Predictive inference for time series: why is split conformal effective despite temporal dependence?

Rina Foygel Barber RINA@UCHICAGO.EDU
Department of Statistics, University of Chicago

Ashwin Pananjady ASHWINPM@GATECH.EDU
Schools of Industrial and Systems Engineering and Electrical and Computer Engineering, Georgia Tech

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

We consider the problem of uncertainty quantification for prediction in a time series: if we use past data to forecast the next time point, can we provide valid prediction intervals around our forecasts? To avoid placing distributional assumptions on the data, in recent years the conformal prediction method has been a popular approach for predictive inference, since it provides distribution-free coverage for any iid or exchangeable data distribution. However, in the time series setting, the strong empirical performance of conformal prediction methods is not well understood, since even short-range temporal dependence is a strong violation of the exchangeability assumption. Using predictors with "memory"—i.e., predictors that utilize past observations, such as autoregressive models—further exacerbates this problem. In this work, we examine the theoretical properties of split conformal prediction in the time series setting, including the case where predictors may have memory. Our results bound the loss of coverage of these methods in terms of a new "switch coefficient", measuring the extent to which temporal dependence within the time series creates violations of exchangeability. Our characterization of the coverage probability is sharp over the class of stationary, β-mixing processes. Along the way, we introduce tools that may prove useful in analyzing other predictive inference methods for dependent data.

Keywords: conformal prediction, time series, mixing, relaxations of exchangeability

1. Introduction

Quantifying uncertainty in forecasts is important across many fields, including climate and weather prediction (Eyring et al., 2024), power systems (Cochran et al., 2015), and supply chain management (Wen et al., 2017). At one extreme, traditional approaches can provide strong theoretical guarantees under parametric assumptions (Box et al., 2015); however, these approaches can yield misleading conclusions when used alongside black-box ML models, which have become state-of-the-art prediction methods in many time series applications (e.g. Hwang et al., 2019). At the other extreme, there exist several black-box uncertainty quantification approaches for time series (Salinas et al., 2020; Borovykh et al., 2017), but these are difficult to equip with theoretical guarantees.

Conformal prediction methods (Vovk et al., 2005; Shafer and Vovk, 2008) occupy a happy medium between these two extremes, and are often preferred for uncertainty quantification in black-box settings because they are easy to "wrap around" any existing prediction model while also providing theoretical coverage guarantees (Angelopoulos and Bates, 2023). In addition to accommodating black-box prediction models, these methods make weak assumptions on the data-generating process, requiring only that the data be exchangeable. Time series data, however, clearly violate these exchangeability assumptions, and a significant body of work has aimed to develop variants of

conformal prediction methods that are adapted for the time series setting (e.g. Chernozhukov et al., 2018; Xu and Xie, 2023a; Gibbs and Candès, 2024).

In spite of these developments, the vanilla split conformal algorithm (Papadopoulos et al., 2002; Lei et al., 2018)—without any modifications or constraints on its implementation—remains an appealing choice for uncertainty quantification in time series models because of its low computational cost and effective practical performance (Chernozhukov et al., 2018; Xu and Xie, 2023b; Oliveira et al., 2024). On the face of it, this may seem quite surprising: due to temporal dependence, time series data is generally far from exchangeable, so how can a framework whose justification relies on exchangeability perform so well? The purpose of this paper is to explain the (often) strong performance of this algorithm in the time series setting.

1.1. The predictive inference problem

To be concrete, suppose we have a time series of covariate-response data $\mathbf{Z} = (Z_1, \dots, Z_{n+1})$, with data points $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$, where X_i is the feature and Y_i is the response. The data point at index $n+1$ is considered to be the "test point", with X_{n+1} observed but Y_{n+1} unobserved, while for $i \in [n] := \{1, \dots, n\}$ we observe the labeled point (X_i, Y_i). We wish to perform uncertainty quantification on the test response Y_{n+1}, by providing a prediction interval around some estimated value. For instance, given a pretrained predictive model $\widehat{f}$ (where $\widehat{f}(X_{n+1})$ is our point prediction for Y_{n+1}), how can we use the available data $(X_i, Y_i)_{i\in[n]}$ to construct a prediction interval around $\widehat{f}(X_{n+1})$ that is likely to contain the target, Y_{n+1}—and, how can we do so without placing overly strong assumptions on the distribution of the data?

Split conformal prediction (Papadopoulos et al., 2002; Vovk et al., 2005) addresses this problem with the following method. Suppose we have a score function $s : \mathcal{Z} \to \mathbb{R}$ that we can evaluate on our data points. Assume for the moment that s is pretrained—that is, the definition of s does not depend on $\mathbf{Z}$. Treating our first n data points as calibration data, we observe that if all data points are iid, the score evaluated at the test point, $s(Z_{n+1})$, must *conform* to the scores of the calibration data points, $(s(Z_i))_{i\in[n]}$ (in that it must be drawn from the same distribution). If we wish to guarantee coverage with probability at least $1-\alpha$, the split conformal prediction set is then given by

$$\widehat{C}_n(X_{n+1}) = \left\{ y \in \mathcal{Y} : s(X_{n+1}, y) \le \text{Quantile}_{(1-\alpha)(1+1/n)}(s(Z_1), \dots, s(Z_n)) \right\}, \tag{1}$$

where the correction factor $1+1/n$ to the coverage is to account for the fact that we can only compute the quantile on the n training points without including the test point.[1] A canonical example in the setting of a real-valued response ($\mathcal{Y} = \mathbb{R}$) is the regression score, $s(z) = |y - \widehat{f}(x)|$ where $z = (x, y)$ and $\widehat{f}$ is a pretrained regression model. This leads to a prediction set of the form $\widehat{C}_n(X_{n+1}) = \widehat{f}(X_{n+1}) \pm \text{Quantile}_{(1-\alpha)(1+1/n)}(s(Z_1), \dots, s(Z_n))$. However, the split conformal method may be implemented with any score function.

In practice, however, the score function is generally not independent of all observed data. For instance, in the setting of the residual score, the regression model $\widehat{f}$ must itself be estimated, which requires data. In such cases, split conformal prediction is based on training a score function s on a portion of the first n data points, and calibrating it on the remaining portion. In particular, letting

1. To formally define the notation $\text{Quantile}(\cdot)$, which computes the quantile of a finite list of values, for any $v \in \mathbb{R}^m$ we use $\text{Quantile}_b(v)$ to denote the $\lceil bm \rceil$-th order statistic of the vector, i.e., $v_{(\lceil bm \rceil)}$ where $v_{(1)} \le \cdots \le v_{(m)}$. We will use the convention that $\text{Quantile}_b(v) = \infty$ if $b > 1$, and $\text{Quantile}_b(v) = -\infty$ if $b \le 0$.

$\mathcal{A}$ denote the (black-box) algorithm used to train the score on the first n_0 data points, the prediction set is given by

$$\widehat{C}_n(X_{n+1}) = \left\{ y \in \mathcal{Y} : s(X_{n+1}, y) \leq \text{Quantile}_{(1-\alpha)(1+1/n_1)}(s(Z_{n_0+1}), \ldots, s(Z_n)) \right\} \quad \text{where } s = \mathcal{A}(Z_1, \ldots, Z_{n_0}) \text{ and } n_1 = n - n_0. \tag{2}$$

In the setting of exchangeable data, split conformal prediction (with any score function, either pretrained as in (1) or data-dependent as in (2)) is guaranteed to cover Y_{n+1} with probability at least $1 - \alpha$ (Papadopoulos et al., 2002; Vovk et al., 2005).

Throughout the paper, we will use the term "pretrained" to describe the setting where the function s is independent of the data $\mathbf{Z}$ (for instance, s uses a model that was trained on an entirely separate dataset), to distinguish it from the scenario where s is trained on $Z_1, \ldots, Z_{n_0}$, as in (2). In the setting of iid data, there is essentially no distinction between the pretrained construction (1) and the split conformal construction (2) (aside from having n versus n_1 many calibration points), since either way, the score function s is independent of the calibration data. In contrast, for a time series setting, this is no longer the case: the first few calibration points, Z_{n_0+i} for small i, may have high dependence with the score function s, since s itself is dependent on all data up to time n_0. For this reason, the split conformal setting will require a more careful analysis.

1.2. A motivating numerical experiment

Despite the theoretical requirement of exchangeability, split conformal prediction has often been observed to perform well on time series data, where the standard exchangeability assumptions are strongly violated. This phenomenon has been observed repeatedly in the literature, where split conformal is often observed to be competitive with other uncertainty quantification methods for time series (see, e.g. Oliveira et al., 2024; Xu and Xie, 2023a).

For concreteness, let us illustrate the coverage of split conformal on a toy example. Let $(W_j)_{j\in\mathbb{Z}}$ denote a collection of standard Gaussian variables, and for each $i \in [n+1]$, set $\epsilon_i = \sum_{j=i-t}^{i} W_j$ to be a moving average process of order t with unit coefficients; denote the joint distribution of $(\epsilon_i)_{i\in[n+1]}$ by $\mathsf{MA}(t; \mathbf{1})$. Suppose we have a time series of data $(X_i, Y_i)_{i\in[n+1]}$ generated from the standard regression model

$$Y_i = f(X_i) + \epsilon_i, \quad \text{where } (\epsilon_i)_{i\in[n+1]} \sim \mathsf{MA}(t; \mathbf{1}). \tag{3}$$

Now suppose that as a pretrained (and memoryless) predictor, we are given access to the true function f, and we use the absolute residual as the score function, i.e. $s(X, Y) = |Y - f(X)|$. With the goal of achieving coverage with probability at least $1 - \alpha$, we then output the pretrained prediction set (1); note that with our choice of score function, this set is an interval.

In Figure 1, we plot the coverage achieved by this prediction interval. Clearly, the prediction interval achieves the desired coverage if the MA process has order $t = 0$, in which case the process is iid, but for all other settings it suffers from a modest loss of coverage. Based on these plots, we might conjecture that the loss in coverage for split conformal prediction is proportional to t/n. But can we guarantee that the coverage loss is always bounded in this fashion? This paper will provide an affirmative answer to this question for a larger class of time series models, accommodating not just pretrained scores and memoryless predictors but also the split conformal approach (2) and predictors with memory, which we introduce next.

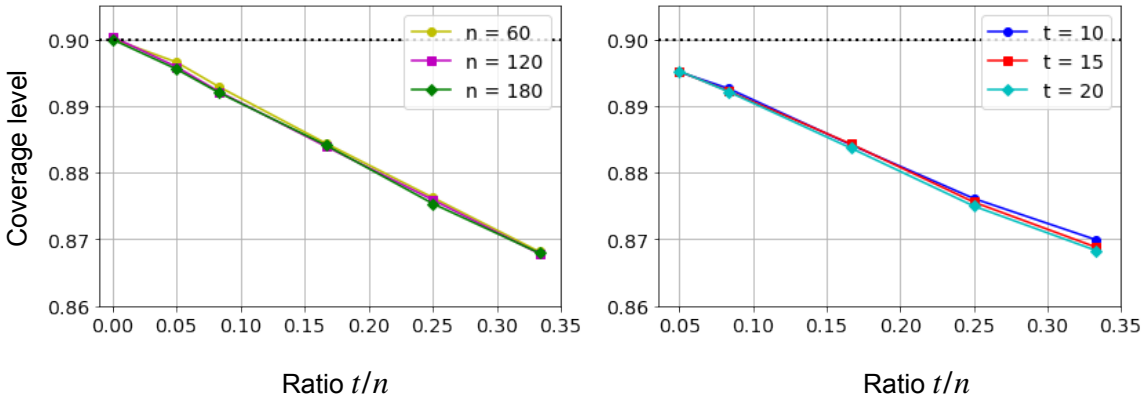

Figure 1: Coverage of the pretrained conformal prediction set (1) on a sequence of length n from the moving average process (3) of order t. The desired (i.e. nominal) coverage is 90% in both experiments, and is denoted by a dotted line. Each point is generated by averaging over 10^6 empirical trials.

