt{\frac{\log(2(m+1)^d)}{m}} + 2\sqrt{\frac{1}{2m}\log\left(\frac{8}{\delta - 8(m-1)\beta(a)}\right)}\right) < 1$$

so Lemma 22 tells us that $\mathbb{E}_*\left[\mathbb{P}_*\left[B\left(r, \varepsilon' - \frac{2(r-1)}{\gamma n_{\text{cal}}}\right) \mid (X_i, Y_i)_{i \in I_{\text{train}}}\right]\right] \leq \delta$ and Lemma 21 tells us $1 - \mathbb{P}[C] \leq \delta$. That is, $\mathbb{P}[E(1,\varepsilon')] \leq 2\delta + \beta(r)$, which is equivalent to $\mathbb{P}[E(1,\varepsilon)] \leq \delta$, if ε is as in the proposition statement. ■

Proposition 24 *Condition (10) holds with* $\varepsilon_{\text{train}} = \beta(k - n_{\text{train}})$.

Proof Given $k \in I_{\text{test}}$, note that we can decompose $\mathbb{P}_* = \mathbb{P}_1^{n_{\text{train}}} \otimes \mathbb{P}_k^k$, so under $\mathbb{P}_*$ we have that (X_k, Y_k) is independent of $(X_i, Y_i)_{i \in I_{\text{train}}}$. Then, defining $\beta_k = \beta(k - n_{\text{train}})$ we have for all $A \in \mathcal{A}$,

$$\beta_k \geq |\mathbb{P}[\widehat{s}_{\text{train}}(X_k, Y_k) \leq q_{\text{train}}(A), X_k \in A] - \mathbb{P}_*[\widehat{s}_{\text{train}}(X_k, Y_k) \leq q_{\text{train}}(A), X_k \in A]|$$

where the β_k penalty follows from Proposition 10. But then, by a conditioning argument,

$$\beta_k \geq |\mathbb{P}[\widehat{s}_{\text{train}}(X_k, Y_k) \leq q_{\text{train}}(A), X_k \in A] - \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq q_{\text{train}}(A), X_* \in A]|\,.$$

Since $\frac{\beta_k}{\mathbb{P}[X_k \in A]} \geq \beta_k$, dividing by $\mathbb{P}[X_k \in A] = \mathbb{P}[X_* \in A]$ yields the result. ■

Proposition 25 *Define*

$$\varepsilon = \inf_{(a,m,s)\in G_{\text{test}}} \left\{ \frac{2}{\gamma} \left(\varepsilon_0 \left(a, m, \frac{\delta - \beta(n_{\text{cal}})}{2} \right) \right) + \frac{s}{n_{\text{test}}} \right\}$$

where $G_{\text{test}} = \{(a,m) \in \mathbb{N}^2_{>0} \ : \ s + 2ma = n_{\text{test}}, \delta > 8(m-1)\beta(a) + \beta(n_{\text{cal}})\}$. *If* $\varepsilon < 1$, *then condition (19) holds with* $\varepsilon_{test} = \varepsilon$.

Proof The proof is similar to the proof of Proposition 23. Let the event $E(\varepsilon)$ be

$$E(\varepsilon) = \left\{ \inf_{A\in\mathcal{A}} \left| P_{q,\text{train}}(A) - \frac{\sum_{i\in I_{\text{test}}(A)} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \le q_{\text{train}}]\}}{n_{\text{test}}(A)} \right| > \varepsilon \right\},$$

Define, $\mathbb{P}_* = \mathbb{P}_1^{n_{\text{train}}} \otimes \mathbb{P}_{n_{\text{train}}+n_{\text{cal}}}^{n_{\text{train}}+n_{\text{cal}}+n_{\text{test}}}$, so under $\mathbb{P}_*$ we have that $(X_i, Y_i)_{i\in I_{\text{train}}}$ and $(X_i, Y_i)_{i\in I_{\text{test}}}$ are independent. By Proposition 10 we have

$$\mathbb{P}[E(\varepsilon)] \le \mathbb{P}_*[E(\varepsilon)] + \beta(n_{\text{cal}}) = \mathbb{E}_*[\mathbb{P}_*[E(\varepsilon) \mid (X_i, Y_i)_{i\in I_{\text{train}}}]] + \beta(n_{\text{cal}}).$$

