Conformal Prediction Methods in Finance

João Vitor Romano
Advisor: Paulo Orenstein

Instituto de Matemática Pura e Aplicada
July 2022

Abstract

Machine learning models are widely used in a variety of fields due to their predictive power, often achieving state-of-the-art results. However, most models generate either point predictions that do not conceive any notion of uncertainty or prediction intervals without finite-sample statistical guarantees. In typical finance problems, the decision-making process can be fragilized due to a lack of rigorous uncertainty quantification. Conformal prediction is a method that allows one to transform point predictions coming from any model in prediction intervals with nonasymptotic guarantees and without strong hypotheses on the data distribution. We present recent results from conformal prediction that justifies its use for dependent data. Through theoretical discussions and empirical experiments, we show the potential of conformal prediction for the field of finance, a largely unexplored endeavour.

Acknowledgments

A master's is marked by classes, exams, projects and research, but more so by the people that take part in the journey. I would like to start by thanking my advisor, Paulo Orenstein, whose good taste in statistics, mathematics and programming is inspiring and one of the many reasons that make it such a pleasure to work with him. Paulo's easy-going and friendly personality blends well with his immense commitment, attention to detail and thirst for success, in a way that draws the best out of people. Such qualities are shared by Roberto Imbuzeiro Oliveira, to whom I am grateful for giving me so many opportunities over the past two years. Roberto's sense of humor and sheer amount of knowledge ensures that every discussion leads to a good time and research progress. The work presented here would also not have been possible without Thiago Ramos, whose creativity and determination find ways through the hardest problems. I extend my gratitude to all the other great students I had the honour of knowing, learning from and collaborating with in the ExactBoost and AmnioML projects: Carolina Piazza, Daniel Csillag, Lucas Monteiro and Rodrigo Schuller. From the finance programme, I thank my fellow classmates Bernardo Cassar, Francis Araújo and Lucas Xisto, who were present from start to finish, as well as professors Alcides Lins Neto, Edison Tito, Fernando Aiube, Luciano Irineu de Castro, Milton Jara, Roberto Velho, Rodrigo Targino, Vinicius Albani and Yuri Saporito. Special thanks go to Yuri for participating in my committee, carefully evaluating this work and providing valuable suggestions. IMPA's staff has always been helpful and I appreciated it; in particular, Roberto Beauclair always ensured smooth sailing during our projects. I thank Bruno Licht, Carlos Rocha and Sérgio Werlang for financially supporting part of my degree as well as all other colleagues from Sarpen that encouraged my studies and understood all the missed lunches. Last but not least, I thank all my friends and my family for their support, in particular my grandparents Araken, Vânia, Lauro and Lia.

Contents

Chapter 1

Introduction

1.1 Motivation and related work

Breiman (2001) put into perspective the perceived dichotomy between two statistical cultures: the *data modeling* culture and the *algorithmic modeling* culture. Although the objective of reaching conclusions from data is common between the cultures, their approaches are fundamentally different. While the former assumes a parametric stochastic model for data, the latter considers the data-generating process as unknown and employs algorithmic models to make predictions. Breiman made the case that good predictions should be the driving force of the field and that restricting oneself to less flexible models was detrimental, but acknowledged the dilemma brought by reduced interpretability of complex models. Leo Breiman's article prompted responses from statisticians David Cox, Bradley Efron, Bruce Hoadley, Emanuel Parzen and many others, both from industry and academia. Parzen (2001) and Hoadley (2001) were mainly in agreement with the importance of algorithmic modeling, but Efron (2001) and especially Cox (2001) presented more fundamental disagreements, that were subsequently addressed. The topic was warmly welcomed and debated at the time, with heterogeneous points of view making for a rich and insightful discussion. Two decades later, the seminal article of Breiman continues to be influential.

Tibshirani and Hastie (2021) argue for a statistics *melting pot* as a metaphor for a merger of the two cultures. The use of both data and algorithmic models, oftentimes on the same problem, is cast in a good light under the argument that each approach has its advantages and, when used in conjunction, can give richer solutions. Additionally, data modeling is shown to illuminate algorithmic models by providing explanations to certain phenomena, such as double descent (Hastie et al. 2022). The main message is that the coexistence of data and algorithmic models should be embraced and using both in harmony has a lot to offer.

Conformal prediction (CP) is a method for uncertainty quantification that fits aptly into the "melting pot" metaphor. Any prediction model — no matter how opaque, complex or misspecified — can be wrapped by CP, under extremely mild assumptions, to yield prediction sets whose sizes represent the model's uncertainty. Moreover, the true value of the prediction is guaranteed to belong to the set with a prescribed probability. In all its generality, conformal prediction allows one to first fully focus on algorithmic modeling by selecting the most powerful model (or ensemble of models) to excel in the given task, and then build prediction sets seamlessly.

Finance represents a prime example of a field in which high-stakes decisions call for both good predictions and uncertainty quantification. Dixon, Halperin, and Bilokon (2020) mention that frequentist machine learning models providing point estimates can be unsuitable for certain financial applications

and that Gaussian processes can be chosen instead since they generate prediction intervals. Indeed, usage of Gaussian processes in finance entails option pricing (Tegner and Roberts 2021), volatility forecasting (Liu, Kiskin, and Roberts 2020) and term-structure interpolation (Cousin, Maatouk, and Rullière 2016), to name a few examples. Although frequentist machine learning models' point predictions could be made into prediction sets via CP, much less work in the finance literature considered this avenue so far.

Dewolf, Baets, and Waegeman (2022) compared Bayesian methods (such as Gaussian processes), ensemble methods, direct interval estimation methods and conformal prediction as general classes of uncertainty quantification methodologies. Gaussian processes, although widely used in finance, assume that the data is (conditionally) normally distributed and were shown to lead to invalid models if the assumption is not met. Contrastingly, conformal prediction is guaranteed to work without any assumption on data distribution. Traditionally, data is assumed independent and identically distributed (iid) in the machine learning literature, but such premise is generally not observed in many applications in finance.

Oliveira et al. (2022) proved that a CP method known as split conformal prediction can be applied to dependent data, which is often the case with financial time series, and retain theoretical guarantees. Our aim is to reproduce and extend synthetic and real experiments from Oliveira et al. (2022), as well as provide novel applications and insights for finance.

Applications of conformal prediction to financial data are not unheard of, but are scarce. Wisniewski, Lindsay, and Lindsay (2020) make use of CP to generate prediction sets for a market maker's net position over time. Kath and Ziel (2021) forecast short-term electricity prices and evaluate conformal prediction applied to that end. Chernozhukov, Wüthrich, and Zhu (2021) employ a CP method to predict stock returns based on their realized volatility. Gibbs and Candès (2021) develop an adaptive method inspired by conformal prediction that works under distribution shift and showcase how it can be used to predict market volatility.

1.2 Organization

The remainder of the work is organized as follows. Chapter 2 presents the traditional theory of conformal prediction pioneered by Vladimir Vovk and collaborators, capable of generating prediction sets that are guaranteed to include the unseen label of new data points under the exchangeability assumption, a concept similar to, but weaker than, independence. Chapter 3 discusses a novel framework of conformal prediction due to Oliveira et al. (2022) that employs concentration of measure and decoupling inequalities to generalize fundamental results beyond the exchangeability assumption. Chapter 4 gives a brief overview of stochastic processes and some important properties, as well as examples of different processes. Chapter 5 applies the theory from Chapter 3 to β-mixing stochastic processes, a notion mathematically defined in Chapter 4, but that informally means that two observations from a stochastic process become less dependent the further they are apart. Chapter 6 showcases synthetic and financial experiments that corroborate the theory presented up to that point and empirically demonstrate that conformal prediction can indeed be used for dependent data. In the same chapter, we also show how uncertainty quantification via conformal prediction can be used for a simple trading strategy. Chapter 7 concludes by summarizing results and discussing possible future developments.

Chapter 2

Conformal Prediction

Machine learning models often achieve state-of-the-art performance on a range of learning tasks and across many fields. However, uncertainty in predictions is not always accounted for, albeit of tremendous importance for some areas, such as medicine and finance. One could argue that the ideal technique for uncertainty quantification would: (i) work for any model; (ii) provide finite-sample statistical guarantees; (iii) work for data coming from any distribution; (iv) be simple to implement and (v) generalize to a wide range of learning tasks. Conformal prediction is a modern technique for distribution-free uncertainty quantification that possesses all those qualities. In this chapter, we give an algorithmic description of conformal prediction and provide proofs of main theoretical results.

The traditional theory of conformal prediction crucially relies on data exchangeability and a property of quantiles, namely that with probability at least $\phi \in (0, 1)$, a random variable from an exchangeable sequence is at most equal to the adjusted ϕ-quantile taken over all other variables from the sequence, where the adjustment is due to the finiteness of the sequence (Lemma 2.1.6). We start by defining what it means for random variables to be exchangeable and how that relates to more widespread concepts such as independence and correlation; we then define population and empirical quantiles and prove the fundamental Lemma 2.1.6, which paves the way forward. Next, full conformal prediction is thoroughly presented and proven to yield prediction sets with valid marginal coverage, that is, the unseen label of a new data point will belong to the prediction set with probability at least equal to a prespecified value. Split conformal prediction and conformalized quantile regression are discussed as particular cases of full CP and their existences are well motivated. Finally, we discuss the importance of well-chosen nonconformity scores to attain small sets and present different notions of coverage.

2.1 Preliminaries

One of the cornerstones of standard conformal prediction theory is data exchangeability.

Definition 2.1.1 (Exchangeability)**.** *A finite sequence of random variables $(Z_1, \ldots, Z_n)$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is exchangeable if, for any permutation function $\pi \colon [n] \to [n]$,*

$$(Z_1, \ldots, Z_n) \stackrel{d}{=} (Z_{\pi(1)}, \ldots, Z_{\pi(n)}).$$

In words, a sequence of random variables whose joint distribution is invariant under rearrangement of the variables is exchangeable, that is, the joint distribution $F_{Z_1,\ldots,Z_n}(z_1, \ldots, z_n)$ is symmetric in its arguments. An infinite sequence $(Z_i)_{i\in\mathbb{N}}$ is exchangeable if every finite subsequence is exchangeable.

Exchangeability may be thought of as a property of the underlying distribution of the random variables instead of a sequence, since a particular sequence being exchangeable implies any other sequence formed from the same random variables will share such property. Therefore, it makes sense to use the term "exchangeable random variables". Less popular synonyms include "symmetrically dependent" and "interchangeable" random variables (Aldous 1985).