1.3. Pretrained and split conformal for predictors with memory

Note that it is typical in time series models to use a prediction for response Y_i that does not only depend on the covariate X_i at time i, but also on the most recently observed L points. Indeed, equipping a predictor with memory is likely to be effective (i.e., to yield more accurate predictions) precisely when there are dependencies in the time series. In such cases, however, the score function can no longer be thought of as a map from $\mathcal{Z} \to \mathbb{R}$, since it is computed using a memory-L predictor. Instead, abusing notation slightly, the score function is now given by a higher dimensional map, $s : \mathcal{Z}^{L+1} \to \mathbb{R}$—for instance, if we have a predictive model $\widehat{f}(x; z_{-1}, \ldots, z_{-L})$ that predicts the response y given the current feature x in addition to the data from the preceding L time points, we might choose a residual score, $s(z; z_{-1}, \ldots, z_{-L}) = |y - \widehat{f}(x; z_{-1}, \ldots, z_{-L})|$, where $z = (x, y)$. The pretrained conformal prediction set is then given by

$$\begin{aligned}&\widehat{C}_n(X_{n+1}; Z_n, \ldots, Z_{n-L+1}) = \\ &\quad \left\{ y \in \mathcal{Y} : s((X_{n+1}, y); Z_n, \ldots, Z_{n-L+1}) \le \text{Quantile}_{(1-\alpha)(1+\frac{1}{n-L})}(S_{L+1}, \ldots, S_n) \right\}\end{aligned} \tag{4a}$$

where, for each $i = L+1, \ldots, n$,

$$S_i = s(Z_i; Z_{i-1}, \ldots, Z_{i-L}) \tag{4b}$$

is the score for prediction at time i using the previous L time points. In the case $L = 0$, this simply reduces back to the original construction (1). On the other hand, for $L \ge 1$, note that our calibration set only yields $n - L$ many scores $S_{L+1}, \ldots, S_n$, rather than n scores as before—this is because we cannot evaluate the conformity score for any data point at time $i \le L$, since we do not have L preceding time points available to make a prediction.

Analogously, the split conformal prediction set is given by

$$\begin{aligned}&\widehat{C}_n(X_{n+1}; Z_n, \ldots, Z_{n-L+1}) = \\ &\left\{ y \in \mathcal{Y} : s((X_{n+1}, y); Z_n, \ldots, Z_{n-L+1}) \le \text{Quantile}_{(1-\alpha)(1+\frac{1}{n_1-L})}(S_{n_0+L+1}, \ldots, S_n) \right\},\end{aligned} \tag{5}$$

where $n_1 = n - n_0$ and the trained score function is given by $s = \mathcal{A}(Z_1, \ldots, Z_{n_0})$ and where S_i is defined as in (4b) for each $i = n_0 + L + 1, \ldots, n$. Here again, we have abused notation in defining $\mathcal{A}$ to be a training algorithm that outputs a score function having memory L.

1.4. Related work

The conformal prediction literature is vast; we refer to the books (Vovk et al., 2005; Angelopoulos and Bates, 2023; Angelopoulos et al., 2024) for a comprehensive treatment of the broader literature, and focus this section only on theoretically grounded conformal prediction methods for time series.

Existing results explaining conformal prediction on time series. Since our focus is on explaining why split conformal is effective on time series data, we begin by surveying existing explanations for why conformal prediction methods more generally can be effective beyond exchangeability. Most of these explanations are based on defining explicit deviations from exchangeability (Barber and Tibshirani, 2025). For example, Barber et al. (2023) defined a measure motivated by settings with distribution shift—however, this measure of deviation from exchangeability can be large for time series, since it relies on the time series $\mathbf{Z}$ having approximately the same distribution if we swap the last data point with an earlier data point, $(Z_1, \ldots, Z_{k-1}, Z_{n+1}, Z_{k+1}, \ldots, Z_n, Z_k)$ (which, under strong short-term temporal dependence, might in fact substantially change the joint distribution). Other deviations from exchangeability include assumptions that the scores are strongly mixing (Xu and Xie, 2023a), but theoretical guarantees are only provided under the additional condition that the predictor is consistent. Note that we may not have consistent prediction in black-box settings, but would still like valid coverage. Closely related to our work is the recent paper by Oliveira et al. (2024), who also study split conformal prediction in time series. Among other results, they show using concentration inequalities for empirical processes that split conformal prediction incurs a loss of coverage on the order $(\mathsf{t}_{\mathsf{mix}}/n)^{1/2}$ for a β-mixing process with mixing time $\mathsf{t}_{\mathsf{mix}}$. While this shows that the coverage loss is asymptotically vanishing in n, it does not explain the type of behavior seen in Figure 1, where the loss of coverage appears to decay proportionally to $1/n$, and to increase linearly in the proxy t for the mixing time. In that sense, our results should be viewed as yielding sharper analogues of the results in Oliveira et al. (2024).

Modifying conformal methods for the time series setting. Moving beyond split conformal, other methods have been specifically developed for the time series setting (and more broadly for non-exchangeable settings). Notable examples are conformal prediction algorithms due to Chernozhukov et al. (2018, 2021), which rely on approximate block exchangeability of time series data and ensemble methods due to Xu and Xie (2023a), which are proven to work when we have a consistent predictor. Other methods are based on weighted versions of conformal prediction (Tibshirani et al., 2019; Fannjiang et al., 2022; Prinster et al., 2024), but these approaches involve correcting for a known distribution shift—information that is not typically available for most time series data. A final family of methods is derived from online learning (e.g., Gibbs and Candès, 2021, 2024), and views the construction of uncertainty sets as a game between nature and the statistician.

1.5. Contributions and organization

Our contributions can be summarized as follows:

- We introduce the notion of a *switch coefficient* for a dependent stochastic process, which measures the total variation distance when we swap certain subvectors of the time series. We

show that the switch coefficients can be bounded for β-mixing processes—and consequently, processes such as the one in the motivating example (3) are covered by our theory.

- We bound the coverage loss of pretrained conformal prediction by a function of the switch coefficient of the score process. For the MA process and its relatives, this result theoretically confirms the empirical observation made in Figure 1, and holds over a more general class of stochastic processes while accommodating predictors with memory. Moreover, we show that our characterization is tight over the class of stationary, β-mixing sequences.
- We extend these findings to split conformal prediction, showing that even here, the coverage loss is bounded by a related switch coefficient.

The rest of this paper is organized as follows. In Section 2, we introduce the switch coefficient of a stochastic process, and show how this relates to standard notions of mixing. Section 3 presents our main results for both pretrained and split conformal prediction. We conclude the main paper with a discussion in Section 4 and postpone our proofs to Appendix A.

2. Quantifying dependence in the time series

In this section, we examine the distribution of the time series of data points $\mathbf{Z} = (Z_1, \ldots, Z_{n+1})$, and define coefficients that measure the extent to which the data violates the exchangeability assumption due to temporal dependence.

2.1. The switch coefficients

To begin, we need to define notation for deleting a block of entries from a vector.

Definition 1 (The deletion operation) *Fix any $m \geq k \geq 1$, and any $\tau \in \{0, \ldots, m-1\}$. Let $\mathbf{w} = (w_1, \ldots, w_m)$ be a vector of length m (taking values in any space). We define $\Delta^0_{k,\tau}(\mathbf{w})$ and $\Delta^1_{k,\tau}(\mathbf{w})$, which are each subvectors of $\mathbf{w}$ obtained by deleting τ many entries, as follows. If $1 \leq k \leq m-1-\tau$, we define*

$$\Delta^0_{k,\tau}(\mathbf{w}) = (w_1, \ldots, w_{m-\tau-k}, w_{m-k+1}, \ldots, w_m),$$

which is the subvector consisting of the first $m-\tau-k$ *entries of* $\mathbf{w}$ *followed by the* last k *entries of* $\mathbf{w}$*, and is obtained by deleting a block of τ many entries after position $m-\tau-k$. Similarly, define*

$$\Delta^1_{k,\tau}(\mathbf{w}) = (w_{k+\tau+1}, \ldots, w_m, w_1, \ldots, w_k),$$

which is the subvector consisting of the last $m-\tau-k$ *entries of* $\mathbf{w}$ *followed by the* first k *entries of* $\mathbf{w}$*. If instead $m-\tau \leq k \leq m$, then we define*

$$\Delta^0_{k,\tau}(\mathbf{w}) = (w_{\tau+1}, \ldots, w_m) \text{ and } \Delta^1_{k,\tau}(\mathbf{w}) = (w_{k-m+\tau+1}, \ldots, w_k),$$

which each consist of $m-\tau$ consecutive entries of $\mathbf{w}$*.*

See Figures 2 and 3 for an illustration of these definitions. In particular, for every k, we note that $\Delta^0_{k,\tau}(\mathbf{w})$ is defined so that the last entry of $\mathbf{w}$ (i.e., w_m) is in the last position, while $\Delta^1_{k,\tau}(\mathbf{w})$ is defined so that w_k is in the last position.

$$\mathbf{w} = \big(\boxed{w_1 \, , w_2} \, , w_3 \, , w_4 \, , w_5 \, , w_6 \, , w_7 \, , \boxed{w_8 \, , w_9 \, , w_{10}} \big) \rightsquigarrow \Delta^0_{3,5}(\mathbf{w}) = \big(\boxed{w_1 \, , w_2} \, , \boxed{w_8 \, , w_9 \, , w_{10}} \big)$$
$$\mathbf{w} = \big(\boxed{w_1 \, , w_2 \, , w_3} \, , w_4 \, , w_5 \, , w_6 \, , w_7 \, , w_8 \, , \boxed{w_9 \, , w_{10}} \big) \rightsquigarrow \Delta^1_{3,5}(\mathbf{w}) = \big(\boxed{w_9 \, , w_{10}} \, , \boxed{w_1 \, , w_2 \, , w_3} \big)$$

Figure 2: Illustration of the definition of the subvectors $\Delta^0_{k,\tau}(\mathbf{w})$ (top) and $\Delta^1_{k,\tau}(\mathbf{w})$ (bottom), for a vector $\mathbf{w}$ of length $m = 10$, in the case $k = 3$, $\tau = 5$.

$$\mathbf{w} = \big(w_1 \, , w_2 \, , w_3 \, , w_4 \, , w_5 \, , \boxed{w_6 \, , w_7 \, , w_8 \, , w_9 \, , w_{10}} \big) \rightsquigarrow \Delta^0_{8,5}(\mathbf{w}) = \big(\boxed{w_6 \, , w_7 \, , w_8 \, , w_9 \, , w_{10}} \big)$$
$$\mathbf{w} = \big(w_1 \, , w_2 \, , w_3 \, , \boxed{w_4 \, , w_5 \, , w_6 \, , w_7 \, , w_8} \, , w_9 \, , w_{10} \big) \rightsquigarrow \Delta^1_{8,5}(\mathbf{w}) = \big(\boxed{w_4 \, , w_5 \, , w_6 \, , w_7 \, , w_8} \big)$$

Figure 3: Illustration of the definition of the subvectors $\Delta^0_{k,\tau}(\mathbf{w})$ (top) and $\Delta^1_{k,\tau}(\mathbf{w})$ (bottom), for a vector $\mathbf{w}$ of length $m = 10$, in the case $k = 8$, $\tau = 5$.