Now we can apply Lemma 22 and conclude that if $\varepsilon < 1$, for any $m, a \in \mathbb{N}_+$ with $n_{\text{test}} = 2ma$ and $\delta_{\text{test}} > 8(m-1)\beta(a)) + \beta(n_{\text{cal}})$, it is true that $\mathbb{P}[E] \le \delta$, where

$$\varepsilon = \frac{1}{\gamma} \left(4\sqrt{\frac{\log(2(m+1)^d)}{m}} + 2\sqrt{\frac{1}{2m} \log\left(\frac{8}{\delta - 8(m-1)\beta(a) - \beta(n_{\text{cal}})} \right)} + \frac{s}{n_{\text{test}}} \right).$$

Finally, since this is true for any choice of $a, m, s \in \mathbb{N}_{>0}$ with $s + 2ma = n_{\text{test}}$ and $\delta_{\text{test}} > 8(m-1)\beta(a) + \beta(n_{\text{cal}})$, we can choose a, m, s optimally to minimize ε. ■

Proof [Theorem 16] Follows from applying Propositions 23 and 24 in Theorem 3. ■

Proof [Theorem 17] Follows from applying Propositions 23 and 25 in Theorem 9. ■

Appendix D. Proofs of Section 3.4 and Further Results

D.1 Non-Stationary Data

Proof [Theorem 6] This proof is similar to that of Theorem 1, but the notation is somewhat more complicated due to nonstationarity.

Consider N_{cal} as in the statement of the present Theorem. For $\phi \in (0,1)$, define the conditional ϕ-quantile of $\widehat{s}_{\text{train}}(X_{*,N_{\text{cal}}}, Y_{*,N_{\text{cal}}})$ given the training data:

$$q_{\phi,\text{train}} \coloneqq \inf\{t \in \mathbb{R} \ : \ \mathbb{P}[\widehat{s}_{\text{train}}(X_{*,N_{\text{cal}}}, Y_{*,N_{\text{cal}}}) \le t \mid \{(X_i, Y_i)\}_{i\in I_{\text{train}}}] \ge \phi\},$$

or alternatively,

$$q_{\phi,\text{train}} \coloneqq \inf \left\{ t \in \mathbb{R} \ : \ \frac{1}{n_{\text{cal}}} \sum_{j\in N_{\text{cal}}} \mathbb{P}[\widehat{s}_{\text{train}}(X_{*,j}, Y_{*,j}) \le t \mid \{(X_i, Y_i)\}_{i\in I_{\text{train}}}] \ge \phi \right\}.$$

For each $j \in I_{\text{cal}}$, set $p^{(j)}_{\phi,\text{train}} \coloneqq \mathbb{P}[\widehat{s}_{\text{train}}(X_{*,j}, Y_{*,j}) \leq q_{\phi_\ell,\text{train}} \mid \{(X_i, Y_i)\}_{i \in I_{\text{train}}}]$. Fix $i \in I_{\text{test}}$. Following the proof of Theorem 1, we consider values of ϕ of the form:

$$\phi_\ell \coloneqq 1 - \alpha - \varepsilon_{\text{cal}} - \delta^{(i)} - 1/\ell \text{ for } \ell = 1, 2, 3 \ldots$$

to obtain that $\mathbb{P}[F] \geq 1 - \delta_{\text{cal}}$, where $F \coloneqq \{q_{1-\alpha-\varepsilon_{\text{cal}},\text{cal}} \leq \widehat{q}_{1-\alpha,\text{cal}}\}$. Now,

$$\begin{aligned}\mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \leq \widehat{q}_{1-\alpha,\text{cal}}] \geq& \mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{1-\alpha-\varepsilon_{\text{cal}}}] - \mathbb{P}[F^c] \\ \geq& \mathbb{P}[\widehat{s}_{\text{train}}(X_{*,i}, Y_{*,i}) \leq q_{1-\alpha-\varepsilon_{\text{cal}}}] - \delta_{\text{cal}} - \varepsilon_{\text{test}}.\end{aligned}$$