Proposition 2.1.2 (Independent and identically distributed random variables are exchangeable)**.** *Let $Z_1, \ldots, Z_n$ be iid random variables. Then $Z_1, \ldots, Z_n$ are exchangeable.*

Proof. As the random variables are independent, the joint cdf equals the product of the individual marginals. Therefore, for any permutation π,

$$F_{Z_{\pi(1)},\ldots,Z_{\pi(n)}}(z_1, \ldots, z_n) = \prod_{i=1}^{n} F_{Z_{\pi(i)}}(z_i).$$

However, since the random variables are identically distributed, the marginal cumulative distribution functions are all the same. Thus, $F_{Z_{\pi(1)}} = \ldots = F_{Z_{\pi(n)}} =: F$ and

$$F_{Z_{\pi(1)},\ldots,Z_{\pi(n)}}(z_1, \ldots, z_n) = \prod_{i=1}^{n} F(z_i),$$

indicating that the distribution is invariant under permutation, since the right-hand side does not depend on π. □

Proposition 2.1.3 (Exchangeable random variables are identically distributed)**.** *Let $Z_1, \ldots, Z_n$ be exchangeable random variables. Then $F_{Z_i} = F_{Z_j}$ for all $i, j \in [n]$.*

Proof. As the random variables are exchangeable, they share the same distribution function under permutation,

$$F_{Z_1,\ldots,Z_n}(z_1, \ldots, z_n) = F_{Z_{\pi(1)},\ldots,Z_{\pi(n)}}(z_1, \ldots, z_n),$$

which may be rewritten as

$$\mathbb{P}[Z_1 \leq z_1, \ldots, Z_n \leq z_n] = \mathbb{P}[Z_{\pi(1)} \leq z_1, \ldots, Z_{\pi(n)} \leq z_n].$$

Fixing an arbitrary z_j, marginal distributions can be recovered by nullifying the effect of all other random variables:

$$\mathbb{P}[Z_j \leq z_j] = \lim_{\substack{z_i \to \infty \\ i \in [n] \setminus j}} \mathbb{P}[Z_1 \leq z_1, \ldots, Z_n \leq z_n] = \lim_{\substack{z_i \to \infty \\ i \in [n] \setminus j}} \mathbb{P}[Z_{\pi(1)} \leq z_1, \ldots, Z_{\pi(n)} \leq z_n] = \mathbb{P}[Z_{\pi(j)} \leq z_j],$$

that is, $F_{Z_j} = F_{Z_{\pi(j)}}$. As this is true for any permutation function π, we can substitute $\pi(j)$ for any other index i to conclude that

$$\forall i, j \in [n] \colon F_{Z_i} = F_{Z_j},$$

proving as desired that all random variables have the same distribution. □

Besides the propositions above relating exchangeability with similar concepts, we provide counterexamples based on Romano and Siegel (1986, §4.17–§4.20) below to illustrate fallacious implications and provide further intuition. There exist methods for testing exchangeability of real datasets (Vovk,

Nouretdinov, and Gammerman 2003; Fedorova et al. 2012; Vovk 2021), but we will not delve into them. Instead, we will show in Chapter 3 how to extend standard conformal prediction results to nonexchangeable data. Out of curiosity, we mention that Fedorova et al. (2012) developed a method based on martingales for testing exchangeability online and found that: (i) the popular USPS dataset of handwritten digits (Hull 1994) does not appear to be exchangeable and (ii) the Statlog Satellite dataset (Srinivasan 1993) hosted at UCI Machine Learning Repository (Dua and Graff 2017), which consists of 36 features processed from satellite images and a discrete label indicating the type of soil in the center of the image, appears to be exchangeable.

Example 1 (Exchangeable random variables that are not independent)**.** *We showed in Proposition 2.1.2 that iid random variables are exchangeable; now we give a counterexample that prevents the converse from being true, even though exchangeable random variables are identically distributed (cf. Proposition 2.1.3). Let the random vector* (X, Y) *assume values* $(0, 1)$*,* $(0, -1)$*,* $(1, 0)$ *and* $(-1, 0)$ *equiprobably. It is evident that* (X, Y) *and* (Y, X) *have the same distribution due to symmetry, which makes the random variables exchangeable. However, note that* $\mathbb{P}[X = 1] = 1/4$ *and* $\mathbb{P}[Y = 0] = 1/2$*, which yields* $\mathbb{P}[X = 1] \cdot \mathbb{P}[Y = 0] = 1/8$*, but* $\mathbb{P}[X = 1, Y = 0] = 1/4$*. Thus,* X *and* Y *are not independent.*

Example 2 (Independent random variables that are not exchangeable)**.** *Proposition 2.1.2 showed that independent and identically distributed random variables are exchangeable. We now show that independence by itself is not enough to ensure the same implication and that it is important for variables to have the same law. Let* $X \sim \text{Poisson}(2)$ *and* $Y \sim \text{Normal}(0, 1)$ *be independent. As the marginal distributions are different, the random variables are not exchangeable. The same conclusion is reached by taking the contrapositive of Proposition 2.1.3.*

Example 3 (Identically distributed random variables that are not exchangeable)**.** *Although Proposition 2.1.3 showed that exchangeable random variables are identically distributed, the converse is not true. In order to exemplify this assertion, let* $X \sim \text{Normal}(0, 1)$ *and* $Y \sim \text{Normal}(0, 1)$ *be independent and set a copy* $Z \equiv Y$*. All three random variables have the same distribution, but are not exchangeable because the joint distributions from* (X, Y, Z) *and* (Y, Z, X) *are different since the former has zero correlation between the first two components while the latter has a correlation of one between its first two components.*

Example 4 (Exchangeable random variables that are not uncorrelated)**.** *Let* $X, Y, Z \sim \text{Normal}(0, 1)$*. Although* $X + Y, X + Z$ *are clearly exchangeable, they have positive correlation.*

We summarize below the conclusions from Propositions 2.1.2 and 2.1.3 and Examples 1, 2, 3 and 4.

$$
\begin{aligned}
\text{Independence and identical distribution} &\implies \text{Exchangeability;} \\
\text{Exchangeability} &\implies \text{Identical distribution;} \\
\text{Exchangeability} &\not\implies \text{Independence;} \\
\text{Independence} &\not\implies \text{Exchangeability;} \\
\text{Identical distribution} &\not\implies \text{Exchangeability;} \\
\text{Exchangeability} &\not\implies \text{Uncorrelation.}
\end{aligned}
$$

The fact that iid data is necessarily exchangeable but the converse is not true makes the point that exchangeability is weaker and thus preferred as an assumption. Conformal prediction being valid for exchangeable data automatically makes it valid for iid data, and more.

We now proceed to the second cornerstone of standard conformal prediction theory: quantiles.

Definition 2.1.4 (Population quantile)**.** *Given $\phi \in [0,1)$, let $q_\phi(Z)$ denote the population ϕ-quantile of Z; that is:*

$$q_\phi(Z) \coloneqq \inf\{t \in \mathbb{R} : \mathbb{P}[Z \leq t] \geq \phi\}.$$

When the random variable is clear from the context, q_ϕ may be used instead.

Definition 2.1.5 (Empirical quantile)**.** *Given $\phi \in [0,1)$, let $\widehat{q}_\phi(Z_{1:n})$ denote the empirical ϕ-quantile of $Z_{1:n} \coloneqq \{Z_i\}_{i=1}^n$; that is:*

$$\widehat{q}_\phi(Z_{1:n}) \coloneqq \inf\left\{t \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Z_i \leq t\} \geq \phi\right\}.$$

When the sample is clear from the context, $\widehat{q}_{\phi,I}$ may be used instead, where I is the set of indices considered. The above definition, for example, could be rewritten as $\widehat{q}_{\phi,[n]}$ in this notation.

Assuming that data is exchangeable reveals a property of quantiles that will later prove extremely useful.

Lemma 2.1.6 (Quantile property under exchangeability. Tibshirani, Barber, et al. 2019)**.** *If $Z_1, \ldots, Z_{n+1}$ are exchangeable random variables, then for any $\phi \in (0,1)$, we have*

$$\mathbb{P}\left[Z_{n+1} \leq \widehat{q}_{(1+\frac{1}{n})\phi}(Z_{1:n})\right] \geq \phi. \tag{2.1}$$

Moreover, if the random variables are almost surely distinct, then

$$\mathbb{P}\left[Z_{n+1} \leq \widehat{q}_{(1+\frac{1}{n})\phi}(Z_{1:n})\right] \leq \phi + \frac{1}{n+1}. \tag{2.2}$$

Proof. If Z_{n+1} is assumed to be greater than the empirical quantile $\widehat{q}_\phi(Z_{1:n+1})$, swapping Z_{n+1} for any other larger value (in particular, $+\infty$) should not alter the quantile. Conversely, if Z_{n+1} is greater than the empirical quantile calculated over the sample with Z_{n+1} swapped for $+\infty$, the substitution could be reverted without altering the quantile value. Formally, we have

$$Z_{n+1} > \widehat{q}_\phi(Z_{1:n+1}) \iff Z_{n+1} > \widehat{q}_\phi(Z_{1:n} \cup \{\infty\}).$$

The contrapositive of the biconditional statement above, naturally, is also true:

$$Z_{n+1} \leq \widehat{q}_\phi(Z_{1:n+1}) \iff Z_{n+1} \leq \widehat{q}_\phi(Z_{1:n} \cup \{\infty\}).$$

Letting $Z_{(1)} \leq \ldots \leq Z_{(n+1)}$ denote the order statistics of $Z_1, \ldots, Z_{n+1}$, it becomes evident that $Z_{n+1} \leq \widehat{q}_\phi(Z_{1:n+1})$ is equivalent to $Z_{n+1} \in \{Z_{(1)}, \ldots, Z_{(\lceil \phi(n+1) \rceil)}\}$, which, considering the possibility of ties, has probability at least $\lceil \phi(n+1) \rceil / (n+1) \geq \phi$ of occurring due to exchangeability of the random variables. Therefore,

$$\mathbb{P}[Z_{n+1} \leq \widehat{q}_\phi(Z_{1:n+1})] \geq \phi,$$

which implies,

$$\mathbb{P}[Z_{n+1} \leq \widehat{q}_\phi(Z_{1:n} \cup \{\infty\})] \geq \phi.$$

Note that the quantile can be recast into the desired form via

$$\begin{aligned}
\widehat{q}_\phi(Z_{1:n} \cup \{\infty\}) &= \inf\left\{t \in \mathbb{R} : \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{1}\{Z_i \leq t\} \geq \phi\right\} && \text{(Definition 2.1.5)} \\
&= \inf\left\{t \in \mathbb{R} : \frac{1}{n+1}\sum_{i=1}^{n} \mathbb{1}\{Z_i \leq t\} \geq \phi\right\} && (Z_{n+1} > t) \\
&= \inf\left\{t \in \mathbb{R} : \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Z_i \leq t\} \geq \frac{n+1}{n} \cdot \phi\right\} \\
&= \widehat{q}_{(1+\frac{1}{n})\phi}(Z_{1:n}), && \text{(Definition 2.1.5)}
\end{aligned}$$

thus yielding the lower bound:

$$\mathbb{P}\left[Z_{n+1} \leq \widehat{q}_{(1+\frac{1}{n})\phi}(Z_{1:n})\right] \geq \phi.$$

If we assume that the random variables are almost surely distinct, the event $Z_{n+1} \in \{Z_{(1)}, \ldots, Z_{(\lceil \phi(n+1) \rceil)}\}$ happens with probability exactly $\lceil \phi(n+1) \rceil/(n+1) \leq (\phi(n+1)+1)/(n+1) = \phi + 1/(n+1)$, which yields the upper bound:

$$\mathbb{P}\left[Z_{n+1} \leq \widehat{q}_{(1+\frac{1}{n})\phi}(Z_{1:n})\right] \leq \phi + \frac{1}{n+1}. \qquad \square$$

Remark 2.1.7. *Taking the ϕ-quantile instead of the adjusted version thereof in Equation* (2.1) *would result in a miscoverage on the order of $1/n$:*

$$\mathbb{P}\left[Z_{n+1} \leq \widehat{q}_\phi(Z_{1:n})\right] \geq \frac{\phi}{1+1/n} = \phi\frac{n}{n+1} = \phi\frac{(n+1)-1}{n+1} = \phi - \frac{\phi}{n+1} = \phi - O(1/n).$$

Therefore, adjusting the quantile by a multiplicative factor of $1+1/n$ can be seen as a way of guaranteeing the desired lower bound of exactly ϕ. Note that the adjustment is particularly important for small samples and becomes less relevant when n increases.

Understanding when exchangeability is preserved after transformations of originally exchangeable random variables is important for a wider use of Lemma 2.1.6. Indeed, it will play a crucial role in the proof of conformal prediction (Theorem 2.2.1).

Lemma 2.1.8 (Exchangeability-preserving transformations. Kuchibhotla 2020, Theorem 3)**.** *Let $Z = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$ be a vector of exchangeable random variables and $G\colon \mathcal{Z}^n \to (\mathcal{Z}')^m$ an arbitrary transformation. Moreover, suppose that, for each possible permutation $\pi_1\colon [m] \to [m]$, there exists a permutation $\pi_2\colon [n] \to [n]$ such that*

$$\pi_1 G(Z) \overset{d}{=} G(\pi_2 Z).$$

Then, the transformation G preserves exchangeability.