In the results developed in this paper, in order to quantify the extent to which a time series $\mathbf{Z} \in \mathcal{Z}^{n+1}$ fails to satisfy the exchangeability assumption, we will be comparing the distributions of the subvectors $\Delta^0_{k,\tau}(\mathbf{Z})$ and $\Delta^1_{k,\tau}(\mathbf{Z})$. Indeed, in the simple case where the data values Z_i are exchangeable, these subvectors have the same distribution. For instance, if $Z_1, \ldots, Z_{n+1} \overset{\text{iid}}{\sim} P$ for some distribution P, then both have the same distribution, $P^{n+1-\tau}$. In a time series setting, however, the distributions of these subvectors may differ. The following definition establishes the *switch coefficients*, which compares the distributions of these subvectors—and, as we will see later, characterizes the performance guarantees of split conformal prediction in the time series setting.

Definition 2 (The switch coefficients) *Let $n \geq 1$, and let $\mathbf{Z} \in \mathcal{Z}^{n+1}$ be a time series. For each $k \in [n+1]$, define*

$$\Psi_{k,\tau}(\mathbf{Z}) = \mathrm{d_{TV}}\big(\Delta^0_{k,\tau}(\mathbf{Z}), \Delta^1_{k,\tau}(\mathbf{Z})\big),$$

where $\mathrm{d_{TV}}$ denotes the total variation distance, and define

$$\bar{\Psi}_\tau(\mathbf{Z}) = \frac{1}{n+1} \sum_{k=1}^{n+1} \Psi_{k,\tau}(\mathbf{Z}).$$

Note that while $\Delta^0_{k,\tau}(\mathbf{Z})$ and $\Delta^1_{k,\tau}(\mathbf{Z})$ are random variables (they each consist of entries of the time series $\mathbf{Z}$), the switch coefficient $\Psi_{k,\tau}(\mathbf{Z})$ is instead a fixed quantity—it is a function of the distribution of $\mathbf{Z}$, rather than the random variable $\mathbf{Z}$ itself.

In many practical settings, we might hope that the switch coefficient $\bar{\Psi}_\tau(\mathbf{Z})$ will be small as long as τ is sufficiently large—that is, while dependence might be strong between consecutive time points, it is plausible that dependence could be relatively weak over a time gap of length $\geq \tau$.

2.2. Connection to mixing coefficients

While the switch coefficients are different than the usual conditions appearing in the time series literature, it is straightforward to relate them to a standard mixing condition. Specifically, for a time

series $\mathbf{Z} \in \mathcal{Z}^{n+1}$, the β-mixing coefficient with lag τ is defined as follows (Doukhan, 1994):

$$\beta(\tau) := \max_{1 \le k \le n-\tau} \mathrm{d_{TV}}\big((Z_1, \ldots, Z_k, Z_{k+\tau+1}, \ldots, Z_{n+1}), (Z_1, \ldots, Z_k, Z'_{k+\tau+1}, \ldots, Z'_{n+1})\big),$$

where $\mathbf{Z}' = (Z'_1, \ldots, Z'_{n+1}) \in \mathcal{Z}^{n+1}$ denotes an iid copy of $\mathbf{Z}$. In other words, if $\beta(\tau)$ is small, this means that the subvectors $(Z_1, \ldots, Z_k)$ and $(Z_{k+\tau+1}, \ldots, Z_{n+1})$ are approximately independent. Note that the deletion operation (Definition 1), is closely connected to β-mixing: the β-mixing condition is defined in terms of measuring dependence between $(Z_1, \ldots, Z_k)$ and $(Z_{k+\tau+1}, \ldots, Z_{n+1})$, which involves deleting τ consecutive data points in the time series.

Next we relate β-mixing coefficients to the switch coefficients defined above.

Proposition 3 *Suppose $\mathbf{Z} \in \mathcal{Z}^{n+1}$ is a stationary time series, with β-mixing coefficient $\beta(\tau)$. Then we have the following bound on the switch coefficients of $\mathbf{Z}$:*

$$\begin{cases} \Psi_{k,\tau}(\mathbf{Z}) \le 2\beta(\tau), & \text{for } 1 \le k \le n-\tau, \\ \Psi_{k,\tau}(\mathbf{Z}) = 0, & \text{for } n-\tau < k \le n+1. \end{cases}$$

We prove Proposition 3 in Section A.5. This result guarantees that any time series with small β-mixing coefficients must also have small switch coefficients. However, the converse is not true: in particular, as mentioned above, any exchangeable distribution on $\mathbf{Z}$ ensures $\Psi_{k,\tau}(\mathbf{Z}) = 0$ for all k, τ; however, $\beta(\tau)$ may be large for data that is exchangeable but not iid.

2.3. Switching data points, or switching scores?

Suppose we are working with a pretrained score function s. Since the prediction set $\widehat{C}_n$ depends on the data points only through their scores, we may ask whether the time series of scores is approximately exchangeable. How does this question relate to the properties of the data time series $\mathbf{Z}$?

First, consider the simple case $L = 0$, with memoryless prediction. Write $\mathbf{S} = (S_1, \ldots, S_{n+1})$ where $S_i = s(Z_i)$ for each $i \in [n+1]$. Since each score S_i is computed as a function of the corresponding data point Z_i, it follows by the data processing inequality (see, e.g., Polyanskiy and Wu, 2025, Chapter 7) that

$$\Psi_{k,\tau}(\mathbf{S}) = \mathrm{d_{TV}}(\Delta^0_{k,\tau}(\mathbf{S}), \Delta^1_{k,\tau}(\mathbf{S})) \le \mathrm{d_{TV}}(\Delta^0_{k,\tau}(\mathbf{Z}), \Delta^1_{k,\tau}(\mathbf{Z})) = \Psi_{k,\tau}(\mathbf{Z}).$$

Consequently

$$\bar{\Psi}_\tau(\mathbf{S}) \le \bar{\Psi}_\tau(\mathbf{Z}).$$

In other words, the deviation from exchangeability among the scores, as measured by the averaged switch coefficient $\bar{\Psi}_\tau(\mathbf{S})$, cannot be higher than the deviation from exchangeability within the time series of data points. Note that in general, it is likely that there is much more dependence among the potentially high-dimensional data points Z_i than among their scores, which are one-dimensional and capture only a limited amount of the information contained in the data. Consequently, in practice $\bar{\Psi}_\tau(\mathbf{S})$ could be significantly smaller than $\bar{\Psi}_\tau(\mathbf{Z})$.

In contrast, in the general case with memory $L \ge 0$, the situation is somewhat more complicated. For example, even if the data points Z_i are exchangeable, the scores are *not* exchangeable when the memory L is positive, and indeed may have strong temporal dependence. In particular, writing $\mathbf{S} = (S_{L+1}, \ldots, S_{n+1})$ where $S_i = s(Z_i; Z_{i-1}, \ldots, Z_{i-L})$, we may have $\bar{\Psi}_\tau(\mathbf{Z}) = 0$ but $\bar{\Psi}_\tau(\mathbf{S}) > 0$, unlike in the memoryless case. Nonetheless, we can still relate the switch coefficients of the scores to those of the data, as shown in the following proposition.

Proposition 4 *Let $s : \mathcal{Z}^{L+1} \to \mathbb{R}$ be a pretrained score function with memory $L \geq 0$, and let $\mathbf{Z} \in \mathcal{Z}^{n+1}$ and $\mathbf{S} \in \mathbb{R}^{n-L+1}$ be defined as above. Then*

$$\Psi_{k,\tau}(\mathbf{S}) \leq \Psi_{k+L,\tau-L}(\mathbf{Z})$$

for all $k \in [n-L+1]$ and $\tau \in \{L, \ldots, n-L\}$, and consequently,

$$\bar{\Psi}_\tau(\mathbf{S}) \leq \frac{n+1}{n-L+1}\bar{\Psi}_{\tau-L}(\mathbf{Z}).$$

We prove this proposition in Section A.6. Of course, in the memoryless case ($L = 0$), it reduces to the above bounds $\Psi_{k,\tau}(\mathbf{S}) \leq \Psi_{k,\tau}(\mathbf{Z})$ and $\bar{\Psi}_\tau(\mathbf{S}) \leq \bar{\Psi}_\tau(\mathbf{Z})$.

3. Main results

In this section we will present our main results on the coverage properties of conformal prediction in the time series setting. We will begin by analyzing the setting of a pretrained score function s, with the main coverage guarantee presented in Section 3.1, and with some related results explored in Sections 3.2 and 3.3. Then, in Section 3.4, we will adapt our coverage guarantee to handle the split conformal setting, where the score function s is trained on a portion of the data. In both cases, our results allow for a memory window of any length $L \geq 0$.

3.1. Coverage guarantee for the pretrained setting

We begin by considering pretrained conformal prediction, i.e., the prediction set defined in (4). The following theorem shows that this prediction set cannot undercover if the switch coefficients of the scores are small.

Theorem 5 *Let $\mathbf{Z} \in \mathcal{Z}^{n+1}$ be a time series of data points, and let $s : \mathcal{Z}^{L+1} \to \mathbb{R}$ be a pretrained score function with memory L, for some $n \geq L \geq 0$. Then the prediction set $\widehat{C}_n$ defined in (4) satisfies*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}; Z_n, \ldots, Z_{n-L+1})\right\} \geq 1 - \alpha - \min_{\tau \in \{0,\ldots,n-L\}} \left\{\frac{\tau}{n-L+1} + \bar{\Psi}_\tau(\mathbf{S})\right\},$$

where $\mathbf{S} = (S_{L+1}, \ldots, S_{n+1})$, for $S_i = s(Z_i; Z_{i-1}, \ldots, Z_{i-L})$.

Theorem 5 is proved in Section A.1. While Theorem 5 is stated in terms of the switch coefficients of the scores, combining this result with Propositions 3 and 4 immediately yields the following corollary, which characterizes the coverage in terms of the properties of the time series $\mathbf{Z}$.