By assumption, the law of $(X_{*,i}, Y_{*,i})$ is $\delta^{(i)}$-close in total variation to that of $(X_{*,N_{\text{cal}}}, Y_{*,N_{\text{cal}}})$. Since the event $\{\widehat{s}_{\text{train}}(X_{*,i}, Y_{*,i}) \leq q_{1-\alpha-\delta^{(i)}-\varepsilon_{\text{cal}}}\}$ depends on $(X_{*,i}, Y_{*,i})$ and on the independent process $(X_j, Y_j)_{j \in [n]}$, and $q_{1-\alpha-\varepsilon_{\text{cal}},\text{train}}$ is the conditional $(1 - \alpha - \varepsilon_{\text{cal}})$-quantile of $\widehat{s}_{\text{train}}(X_{*,N_{\text{cal}}}, Y_{*,N_{\text{cal}}})$ given the training data, one may conclude:

$$\begin{aligned}\mathbb{P}[\widehat{s}_{\text{train}}(X_{*,i}, Y_{*,i}) \leq q_{1-\alpha-\varepsilon_{\text{cal}}}] \geq& \mathbb{P}[\widehat{s}_{\text{train}}(X_{*,N_{\text{cal}}}, Y_{*,N_{\text{cal}}}) \leq q_{1-\alpha-\varepsilon_{\text{cal}}}] - \delta^{(i)} \\ \geq& 1 - \alpha - \varepsilon_{\text{cal}} - \delta^{(i)}.\end{aligned}$$

Plugging this back into the previous display finishes the proof. ■

D.2 Risk-Controlling Prediction Sets

Risk-controlling prediction sets (RCPS), introduced by Bates et al. (2021), give a general methodology for CP that applies in a variety of settings, including regression, multiclass classification and image segmentation. Importantly, RCPS does not involve nonconformity scores, but rather, the construction of nested sets. While the original theory of RCPS assumes independent data, we now show it also applies within our framework.

Suppose $\mathcal{Y}'$ is a family of sets, $\Lambda \subset \mathbb{R} \cup \{+\infty\}$ is a closed set, and a map $\mathcal{T} \colon (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}} \times \mathcal{X} \times \Lambda \to \mathcal{Y}'$ is given with the following property: for all choices of $(x_i, y_i)_{i=1}^{n_{\text{cal}}} \in (\mathcal{X} \times \mathcal{Y})^{n_{\text{cal}}}$, $x \in \mathcal{X}$ and $\lambda_1, \lambda_2 \in \Lambda$: if $\lambda_1 \leq \lambda_2$, then $\mathcal{T}((x_i, y_i)_{i=1}^{n_{\text{cal}}}, x, \lambda_1) \subset \mathcal{T}((x_i, y_i)_{i=1}^{n_{\text{cal}}}, x, \lambda_2)$.

For $(x, \lambda) \in \mathcal{X}$, we use the notation

$$\widehat{\mathcal{T}}_{\lambda,\text{train}}(x) \coloneqq \mathcal{T}((X_i, Y_i)_{i \in I_{\text{train}}}, x, \lambda)$$

to denote the values of $\mathcal{T}$ when the first n_{train} pairs in the input correspond to the training data. We call $\widehat{\mathcal{T}}_{\lambda,\text{train}}(\cdot)$ a trained tolerance region. Finally, $L \colon \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}$ is a loss function that is decreasing in the $\mathcal{Y}'$ component. The goal of RCPS is to compute a value $\widehat{\lambda}$ from the calibration data that achieves (conditional) risk smaller than a prespecified level $\alpha > 0$.

To define the conditional risk, assume that the map from $\lambda \in \Lambda$ to $\mathbb{E}[L(Y_*, \widehat{\mathcal{T}}_\lambda(X_*)) \mid (X_i, Y_i)_{i \in I_{\text{train}}}]$ almost surely is continuous and achieves arbitrarily small positive values. Given a measurable $\ell \colon (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}} \to \Lambda$, we let $\ell_{\text{train}} \coloneqq \ell((X_i, Y_i)_{i \in I_{\text{train}}})$, define the conditional expected risk as

$$R(\ell) \coloneqq \mathbb{E}[L\big(Y_*, \widehat{\mathcal{T}}_{\ell_{\text{train}}}(X_*)\big) \mid (X_i, Y_i)_{i \in I_{\text{train}}}].$$

Also, define the empirical risk over the calibration data as

$$\widehat{R}_{\text{cal}}(\lambda) \coloneqq \frac{1}{n_{\text{cal}}} \sum_{i \in I_{\text{cal}}} L(Y_i, \mathcal{T}_\lambda(X_i)).$$