Proof. Given that Z is an exchangeable vector, $\pi_2 Z$ and Z have the same joint distribution. Therefore, $G(\pi_2 Z)$ and $G(Z)$ also have the same distribution. Allying this observation with the assumption that $\pi_1 G(Z) \overset{d}{=} G(\pi_2 Z)$ implies the desired result: $G(Z) \overset{d}{=} \pi_1 G(Z)$. $\square$

Lemma 2.1.9 (Symmetric functions preserve exchangeability. Kuchibhotla 2020, Proposition 4)**.** *Let $Z_1, \ldots, Z_n$ be exchangeable random variables taking values in $\mathcal{Z}$ and $\widehat{f}\colon \mathcal{Z}^n \times \mathcal{Z} \to \mathcal{Z}'$ any function symmetric in its first n arguments. Then, $\widehat{f}(Z_1, \ldots, Z_n; Z_1), \ldots, \widehat{f}(Z_1, \ldots, Z_n; Z_n)$ are also exchangeable.*

Proof. Define $G(Z) \coloneqq (\widehat{f}(Z_1, \dots, Z_n; Z_1), \dots \widehat{f}(Z_1, \dots, Z_n; Z_n))$, where G and Z are as in Lemma 2.1.8. Given that $\widehat{f}$ is symmetric in its first n arguments, we have, for any permutation function $\pi \colon [n] \to [n]$,

$$\begin{aligned} G(\pi Z) &= (\widehat{f}(Z_{\pi(1)}, \dots, Z_{\pi(n)}; Z_{\pi(1)}), \dots, \widehat{f}(Z_{\pi(1)}, \dots, Z_{\pi(n)}; Z_{\pi(n)})) \\ &= (\widehat{f}(Z_1, \dots, Z_n; Z_{\pi(1)}), \dots, \widehat{f}(Z_1, \dots, Z_n; Z_{\pi(n)})) \\ &= \pi G(Z). \end{aligned}$$

Applying Lemma 2.1.8 with $\pi_1 \equiv \pi_2 =: \pi$ completes the proof. □

Lemma 2.1.10 (Functions that preserve exchangeability under data splitting. Kuchibhotla 2020, Proposition 3)**.** *Let $Z_1, \dots, Z_n, Z_{n+1}$ be exchangeable random variables taking values in $\mathcal{Z}$. Consider a partition of the indices into I_1, I_2 and $\{n+1\}$ such that $I_1 \subset [n]$, $I_2 \coloneqq [n] \setminus I_1$ and, naturally, $I_1 \sqcup I_2 \sqcup \{n+1\} = [n+1]$. Set $n_1 \coloneqq \#I_1$ and $n_2 \coloneqq \#I_2$ so that $n_1 + n_2 = n$. Then, for any function $\widehat{f}_{I_1}$ that depends arbitrarily on $\{Z_i : i \in I_1\}$, the random variables*

$$\{\widehat{f}_{I_1}(Z_j) : j \in I_2 \sqcup \{n+1\}\}$$

are exchangeable.

Proof. Without loss of generality, assume that $I_1 = \{1, \dots, n_1\}$. As the random variables are exchangeable, they can be rearranged to make that true. Next, define the function $G(Z_1, \dots, Z_{n+1}) \coloneqq (\widehat{f}_{I_1}(Z_1, \dots, Z_{n_1}; Z_{n_1+1}), \dots, \widehat{f}_{I_1}(Z_1, \dots, Z_{n_1}; Z_{n+1}))$. Then, for any permutation $\pi \colon \{n_1+1, \dots, n+1\} \to \{n_1+1, \dots, n+1\}$,

$$\begin{aligned} \pi G(Z) &= (\widehat{f}_{I_1}(Z_1, \dots, Z_{n_1}; Z_{\pi(n_1+1)}), \dots, \widehat{f}_{I_1}(Z_1, \dots, Z_{n_1}; Z_{\pi(n+1)})) \\ &= G(Z_1, \dots, Z_{n_1}, \pi(Z_{n_1+1}, \dots, Z_{n+1})). \end{aligned}$$

By defining an auxiliary permutation $\pi_1 \colon [n+1] \to [n+1]$ such that $\pi_1(j) = j$ for $j \in [n_1]$ and $\pi_1(j) = \pi(j)$ for $j > n_1$, we have

$$\pi G(Z) = G(\pi_1 Z),$$

from which the results follow due to Lemma 2.1.8. □

Although not directly relevant to the theory of conformal prediction, a discussion of exchangeability would not be complete without the celebrated de Finetti's theorem, first stated for Bernoulli random variables (de Finetti 1931; de Finetti 1937) and later generalized by Hewitt and Savage (1955), relating infinite exchangeability to conditional independence.

Theorem 2.1.11 (de Finetti's theorem. Hewitt–Savage generalization.)**.** *An infinite sequence of random variables $(Z_i)_{i \in \mathbb{N}}$ is exchangeable if and only if there exists a σ-algebra conditional on which the random variables are independent and identically distributed.*

2.2 Full conformal prediction

Conformal prediction has its roots in the works by Gammerman, Saunders, Vapnik and Vovk in the late 1990s on transduction and randomness (Gammerman, Vovk, and Vapnik 1998; Saunders, Gammerman, and Vovk 1999; Vovk, Gammerman, and Saunders 1999). The method we will first describe — full conformal prediction — has historically gone by the names of *transductive confidence machines* and

transductive conformal prediction, with the latter still in use. Full CP was the bedrock for subsequent methods, such as split conformal prediction (Papadopoulos et al. 2002; Lei, Rinaldo, and Wasserman 2015), cross-conformal prediction (Vovk 2015; Vovk, Nouretdinov, Manokhin, et al. 2018), the jackknife+ (Barber et al. 2021), and many others.

Full conformal prediction is a theoretically sound methodology for generation of valid prediction sets without any assumption on the data distribution or underlying prediction model, relying exclusively on exchangeability. That is, with probability at least $1-\alpha$, for any miscoverage level α of the user's choosing, full CP provides finite-sample guarantees that a true value y will be contained in a prediction set C (Theorem 2.2.1).

Algorithm 1 describes the full conformal prediction method. Given labeled exchangeable data pairs $\{(X_i, Y_i)\}_{i=1}^n$ and a desirable nominal coverage $1-\alpha$, the objective is to generate prediction sets $C_{\text{full}}(x)$ for new unlabelled data points x such that the true response value belongs to the set with probability $1-\alpha$.

The first step is to fix an algorithm $\mathcal{A}$ that maps an arbitrary amount $m \in \mathbb{N}_{>0}$ of data pairs to a prediction function $\widehat{\mu}$ representing our trained model:

$$\mathcal{A}\colon \bigcup_{m\geq 1} (\mathcal{X}\times\mathcal{Y})^m \to \{\text{Measurable functions } \widehat{\mu}\colon \mathcal{X}\to\mathcal{Y}\}.$$

For a reason that will be made clear in Theorem 2.2.1, algorithm $\mathcal{A}$ is also required to treat data exchangeably, i.e., for any permutation $\pi\colon [m]\to[m]$,

$$\mathcal{A}(\{(x_1,y_1),\ldots,(x_m,y_m)\}) = \mathcal{A}(\{(x_{\pi(1)},y_{\pi(1)}),\ldots,(x_{\pi(m)},y_{\pi(m)})\}), \tag{2.3}$$

If the algorithm employs randomness in its procedure to create the prediction function $\widehat{\mu}$, such as neural networks and random forests, it suffices for the equality to hold in distribution:

$$\mathcal{A}(\{(x_1,y_1),\ldots,(x_m,y_m)\}) \overset{d}{=} \mathcal{A}(\{(x_{\pi(1)},y_{\pi(1)}),\ldots,(x_{\pi(m)},y_{\pi(m)})\}). \tag{2.4}$$

Then, for each new covariate $x \in \{X_{n+1}, X_{n+2}, \ldots\}$ that we wish to predict, models must be trained for all possible y values in the label space $\mathcal{Y}$, considering the original dataset $\{(X_i,Y_i)\}_{i=1}^n$ augmented by (x,y):

$$\widehat{\mu}_y = \mathcal{A}\big(\{(X_1,Y_1),\ldots,(X_n,Y_n),(x,y)\}\big),$$

such that m equals $n+1$ in this case.

A nonconformity score function $\widehat{s}_{\widehat{\mu}_y}$ is assigned to measure the discrepancy of a data point in relation to the fitted data, outputting small values when the point conforms to data and large values otherwise:

$$\widehat{s}_{\widehat{\mu}_y}\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R}.$$

In order to ease notation, we let $\widehat{s}_y$ represent $\widehat{s}_{\widehat{\mu}_y}$ from now on, with the dependence on the fitted model $\widehat{\mu}_y$ implicit.

The choice of a nonconformity score function is completely arbitrary from a theoretical point of view — all results hold independently of the chosen score function — but there are practical implications that will be discussed in Section 2.5. A common option is to use the absolute residuals, i.e., $\widehat{s}_y = |y - \widehat{\mu}_y(x)|$.

Finally, the prediction set for the new data point x is generated by taking all y values whose score

$\widehat{s}_y(x,y)$ is no larger than the adjusted $(1-\alpha)$-quantile of all scores in the original dataset $\{(X_i,Y_i)\}_{i=1}^n$:

$$C_{\text{full}}(x) = \{y \in \mathcal{Y} : \widehat{s}_y(x,y) \leq \widehat{q}_{(1+1/n)(1-\alpha)}(\{\widehat{s}_y(X_i,Y_i) : i \in [n]\})\}. \tag{2.5}$$

Algorithm 1: Full conformal prediction.

Input
Data $(X_i,Y_i) \in \mathcal{X} \times \mathcal{Y}$, for each $i \in [n]$.
Nominal coverage level $1-\alpha \in (0,1)$.
Symmetric algorithm $\mathcal{A}$.
Nonconformity score function $\widehat{s}_y$.
Test points $\mathcal{X}_{\text{new}} = \{X_{n+1}, X_{n+2}, \ldots, X_N\}$.

Procedure
for $x \in \mathcal{X}_{\text{new}}$ **do**
  for $y \in \mathcal{Y}$ **do**
    Train model $\widehat{\mu}_y = \mathcal{A}\big(\{(X_1,Y_1),\ldots,(X_n,Y_n),(x,y)\}\big)$.
    Evaluate score function $\widehat{s}_y(x,y)$. `// e.g.,` $|y - \widehat{\mu}_y(x)|$.
  end
  $C_{\text{full}}(x) = \{y \in \mathcal{Y} : \widehat{s}_y(x,y) \leq \widehat{q}_{(1+1/n)(1-\alpha)}(\{\widehat{s}_y(X_i,Y_i) : i \in [n]\})\}$.
end

Output
Prediction sets $C_{\text{full}}(x)$ for each $x \in \mathcal{X}_{\text{new}}$.

Theorem 2.2.1 (Validity of conformal sets. Vovk, Gammerman, and Shafer 2005)**.** *Given exchangeable data $\{(X_i,Y_i)\}_{i=1}^n$ and a miscoverage level $\alpha \in (0,1)$,*

$$\mathbb{P}[Y_{n+1} \in C_{\text{full}}(X_{n+1})] \geq 1-\alpha,$$

for any new exchangeable pair (X_{n+1},Y_{n+1}), with prediction set C_{full} constructed as in Equation (2.5).