Corollary 6 *In the setting of Theorem 5, it holds that*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}; Z_n, \ldots, Z_{n-L+1})\right\} \geq 1 - \alpha - \min_{\tau \in \{0,\ldots,n-2L\}} \left\{\frac{\tau+L}{n-L+1} + \frac{n+1}{n-L+1} \cdot \bar{\Psi}_\tau(\mathbf{Z})\right\}.$$

Moreover, if we also assume that $\mathbf{Z}$ is stationary and has β-mixing coefficients $\beta(\tau)$, then

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}; Z_n, \ldots, Z_{n-L+1})\right\} \geq 1 - \alpha - \min_{\tau \in \{0,\ldots,n-2L\}} \left\{\frac{\tau+L}{n-L+1} + 2\beta(\tau)\right\}. \quad (6)$$

At a high level, we can interpret these results as guaranteeing that if the memory satisfies $L \ll n$ and temporal dependence is weak for some $\tau \ll n$, then the prediction set is guaranteed to have coverage at nearly the nominal level $1-\alpha$. We emphasize that this result does not require any modifications to the conformal prediction method; it simply explains why the method might perform reasonably well even when substantial temporal dependence is present, as illustrated in Figure 1.

In the special case of iid data, the minimum is achieved for $\tau = 0$ since $\beta(0) = 0$. We thus recover the marginal coverage guarantee $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha - \frac{L}{n-L+1}$, and in particular for the memoryless case, coverage is at least $1-\alpha$. In a similar fashion, one can recover the standard conformal guarantee for exchangeable data in the memoryless ($L = 0$) setting: setting $\tau = 0$ and noting that $\Psi_{k,\tau}(\mathbf{Z}) = 0$ for all $k \in [n+1]$, here again we obtain the familiar guarantee $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha$.

To compare with existing results, we begin by noting that standard results for conformal prediction (Shafer and Vovk, 2008; Lei et al., 2018; Angelopoulos et al., 2024) do not allow for memory-based predictors even when the process $\mathbf{Z}$ is exchangeable, since memory renders the score process $\mathbf{S}$ non-exchangeable. Thus, there is no analogue of Theorem 5 in the classical literature on pretrained conformal prediction. Among existing results for pretrained conformal prediction in time series settings, the closest to ours are those of Oliveira et al. (2024, Theorem 4), who show that if the pretrained predictor is memoryless, then its coverage loss on a β-mixing process is bounded on the order[2] $\min_\tau\{\sqrt{\tau/n} + 2\beta(\tau)\}$, up to logarithmic factors. Comparing with Corollary 6 above, note that we replace the first term with the "fast rate" τ/n, leading to a stronger guarantee. Concretely, our improvement is obtained by eschewing arguments based on empirical processes and blocking techniques (Yu, 1994; Mohri and Rostamizadeh, 2010) and instead introducing a new technique that exploits the stability of the quantile function upon adding and deleting score values.

3.2. A matching lower bound

Our main results provide a guarantee that the loss of coverage, as compared to the nominal level $1-\alpha$, can be bounded by the switch coefficients of the scores—and in turn, can therefore be bounded by the β-mixing coefficients of the time series, as in (6). A natural question in light of the comparison with prior work given above is whether our bound on the loss of coverage is tight. In the following result, we provide a matching lower bound; for simplicity, we will work in the memoryless setting ($L = 0$), and will assume $(1-\alpha)(n+1)$ is an integer.

Theorem 7 *Fix any $\alpha \in (0,1)$, data space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and sample size $n \geq 1$, where $(1-\alpha)(n+1)$ is an integer. For any constant $b \in [0,1]$, there exists a stationary time series $\mathbf{Z} \in \mathcal{Z}^{n+1}$ and a pretrained score function $s : \mathcal{Z} \to \mathbb{R}$, for which it holds that*

$$\min_{\tau \in \{0,\ldots,n\}} \left\{ \frac{\tau}{n+1} + 2\beta(\tau) \right\} \leq b,$$

and the prediction set $\widehat{C}_n$ defined in (1) satisfies

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \leq \left(1 - \frac{b}{4}\right) \cdot (1-\alpha) + \frac{n(n+1)}{2|\mathcal{Z}|}.$$

2. Note that the result of Oliveira et al. (2024) is stated with more parameters, but here we have stated a simplified corollary of their result for the pretrained setting, emphasizing dependence on the pair (τ, n).

Theorem 7 is proved in Section A.2. In particular, if $|\mathcal{Z}| = \infty$ (i.e., at least one of the spaces $\mathcal{X}$ and $\mathcal{Y}$ has infinite cardinality), then we obtain the upper bound

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \leq (1-\alpha) - \frac{1-\alpha}{4} \cdot b \leq (1-\alpha) - \frac{1-\alpha}{4} \cdot \min_{\tau \in \{0,\ldots,n\}} \left\{\frac{\tau}{n+1} + 2\beta(\tau)\right\}.$$

This implies that the coverage gap in (6) (and hence, the guarantee given in Theorem 5) is tight up to a factor $\frac{1-\alpha}{4}$. Since it is typical to take $\alpha \leq 1/2$, this factor should be viewed as a universal constant in the range $[1/8, 1]$.

3.3. Can the conformal prediction set overcover?

Our results above prove that the switch coefficients of $\mathbf{S}$ (and consequently, the β-mixing coefficients of $\mathbf{Z}$) can be used to bound the loss of coverage of the conformal prediction set—and, moreover, these bounds are tight up to a constant, meaning that there exist settings for which the loss of coverage can indeed be this large. But is it possible that, in other settings, the conformal prediction set can *overcover* rather than undercover? That is, in the time series setting, might conformal prediction lead to sets that are too conservative?

We will now see that the switch coefficients can also be used to provide an upper bound on the coverage probability, to guarantee that the conformal prediction set is not overly conservative.

Theorem 8 *In the setting of Theorem 5, assume also that the scores $S_{L+1}, \ldots, S_{n+1}$ are distinct almost surely. Then*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}; Z_n, \ldots, Z_{n-L+1})\right\} \leq \frac{\lceil (1-\alpha)(n-L+1) \rceil}{n-L+1} + \min_{\tau \in \{0,\ldots,n-L\}} \left\{\frac{\tau}{n-L+1} + \bar{\Psi}_\tau(\mathbf{S})\right\}.$$

Theorem 8 is proved in Section A.3. We note that as a corollary, upper bounds in terms of the properties of the time series $\mathbf{Z}$ (analogous to Corollary 6) also follow in this case. While the coverage upper bound in Theorem 8 may be conservative for certain settings, it is possible—analogously to Theorem 7—to construct an example that essentially achieves this upper bound.

3.4. Coverage guarantee for the split conformal setting

Next, we turn to split conformal prediction, where the score function s is now trained on a portion of the available data, as in (5). Throughout, we will assume that the sample size is split as $n = n_0 + n_1$, where $n_0, n_1 \geq 1$ and $n_1 \geq L$. Write $\mathbf{S} = (S_{n_0+L+1}, \ldots, S_{n+1})$, the vector of scores on the calibration set together with the test point score, $S_{n+1} = s(Z_{n+1}; Z_n, \ldots, Z_{n-L+1})$. Define also

$$\mathbf{S}_{\text{split},\tau_*} = (S_{n_0+L+\tau_*+1}, \ldots, S_{n+1}),$$

which deletes the first τ_* scores for some $\tau_* \geq 0$. The motivation for working with this subvector is that, by deleting the first τ_* scores, we have removed those scores that may have high dependence with $Z_1, \ldots, Z_{n_0}$ (and thus, may have high dependence with the trained score function s). Now we state our main result for coverage in this setting.

Theorem 9 *Consider the split conformal setting, with the first n_0 data points used for training the score function and the remaining $n_1 = n - n_0$ points used for calibration. Then the prediction set $\widehat{C}_n$ defined in (5) satisfies*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}; Z_n, \dots, Z_{n-L+1})\right\} \geq 1 - \alpha - \min_{\substack{\tau, \tau_* \geq 0 \\ \tau + \tau_* \leq n_1 - L}} \left\{\frac{\tau + \alpha\tau_*}{n_1 - \tau_* - L + 1} + \bar{\Psi}_\tau(\mathbf{S}_{\text{split},\tau_*})\right\}.$$

Theorem 9 is proved in Section A.4. One might ask why we need to work with $\mathbf{S}_{\text{split},\tau_*}$, rather than $\mathbf{S}$. Indeed, by choosing $\tau_* = 0$, we simply have $\mathbf{S}_{\text{split},\tau_*} = \mathbf{S}$, and this result yields

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}; Z_n, \dots, Z_{n-L+1})\right\} \geq 1 - \alpha - \min_{\tau \in \{0,\dots,n_1-L\}} \left\{\frac{\tau}{n_1 - L + 1} + \bar{\Psi}_\tau(\mathbf{S})\right\},$$

which is identical to the bound established in Theorem 5 for the pretrained setting except with n_1 in place of n. But, importantly, in the setting of split conformal, this result is no longer meaningful. This is because $\bar{\Psi}_\tau(\mathbf{S})$ may be large in the time series setting, for any choice of τ. For example, taking the memoryless case $L = 0$ for simplicity, for any $k \leq n_1 - \tau$ we have

$$\Psi_{k,\tau}(\mathbf{S}) = \mathrm{d_{TV}}\big(\Delta^0_{k,\tau}(\mathbf{S}), \Delta^1_{k,\tau}(\mathbf{S})\big) \geq \mathrm{d_{TV}}(S_{n_0+1}, S_{n_0+k+\tau+1}) = \mathrm{d_{TV}}(s(Z_{n_0+1}), s(Z_{n_0+k+\tau+1})),$$

where the inequality holds since S_{n_0+1} is the first entry of $\Delta^0_{k,\tau}(\mathbf{S})$ while $S_{n_0+k+\tau+1}$ is the first entry of $\Delta^1_{k,\tau}(\mathbf{S})$. Since the data point Z_{n_0+1} comes immediately after the data $Z_1, \dots, Z_{n_0}$ used for training s, it may be the case that s has higher dependence with Z_{n_0+1} than with a data point $Z_{n_0+k+\tau+1}$ appearing much later in time—and therefore, the total variation distance between these two data points' scores might be large.

To further explore this point, we will now see how this result can be connected to the β-mixing coefficients of the time series $\mathbf{Z}$. This next result is the analogue of Propositions 3 and 4, modified for the split conformal setting.

Proposition 10 *In the setting of Theorem 9, assume also that $\mathbf{Z}$ is a stationary time series with β-mixing coefficients $\beta(\tau)$. Then for each k, τ, τ_* with $\tau_* \geq 0$, $L \leq \tau \leq n_1 - \tau_*$, and $1 \leq k \leq n_1 - L + 1 - \tau_*$, it holds that*

$$\Psi_{k,\tau}(\mathbf{S}_{\text{split},\tau_*}) \leq \begin{cases} 2\beta(\tau_*) + 2\beta(\tau - L), & \text{for } 1 \leq k \leq n_1 - \tau - \tau_*, \\ 2\beta(\tau_*), & \text{for } n_1 - \tau - \tau_* < k \leq n_1 - L + 1 - \tau_*. \end{cases}$$

We prove Proposition 10 in Section A.7. As we will see in the proof, the key step is to bound $\Psi_{k,\tau}(\mathbf{S}_{\text{split},\tau_*})$ using total variation distances of certain subvectors of $\mathbf{Z}$ (a more complex form of the switch coefficient). Combining this result with Theorem 9, we obtain the following corollary.