Now, a threshold $\widehat{\lambda}$ must be chosen from calibration data to control the risk. In Bates et al. (2021), this requires finding a function $\lambda \mapsto \widehat{R}_{\text{UCB}}(\lambda)$ that gives a pointwise high-probability upper bound on $R(\lambda)$. In our case, we can allow for a $\widehat{R}(\lambda)$ that bounds the risk up to a small error; for us, the empirical risk will play this role. Thus, consider the empirical threshold

$$\widehat{\lambda}_{\alpha,\text{cal}} \coloneqq \inf \left\{ \lambda \in \Lambda \,:\, \forall \lambda' \in \Lambda, \lambda' > \lambda \implies \widehat{R}_{\text{cal}}(\lambda) < \alpha \right\}.$$

Finally, we give conditions that guarantee that $\widehat{\lambda}_{\alpha,\text{cal}}$ controls the risk with high probability. First, assume that there exist $\varepsilon_{\text{cal}} > 0$, $\delta_{\text{cal}} \in (0,1)$ such that, for any $\ell, \ell_{\text{train}}$,

$$\mathbb{P}\left[|\widehat{R}_{\text{cal}}(\ell_{\text{train}}) - R(\ell)| \le \varepsilon_{\text{cal}} \right] \ge 1 - \delta_{\text{cal}}. \tag{26}$$

Also, assume there exists a $\varepsilon_{\text{test}}$ such that for all $i \in I_{\text{test}}$ and all ℓ,

$$|\mathbb{E}[L(Y_i, \mathcal{T}_{\ell_{\text{train}}}(X_i))] - \mathbb{E}[R(\ell)]| \le \varepsilon_{\text{test}}. \tag{27}$$

Then, the following result on the performance of RCPS over a single test point holds.

Theorem 26 (Approximate risk control for $\widehat{\lambda}_{\alpha,\text{cal}}$) *Assume (26) and (27). Then,*

$$\mathbb{P}\left[\mathbb{E}[L(Y_*, \mathcal{T}_{\widehat{\lambda}_{\alpha,\text{cal}}}(X_*))] \le \alpha + \varepsilon_{\text{cal}} \right] \ge 1 - \delta_{\text{cal}}.$$

Moreover, if L is uniformly bounded, we have the following for all $i \in I_{\text{test}}$:

$$\mathbb{E}[L(Y_i, \mathcal{T}_{\widehat{\lambda}_{\alpha,\text{cal}}}(X_i))] \le \alpha + \varepsilon_{\text{test}} + \varepsilon_{\text{cal}} + \|L\|_\infty \delta_{\text{test}}.$$

Proof The proof is a combination of our arguments for Theorem 1 with the reasoning in Bates et al. (2021). Recall that for any function $\ell : (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}} \to \Lambda$, if $\ell_{\text{train}} \coloneqq \ell((X_i, Y_i)_{i \in I_{\text{train}}})$,

$$R(\ell) \coloneqq \mathbb{E}[L(Y_*, \mathcal{T}_{\ell_{\text{train}}}(X_*)) \mid (X_i, Y_i)_{i \in I_{\text{train}}}].$$

Make the specific choice $\ell_{\text{train}} \coloneqq \inf\{\lambda \in \Lambda \,:\, R(\lambda) \le \alpha + \varepsilon_{\text{cal}}\}$, so that, by right-continuity of $\lambda \mapsto \mathbb{E}[L(Y_*, \mathcal{T}_\lambda(X_*)) \mid (X_i, Y_i)_{i \in I_{\text{train}}}]$,

$$R(\ell_{\text{train}}) \le \alpha + \varepsilon_{\text{cal}} \tag{28}$$

while at the same time $R(\ell_{\text{train}} - 1/k) > \alpha + \varepsilon_{\text{cal}}$ for all $k \in \mathbb{N}$. The definition of $\widehat{\lambda}_{\alpha,\text{cal}} = \inf\{\lambda \in \Lambda \,:\, \widehat{R}_{\text{cal}}(\lambda) < \alpha\}$ and the fact that our risk decreases with λ imply:

$$\mathbb{P}[\widehat{\lambda}_{\alpha,\text{cal}} \ge \ell_{\text{train}} - 1/k] \ge \mathbb{P}[\widehat{R}_{\text{cal}}(\ell_{\text{train}} - 1/k) > \alpha] \ge 1 - \delta_{\text{cal}}$$

where the last step uses assumption (26). Letting $k \to +\infty$ gives:

$$\mathbb{P}[\widehat{\lambda}_{\alpha,\text{cal}} \ge \ell_{\text{train}}] \ge 1 - \delta_{\text{cal}} \tag{29}$$