Proof. Let $\widehat{s}_j \coloneqq \widehat{s}_{Y_j}(X_j,Y_j)$ be the nonconformity score evaluated at (X_j,Y_j) for $j \in [n+1]$. It follows from the construction of the prediction set that

$$Y_{n+1} \in C_{\text{full}}(X_{n+1}) \iff \widehat{s}_{n+1} \leq \widehat{q}_{(1+1/n)(1-\alpha)}(\widehat{s}_{1:n}).$$

Since data is assumed exchangeable and we required algorithms to be symmetric (Equations (2.3) and (2.4)), nonconformity scores inherit the exchangeability property, as demonstrated by Lemma 2.1.9. Then, applying Lemma 2.1.6 yields

$$\mathbb{P}[\widehat{s}_{n+1} \leq \widehat{q}_{(1+1/n)(1-\alpha)}(\widehat{s}_{1:n})] \geq 1-\alpha,$$

which implies

$$\mathbb{P}[Y_{n+1} \in C_{\text{full}}(X_{n+1})] \geq 1-\alpha. \qquad \square$$

Theorem 2.2.2 (Anti-conservativeness of conformal sets. Lei, G'Sell, et al. 2018, Theorem 2.1)**.** *If the nonconformity scores are almost surely distinct, then the prediction set can be upper bounded:*

$$\mathbb{P}[Y_{n+1} \in C_{\text{full}}(X_{n+1})] \leq 1-\alpha+\frac{1}{n+1}.$$

Proof. Applying Lemma 2.1.6 as in Theorem 2.2.1 but with the further assumption that nonconformity scores are almost surely distinct gives us

$$\mathbb{P}[\widehat{s}_{n+1} \leq \widehat{q}_{(1+1/n)(1-\alpha)}(\widehat{s}_{1:n})] \leq 1 - \alpha + \frac{1}{n+1},$$

which implies

$$\mathbb{P}[Y_{n+1} \in C_{\text{full}}(X_{n+1})] \leq 1 - \alpha + \frac{1}{n+1}. \qquad \square$$

Remark 2.2.3. *The condition that nonconformity scores are almost surely distinct in Theorem 2.2.2 and Lemma 2.1.6 could be accomplished by assuming the joint distribution of the scores is continuous. Alternatively, Tibshirani, Barber, et al. (2019) mention that if a random tie-breaking rule is employed, results hold in general without any such assumption.*

Perhaps the most glaring shortcoming of full conformal prediction is the computational cost: the underlying model must be retrained for every single prediction point x and possible label y. For classification tasks (discrete label space $\mathcal{Y}$), the problem can soon become too resource intensive as the number of classes increases. For regression tasks (continuous label space $\mathcal{Y}$), it is impossible to employ Algorithm 1 exactly and an approximation must be made by taking a fine grid of $\mathcal{Y}$, which makes the procedure cumbersome and nonviable in many situations[1]. Although there are uses for full CP, it can be computationally intractable or prohibitively expensive for many real-world problems.

Attempts have been made to remedy this resource inefficiency of full conformal prediction. Burnaev and Vovk (2014) developed a procedure to construct prediction sets efficiently for linear and ridge regression. Lei (2019) provided an exact and tractable conformalization of the lasso and the elastic net. Ndiaye and Takeuchi (2019) introduced an approximate homotopy algorithm capable of wrapping a wider class of regressors than previous approaches. Abad et al. (2022) used influence functions to efficiently approximate full conformal prediction, but focused solely on classification tasks.

Another shortcoming of full CP lies in the fact that nonconformity scores are calculated in-sample, so a sufficiently complex model may interpolate all training data points and scores would be equal to zero. Consequently, prediction sets would be completely uninformative and of no practical use.

Fortunately, full conformal's main drawbacks can be solved without parting ways with the method. Although full CP may be impractical in its most general form (Algorithm 1), an ingenious split of the data to pretrain a fixed model simplifies the algorithm and reduces computational costs tremendously. Due to its usefulness, this particular case of full CP received a name of its own. Split conformal prediction (Papadopoulos et al. 2002; Lei, Rinaldo, and Wasserman 2015; Lei, G'Sell, et al. 2018) is a computationally efficient method for conformal inference that relies on splitting data into a training set and a calibration set. By calculating nonconformity scores on out-of-sample data, the calibration set, it also overcomes the aforementioned issue with interpolating algorithms. The simplicity and effectiveness of split CP makes it one of the most successful and widely used conformal prediction methods.

2.3 Split conformal prediction

Data splitting in statistics can be traced back to at least Larson (1931), who observed that the coefficient of multiple correlation used to shrink from in-sample predictions to out-of-sample predictions. Moran (1973) gives one of the earliest accounts of data splitting for a different purpose: by partitioning a

[1]Chen, Chun, and Barber (2018) showed that coverage guarantees can be attained after discretization, so taking a fine grid of $\mathcal{Y}$ does not invalidate conformal prediction.

dataset into two disjoint samples, the first sample could be used to select a test statistic for hypothesis testing, while the second sample could serve to assess statistical significance.

Split conformal prediction was initially developed under the name of inductive conformal prediction (Papadopoulos et al. 2002; Vovk, Gammerman, and Shafer 2005) in the online learning literature and later presented in the statistics community as split CP (Lei, Rinaldo, and Wasserman 2015; Lei, G'Sell, et al. 2018). It is a particular case of full CP and follows as a natural simplification when one considers data splitting. Suppose the model $\widehat{\mu}_y$ in Algorithm 1 were pretrained in a subset $I_{\text{train}} \subset [n]$ and for each $y \in \mathcal{Y}$ the model ignored further input and simply remained unchanged. Evidently, both the model $\widehat{\mu}_y$ and the nonconformity score function $\widehat{s}_y$ would become independent of y. The model would be better named $\widehat{\mu}_{\text{train}}$ and the score function $\widehat{s}_{\text{train}}$ to emphasize they were built from I_{train} and then fixed. The last adjustment to be made is on the quantile, that should be calculated over $I_{\text{cal}} \coloneqq [n] \setminus I_{\text{train}}$ given that the training set's sole purpose was already fulfilled and the set thereafter discarded. Note that after these modifications, a test point x has no effect on the procedure so far. Therefore, instead of individual prediction sets as in the case of full conformal, we are now able to define a prediction band $C_{\text{split}} \subseteq \mathcal{X} \times \mathbb{R}$ valid for the entire feature space. Algorithm 2 gives a standalone presentation of split CP.

Algorithm 2: Split conformal prediction.

Input

Training indices I_{train}.
Calibration indices I_{cal}.
Data $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, for each $i \in I_{\text{train}} \sqcup I_{\text{cal}}$.
Nominal coverage level $1 - \alpha \in (0, 1)$.
Algorithm $\mathcal{A}$.

Procedure

Train model $\widehat{\mu} = \mathcal{A}\big(\{(X_i, Y_i) : i \in I_{\text{train}}\}\big)$.
Set nonconformity score function $\widehat{s}_{\text{train}}(x, y)$. `// e.g.,` $|y - \widehat{\mu}(x)|$.
Calculate quantile $d = \widehat{q}_{(1+1/n_{\text{cal}})(1-\alpha)}(\{\widehat{s}_{\text{train}}(X_i, Y_i) : i \in I_{\text{cal}}\})$.

Output

Prediction band $C_{\text{split}}(x) = \{y \in \mathcal{Y} : \widehat{s}_{\text{train}}(x, y) \leq d\}$, for all $x \in \mathcal{X}$.

We have shown that split CP is a special case of full CP and how to go from the latter to the former. For the sake of completeness and self-containment, we now describe split conformal prediction (Algorithm 2) on its own, as done in Section 2.2.

Given unlabeled data $\{(X_i, Y_i)\}_{i=1}^n$, the procedure starts by splitting it into two disjoint sets, I_{train} and I_{cal}, such that $I_{\text{train}} \sqcup I_{\text{cal}} = [n]$. Let $n_{\text{train}} \geq 1$ denote the cardinality of set I_{train} and $n_{\text{cal}} \geq 1$ the cardinality of set I_{cal}, with $n_{\text{train}} + n_{\text{cal}} = n$, evidently. We proceed by selecting an algorithm $\mathcal{A}$ that maps the training data $\{(X_i, Y_i)\}_{i \in I_{\text{train}}}$ to a prediction function $\widehat{\mu}$:

$$\mathcal{A}\colon (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}} \mapsto \widehat{\mu}\colon \mathcal{X} \to \mathcal{Y}.$$

A model is then trained one single time on I_{train},

$$\widehat{\mu} = \mathcal{A}\big(\{(X_i, Y_i) : i \in I_{\text{train}}\}\big),$$

and a nonconformity score function $\widehat{s}_{\text{train}}\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that depends solely on training data is assigned to measure the discrepancy of new data pairs (x, y).

Applying Lemma 2.1.10 with I_1 set to I_{train} and $\widehat{f}_{I_1}$ set to $\widehat{s}_{\text{train}}$ shows that the nonconformity scores preserve exchangeability. Note that full CP's requirement of a symmetric algorithm is relaxed for split CP and the algorithm $\mathcal{A}$ can treat data arbitrarily. This can be useful for time series in general, and financial time series in particular, if one wants to increase the weight of more recent observations during training, under the premise that the distant past is less important than recent events.

The adjusted $(1-\alpha)$-quantile of the trained score function evaluated over the calibration set I_{cal} can then be calculated

$$d = \widehat{q}_{(1+1/n_{\text{cal}})(1-\alpha)}(\{\widehat{s}_{\text{train}}(X_i, Y_i) : i \in I_{\text{cal}}\}).$$

Now, it is possible to create valid prediction sets for any $x \in \mathcal{X}$ via

$$C_{\text{split}}(x) = \{y \in \mathcal{Y} : \widehat{s}_{\text{train}}(x, y) \leq d\}. \tag{2.6}$$

Note that the model was trained only once for the entire procedure and the resulting prediction band C_{split} can be used for any new prediction, without the need to retrain the model, making split CP highly computationally efficient.

Theorem 2.3.1 (Validity of conformal sets. Vovk, Gammerman, and Shafer 2005)**.** *Given exchangeable data $\{(X_i, Y_i)\}_{i=1}^n$ and a miscoverage level $\alpha \in (0, 1)$,*

$$\mathbb{P}[Y_{n+1} \in C_{\text{split}}(X_{n+1})] \geq 1 - \alpha,$$

for any new exchangeable pair (X_{n+1}, Y_{n+1}), with prediction band C_{split} constructed as in Equation (2.6)*.*

Proof. The result follows trivially once we cast split conformal prediction as a particular case of full conformal prediction, as done above, and apply Theorem 2.2.1. □

Theorem 2.3.2 (Anti-conservativeness of conformal sets. Lei, G'Sell, et al. 2018, Theorem 2.2)**.** *If the nonconformity scores are almost surely distinct, then the prediction set can be upper bounded:*

$$\mathbb{P}[Y_{n+1} \in C_{\text{split}}(X_{n+1})] \leq 1 - \alpha + \frac{1}{n_{\text{cal}} + 1}.$$

Proof. Applying Theorem 2.2.2 proves the anti-conservativeness of split CP's conformal sets, since split conformal prediction is a particular case of full conformal prediction. □

Remark 2.3.3. *Choosing how to split available data into training and calibration sets may be nontrivial. On the one hand, more training data usually translates to more accurate base models and to smaller prediction intervals, which is desirable in practice. On the other hand, the coverage of conformal intervals holds on average over the randomness of the calibration set, so more calibration data should lead to a coverage distribution more concentrated around $1 - \alpha$ (Angelopoulos and Bates 2021). When data is abundant, there is not much concern about data splitting. However, if data is not plentiful, one should carefully consider the trade-offs when defining the sizes of training and calibration sets. In case data is really scarce, it makes sense to assess the viability of using full conformal prediction instead, as it avoids data splitting and becomes less computationally intensive the less data there is. A rule of thumb given by Angelopoulos and Bates (2021) states that 1000 calibration points should usually suffice.*

2.4 Conformalized quantile regression

Before the advent of finite-sample valid distribution-free prediction through CP, specific methodologies existed for generating prediction intervals. Quantile regression, pioneered by Koenker and Bassett (1978) in the form of quantile least squares, aims to estimate conditional quantiles instead of the conditional mean of the response variable. Besides being more robust to outliers, the flexibility in estimating as many quantiles as desired in $(0, 1)$ allows one to achieve not only a measure of tendency, but also of dispersion. Furthermore, constructing prediction intervals may be as easy as running quantile regression for a lower quantile ϕ_{lo} and a higher quantile ϕ_{hi} and setting the former's estimate as the lower bound and the latter's as the upper bounds. Depending on the quantile estimator, the lower-quantile estimate may end up being higher than that of the higher-quantile, an unfortunate phenomenon called quantile crossing (Bassett Jr and Koenker 1982). However, there are techniques to deal with this situation and there also exists models that are not susceptible to it. Compared to split conformal prediction, quantile regression's main disadvantage lies on the fact that coverage validity is guaranteed only asymptotically and for specific models. One important property of quantile regression, specially for heteroscedastic data, is that intervals can vary in length depending on the covariate.