Corollary 11 *In the setting of Theorem 9, if $\mathbf{Z}$ is stationary and has β-mixing coefficients $\beta(\tau)$, then*

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}; Z_n, \dots, Z_{n-L+1})\right\} \geq 1 - \alpha - \min_{\substack{\tau, \tau_* \geq 0 \\ \tau + \tau_* \leq n_1 - 2L}} \left\{\frac{\tau + \alpha\tau_* + L}{n_1 - \tau_* - L + 1} + 2\beta(\tau) + 2\beta(\tau_*)\right\}.$$

For this result to give a meaningful coverage guarantee in the presence of temporal dependence, we see that we need both τ and τ_* to be sufficiently large, so that dependence (as captured by the β-mixing coefficients) is low.

Let us compare again with the result of Oliveira et al. (2024, Theorem 4), who show that if the score function is memoryless, then its coverage loss for split conformal prediction on a β-mixing process is bounded (in our notation and up to logarithmic factors) by a term of the order $\min_{\tau,\tau_*}\{\sqrt{\tau/n} + \sqrt{\tau_*/n} + 2\beta(\tau) + 2\beta(\tau_*)\}$. As before, comparing with Corollary 11 above, note that our bound on the coverage loss is tighter, scaling linearly in τ/n and τ_*/n. Once again, this improvement is a consequence of our new proof technique.

4. Discussion

Motivated by the question of why pretrained and split conformal prediction are effective in spite of temporal dependence, we introduced a new "switch coefficient" to measure the deviation of scores from exchangeability, and showed that the loss of coverage is bounded whenever the score process has small switch coefficient. This covers the class of β-mixing processes, and improves upon previous characterizations of the coverage loss. We also showed that our characterization of the coverage loss is tight, and can accurately reflect empirically observed behavior in canonical time series models. An important open question is whether these bounds on under- and over-coverage are optimal for *any* predictive method—that is, whether split conformal is optimal for this class of problems, or whether alternative methods (that may use knowledge of the mixing properties of the time series) could improve these bounds.

We believe that our definitions and proof techniques can find broader applications to other conformal prediction methods. In particular, we expect that the switch coefficient of a process can characterize the coverage loss of other methods when applied to time series data. It is also a natural object in its own right, worth studying for general stochastic processes. Our proof technique, which exploits the stability of the quantile function to the addition or deletion of score values, may also lead to a sharp analysis of other conformal prediction methods. It offers an alternative to blocking techniques (Yu, 1994), which have seen extensive use in analyzing many statistical estimation and inference methods (beyond uncertainty quantification) in other dynamic settings (Mohri and Rostamizadeh, 2010; Yang et al., 2017; Mou et al., 2024; Nakul et al., 2025).

Acknowledgments

R.F.B. was partially supported by the National Science Foundation via grant DMS-2023109, and by the Office of Naval Research via grant N00014-24-1-2544. A.P. was supported in part by the National Science Foundation through grants CCF-2107455 and DMS-2210734, and by research awards from Adobe, Amazon, Mathworks and Google. The authors thank Hanyang Jiang, Ryan Tibshirani, and Yao Xie for helpful feedback.

References

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

Rina Foygel Barber and Ryan J Tibshirani. Unifying different theories of conformal prediction. *arXiv preprint arXiv:2504.02292*, 2025.

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.

Anastasia Borovykh, Sander Bohte, and Cornelis W. Oosterlee. Conditional time series forecasting with convolutional neural networks. In *International Conference on Artificial Neural Networks, ICANN 2017*. Springer, 2017.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

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

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

Jaquelin Cochran, Paul Denholm, Bethany Speer, and Mackay Miller. Grid integration and the carrying capacity of the US grid to incorporate variable renewable energy. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2015.

Paul Doukhan. *Mixing: Properties and examples*, volume 85 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1994. ISBN 0-387-94214-9. doi: 10.1007/978-1-4612-2642-0. URL https://doi.org/10.1007/978-1-4612-2642-0.

Veronika Eyring, William D Collins, Pierre Gentine, Elizabeth A Barnes, Marcelo Barreiro, Tom Beucler, Marc Bocquet, Christopher S Bretherton, Hannah M Christensen, and Katherine Dagon. Pushing the frontiers in climate modelling and analysis with machine learning. *Nature Climate Change*, 14(9):916–928, 2024.

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

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

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.

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.

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.

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

Wenlong Mou, Ashwin Pananjady, Martin J Wainwright, and Peter L Bartlett. Optimal and instance-dependent guarantees for Markovian linear stochastic approximation. *Mathematical Statistics and Learning*, 7(1):41–153, 2024.

Milind Nakul, Vidya Muthukumar, and Ashwin Pananjady. Estimating stationary mass, frequency by frequency. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 4359–4359. PMLR, 30 Jun–04 Jul 2025.

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

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.

Drew Prinster, Samuel Don Stanton, Anqi Liu, and Suchi Saria. Conformal validity guarantees exist for any data distribution (and how to find them). In *International Conference on Machine Learning*, pages 41086–41118. PMLR, 2024.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3): 1181–1191, 2020.

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

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

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

Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. A multi-horizon quantile recurrent forecaster. *arXiv preprint arXiv:1711.11053*, 2017.

Chen Xu and Yao Xie. Conformal prediction for time series. *IEEE transactions on pattern analysis and machine intelligence*, 45(10):11575–11587, 2023a.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, 2023b.

Fanny Yang, Sivaraman Balakrishnan, and Martin J Wainwright. Statistical and computational guarantees for the Baum–Welch algorithm. *Journal of Machine Learning Research*, 18(125): 1–53, 2017.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Appendix A. Proofs

We prove our four main theorems in the first four subsections of this appendix. Proofs of propositions and lemmas can be found in the later subsections.

A.1. Proof of Theorem 5

By definition of the prediction set (4), the coverage event $Y_{n+1} \in \widehat{C}_n(X_{n+1}; Z_n, \ldots, Z_{n-L+1})$ holds if and only if

$$S_{n+1} \leq \text{Quantile}_{(1-\alpha)(1+\frac{1}{n-L})}(S_{L+1}, \ldots, S_n).$$

By properties of the quantile of a finite list (see, e.g., Angelopoulos et al. (2024, Lemma 3.4)), this event can equivalently be written as

$$S_{n+1} \leq \text{Quantile}_{1-\alpha}(\mathbf{S}).$$

Now fix any $\tau \in \{0, \ldots, n-L\}$. Below, we will show that for each $i \in \{L+1, \ldots, n+1\}$, it holds that

$$\mathbb{P}\left\{S_{n+1} \leq \text{Quantile}_{1-\alpha}(\mathbf{S})\right\} \geq \mathbb{P}\left\{S_i \leq \text{Quantile}_{1-\alpha-\frac{\tau}{n-L+1}}(\mathbf{S})\right\} - \Psi_{i-L,\tau}(\mathbf{S}). \tag{7}$$

Assuming for the moment that this is true, we then calculate

$$\begin{aligned}
&\mathbb{P}\left\{S_{n+1} \leq \text{Quantile}_{1-\alpha}(\mathbf{S})\right\} \\
&\geq \frac{1}{n-L+1}\sum_{i=L+1}^{n+1}\left[\mathbb{P}\left\{S_i \leq \text{Quantile}_{1-\alpha-\frac{\tau}{n-L+1}}(\mathbf{S})\right\} - \Psi_{i-L,\tau}(\mathbf{S})\right] \\
&= \mathbb{E}\left[\frac{1}{n-L+1}\sum_{i=L+1}^{n+1}\mathbb{1}\left\{S_i \leq \text{Quantile}_{1-\alpha-\frac{\tau}{n-L+1}}(\mathbf{S})\right\}\right] - \frac{1}{n-L+1}\sum_{i=L+1}^{n+1}\Psi_{i-L,\tau}(\mathbf{S}) \\
&\overset{(i)}{\geq} \left(1-\alpha-\frac{\tau}{n-L+1}\right) - \frac{1}{n-L+1}\sum_{i=L+1}^{n+1}\Psi_{i-L,\tau}(\mathbf{S}) \\
&= \left(1-\alpha-\frac{\tau}{n-L+1}\right) - \bar{\Psi}_\tau(\mathbf{S}),
\end{aligned}$$

where step (i) holds since, for any vector $\mathbf{w} = (w_1, \ldots, w_m) \in \mathbb{R}^m$ and any $a \in [0,1]$, it must hold that $\frac{1}{m}\sum_{i=1}^m \mathbb{1}\left\{w_i \leq \text{Quantile}_{1-a}(\mathbf{w})\right\} \geq 1-a$, by definition of the quantile. Therefore, we have proved the desired lower bound on coverage.

It remains to be shown that (7) holds, for all i. For every $k \in [n-L+1]$, since $\Delta^0_{k,\tau}(\mathbf{S})$ and $\Delta^1_{k,\tau}(\mathbf{S})$ are each subvectors of $\mathbf{S} \in \mathbb{R}^{n-L+1}$, obtained by deleting exactly τ many entries, it holds surely that

$$\text{Quantile}_{(1-a)\cdot\frac{n-L+1-\tau}{n-L+1}}(\mathbf{S}) \leq \text{Quantile}_{1-a}\left(\Delta^j_{k,\tau}(\mathbf{S})\right) \leq \text{Quantile}_{1-a\cdot\frac{n-L+1-\tau}{n-L+1}}(\mathbf{S}), \tag{8}$$

for each $j = 0, 1$, by definition of the quantile. (Recall that we interpret $\mathrm{Quantile}_t(\mathbf{w})$ as ∞ if $t > 1$.) In other words, the quantile function is stable to insertion and deletion. Therefore, for any k, we may lower bound the probability of coverage as

$$\begin{aligned}
&\mathbb{P}\left\{S_{n+1} \le \mathrm{Quantile}_{1-\alpha}(\mathbf{S})\right\} \\
&\overset{(i)}{\ge} \mathbb{P}\left\{S_{n+1} \le \mathrm{Quantile}_{1-\alpha\cdot\frac{n-L+1}{n-L+1-\tau}}\left(\Delta^0_{k,\tau}(\mathbf{S})\right)\right\} \\
&\overset{(ii)}{\ge} \mathbb{P}\left\{S_{L+k} \le \mathrm{Quantile}_{1-\alpha\cdot\frac{n-L+1}{n-L+1-\tau}}\left(\Delta^1_{k,\tau}(\mathbf{S})\right)\right\} - \mathrm{d_{TV}}\left(\Delta^0_{k,\tau}(\mathbf{S}), \Delta^1_{k,\tau}(\mathbf{S})\right) \\
&\overset{(iii)}{\ge} \mathbb{P}\left\{S_{L+k} \le \mathrm{Quantile}_{1-\alpha-\frac{\tau}{n-L+1}}(\mathbf{S})\right\} - \mathrm{d_{TV}}\left(\Delta^0_{k,\tau}(\mathbf{S}), \Delta^1_{k,\tau}(\mathbf{S})\right).
\end{aligned}$$

Here, steps (i) and (iii) apply (8), while for step (ii), we use the fact that S_{n+1} is the last entry of $\Delta^0_{k,\tau}(\mathbf{S})$ while S_{L+k} is the last entry of $\Delta^1_{k,\tau}(\mathbf{S})$. Concretely, both expressions are calculating the probability of the same event (that the last entry is no larger than the quantile), for $\Delta^0_{k,\tau}(\mathbf{S})$ and for $\Delta^1_{k,\tau}(\mathbf{S})$, which are in turn close in total variation. Finally, taking $k = i - L$, we have verified (7) since $\mathrm{d_{TV}}\left(\Delta^0_{k,\tau}(\mathbf{S}), \Delta^1_{k,\tau}(\mathbf{S})\right) = \Psi_{k,\tau}(\mathbf{S}) = \Psi_{i-L,\tau}(\mathbf{S})$.