Now, $\widehat{\lambda}_{\alpha,\text{cal}} \geq \ell_{\text{train}}$ and (28) together imply:

$$\mathbb{E}[L(Y_*, \mathcal{T}_{\widehat{\lambda}_{\alpha,\text{cal}}}(X_*))] \leq \mathbb{E}[L(Y_*, \mathcal{T}_{\ell_{\text{train}}}(X_*))] = \mathbb{E}[R(\ell_{\text{train}})] \leq \alpha + \varepsilon_{\text{cal}}, \tag{30}$$

so the first assertion in the Theorem follows from (29). The second assertion follows from:

$$\mathbb{E}[L(Y_i, \mathcal{T}_{\widehat{\lambda}_{\alpha,\text{cal}}}(X_i))] \leq \mathbb{E}[L(Y_i, \mathcal{T}_{\ell_{\text{train}}}(X_i))] + \|L\|_\infty \mathbb{P}[\widehat{\lambda}_{\alpha,\text{cal}} < \ell_{\text{train}}]$$

combined with (27) and (30). ■

Thus the expected loss at any test point is controlled by α plus an error term that can be shown to be small, even for non-exchangeable data. Importantly, the result is achieved via assumptions that only bound the behavior of the loss over individual thresholds ℓ_{train} obtained from the training data. In particular, there is no need to require uniform control of the loss over a range of ℓ, which would require stronger (and looser) concentration bounds. The uniform bound on L can be replaced by a moment assumption, at the cost of a less clean bound.

D.3 Rank-One-Out Conformal Prediction

Rank-one-out (ROO) conformal prediction, introduced by Lei et al. (2018), is different from split CP in that the method calibrates the quantile used for each test data point by looking at the remaining test points. This requires adapting the above setup as follows: partition the data indices as $[n] = I_{\text{train}} \sqcup I_{\text{test}}$, and for each $i \in I_{\text{test}}$ the calibration set is $I^{(i)}_{\text{cal}} = I_{\text{test}} \setminus \{i\}$. Also, define the empirical quantiles as follows: given $\phi \in [0, 1)$ and $i \in I_{\text{test}}$, let $\widehat{q}^{(i)}_{\phi,\text{cal}}$ denote the empirical ϕ-quantile

$$\widehat{q}^{(i)}_{\phi,\text{cal}} \coloneqq \inf \left\{ t \in \mathbb{R} \,:\, \frac{1}{n_{\text{test}} - 1} \sum_{j \in I^{(i)}_{\text{cal}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_j, Y_j) \leq t\} \geq \phi \right\}.$$

For $x \in \mathcal{X}$, the rank-one-out prediction set for $i \in I_{\text{test}}$ is then defined via:

$$C^{(i)}_\phi(x) \coloneqq \{y \in \mathcal{Y} \,:\, \widehat{s}_{\text{train}}(x, y) \leq \widehat{q}^{(i)}_{\phi,\text{cal}}\}.$$

We can then adapt the concentration and decoupling hypotheses. Indeed, we assume there exist $\varepsilon_{\text{test}} \in (0, 1)$, $\{\varepsilon_{\text{test}}(i)\}_{i \in I_{\text{test}}} \subset (0, 1)$ and $\delta_{\text{test}} \in (0, 1)$ such that, for any $i \in I_{\text{test}}$,

$$|\mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}] - \mathbb{P}[\widehat{s}_{\text{train}}(X_*, Y_*) \leq q_{\text{train}}]| \leq \varepsilon_{\text{test}}(i), \tag{31}$$

and, moreover,

$$\mathbb{P}\left[\left| \frac{1}{n_{\text{test}}} \sum_{i \in I_{\text{test}}} \mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}\} - P_{q,\text{train}} \right| \leq \varepsilon_{\text{test}} \right] \geq 1 - \delta_{\text{test}}. \tag{32}$$

Then, the analogue of Theorems 1 and 2 still hold for ROO.