Conformalized quantile regression (CQR) is a particular case of split CP that leverages the best properties of quantile regression and conformal prediction to yield finite-sample valid distribution-free prediction intervals of variable length. Intuitively, easy predictions should enjoy shorter intervals, while harder predictions should accompany higher uncertainty, hence larger intervals. Other nonconformity scores, such as weighted regression residuals (Lei, G'Sell, et al. 2018), also achieve locally adaptive intervals, but they separately estimate the mean absolute deviation $|y - \widehat{\mu}(x)|$ and the conditional mean. In contrast, CQR builds on the idea that estimating quantiles instead is a natural and more direct way of producing variable-length prediction intervals.

Although full CP and split CP work for many learning tasks, such as classification and regression, we will focus on regression from now on through CQR. Algorithm 3 describes conformalized quantile regression. Procedurally, it is split conformal prediction with a specific nonconformity score and two base models instead of one. Therefore, all theorems remain valid since CQR is a particular case of split CP. First, an interval without finite-sample guarantees is generated via quantile regression models $\widehat{\mu}_{\text{lo}}$ and $\widehat{\mu}_{\text{hi}}$, where the subscripts indicate the former model estimates a lower quantile and the latter a higher quantile. Then, by making use of a specially crafted nonconformity score function, the plug-in prediction interval error,

$$\widehat{s}_{\text{train}}(x, y) = \max\{\widehat{\mu}_{\text{lo}}(x) - y, y - \widehat{\mu}_{\text{hi}}(x)\},$$

the interval is conformalized, yielding all guarantees. The resulting prediction intervals are locally adaptive due to quantile regression underlying models and the finite-sample coverage guarantees are valid due to CP.

Note that the quantile levels ϕ_{lo} and ϕ_{hi} could be taken as any value between zero and one. Although a natural choice would be to set them to $\alpha/2$ and $1 - \alpha/2$, that is not a requirement. Indeed, one could even treat the base models' quantiles as hyperparameters and aim to minimize the size of prediction intervals.

CQR can be used with any base model that estimates quantiles. We will consider in Chapter 6 the following models:

- Quantile regression forests (Meinshausen 2006)

Algorithm 3: Conformalized quantile regression.

Input

Training indices I_{train}.
Calibration indices I_{cal}.
Data $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, for each $i \in I_{\text{train}} \sqcup I_{\text{cal}}$.
Nominal coverage level $1 - \alpha \in (0, 1)$.
Quantile regression algorithm $\mathcal{A}_\phi$.
Quantiles ϕ_{lo} and ϕ_{hi}.

Procedure

Train low-quantile model $\widehat{\mu}_{\text{lo}} = \mathcal{A}_{\phi_{\text{lo}}}\big(\{(X_i, Y_i) : i \in I_{\text{train}}\}\big)$.
Train high-quantile model $\widehat{\mu}_{\text{hi}} = \mathcal{A}_{\phi_{\text{hi}}}\big(\{(X_i, Y_i) : i \in I_{\text{train}}\}\big)$.
Set nonconformity score function $\widehat{s}_{\text{train}}(x, y) = \max\{\widehat{\mu}_{\text{lo}}(x) - y, y - \widehat{\mu}_{\text{hi}}(x)\}$.
Calculate quantile $d = \widehat{q}_{(1+1/n_{\text{cal}})(1-\alpha)}(\{\widehat{s}_{\text{train}}(X_i, Y_i) : i \in I_{\text{cal}}\})$.

Output

Prediction band $C_{\text{cqr}}(x) = \{y \in \mathcal{Y} : \widehat{s}_{\text{train}}(x, y) \leq d\}$, for all $x \in \mathcal{X}$.

  - Same training as usual random forest algorithm;
  - At inference, calculates weighted quantiles on the ensemble of all predicted leafs instead of taking the average;
  - Guaranteed not to suffer from quantile crossing.

- Quantile k-nearest neighbors
  - Same training as usual k-nearest neighbors;
  - Quantiles calculated at prediction time;
  - No crossing issues.

- Linear quantile regression
  - Linear regression that minimizes the pinball loss (Koenker and Bassett 1978);
  - May suffer from quantile crossing.

- Gradient boosting quantile regressor
  - Gradient Boosting that minimizes the pinball loss;
  - May suffer from quantile crossing.

- Neural network quantile regressor
  - Neural Network that minimizes the pinball loss;
  - Quantile crossing can be dealt with though an "uncrossing" layer.

Some quantile models may achieve validity under regularity and asymptotic conditions. We emphasize conformal prediction achieves model-agnostic finite-sample validity.

2.5 Nonconformity scores

The choice of nonconformity scores is completely arbitrary and all coverage guarantees should hold regardless of the score. Therefore, *validity* of conformal prediction is score-agnostic.

However, the size of prediction sets is highly dependent on the nonconformity score. This property is known as *efficiency*: a task is deemed efficient if the prediction sets are reasonably small.

Some examples of scores that have been considered in the literature include:

- *Regression residuals (Lei, G'Sell, et al. 2018):* here $\mathcal{Y} = \mathbb{R}$ and $\widehat{s}_{\text{train}}(x, y) = |y - \widehat{\mu}(x)|$, where $\widehat{\mu}$ is some regression model trained on $(X_i, Y_i)_{i \in I_{\text{train}}}$;
- *Weighted regression residuals (Lei, G'Sell, et al. 2018):* $\widehat{s}_{\text{train}}(x, y) = |y - \widehat{\mu}(x)|/\widehat{\rho}(x)$ where $\widehat{\rho}$ is an estimate of the mean absolute deviation $|y - \widehat{\mu}(x)|$ that was learned from the training data;
- *Increasing sets (Hechtlinger, Póczos, and Wasserman 2019; Angelopoulos, Bates, et al. 2021):* in classification tasks, learn from the training data a map

$$(x, t) : \mathcal{X} \times [0, 1] \mapsto \widehat{S}_{\text{train}}(x; t) \subset \mathcal{Y}$$

  where $\widehat{S}_{\text{train}}(x; t)$ increases with t and $\widehat{S}_{\text{train}}(x; 1) = \mathcal{Y}$. Then take:

$$\widehat{s}_{\text{train}}(x, y) \coloneqq \inf\{t \in [0, 1] \, : \, y \in \widehat{S}_{\text{train}}(x; t)\}.$$

- *Plug-in prediction interval error (Romano, Patterson, and Candès 2019):* given $\widehat{\mu}_\phi$ any regression model trained to estimate the conditional ϕ-quantile, set $\widehat{s}_{\text{train}}(x, y) = \max\{\widehat{\mu}_{\alpha/2}(x) - y, y - \widehat{\mu}_{1-\alpha/2}(x)\}$.

Given that validity is automatically achieved for conformal prediction, the main concern of practical deployment is to generate small prediction intervals that are informative. The choice of nonconformity score being crucial towards this goal points to the need of careful consideration and evaluation of different scores. For finance problems, a good starting point is CQR due to its local adaptiveness, capable of dealing with heteroscedasticity. However, nonconformity scores tailored exactly to the problem at hand can be developed in search of smaller intervals.

2.6 Beyond marginal coverage

In this section, we discuss possible shortfalls of marginal coverage guarantees from Theorems 2.2.1 and 2.3.1 and present results from the literature that might be of interest, in particular to finance. Up until now, we have been dealing with coverages of the form

$$\mathbb{P}[Y_{n+1} \in C(X_{n+1})] \geq 1 - \alpha, \tag{2.7}$$

for a given coverage level $1 - \alpha$ and test pair (X_{n+1}, Y_{n+1}). Notice, however, that this is a marginal probability statement, holding on average over the randomness of all $n + 1$ points. Moreover, one single new test point is addressed at a time, even though data may come in batches, in which case guaranteeing coverage for the entire test set could be of interest. Conditional coverage will tackle the former limitation and empirical coverage will tackle the latter.

2.6.1 Conditional coverage

Although marginal coverage guarantees (Equation (2.7)) are of great practical value in general, they can come short on situations that require finer-grained control over specific points. Consider a bank interested in generating prediction intervals for the future return of commodity contracts. A set of informative covariates is defined and a coverage level $1-\alpha$ that adheres to the bank's policy is chosen. If the bank also requires incurred risk to be constant across assets, markets or time periods, conformal prediction's marginal coverage would be insufficient. As an example, assume $1-\alpha = 0.95$ and market volatility is one of the features used for prediction. It might be the case that the market has high volatility 10% of the time on average and the other 90% is less volatile. Equation (2.7) would still be valid if coverage were of 50% during high-volatility periods and 100% otherwise: $1 \cdot 0.9 + 0.5 \cdot 0.1 = 0.95$. It is clear, however, that the bank is taking much more risk during turbulent periods if coverage is assumed to always be of 95%. The ultimate goal would be to develop prediction intervals with pointwise coverage guarantees for every possible x,

$$\mathbb{P}\left[Y_{n+1} \in C_{1-\alpha}(X_{n+1}) \mid X_{n+1} = x\right] \geq 1-\alpha, \tag{2.8}$$

but this is unachievable in general, as shown by Barber et al. (2020). Instead, we will focus on conditional coverage of the form

$$\mathbb{P}\left[Y_{n+1} \in C_{1-\alpha}(X_{n+1}; K) \mid X_{n+1} \in K\right] \geq 1-\alpha, \tag{2.9}$$

where K is a set large enough for theory to hold but small enough to be informative and $C_{1-\alpha}(X_i; K)$ indicates that calibration data was conditioned to K during the construction of prediction sets, with training data still being used unconditionally for training the algorithm.

Evidently, Equation (2.8) implies Equation (2.9) which implies Equation (2.7). Therefore, conditional coverage may be thought as a stronger notion of coverage.

2.6.2 Empirical coverage

A limitation of both marginal and conditional guarantees presented so far is that they hold for a single test point at a time, while in practice a batch of testing data may be made available. In this scenario, we would be interested in marginal coverages of the form

$$\mathbb{P}\left[\frac{1}{n_{\text{test}}} \sum_{i \in I_{\text{test}}} \mathbb{1}\{Y_i \in C_{1-\alpha}(X_i)\} \geq 1-\alpha\right] \geq 1-\delta, \tag{2.10}$$

and, taking $I_{\text{test}}(K)$ as the subset of the original test set such that the covariates belong to K and $n_{\text{test}}(K)$ as the cardinality of $I_{\text{test}}(K)$, we would be interested in conditional coverages of the form

$$\mathbb{P}\left[\inf_{K \in \mathcal{K}} \frac{1}{n_{\text{test}}(K)} \sum_{i \in I_{\text{test}}(K)} \mathbb{1}\{Y_i \in C_{1-\alpha}(X_i; K)\} \geq 1-\alpha\right] \geq 1-\delta. \tag{2.11}$$

Lei, G'Sell, et al. (2018) have proven that Equation (2.10) holds for iid data. Their proof relies on McDiarmid's inequality (Theorem 3.1.14), so independence is needed. On the other hand, Oliveira et al. (2022) have shown Equation (2.11) to be valid under their novel approach to conformal prediction, under much milder assumptions.

Chapter 3

A Concentration of Measure Approach to Split Conformal Prediction

Conformal prediction as presented in Chapter 2 crucially relies on exchangeability of the data. Theorems 2.2.1 and 2.2.2 and consequently Theorems 2.3.1 and 2.3.2 all break down if data is nonexchangeable. Empirical coverage guarantees from Equations (2.10) and (2.11) even require the stronger assumption of independent and identically distributed data.

Methodological work has been done to deal with lack of exchangeability, mainly due to nonstationarity and to a lesser degree due to dependence. Gibbs and Candès (2021) propose an adaptive algorithm with no distributional assumptions, but quite different to traditional conformal prediction. Tibshirani, Barber, et al. (2019) deal with nonstationarity due to covariate shift and Barber et al. (2022) develop a new CP method to handle more general cases of nonexchangeability. Chernozhukov, Wüthrich, and Zhu (2018) and Xu and Xie (2021) treat time series specifically, but also deviate significantly from standard CP algorithms.