A.2. Proof of Theorem 7

Choose a positive integer $K \le |\mathcal{Z}|$, and let $z_0, \ldots, z_{K-1}$ be distinct points in $\mathcal{X} \times \mathcal{Y}$. We first define two distributions:

- Let P_{cyclic} be a distribution on $\mathcal{Z}^{n+1}$, defined as follows. Sample $J_1 \sim \mathrm{Unif}(\{0, \ldots, K-1\})$, and let $J_{i+1} = (J_i + 1) \mod K$, for each $i = 1, \ldots, n$, then return the sequence $(z_{J_1}, \ldots, z_{J_{n+1}})$.
- Let Q denote the uniform distribution on $\{z_0, \ldots, z_{K-1}\}$.

Now we define our distribution on the time series $\mathbf{Z} \in \mathcal{Z}^{n+1}$. We draw from the mixture distribution

$$\mathbf{Z} \sim \frac{b}{4} \cdot P_{\text{cyclic}} + \left(1 - \frac{b}{4}\right) \cdot Q^{n+1}.$$

In words, we sample $\mathbf{Z}$ from P_{cyclic} with probability $b/4$; otherwise, we sample each of the $n+1$ data points independently and uniformly at random from the set $\{z_0, \ldots, z_{K-1}\}$.

First, observe that this distribution is stationary by construction. Next, for any $\tau \ge 0$, we bound the β-mixing coefficient. Fix any $k \in [n - \tau]$, and as usual let $\mathbf{Z}'$ denote an iid copy of $\mathbf{Z}$. Let P_0, P_1, P_2 denote the marginal distribution of the subvectors $(Z_1, \ldots, Z_k, Z_{k+\tau+1}, \ldots, Z_{n+1})$, $(Z_1, \ldots, Z_k)$, and $(Z_{k+\tau+1}, \ldots, Z_{n+1})$, respectively, under the joint distribution $\mathbf{Z} \sim P_{\text{cyclic}}$. Then, we have

$$\begin{aligned}
(Z_1, \ldots, Z_k, Z_{k+\tau+1}, \ldots, Z_{n+1}) &\sim \frac{b}{4} \cdot P_0 + \left(1 - \frac{b}{4}\right) \cdot Q^{n+1-\tau}, \\
(Z_1, \ldots, Z_k) &\sim \frac{b}{4} \cdot P_1 + \left(1 - \frac{b}{4}\right) \cdot Q^{k}, \text{ and} \\
(Z'_{k+\tau+1}, \ldots, Z'_{n+1}) &\sim \frac{b}{4} \cdot P_2 + \left(1 - \frac{b}{4}\right) \cdot Q^{n+1-\tau-k}.
\end{aligned}$$

Therefore,

$$(Z_1, \dots, Z_k, Z'_{k+\tau+1}, \dots, Z'_{n+1}) \sim \left(\frac{b}{4} \cdot P_1 + \left(1 - \frac{b}{4}\right) \cdot Q^k\right) \times \left(\frac{b}{4} \cdot P_2 + \left(1 - \frac{b}{4}\right) \cdot Q^{n+1-\tau-k}\right) = \left(1 - \left(1 - \frac{b}{4}\right)^2\right) \cdot P_3 + \left(1 - \frac{b}{4}\right)^2 \cdot Q^{n+1-\tau},$$

for an appropriately defined distribution P_3. Consequently, we have

$$\mathrm{d_{TV}}\big((Z_1, \dots, Z_k, Z_{k+\tau+1}, \dots, Z_{n+1}), (Z_1, \dots, Z_k, Z'_{k+\tau+1}, \dots, Z'_{n+1})\big) \leq \left(1 - \left(1 - \frac{b}{4}\right)^2\right).$$

Since this is true for all $k \in [n-\tau]$, the mixing coefficient is bounded as $\beta(\tau) \leq (1-(1-\frac{b}{4})^2) \leq \frac{b}{2}$, for any $\tau \geq 0$. Thus $\min_\tau\{\frac{\tau}{n+1} + 2\beta(\tau)\} \leq b$.

Next, we prove the bound on coverage. We first need to specify the score function: define

$$s(z) = \sum_{k=0}^{K} k \cdot \mathbb{1}_{z=z_k}.$$

In other words, $s(z_k) = k$ for each $k \in \{0, \dots, K-1\}$. We are now ready to calculate the coverage probability when the prediction set is constructed with this pretrained score function.

- With probability $b/4$, we draw $\mathbf{Z}$ from P_{cyclic}, meaning that $Z_i = z_{J_i}$ for each i (so that $s(Z_i) = J_i$), with the indices J_i defined via the cyclic construction. If $J_1 \leq K - 1 - n$, then we have $J_{i+1} = J_i + 1$ for all $i \in [n]$, i.e., J_{n+1} is the largest among all the J_i's—and therefore, $s(Z_{n+1}) > \max_{i\in[n]} s(Z_i)$, which implies coverage does not hold. Therefore, on this event, the probability of coverage is at most $\frac{n}{K}$ (i.e., the probability that, when we sample $J_1 \in \{0, \dots, K-1\}$ uniformly at random, we draw $J_1 > K - 1 - n$).

- With probability $1 - b/4$, we draw $\mathbf{Z}$ from Q^{n+1}. In this case, by construction, we have $s(Z_1), \dots, s(Z_{n+1}) \stackrel{\text{iid}}{\sim} \mathrm{Unif}(\{0, \dots, K-1\})$. On the event that all $n+1$ scores are distinct, by exchangeability the coverage probability is exactly $1 - \alpha$ (recalling that we have assumed that $(1-\alpha)(n+1)$ is an integer). And, the event that there is at least one repeated value has probability bounded by $\frac{n(n+1)}{2K}$. In total, therefore, the probability of coverage in this case is bounded by $1 - \alpha + \frac{n(n+1)}{2K}$.

Combining the cases, then,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \leq \frac{b}{4} \cdot \frac{n}{K} + \left(1 - \frac{b}{4}\right) \cdot \left(1 - \alpha + \frac{n(n+1)}{2K}\right).$$

Since $\frac{n(n+1)}{2K} \geq \frac{n}{K}$, this completes the proof.

A.3. Proof of Theorem 8

The proof follows essentially the same argument as the lower bound on coverage, Theorem 5. Fix any $\tau \in \{0, \ldots, n-L\}$. For each $i \in \{L+1, \ldots, n+1\}$, it holds that

$$\mathbb{P}\left\{S_{n+1} \leq \text{Quantile}_{1-\alpha}(\mathbf{S})\right\} \leq \mathbb{P}\left\{S_i \leq \text{Quantile}_{1-\alpha+\frac{\tau}{n-L+1}}(\mathbf{S})\right\} + \Psi_{i-L,\tau}(\mathbf{S}). \tag{9}$$

The proof of this bound is essentially identical to the proof of the analogous bound (7) in the proof of Theorem 5, so we omit the details. With this bound in place, we calculate

$$\begin{aligned}
&\mathbb{P}\left\{S_{n+1} \leq \text{Quantile}_{1-\alpha}(\mathbf{S})\right\} \\
&\leq \frac{1}{n-L+1}\sum_{i=L+1}^{n+1}\left[\mathbb{P}\left\{S_i \leq \text{Quantile}_{1-\alpha+\frac{\tau}{n-L+1}}(\mathbf{S})\right\} + \Psi_{i-L,\tau}(\mathbf{S})\right] \\
&= \mathbb{E}\left[\frac{1}{n-L+1}\sum_{i=L+1}^{n+1}\mathbb{1}\left\{S_i \leq \text{Quantile}_{1-\alpha+\frac{\tau}{n-L+1}}(\mathbf{S})\right\}\right] + \frac{1}{n-L+1}\sum_{i=L+1}^{n+1}\Psi_{i-L,\tau}(\mathbf{S}) \\
&= \mathbb{E}\left[\frac{1}{n-L+1}\sum_{i=L+1}^{n+1}\mathbb{1}\left\{S_i \leq \text{Quantile}_{1-\alpha+\frac{\tau}{n-L+1}}(\mathbf{S})\right\}\right] + \bar{\Psi}_{\tau}(\mathbf{S}).
\end{aligned}$$

For any vector $\mathbf{w} = (w_1, \ldots, w_m) \in \mathbb{R}^m$ and any $a \in [0,1]$, if $w_1, \ldots, w_m$ are distinct, it must hold that $\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}\left\{w_i \leq \text{Quantile}_{1-a}(\mathbf{w})\right\} \leq \frac{\lceil (1-a)m \rceil}{m}$, by definition of the quantile. Therefore, since we have assumed that the scores $S_{L+1}, \ldots, S_{n+1}$ are distinct almost surely,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n-L+1}\sum_{i=L+1}^{n+1}\mathbb{1}\left\{S_i \leq \text{Quantile}_{1-\alpha+\frac{\tau}{n-L+1}}(\mathbf{S})\right\}\right] &\leq \frac{\left\lceil \left(1-\alpha+\frac{\tau}{n-L+1}\right)(n-L+1)\right\rceil}{n-L+1} \\
&= \frac{\lceil (1-\alpha)(n-L+1)\rceil}{n-L+1} + \frac{\tau}{n-L+1},
\end{aligned}$$

which completes the proof.