Theorem 27 (Marginal and empirical coverage over test data for ROO) *Given a level $\alpha \in (0,1)$, if (31) and (32) hold, then, for all $i \in I_{\text{test}}$:*

$$\mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \leq \widehat{q}_{1-\alpha,\text{cal}}] \geq 1 - \alpha - \varepsilon_{\text{test}}(i) - \varepsilon_{\text{test}} - \delta_{\text{cal}} - \frac{1}{n_{\text{test}}}.$$

Moreover, it holds that

$$\mathbb{P}\left[\frac{1}{n_{\text{test}}} \sum_{i \in I_{\text{test}}} \mathbb{1}\left\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq \widehat{q}^{(i)}_{1-\alpha,\text{cal}}\right\} \geq 1 - \alpha - \varepsilon_{\text{test}} - \frac{1}{n_{\text{test}}}\right] \geq 1 - 2\delta_{\text{test}}.$$

Proof If we consider $I_{\text{cal}} \coloneqq I_{\text{test}}$ in the proof of Theorem 1, the event $F = \{q_{1-\alpha-\varepsilon_{\text{cal}},\text{train}} \leq \widehat{q}_{1-\alpha,\text{cal}}\}$ satisfies $\mathbb{P}[F] \geq 1 - \delta_{\text{cal}}$. But since $\widehat{q}^{(i)}_{1-\alpha,\text{cal}} \geq \widehat{q}_{1-\alpha-1/n_{\text{test}},\text{cal}}$, it is also true that the event $F' = \{q_{1-\alpha-\varepsilon_{\text{cal}}-1/n_{\text{test}},\text{train}} \leq \widehat{q}^{(i)}_{1-\alpha,\text{cal}}\}$ also satisfies $\mathbb{P}[F'] \geq 1 - \delta_{\text{cal}}$. The rest of the proof follows the same strategy as in Theorem 1 using $\widehat{q}^{(i)}_{1-\alpha,\text{cal}}$ instead of $\widehat{q}_{1-\alpha,\text{cal}}$. ■

One can adapt the analysis in Section 3.3 to bound the parameters δ_{test}, $\varepsilon_{\text{test}}$ and $\varepsilon_{\text{test}}(i)$ for β-mixing data. In particular, one may take $\varepsilon_{\text{test}}(i) = \beta(i - n_{\text{cal}})$, and $\varepsilon_{\text{test}}, \delta_{\text{test}}$ equal to the respective parameters $\varepsilon_{\text{cal}}, \delta_{\text{cal}}$ in that section, but with n_{test} replacing n_{cal} (since the calibration set for each point of rank-one-out is essentially equal to the test set).

On the other hand, we note that marginal coverage might suffer somewhat over the first few test data, since $\varepsilon_{\text{test}}(i) = \beta(i - n_{\text{cal}})$ may be large for small values $i - n_{\text{cal}}$. In contrast to split CP, there is no gap in ROO between training and test data so the first test points may be strongly correlated with the training data.

Appendix E. Details of Section 4

For the experiment comparing split CP and EnbPI (cf. Figure 3), split CP's underlying random forest model comprised 100 trees; mean squared error was used as the split criterion; no maximum tree depth was set, so nodes are expanded until all leaves contain less than 2 samples; all features were considered for splitting. EnbPI's random forest was exactly the same and the length of the blocks in the EnbPI's block bootstrap procedure was set to 8 with the number of resamplings set to 30. For the AR(1) experiment (cf. Figure 1) and Examples 3 and 4, gradient boosting was set to boost 100 trees with a learning rate of 0.1 and pinball loss; trees of any depth were allowed; the minimal number of data in one leaf was 20; the minimal sum hessian in one leaf was 0.001; no minimal gain to perform a split was required; no more than 31 leaves were allowed per tree; no regularization was set. The neural network consisted of three fully connected layers with ReLU activation; the number of output units were 128, 64 and 2, respectively, where the final output of 2 units represents the low and high quantiles being estimated; AdamW with learning rate of 10^{-3} and weight decay of 10^{-6} was used; training was over 100 epochs with batches of size 64; pinball loss was used. The random forest model (quantile regression forest) comprised 10 trees; mean squared error was used as the split criterion; no maximum tree depth was set, so nodes are expanded until all leaves contain less than 2 samples; all features were considered for splitting. Quantile regressors making use of the pinball loss $L_\tau(y, \hat{y}) = \tau(y - \hat{y})\,\mathbb{1}\{y \geq \hat{y}\} + (1 - \tau)(\hat{y} - y)\,\mathbb{1}\{y < \hat{y}\}$ had τ set to $\alpha/2$ and $1 - \alpha/2$, with α the acceptable miscoverage level.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023. ISSN 1935-8237. doi: 10.1561/2200000101.