On the other hand, Oliveira et al. (2022) developed a different approach to theoretically analyze conformal prediction, based on concentration of measure and decoupling inequalities instead of exchangeability, that is applicable more generally, with exchangeable data being a particular case. Their approach allows one to use split conformal prediction exactly as is, without any methodological modification, and retain all desired coverage guarantees upon the addition of a small penalty.

In this chapter, we will introduce the conformal prediction framework of Oliveira et al. (2022). We start with a primer on concentration of measure results that underpin their novel conformal prediction framework. All proofs for Sections 3.3, 3.4 and 3.5 are omitted and can be found in Oliveira et al. (2022).

3.1 Preliminaries

Large deviation theory seeks to control the probability of a random variable Z deviating from its mean $\mathbb{E}[Z]$ or some other measure of central tendency, such as the median. Formally, for $\delta \in (0,1)$ and suitable $\varepsilon \in \mathbb{R}_{\geq 0}$, one seeks bounds of the form

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq \varepsilon] \leq \delta. \tag{3.1}$$

Small deviation theory seeks to control the probability of Z being very small. For a given ε, the

objective is to find bounds of the form

$$\mathbb{P}[|Z| \geq \varepsilon] \leq \delta. \tag{3.2}$$

Probabilities such as those from Equation (3.1) are also known as *tail probabilities*, while those similar to Equation (3.2) are sometimes called *small ball probabilities*.

Concentration of measure phenomenon was observed early by Levy, who studied a concentration property on the sphere which could be described equivalently on functions. Milman popularized the concept in his investigation of asymptotic geometric analysis.

A cornerstone of finite-sample probability, statistics and machine learning theory, concentration of measure inequalities are also useful in many contexts, ranging from combinatorics, functional analysis, information theory, geometry and statistical physics, to name a few.

We present below some classical inequalities, following the treatment from Boucheron, Lugosi, and Massart (2013), with proofs oftentimes deliberately more detailed to cater to a wider audience. Azuma's (Theorem 3.1.12) and McDiarmid's (Theorem 3.1.14) inequalities, however, are based on Mohri, Rostamizadeh, and Talwalkar (2018), with an alternative proof of McDiarmid's via entropy provided in Appendix B.

We consider the most interesting case of distribution-free inequalities, which do not assume the random variables follow any specific distribution. Some requirements may be needed in terms of finite moments, but nothing about the distribution per se.

Theorem 3.1.1 (Markov's inequality)**.** *For any nonnegative random variable Z and $t \in \mathbb{R}_{>0}$,*

$$\mathbb{P}[Z \geq t] \leq \frac{\mathbb{E}[Z]}{t}.$$

Proof. We start with the observation that

$$Z\,\mathbb{1}\{Z \geq t\} \geq t\,\mathbb{1}\{Z \geq t\}$$

and take expectations:

$$\begin{aligned}\mathbb{E}[Z\,\mathbb{1}\{Z \geq t\}] &\geq \mathbb{E}[t\,\mathbb{1}\{Z \geq t\}] \\ &= t\mathbb{E}[\mathbb{1}\{Z \geq t\}] \\ &= t\mathbb{P}[Z \geq t].\end{aligned}$$

Rearranging terms,

$$\mathbb{P}[Z \geq t] \leq \frac{\mathbb{E}[Z\,\mathbb{1}\{Z \geq t\}]}{t}.$$

It is easy to see that $Z\,\mathbb{1}\{Z \geq t\} \leq Z$, with equality holding only when Z is greater than t, concluding the proof:

$$\mathbb{P}[Z \geq t] \leq \frac{\mathbb{E}[Z]}{t}.$$ □

Theorem 3.1.2 (Generalized Markov's inequality)**.** *Let ϕ be a nondecreasing and nonnegative function defined on $S \subseteq \mathbb{R}$. For any random variable Z taking values in S and for every $t \in S$ with $\phi(t) \in \mathbb{R}_{>0}$,*

$$\mathbb{P}[Z \geq t] \leq \mathbb{P}[\phi(Z) \geq \phi(t)] \leq \frac{\mathbb{E}[\phi(Z)]}{\phi(t)}.$$

Proof. Given that the random variable $\phi(Z)$ is nonnegative and $\phi(t) \in \mathbb{R}_{>0}$, Markov's inequality can

be applied, yielding

$$\mathbb{P}[\phi(Z) \geq \phi(t)] \leq \frac{\mathbb{E}[\phi(Z)]}{\phi(t)}. \qquad \square$$

Noting that $\mathbb{P}[Z \geq t] \leq \mathbb{P}[\phi(Z) \geq \phi(t)]$ due to the nondecreasing and nonnegative properties of ϕ conclude the proof.

Theorem 3.1.3 (Chebyshev's inequality)**.** *For any random variable Z and $t \in \mathbb{R}_{>0}$,*

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t] \leq \frac{\mathrm{Var}[Z]}{t^2}.$$

Proof. Consider the random variable $|Z - \mathbb{E}[Z]|$ and the function $\phi(t) = t^2$ defined on $\mathbb{R}_{>0}$. Applying the generalized Markov's inequality from Theorem 3.1.2 gives

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t] \leq \frac{\mathbb{E}[|Z - \mathbb{E}[Z]|^2]}{t^2}.$$

However, the definition of variance is precisely $\mathrm{Var}[Z] = \mathbb{E}[(Z - \mathbb{E}[Z])^2]$, so

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t] \leq \frac{\mathrm{Var}[Z]}{t^2}. \qquad \square$$

Theorem 3.1.4 (Generalized Chebyshev's inequality)**.** *For any random variable Z and $t \in \mathbb{R}_{>0}$,*

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t] \leq \frac{\mathbb{E}[|Z - \mathbb{E}[Z]|^q]}{t^q}.$$

Proof. Consider the random variable $|Z - \mathbb{E}[Z]|$ and the function $\phi(t) = t^q$ defined on $\mathbb{R}_{>0}$. Applying the generalized Markov's inequality from Theorem 3.1.2 yields the result. $\square$

Remark 3.1.5. *Suppose one wants to upper bound $\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t]$ and a tight bound is preferred over a loose one, as usual in practical applications. Instead of settling for Chebyshev's inequality from Theorem 3.1.3, one could evaluate the generalized Chebyshev's inequality from Theorem 3.1.4 at different values of q (assuming moments exist) and choose the one that achieves the tightest bound.*

Theorem 3.1.6 (Chernoff's inequality)**.** *For any real-valued random variable Z and $t \in \mathbb{R}_{>0}$,*

$$\mathbb{P}[Z \geq t] \leq \exp(-\psi_Z^*(t)),$$

where ψ_Z^* is the Cramér transform of Z (Cramér 1938), that is,

$$\psi_Z^*(t) \coloneqq \sup_{\lambda \in \mathbb{R}} (\lambda t - \log \mathbb{E}[e^{\lambda Z}]).$$

Proof. Applying the generalized Markov's inequality (Theorem 3.1.2), taking $\phi(t) = e^{\lambda t}$ with $\lambda \geq 0$, gives

$$\begin{aligned} \mathbb{P}[Z \geq t] &\leq \frac{\mathbb{E}[e^{\lambda Z}]}{e^{\lambda t}} \\ &= e^{-\lambda t} \cdot e^{\log \mathbb{E}[e^{\lambda Z}]} \\ &= \exp(-\lambda t + \log \mathbb{E}[e^{\lambda Z}]) \end{aligned} \quad .$$

As this is valid for any $\lambda \geq 0$ and the tightest bound is desirable, the λ that maximizes $\lambda t - \log \mathbb{E}[e^{\lambda Z}]$ should be chosen:

$$\mathbb{P}[Z \geq t] \leq \exp\left(-\sup_{\lambda \geq 0}(\lambda t - \log \mathbb{E}[e^{\lambda Z}])\right). \qquad \square$$

Lemma 3.1.7 (Hoeffding's lemma)**.** *Let Z be a zero-mean random variable defined on the interval $[a, b]$ and let $\psi_Z(\lambda) \coloneqq \log \mathbb{E}[e^{\lambda Z}]$. Then,*

$$\psi_Z''(\lambda) \leq \frac{(b-a)^2}{4},$$

and for every $\lambda \in \mathbb{R}$,

$$\psi_Z(\lambda) \leq \frac{\lambda^2 (b-a)^2}{8}.$$

Proof. Note that a is nonpositive, b is nonnegative and $Z \in [a, b]$, so

$$a - \frac{(b+a)}{2} \leq Z - \frac{(b+a)}{2} \leq b - \frac{(b+a)}{2} \implies \frac{a-b}{2} \leq Z - \frac{(b+a)}{2} \leq \frac{b-a}{2} \implies \left|Z - \frac{(b+a)}{2}\right| \leq \frac{b-a}{2}.$$

Squaring and then taking expectations yields

$$\mathbb{E}\left[\left|Z - \frac{(b+a)}{2}\right|^2\right] \leq \mathbb{E}\left[\left(\frac{b-a}{2}\right)^2\right] = \frac{(b-a)^2}{4}.$$

Taking expectations and then squaring, on the other hand, gives

$$\mathbb{E}\left[\left|Z - \frac{(b+a)}{2}\right|\right]^2 \leq \mathbb{E}\left[\frac{b-a}{2}\right]^2 = \frac{(b-a)^2}{4}.$$

It then follows by the definition of variance,

$$\mathrm{Var}\left[\left|Z - \frac{(b+a)}{2}\right|\right] \coloneqq \mathbb{E}\left[\left|Z - \frac{(b+a)}{2}\right|^2\right] - \mathbb{E}\left[\left|Z - \frac{(b+a)}{2}\right|\right]^2 \leq \frac{(b-a)^2}{4}.$$

Moreover, as variance is invariant under subtraction of a constant and change of sign,

$$\mathrm{Var}[Z] \leq \frac{(b-a)^2}{4}, \tag{3.3}$$

which let us conclude that any random variable taking values in a bounded interval $[a, b]$ has variance at most $\frac{(b-a)^2}{4}$. We now calculate the first and second derivatives of the moment-generating function by making use of the law of the unconscious statistician:

$$\begin{aligned} M_Z(\lambda) &\coloneqq \mathbb{E}[e^{\lambda Z}] = \int_{\mathbb{R}} e^{\lambda z} f_Z(z) dz, \\ M_Z'(\lambda) &= \int_{\mathbb{R}} z e^{\lambda z} f_Z(z) dz = \mathbb{E}[Z e^{\lambda Z}], \\ M_Z''(\lambda) &= \int_{\mathbb{R}} z^2 e^{\lambda z} f_Z(z) dz = \mathbb{E}[Z^2 e^{\lambda Z}]. \end{aligned}$$

The first and second derivatives of the logarithm of the moment-generating function $\psi_Z(\lambda)$ follow

accordingly

$$\begin{aligned}
\psi'_Z(\lambda) &= (\log M_Z)'(\lambda) = \frac{M'_Z(\lambda)}{M_Z(\lambda)}, \\
\psi''_Z(\lambda) &= \frac{M_Z(\lambda)M''_Z(\lambda) - (M'_Z(\lambda))^2}{(M_Z(\lambda))^2} \\
&= \frac{M_Z(\lambda)\mathbb{E}[Z^2e^{\lambda Z}] - \mathbb{E}[Ze^{\lambda Z}]^2}{(M_Z(\lambda))^2} \\
&= \frac{\mathbb{E}[Z^2e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \left(\frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]}\right)^2
\end{aligned}$$

We next define the exponential change of measure

$$\mathbb{E}[f(Q)] \coloneqq \frac{\mathbb{E}[f(Z)e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]}$$

and show it is also bounded in $[a, b]$:

$$\begin{aligned}
\mathbb{P}[a \le Q \le b] &= \mathbb{E}[\mathbb{1}\{a \le Q \le b\}] \\
&= \frac{\mathbb{E}[\mathbb{1}\{a \le Z \le b\}e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} \\
&= \frac{\mathbb{E}[e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} \\
&= 1.
\end{aligned}$$