A.4. Proof of Theorem 9

As in the proof of Theorem 5, the coverage event $Y_{n+1} \in \widehat{C}_n(X_{n+1}; Z_n, \ldots, Z_{n-L+1})$ holds if and only if

$$S_{n+1} \leq \text{Quantile}_{1-\alpha}(\mathbf{S}).$$

And, since the vectors $\mathbf{S}_{\text{split},\tau_*}$ and $\mathbf{S}$ are the same aside from the deleted scores $S_{n_0+L+1}, \ldots, S_{n_0+L+\tau_*}$, it holds surely that

$$\text{Quantile}_{1-\alpha}(\mathbf{S}) \geq \text{Quantile}_{1-\alpha'}(\mathbf{S}_{\text{split},\tau_*}),$$

where $\alpha' = \alpha \cdot \frac{n_1-L+1}{n_1-\tau_*-L+1}$, by a similar calculation to (8) in the proof of Theorem 5. Therefore,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}; Z_n, \ldots, Z_{n-L+1})\right\} \geq \mathbb{P}\left\{S_{n+1} \leq \text{Quantile}_{1-\alpha'}(\mathbf{S}_{\text{split},\tau_*})\right\},$$

and from now on we only need to bound the probability on the right-hand side. The remaining steps are exactly the same as in the proof of Theorem 5, so we omit the details and only summarize briefly. By an argument similar to the one before, we have

$$\mathbb{P}\left\{S_{n+1} \le \mathrm{Quantile}_{1-\alpha'}(\mathbf{S}_{\mathrm{split},\tau_*})\right\} \ge \\ \mathbb{P}\left\{S_i \le \mathrm{Quantile}_{1-\alpha'-\frac{\tau}{n_1-\tau_*-L+1}}(\mathbf{S}_{\mathrm{split},\tau_*})\right\} - \Psi_{i-L-n_0-\tau_*,\tau}(\mathbf{S}_{\mathrm{split},\tau_*})$$

for each $i \in \{n_0 + L + \tau_* + 1, \dots, n+1\}$, and therefore, taking an average over all such indices i,

$$\mathbb{P}\left\{S_{n+1} \le \mathrm{Quantile}_{1-\alpha'}(\mathbf{S}_{\mathrm{split},\tau_*})\right\} \ge 1 - \alpha' - \frac{\tau}{n_1 - \tau_* - L + 1} \\ - \frac{1}{n_1 + 1 - L - \tau_*} \sum_{i=n_0+L+\tau_*+1}^{n+1} \Psi_{i-L-n_0-\tau_*,\tau}(\mathbf{S}_{\mathrm{split},\tau_*}).$$

Substituting for α' in terms of α and simplifying, this yields the desired bound.

A.5. Proof of Proposition 3

First, for any $k > n - \tau$, since the time series is stationary it holds that

$$\Delta^0_{k,\tau}(\mathbf{Z}) = (Z_{\tau+1}, \dots, Z_{n+1}) \overset{\mathrm{d}}{=} (Z_{k+\tau-n}, \dots, Z_k) = \Delta^1_{k,\tau}(\mathbf{Z})$$

(where $\overset{\mathrm{d}}{=}$ denotes equality in distribution), and therefore $\Psi_{k,\tau}(\mathbf{Z}) = \mathrm{d_{TV}}\big(\Delta^0_{k,\tau}(\mathbf{Z}), \Delta^1_{k,\tau}(\mathbf{Z})\big) = 0$.

Now we consider the case $k \le n - \tau$. Let $\mathbf{Z}' = (Z'_1, \dots, Z'_{n+1}) \in \mathcal{Z}^{n+1}$ denote an iid copy of $\mathbf{Z}$, and define

$$\widetilde{\mathbf{Z}}^0 = (Z_1, \dots, Z_{n+1-\tau-k}, Z'_{n+2-k}, \dots, Z'_{n+1})$$

and

$$\widetilde{\mathbf{Z}}^1 = (Z_{k+\tau+1}, \dots, Z_{n+1}, Z'_1, \dots, Z'_k).$$

By the triangle inequality, we have

$$\Psi_{k,\tau}(\mathbf{Z}) = \mathrm{d_{TV}}\big(\Delta^0_{k,\tau}(\mathbf{Z}), \Delta^1_{k,\tau}(\mathbf{Z})\big) \le \mathrm{d_{TV}}\big(\Delta^0_{k,\tau}(\mathbf{Z}), \widetilde{\mathbf{Z}}^0\big) + \mathrm{d_{TV}}\big(\Delta^1_{k,\tau}(\mathbf{Z}), \widetilde{\mathbf{Z}}^1\big) + \mathrm{d_{TV}}\big(\widetilde{\mathbf{Z}}^0, \tilde{\mathbf{Z}}^1\big).$$

Note that by stationarity of $\mathbf{Z}$ and $\mathbf{Z}'$, together with independence $\mathbf{Z} \perp\!\!\!\perp \mathbf{Z}'$, it holds that $\widetilde{\mathbf{Z}}^0 \overset{\mathrm{d}}{=} \widetilde{\mathbf{Z}}^1$, and so the last term in the bound above is zero—that is,

$$\Psi_{k,\tau}(\mathbf{Z}) \le \mathrm{d_{TV}}\big(\Delta^0_{k,\tau}(\mathbf{Z}), \widetilde{\mathbf{Z}}^0\big) + \mathrm{d_{TV}}\big(\Delta^1_{k,\tau}(\mathbf{Z}), \widetilde{\mathbf{Z}}^1\big).$$

But each of these two remaining terms on the right-hand side is bounded by $\beta(\tau)$ by the definition of β-mixing, which completes the proof.

A.6. Proof of Proposition 4

First consider the case $1 \le k \le n - L - \tau$, so that we have

$$\Delta^0_{k,\tau}(\mathbf{S}) = (S_{L+1}, \ldots, S_{n+1-k-\tau}, S_{n+2-k}, \ldots, S_{n+1})$$

and

$$\Delta^1_{k,\tau}(\mathbf{S}) = (S_{L+k+\tau+1}, \ldots, S_{n+1}, S_{L+1}, \ldots, S_{L+k}).$$

Define the function $f_k : \mathcal{Z}^{n+L+1-\tau} \to \mathbb{R}^{n-L+1-\tau}$ as

$$\begin{aligned}(z_1, \ldots, z_{n+1-k-\tau}, z'_1, \ldots, z'_{L+k}) \mapsto \\ \big(s(z_{L+1}; z_L, \ldots, z_1), \ldots, s(z_{n+1-k-\tau}; z_{n-k-\tau}, \ldots, z_{n-k-\tau-L+1}), \\ s(z'_{L+1}; z'_L, \ldots, z'_1), \ldots, s(z'_{L+k}; z'_{L+k-1}, \ldots, z'_k)\big).\end{aligned}$$

Then, by construction, $\Delta^0_{k,\tau}(\mathbf{S}) = f_k\big(\Delta^0_{k+L,\tau-L}(\mathbf{Z})\big)$ and $\Delta^1_{k,\tau}(\mathbf{S}) = f_k\big(\Delta^1_{k+L,\tau-L}(\mathbf{Z})\big)$. Therefore,

$$\Psi_{k,\tau}(\mathbf{S}) = \mathrm{d_{TV}}(\Delta^0_{k,\tau}(\mathbf{S}), \Delta^1_{k,\tau}(\mathbf{S})) \le \mathrm{d_{TV}}(\Delta^0_{k+L,\tau-L}(\mathbf{Z}), \Delta^1_{k+L,\tau-L}(\mathbf{Z})) = \Psi_{k+L,\tau-L}(\mathbf{Z}),$$

where the inequality follows by data processing.

Next, if $n - L - \tau < k \le n - L + 1$, we have

$$\Delta^0_{k,\tau}(\mathbf{S}) = (S_{L+\tau+1}, \ldots, S_{n+1})$$

and

$$\Delta^1_{k,\tau}(\mathbf{S}) = (S_{k+2L+\tau-n}, \ldots, S_{k+L}).$$

In this case, define the function $f_k : \mathcal{Z}^{n+L+1-\tau} \to \mathbb{R}^{n-L+1-\tau}$ as

$$(z_1, \ldots, z_L, z'_1, \ldots, z'_{n+1-\tau}) \mapsto \big(s(z'_{L+1}; z'_L, \ldots, z'_1), \ldots, s(z'_{n+1-\tau}; z'_{n-\tau}, \ldots, z'_{n-L+1-\tau})\big).$$

Then we again have $\Delta^0_{k,\tau}(\mathbf{S}) = f_k\big(\Delta^0_{k+L,\tau-L}(\mathbf{Z})\big)$ and $\Delta^1_{k,\tau}(\mathbf{S}) = f_k\big(\Delta^1_{k+L,\tau-L}(\mathbf{Z})\big)$, and so

$$\Psi_{k,\tau}(\mathbf{S}) = \mathrm{d_{TV}}(\Delta^0_{k,\tau}(\mathbf{S}), \Delta^1_{k,\tau}(\mathbf{S})) \le \mathrm{d_{TV}}(\Delta^0_{k+L,\tau-L}(\mathbf{Z}), \Delta^1_{k+L,\tau-L}(\mathbf{Z})) = \Psi_{k+L,\tau-L}(\mathbf{Z}).$$

Once again, the inequality follows by data processing.

A.7. Proof of Proposition 10

For each $1 \le k \le n_1 - \tau - \tau_*$, define

$$\Delta^{\text{split},0}_{k,\tau,\tau_*}(\mathbf{Z}) = (Z_1, \ldots, Z_{n_0}, Z_{n_0+\tau_*+1}, \ldots, Z_{n+1-k-\tau}, Z_{n+2-k}, \ldots, Z_{n+1})$$

and

$$\Delta^{\text{split},1}_{k,\tau,\tau_*}(\mathbf{Z}) = (Z_1, \ldots, Z_{n_0}, Z_{n_0+\tau+\tau_*+k+1}, \ldots, Z_{n+1}, Z_{n_0+\tau_*+1}, \ldots, Z_{n_0+\tau_*+k}),$$

and for $n_1 - \tau - \tau_* < k \le n_1 + 1 - \tau_*$, define

$$\Delta^{\text{split},0}_{k,\tau,\tau_*}(\mathbf{Z}) = (Z_1, \ldots, Z_{n_0}, Z_{n_0+\tau+\tau_*+1}, \ldots, Z_{n+1})$$

and

$$\Delta^{\text{split},1}_{k,\tau,\tau_*}(\mathbf{Z}) = (Z_1, \ldots, Z_{n_0}, Z_{n_0+k+\tau+2\tau_*-n_1}, \ldots, Z_{n_0+k+\tau_*}).$$

The result of the proposition is then an immediate consequence of the following two lemmas.