Anastasios N. Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eNdiU_DbM9.

Anthony Arguez, Imke Durre, Scott Applequist, Russell S Vose, Michael F Squires, Xungang Yin, Richard R Heim, and Timothy W Owen. NOAA's 1981–2010 US climate normals: an overview. *Bulletin of the American Meteorological Society*, 93(11):1687–1697, 2012.

Rina Foygel Barber. Hoeffding and bernstein inequalities for weighted sums of exchangeable random variables. *arXiv preprint arXiv:2404.06457*, 2024.

Rina Foygel Barber, Emmanuel J. Candès, 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, 2020. ISSN 2049-8772. doi: 10.1093/imaiai/iaaa017.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6): 1–34, 2021.

Marine Carrasco and Xiaohong Chen. Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory*, 18(1):17–39, 2002.

Maxime Cauchois, Suyash Gupta, and John C. Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22(81):1–42, 2021. URL http://jmlr.org/papers/v22/20-753.html.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 732–749. PMLR, 2018. URL http://proceedings.mlr.press/v75/chernozhukov18a.html.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021. doi: 10.1073/pnas.2107794118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2107794118.

Paul Doukhan. *Mixing: properties and examples*, volume 85. Springer Science & Business Media, 2012.

Shai Feldman, Liran Ringel, Stephen Bates, and Yaniv Romano. Risk control for online learning models. *arXiv preprint arXiv:2205.09095*, 2022.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=6vaActvpcp3.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024. URL http://jmlr.org/papers/v25/22-1218.html.

Yotam Hechtlinger, Barnabas Poczos, and Larry Wasserman. Cautious deep learning. *arXiv preprint arXiv:1805.09460*, 2019. URL https://openreview.net/forum?id=SJequsCqKQ.

HistData, 2022. https://www.histdata.com/, Retrieved on 2022-01-27.

Jessica Hwang, Paulo Orenstein, Judah Cohen, Karl Pfeiffer, and Lester Mackey. Improving subseasonal forecasting in the western US with machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2325–2335, 2019.

Vilde Jensen, Filippo Maria Bianchi, and Stian Norman Anfinsen. Ensemble conformalized quantile regression for probabilistic time series forecasting. *arXiv preprint arXiv:2202.08756*, 2022.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017. ISSN 1573-0565. doi: 10.1007/s10994-016-5588-2.

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. doi: 10.1080/01621459.2017.1307116.

Daniel J. McDonald, Cosma Rohilla Shalizi, and Mark Schervish. Estimating beta-mixing coefficients via histograms. *Electronic Journal of Statistics*, 9:2855–2883, 2015. doi: 10.1214/15-EJS1094.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2008/file/7eacb532570ff6858afd2723755ff790-Paper.pdf.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11(2), 2010.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, 2nd edition, 2018. ISBN 978-0-262-03940-6.

Abdelkader Mokkadem. Mixing properties of ARMA processes. *Stochastic processes and their applications*, 29(2):309–315, 1988.

NOAA, 2023. `http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP/.CPC/.temperature/.daily/`, Retrieved on 2023-03-08.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alexander Gammerman. Inductive confidence machines for regression. In *Proceedings of the 13th European Conference on Machine Learning*, ECML '02, pages 345–356. Springer-Verlag, 2002.

William A. Gardner; Antonio Napolitano; Luigi Paura. Cyclostationarity: Half a century of research. *Signal Processing*, 86:639–697, 2006. ISSN 0165-1684. doi: 10.1016/j.sigpro.2005.06.016.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf`.

Philippe Rémy. HistData API. `https://github.com/philipperemy/FX-1-Minute-Data`, 2022.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. URL `http://jmlr.org/papers/v9/shafer08a.html`.

Vianney Taquet, Vincent Blot, Thomas Morzadec, Louis Lacombe, and Nicolas Brunel. Mapie: an open-source library for distribution-free uncertainty quantification, 2022.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, 2005. ISBN 0387001522.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11559–11569. PMLR, 2021. URL `https://proceedings.mlr.press/v139/xu21h.html`.

Chen Xu and Yao Xie. Conformal prediction for time series. *arXiv preprint arXiv:2010.09107*, 2023.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994. doi: 10.1214/aop/1176988849.

Margaux Zaffran, Olivier Feron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25834–25866. PMLR, 2022. URL `https://proceedings.mlr.press/v162/zaffran22a.html`.