Recasting ψ''_Z in terms of Q and using (3.3) to bound its variance, we have

$$\begin{aligned}
\psi''_Z(\lambda) &= \frac{\mathbb{E}[Z^2e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \left(\frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]}\right)^2 \\
&= \mathbb{E}[Q^2] - \mathbb{E}[Q]^2 \\
&= \mathrm{Var}[Q] \\
&\le \frac{(b-a)^2}{4}.
\end{aligned}$$

From Taylor's theorem, there exists $0 \le \theta \le 1$ such that

$$\psi_Z(\lambda) = \psi_Z(0) + \lambda\psi'_Z(0) + \frac{\lambda^2}{2}\psi''_Z(\theta\lambda).$$

Since $\psi_Z(0) = \log(1) = 0$ and $\psi'_Z(0) = \mathbb{E}[Z] = 0$,

$$\begin{aligned}
\psi_Z(\lambda) &= \frac{\lambda^2}{2}\psi''_Z(\theta\lambda) \\
&\le \frac{\lambda^2(b-a)^2}{8}.
\end{aligned}$$

□

Theorem 3.1.8 (Hoeffding's inequality)**.** *Let $Z_1, \ldots, Z_n$ be independent random variables such that, for $i \in [n]$, each Z_i is bounded almost surely on $[a_i, b_i]$. Define $S \coloneqq \sum_{i=1}^n (Z_i - \mathbb{E}[Z_i])$. Then, for every $t > 0$,*

$$\mathbb{P}[S \ge t] \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Proof. In order to simplify notation, let $\widetilde{Z} \coloneqq Z - \mathbb{E}[Z]$. It follows that

$$\begin{aligned}
\psi_S(\lambda) &= \log \mathbb{E}[e^{\lambda S}] \\
&= \log \mathbb{E}\left[e^{\lambda \sum_{i=1}^n \widetilde{Z}_i}\right] \\
&= \log \mathbb{E}\left[\prod_{i=1}^n e^{\lambda \widetilde{Z}_i}\right] \\
&= \log \prod_{i=1}^n \mathbb{E}\left[e^{\lambda \widetilde{Z}_i}\right] \\
&= \log \exp\left(\sum_{i=1}^n \log \mathbb{E}[e^{\lambda \widetilde{Z}_i}]\right) \\
&= \sum_{i=1}^n \log \mathbb{E}[e^{\lambda \widetilde{Z}_i}] \\
&= \sum_{i=1}^n \psi_{\widetilde{Z}_i}(\lambda),
\end{aligned}$$

where the expectation of the product could be written as the product of expectations due to independence of the random variables. Then, by Hoeffding's lemma (Lemma 3.1.7),

$$\psi_S(\lambda) \leq \frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2.$$

Now, if we follow as in the proof of Chernoff's inequality (Theorem 3.1.6) by taking $\phi(t) = e^{\lambda t}$ and applying the generalized Markov's inequality, we get

$$\mathbb{P}[S \geq t] \leq e^{-\lambda t} \mathbb{E}[e^{\lambda S}].$$

However, note that $\mathbb{E}[e^{\lambda S}] = e^{\psi_S(\lambda)}$, so

$$\begin{aligned}
\mathbb{P}[S \geq t] &\leq e^{-\lambda t} e^{\psi_S(\lambda)} \\
&\leq \exp\left(-\lambda t + \frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2\right)
\end{aligned}$$

As this is valid for any $\lambda \geq 0$, we can take $\lambda = \frac{4t}{\sum_{i=1}^n (b_i - a_i)^2}$ to conclude the proof:

$$\begin{aligned}
\mathbb{P}[S \geq t] &\leq \exp\left(-\frac{4t}{\sum_{i=1}^n (b_i - a_i)^2} t + \frac{1}{8}\left(\frac{4t}{\sum_{i=1}^n (b_i - a_i)^2}\right)^2 \sum_{i=1}^n (b_i - a_i)^2\right) \\
&= \exp\left(-\frac{4t^2}{\sum_{i=1}^n (b_i - a_i)^2} + \frac{1}{8} \frac{16t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \\
&= \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)
\end{aligned}$$

□

We now proceed to generalize Hoeffding's inequality to random variables that are not necessarily independent, but form a martingale difference sequence (Definition 3.1.10), which is a milder assumption.

Definition 3.1.9 (Martingale)**.** *A stochastic process $Z_1, Z_2, \ldots$ is a martingale with respect to another stochastic process $W_1, W_2, \ldots$ if, for all $i > 0$, the random variable Z_i is a function of $W_1, \ldots, W_i$,*

$\mathbb{E}[|Z_i|] < \infty$ *and*

$$\mathbb{E}[Z_{i+1}|W_1, \ldots, W_i] = Z_i.$$

Definition 3.1.10 (Martingale difference)**.** *A stochastic process $Z_1, Z_2, \ldots$ is a martingale difference sequence with respect to another stochastic process $W_1, W_2, \ldots$ if, for all $i > 0$, the random variable Z_i is a function of $W_1, \ldots, W_i$, $\mathbb{E}[|Z_i|] < \infty$ and*

$$\mathbb{E}[Z_{i+1}|W_1, \ldots, W_i] = 0.$$

Note that for a given martingale $Z_1, Z_2, \ldots$, the sequence $Z_2 - Z_1, Z_3 - Z_2, \ldots$ will have the properties of a martingale difference sequence, which explains its name.

Lemma 3.1.11 (Conditional Hoeffding's lemma)**.** *Let Z and W be random variables satisfying $\mathbb{E}[Z|W] = 0$ and, for some function f and constant $c \geq 0$, assume that Z takes values in the bounded interval $[f(Z), f(Z) + c]$. Then, for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[e^{\lambda Z}|W] \leq \exp\left(\frac{\lambda^2 c^2}{8}\right).$$

The proof is the same as Hoeffding's lemma (Lemma 3.1.7), taking $a \equiv f(Z)$, $b \equiv f(Z) + c$ and conditional instead of unconditional expectations.

Theorem 3.1.12 (Azuma's inequality)**.** *Let $Z_1, Z_2, \ldots$ form a martingale difference sequence with respect to the random variables $W_1, W_2, \ldots$ and assume that, for all $i > 0$, there exist a constant $c_i \geq 0$ and a function f_i such that Z_i takes values in $[f_i(W_1, \ldots, W_{i-1}), f_i(W_1, \ldots, W_{i-1}) + c_i]$. Define $S_k \coloneqq \sum_{i=1}^{k} Z_i$ for $k \in [n]$. Then, for every $t > 0$,*

$$\mathbb{P}\left[S_n \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right).$$

Proof. Applying the generalized Markov's inequality (Theorem 3.1.2) with $e^{\lambda t}$, we have, for any $\lambda \geq 0$,

$$\begin{aligned}
\mathbb{P}\left[S_n \geq t\right] &\leq e^{-\lambda t}\mathbb{E}[e^{\lambda S_n}] \\
&= e^{-\lambda t}\mathbb{E}[e^{\lambda(S_{n-1}+Z_n)}] \\
&= e^{-\lambda t}\mathbb{E}[e^{\lambda S_{n-1}} e^{\lambda Z_n}] \\
&= e^{-\lambda t}\mathbb{E}[\mathbb{E}[e^{\lambda S_{n-1}} e^{\lambda Z_n}|W_1, \ldots, W_{n-1}]] && \text{(Law of iterated expectations)} \\
&= e^{-\lambda t}\mathbb{E}[e^{\lambda S_{n-1}}\mathbb{E}[e^{\lambda Z_n}|W_1, \ldots, W_{n-1}]] && (S_{n-1} \text{ is a function of } W_1, \ldots, W_{n-1}) \\
&\leq e^{-\lambda t}\mathbb{E}[e^{\lambda S_{n-1}}]e^{\lambda^2 c_n^2/8} && \text{(Lemma 3.1.11)} \\
&\leq e^{-\lambda t}\mathbb{E}[e^{\lambda S_{n-2}}]e^{\lambda^2 c_{n-1}^2/8} e^{\lambda^2 c_n^2/8} && \text{(Iterated argument)} \\
&\leq e^{-\lambda t} e^{\lambda^2 \sum_{i=1}^{n} c_i^2/8}. && \text{(Iterated argument } n-2 \text{ more times)}
\end{aligned}$$

As this is valid for any $\lambda \geq 0$, we minimize $-\lambda t + \lambda^2 \sum_{i=1}^{n} c_i^2/8$ by selecting $\lambda = 4t/\sum_{i=1}^{n} c_i^2$, which concludes the proof by yielding

$$\mathbb{P}\left[S_n \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right). \qquad \square$$

Definition 3.1.13 (Bounded differences)**.** *A function $f\colon \mathcal{Z}^n \to \mathbb{R}$ satisfies the bounded differences condition if there exists nonnegative constants $c_1, \ldots, c_n > 0$ such that, for every $i \in [n]$, changing any single variable does not alter the value of the function by much:*

$$\sup_{z_1,\ldots,z_n,z_i' \in \mathcal{Z}} |f(z_1, \ldots, z_n) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \leq c_i.$$

McDiarmid's inequality is another generalization of Hoeffding's inequality, particularly useful in machine learning theory. While Hoeffding's bounds the sum of independent random variables, McDiarmid's bounds any function of independent random variables, as long as it satisfies the bounded differences condition.

Theorem 3.1.14 (McDiarmid's inequality)**.** *Let f be a function that satisfies the bounded differences condition (Definition 3.1.13) with constants $c_1, \ldots, c_n > 0$ and define*

$$v \coloneqq \frac{1}{4} \sum_{i=1}^{n} c_i^2.$$

For independent random variables $W_1, \ldots, W_n$, set $Z = f(W_1, \ldots, W_n)$. Then, for any $t > 0$,

$$\mathbb{P}[Z - \mathbb{E}[Z] > t] \leq \exp\left(\tfrac{-t^2}{2v}\right),$$

and, by symmetry of the bounded differences assumption,

$$\mathbb{P}[Z - \mathbb{E}[Z] < -t] \leq \exp\left(\tfrac{-t^2}{2v}\right).$$

Proof. Define $V \coloneqq Z - \mathbb{E}[Z]$, $V_1 \coloneqq \mathbb{E}[V|X_1] - \mathbb{E}[V]$ and, for $k \in \{2, \ldots, n\}$, $V_k \coloneqq \mathbb{E}[V|W_1, \ldots, W_k] - \mathbb{E}[V|W_1, \ldots, W_{k-1}]$. Note that $\sum_{k=1}^{n} V_k = \mathbb{E}[V|W_1, \ldots, W_n] - \mathbb{E}[V] = \mathbb{E}[V|W_1, \ldots, W_n] = V$ due to telescoping and the fact that V is a function of $W_1, \ldots, W_n$. Next, conditioning on $W_1, \ldots, W_{k-1}$ and taking expectation gives

$$\mathbb{E}[\mathbb{E}[V|W_1, \ldots, W_k]|W_1, \ldots, W_{k-1}] = \mathbb{E}[V|W_1, \ldots, W_{k-1}],$$

which implies

$$\mathbb{E}[\mathbb{E}[V|W_1, \ldots, W_k] - V|W_1, \ldots, W_{k-1}] = 0,$$

and, consequently,

$$\mathbb{E}[V_k|W_1, \ldots, W_{k-1}] = 0.$$

Therefore, the sequence $\{V_k\}_{k\in[m]}$ is a martingale difference with respect to $\{W_k\}_{k\in[m]}$. By noting that $\mathbb{E}[Z]$ is a scalar, we can write

$$V_k \coloneqq \mathbb{E}[Z|W_1, \ldots, W_k] - \mathbb{E}[Z|W_1, \ldots, W_{k-1}].$$

Next, we define the lower (L_k) and upper (U_k) bounds for V_k as

$$U_k \coloneqq \sup_x \mathbb{E}[Z|W_1, \ldots, W_{k-1}, x] - \mathbb{E}[Z|W_1, \ldots, W_{k-1}]$$
$$L_k \coloneqq \inf_x \mathbb{E}[Z|W_1, \ldots, W_{k-1}, x] - \mathbb{E}[Z|W_1, \ldots, W_{k-1}].$$

However, by the bounded differences assumption, for all $k \in [n]$,

$$U_k - L_k = \sup_{x,x'} \mathbb{E}[Z|W_1, \ldots, W_{k-1}, x] - \mathbb{E}[Z|W_1, \ldots, W_{k-1}, x'] \leq c_k,$$

which implies

$$L_k \leq V_k \leq L_k + c_k.$$

Finally, applying Azuma's inequality (Theorem 3.1.12) to $V = \sum_{k=1}^n V_k$ concludes the proof. □

Remark 3.1.15 (Hoeffding's inequality is a particular case of McDiarmid's inequality)**.** *As previously stated, McDiarmid's inequality generalizes Hoeffding's inequality. In fact, taking* $f\colon (x_1, \ldots, x_n) \mapsto \frac{1}{n}\sum_{i=1}^n x_i$ *as McDiarmid's bounded differences function recovers Hoeffding's.*

The next inequality we want to prove needs the following lemma.