Lemma 12 *Under the notation defined above, for any k, τ, τ_* with $\tau_* \geq 0$, $L \leq \tau \leq n_1 - \tau_*$, and $1 \leq k \leq n_1 - L + 1 - \tau_*$, we have*

$$\Psi_{k,\tau}(\mathbf{S}_{\text{split},\tau_*}) \leq \mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k+L,\tau-L,\tau_*}(\mathbf{Z}), \Delta^{\text{split},1}_{k+L,\tau-L,\tau_*}(\mathbf{Z})\big).$$

Lemma 13 *Under the notation defined above, if we additionally assume that $\mathbf{Z}$ is a stationary time series with β-mixing coefficients $\beta(\tau)$, then for any k, τ, τ_* with $\tau, \tau_* \geq 0$, $\tau + \tau_* \leq n$, and $1 \leq k \leq n_1 + 1 - \tau_*$, we have*

$$\mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k,\tau,\tau_*}(\mathbf{Z}), \Delta^{\text{split},1}_{k,\tau,\tau_*}(\mathbf{Z})\big) \leq \begin{cases} 2\beta(\tau_*) + 2\beta(\tau), & \text{for } 1 \leq k \leq n_1 - \tau - \tau_*, \\ 2\beta(\tau_*), & \text{for } n_1 - \tau - \tau_* < k \leq n_1 + 1 - \tau_*. \end{cases}$$

A.7.1. PROOF OF LEMMA 12

First suppose $1 \leq k \leq n_1 - L - \tau - \tau_*$. Then

$$\Delta^0_{k,\tau}(\mathbf{S}_{\text{split},\tau_*}) = (S_{n_0+L+\tau_*+1}, \dots, S_{n+1-k-\tau}, S_{n+2-k}, \dots, S_{n+1})$$

and

$$\Delta^1_{k,\tau}(\mathbf{S}_{\text{split},\tau_*}) = (S_{n_0+L+\tau_*+k+\tau+1}, \dots, S_{n+1}, S_{n_0+L+\tau_*+1}, \dots, S_{n_0+L+\tau_*+k}).$$

Now define a function $f_k : \mathcal{Z}^{n+L+1-\tau-\tau_*} \to \mathbb{R}^{n_1-L+1-\tau-\tau_*}$ as

$$\begin{aligned}(z_1, \dots, z_{n_0}, z'_1, \dots, z'_{n_1+1-\tau-\tau_*-k}, z''_1, \dots, z''_{k+L}) \mapsto \\ \big(s(z'_{L+1}; z'_L, \dots, z'_1), \dots, s(z'_{n_1+1-\tau-\tau_*-k}; z'_{n_1-\tau-\tau_*-k}, \dots, z'_{n_1-L-\tau-\tau_*-k+1}), \\ s(z''_{L+1}; z''_L, \dots, z''_1), \dots, s(z''_{k+L}; z''_{k+L-1}, \dots, z''_k)\big) \text{ where } s = \mathcal{A}(z_1, \dots, z_{n_0}).\end{aligned}$$

Then we can observe that

$$\Delta^j_{k,\tau}(\mathbf{S}_{\text{split},\tau_*}) = f_k\big(\Delta^{\text{split},j}_{k+L,\tau-L,\tau_*}(\mathbf{Z})\big)$$

for each $j = 0, 1$. Consequently, by the data processing inequality, we have

$$\Psi_{k,\tau}(\mathbf{S}_{\text{split},\tau_*}) = \mathrm{d_{TV}}\big(\Delta^0_{k,\tau}(\mathbf{S}_{\text{split},\tau_*}), \Delta^1_{k,\tau}(\mathbf{S}_{\text{split},\tau_*})\big) \leq \mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k+L,\tau-L,\tau_*}(\mathbf{Z}), \Delta^{\text{split},1}_{k+L,\tau-L,\tau_*}(\mathbf{Z})\big).$$

Next suppose $n_1 - L - \tau - \tau_* < k \leq n_1 - L + 1 - \tau_*$. Then

$$\Delta^0_{k,\tau}(\mathbf{S}_{\text{split},\tau_*}) = (S_{n_0+L+\tau_*+\tau+1}, \dots, S_{n+1})$$

and

$$\Delta^1_{k,\tau}(\mathbf{S}_{\text{split},\tau_*}) = (S_{n_0+2L+2\tau_*+\tau+k-n_1}, \dots, S_{n_0+L+\tau_*+k}).$$

For this case, define the function $f_k : \mathcal{Z}^{n+L+1-\tau-\tau_*} \to \mathbb{R}^{n_1-L+1-\tau-\tau_*}$ as

$$\begin{aligned}(z_1, \dots, z_{n_0}, z'_1, \dots, z'_L, z''_1, \dots, z''_{n_1+1-\tau-\tau_*}) \mapsto \\ \big(s(z''_{L+1}; z''_L, \dots, z''_1), \dots, s(z''_{n_1+1-\tau-\tau_*}; z''_{n_1-\tau-\tau_*}, \dots, z''_{n_1-\tau-\tau_*-L+1})\big) \\ \text{where } s = \mathcal{A}(z_1, \dots, z_{n_0}).\end{aligned}$$

Then we can observe that

$$\Delta^j_{k,\tau}(\mathbf{S}_{\text{split},\tau_*}) = f_k\big(\Delta^{\text{split},j}_{k+L,\tau-L,\tau_*}(\mathbf{Z})\big)$$

for each $j = 0, 1$, and so again by data processing we have

$$\Psi_{k,\tau}(\mathbf{S}_{\text{split},\tau_*}) \leq \mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k+L,\tau-L,\tau_*}(\mathbf{Z}), \Delta^{\text{split},1}_{k+L,\tau-L,\tau_*}(\mathbf{Z})\big).$$

A.7.2. PROOF OF LEMMA 13

First consider the case $1 \leq k \leq n_1 - \tau - \tau_*$. Let $\mathbf{Z}', \mathbf{Z}'' \in \mathcal{Z}^{n+1}$ denote iid copies of $\mathbf{Z}$, and define

$$\widetilde{\mathbf{Z}}^0 = (Z_1, \ldots, Z_{n_0}, Z'_{n_0+\tau_*+1}, \ldots, Z'_{n+1-k-\tau}, Z''_{n+2-k}, \ldots, Z''_{n+1})$$

and

$$\widetilde{\mathbf{Z}}^1 = (Z_1, \ldots, Z_{n_0}, Z'_{n_0+\tau+\tau_*+k+1}, \ldots, Z'_{n+1}, Z''_{n_0+\tau+1}, \ldots, Z''_{n_0+\tau+k}),$$

Then, by the triangle inequality,

$$\mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k,\tau,\tau_*}(\mathbf{Z}), \Delta^{\text{split},1}_{k,\tau,\tau_*}(\mathbf{Z})\big) \leq \mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k,\tau,\tau_*}(\mathbf{Z}), \widetilde{\mathbf{Z}}^0\big) + \mathrm{d_{TV}}\big(\Delta^{\text{split},1}_{k,\tau,\tau_*}(\mathbf{Z}), \widetilde{\mathbf{Z}}^1\big) + \mathrm{d_{TV}}\big(\widetilde{\mathbf{Z}}^0, \widetilde{\mathbf{Z}}^1\big).$$

Since the three time series $\mathbf{Z}, \mathbf{Z}', \mathbf{Z}''$ are mutually independent and are each stationary, it holds that $\widetilde{\mathbf{Z}}^0 \overset{\mathrm{d}}{=} \widetilde{\mathbf{Z}}^1$, and so the last term above is zero. Therefore,

$$\mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k,\tau,\tau_*}(\mathbf{Z}), \Delta^{\text{split},1}_{k,\tau,\tau_*}(\mathbf{Z})\big) \leq \mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k,\tau,\tau_*}(\mathbf{Z}), \widetilde{\mathbf{Z}}^0\big) + \mathrm{d_{TV}}\big(\Delta^{\text{split},1}_{k,\tau,\tau_*}(\mathbf{Z}), \widetilde{\mathbf{Z}}^1\big).$$

Next define

$$\breve{\mathbf{Z}}^0 = (Z_1, \ldots, Z_{n_0}, Z_{n_0+\tau_*+1}, \ldots, Z_{n+1-k-\tau}, Z''_{n+2-k}, \ldots, Z''_{n+1}).$$

Since $\mathbf{Z}''$ is independent of $\mathbf{Z}$ and $\mathbf{Z}'$, we have

$$\begin{aligned}
&\mathrm{d_{TV}}\big(\widetilde{\mathbf{Z}}^0, \breve{\mathbf{Z}}^0\big) \\
&= \mathrm{d_{TV}}\big((Z_1, \ldots, Z_{n_0}, Z'_{n_0+\tau_*+1}, \ldots, Z'_{n+1-k-\tau}), (Z_1, \ldots, Z_{n_0}, Z_{n_0+\tau_*+1}, \ldots, Z_{n+1-k-\tau})\big) \\
&\overset{(i)}{\leq} \beta(\tau_*),
\end{aligned}$$

where step (i) holds by definition of β-mixing. Reasoning similarly, we also have

$$\mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k,\tau,\tau_*}(\mathbf{Z}), \breve{\mathbf{Z}}^0\big) \leq \beta(\tau),$$

again by definition of the β-mixing coefficients. Therefore, again applying the triangle inequality yields

$$\mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k,\tau,\tau_*}(\mathbf{Z}), \widetilde{\mathbf{Z}}^0\big) \leq \mathrm{d_{TV}}\big(\widetilde{\mathbf{Z}}^0, \breve{\mathbf{Z}}^0\big) + \mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k,\tau,\tau_*}(\mathbf{Z}), \breve{\mathbf{Z}}^0\big) \leq \beta(\tau_*) + \beta(\tau).$$

A similar argument yields that $\mathrm{d_{TV}}\big(\mathbf{Z}^1_{\text{split}}(k,\tau), \widetilde{\mathbf{Z}}^1_{\text{split}}(k,\tau)\big) \leq \beta(\tau_*) + \beta(\tau)$, by considering

$$\breve{\mathbf{Z}}^1 = (Z_1, \ldots, Z_{n_0}, Z'_{n_0+\tau+\tau_*+k+1}, \ldots, Z'_{n+1}, Z_{n_0+\tau+1}, \ldots, Z_{n_0+\tau+k})$$

in place of $\breve{\mathbf{Z}}^0$. Therefore we have shown that

$$\mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k,\tau,\tau_*}(\mathbf{Z}), \Delta^{\text{split},1}_{k,\tau,\tau_*}(\mathbf{Z})\big) \leq 2\beta(\tau_*) + 2\beta(\tau).$$

Next we turn to the case that $n_1 - \tau - \tau_* < k \leq n_1 + 1 - \tau_*$. Define

$$\widetilde{\mathbf{Z}}^0 = (Z_1, \ldots, Z_{n_0}, Z'_{n_0+\tau+\tau_*+1}, \ldots, Z'_{n+1})$$

and

$$\widetilde{\mathbf{Z}}^1 = (Z_1, \dots, Z_{n_0}, Z'_{n_0+k+\tau+2\tau_*-n_1}, \dots, Z'_{n_0+k+\tau_*}),$$

where again $\mathbf{Z}'$ denotes an iid copy of $\mathbf{Z}$. Then, as before,

$$\mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k,\tau,\tau_*}(\mathbf{Z}), \Delta^{\text{split},1}_{k,\tau,\tau_*}(\mathbf{Z})\big) \le \mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k,\tau,\tau_*}(\mathbf{Z}), \widetilde{\mathbf{Z}}^0\big) + \mathrm{d_{TV}}\big(\Delta^{\text{split},1}_{k,\tau,\tau_*}(\mathbf{Z}), \widetilde{\mathbf{Z}}^1\big) + \mathrm{d_{TV}}\big(\widetilde{\mathbf{Z}}^0, \widetilde{\mathbf{Z}}^1\big).$$

The first two terms on the right-hand side are each bounded by $\beta(\tau_*)$, by definition of the β-mixing coefficients, while the final term is zero since $\widetilde{\mathbf{Z}}^0 \overset{\mathrm{d}}{=} \widetilde{\mathbf{Z}}^1$ by stationarity of $\mathbf{Z}'$, together with the fact that $\mathbf{Z} \perp\!\!\!\perp \mathbf{Z}'$. Therefore, for this case we have

$$\mathrm{d_{TV}}\big(\Delta^{\text{split},0}_{k,\tau,\tau_*}(\mathbf{Z}), \Delta^{\text{split},1}_{k,\tau,\tau_*}(\mathbf{Z})\big) \le 2\beta(\tau_*),$$

which completes the proof.