Lemma 3.1.16 (Cramér transform of a centered Poisson random variable)**.** *A random variable* W *with probability mass function* $\mathbb{P}[W = k] = e^{-v}v^k/k!$ *for all* $k \in \mathbb{N}_{\geq 0}$ *is named a Poisson random variable with parameter* v*. The corresponding centered variable is defined as* $Z \coloneqq W - v$ *and its Cramér transform for* $t > 0$ *is given by*

$$\psi_Z^*(t) = v \cdot h(t/v),$$

where $h(u) \coloneqq (1+u)\log(1+u) - u$ *for positive* $u > 0$*.*

Proof. The moment-generating function of Z can be directly calculated:

$$\begin{aligned}
\mathbb{E}[e^{\lambda Z}] &= e^{-\lambda v}\sum_{k=0}^{\infty} e^{\lambda k} e^{-v}\frac{v^k}{k!} \\
&= e^{-\lambda v - v}\sum_{k=0}^{\infty}\frac{(ve^{\lambda})^k}{k!} \\
&= e^{-\lambda v - v}e^{ve^{\lambda}}.
\end{aligned}$$

Taking the logarithm yields

$$\begin{aligned}
\psi_Z(t) &\coloneqq \log \mathbb{E}[e^{\lambda Z}] \\
&= -\lambda v - v + ve^{\lambda} \\
&= v(e^{\lambda} - \lambda - 1).
\end{aligned}$$

The Cramér transform $\psi_Z^*(t) \coloneqq \sup_{\lambda\in\mathbb{R}}(\lambda t - \psi_Z(t))$ can then be calculated by first finding the optimal λ, which is generally given by the λ_t such that $\psi_Z'(\lambda_t) = t$. For the centered Poisson process, $\psi_Z'(\lambda_t) = (e^{\lambda_t} - 1)v$ and $(e^{\lambda_t} - 1)v = t$ implies $\lambda_t = \log(1 + t/v)$, so

$$\begin{aligned}
\psi_Z^*(t) &= \log(1+t/v)t - v(e^{\log(1+t/v)} - \log(1+t/v) - 1) \\
&= \log(1+t/v)t - v(1+t/v) + v\log(1+t/v) + v \\
&= \log(1+t/v)t - t + v\log(1+t/v) \\
&= v(\log(1+t/v)t/v - t/v + \log(1+t/v)) \\
&= v((1+t/v)\log(1+t/v) - t/v) \\
&= vh(t/v),
\end{aligned}$$

for h defined as in the lemma's statement. □

Theorem 3.1.17 (Bennett's inequality)**.** *Let $Z_1, \dots, Z_n$ be independent random variables with finite variance. Moreover, assume $Z_i \leq b$ for some positive $b > 0$ almost surely for all $i \in [n]$. Define*

$$S \coloneqq \sum_{i=1}^{n} Z_i - \mathbb{E}[Z_i],$$

and

$$v \coloneqq \sum_{i=1}^{n} \mathbb{E}[Z_i^2].$$

Letting $\phi(p) = e^p - p - 1$ for $p \in \mathbb{R}$, then

$$\forall \lambda > 0 \colon \quad \log \mathbb{E}[e^{\lambda S}] \leq n \log\left(1 + \frac{v}{nb^2} \cdot \phi(b\lambda)\right) \leq \frac{v}{b^2} \cdot \phi(b\lambda),$$

and

$$\forall t > 0 \colon \quad \mathbb{P}[S \geq t] \leq \exp\left(-\frac{v}{b^2} \cdot h\left(\frac{bt}{v}\right)\right),$$

where $h(u) = (1+u)\log(1+u) - u$ for positive $u > 0$.

Proof. Due to homogeneity of the inequalities, we may assume without loss of generality that $b = 1$. We then notice that $\frac{\phi(u)}{u^2}$ is nondecreasing on the real line, which implies for all $\lambda > 0$ and $i \in [n]$,

$$\frac{\phi(\lambda Z_i)}{\lambda^2 Z_i^2} \leq \frac{\phi(\lambda)}{\lambda^2} \implies \phi(\lambda Z_i) \leq \phi(\lambda) Z_i^2 \implies e^{\lambda Z_i} - \lambda Z_i - 1 \leq (e^\lambda - \lambda - 1) Z_i^2,$$

since $Z_i \leq 1$. Taking expectations and rearranging,

$$\mathbb{E}[e^{\lambda Z_i}] \leq \mathbb{E}[Z_i^2]\phi(\lambda) + \lambda \mathbb{E}[Z_i] + 1.$$

We may now take the logarithm and sum all resulting inequalities for $i \in [n]$:

$$\sum_{i=1}^{n} \log \mathbb{E}[e^{\lambda Z_i}] \leq \sum_{i=1}^{n} \log\left(\mathbb{E}[Z_i^2]\phi(\lambda) + \lambda \mathbb{E}[Z_i] + 1\right).$$

Recall that $\psi_S(\lambda)$ can be written as $\sum_{i=1}^n \log \mathbb{E}[e^{\lambda Z_i}] - \lambda \mathbb{E}[Z_i]$, so subtracting $\lambda \mathbb{E}[Z_i]$ yields

$$\psi_S(\lambda) \leq \sum_{i=1}^{n} \log\left(\mathbb{E}[Z_i^2]\phi(\lambda) + \lambda \mathbb{E}[Z_i] + 1\right) - \lambda \mathbb{E}[Z_i].$$

Now, consider that the logarithm function is concave, i.e., $\log((1-c)x + cy) \geq (1-c)\log(x) + c\log(y)$ for any $x, y \in \mathbb{R}_{>0}$ and $c \in [0,1]$. In particular, $\log(x) + \log(y) \leq 2\log((x+y)/2)$, which readily generalizes to $\sum_{i=1}^n \log(z_i) \leq n \log(\sum_{i=1}^n z_i / n)$. Therefore,

$$\psi_S(\lambda) \leq n \left(\log\left(\frac{v}{n}\phi(\lambda) + \lambda \frac{\sum_{i=1}^n \mathbb{E}[Z_i]}{n} + 1\right) - \lambda \frac{\sum_{i=1}^n \mathbb{E}[Z_i]}{n}\right).$$

Using the fact that $\log(1+u) \leq u$,

$$\psi_S(\lambda) \leq v\phi(\lambda),$$

which proves the theorem's first inequality.

Now, note that $v\phi(\lambda) = v(e^\lambda - \lambda - 1)$ is the logarithm of the moment-generating function of a centered Poisson random variable and taking Cramér transforms yield (cf. Lemma 3.1.16)

$$\psi_S^*(t) \geq vh(t/v).$$

Then, by Chernoff's inequality (Theorem 3.1.6),

$$\begin{aligned}\mathbb{P}[S \geq t] &\leq \exp(-\psi_S^*(t)) \\ &\leq \exp(-v \cdot h(t/v)).\end{aligned}$$ □

Bernstein's inequality can then be derived from Bennett's inequality, relying on the fact that, for every $z > 0$,

$$\log(z) \leq \frac{(z-1)(z+5)}{4z+2}, \tag{3.4}$$

which follows straightforwardly: $(z-1)^3 + 1 \geq 1 \implies z^3 - 3z^2 + 3z \geq 1 \implies z^3 + 3z \geq 3z^2 + 1 \implies z^3 + z^2 + 3z \geq 4z^2 + 1 \implies z^3 + z^2 + 7z \geq 4z^2 + 4z + 1 \implies z^3 + z^2 + 7z \geq (2z+1)^2 \implies \frac{z^2+z+7}{(2z+1)^2} \geq \frac{1}{z} \implies \int_1^z \frac{t^2+t+7}{(2t+1)^2}dt \geq \int_1^z \frac{1}{t}dt \implies \frac{(z-1)(z+5)}{4z+2} \geq \log(z)$.

This bound can also be regarded as a good approximation of the logarithm. In fact, it is a *Padé approximant* of the logarithmic function.

Definition 3.1.18 (Padé approximants)**.** *Given a function f with power series representation $f(z) = \sum_{i=0}^{\infty} c_i z^i$, the $[L/M]$ Padé approximant of $f(z)$ is the rational function*

$$[L/M]_f(z) \coloneqq \frac{a_0 + a_1 z + \ldots + a_L z^L}{b_0 + b_1 z + \ldots + b_M z^M},$$

such that $b_0 \equiv 1$ and

$$[L/M]_f(z) - f(z) = O(z^{L+M+1}),$$

which means that the Maclaurin expansion of $[L/M]_f(z)$ agrees with $f(z)$ as far as possible.

The function is named due to Padé (1892), with preceding work by Frobenius (1881) and Jacobi (1846). Baker and Graves-Morris (1996) note that some Padé approximants for the logarithmic function appeared in a letter of Anderson (1740).

Example 5 (A Padé approximant for the logarithmic function)**.** *The natural logarithm of $x+1$ has a power series representation for $-1 < x \leq 1$ given by the Newton-Mercator series:*

$$\log(x+1) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} x^n}{n}.$$

Considering a [2/1] Padé approximant of the form

$$\frac{a_0 + a_1 z + a_2 z^2}{1 + b_1 z},$$

and solving the linear system based on matching coefficients of its Maclaurin expansion with those of the Newton-Mercator series yields $a_0 = 0, a_1 = 1, a_2 = \frac{1}{6}$ and $b = \frac{2}{3}$, so the [2/1] Padé approximant for

$log(x+1)$ is

$$[2/1]_f(x) = \frac{\frac{x^2}{6} + x}{\frac{2}{3}x + 1},$$

and by the change of variable $z = x + 1$,

$$[2/1]_f(z) = \frac{\frac{(z-1)^2}{6} + (z-1)}{\frac{2}{3}(z-1) + 1} = \frac{(z-1)(z+5)}{4z+2}.$$

We now derive Bernstein's inequality from Bennett's inequality.

Theorem 3.1.19 (Bernstein's inequality)**.** *Let $Z_1, \ldots, Z_n$ be independent random variables with finite variance. Moreover, assume $Z_i \leq b$ for some positive $b > 0$ almost surely for all $i \in [n]$. Define*

$$S \coloneqq \sum_{i=1}^{n} Z_i - \mathbb{E}[Z_i],$$

and

$$v \coloneqq \sum_{i=1}^{n} \mathbb{E}[X_i^2].$$

Then

$$\forall t > 0\colon \quad \mathbb{P}[S \geq t] \leq \exp\left(-\frac{t^2}{2(v + bt/3)}\right).$$

Proof. The elementary bound from Equation (3.4) asserts that $-\log(z) = \log(1/z) \geq -\frac{(z-1)(z+5)}{4z+2}$. The change of variable $u = 1/x - 1$ yields $\log(1+u) \geq -\frac{(\frac{1}{1+u}-1)(\frac{1}{1+u}+5)}{\frac{4}{1+u}+2}$ from which we have

$$\begin{aligned} h(u) = (1+u)\log(1+u) - u &\geq \frac{\frac{u}{1+u} + 5u}{\frac{4}{1+u} + 2} - u \\ &= \frac{u + 5u(1+u)}{4 + 2(1+u)} - u \\ &= \frac{6u + 5u^2 - u(6+2u)}{6+2u} \\ &= \frac{3u^2}{6+2u} \\ &= \frac{u^2}{2(1+u/3)}. \end{aligned}$$

The result then promptly follows from Bennett's inequality (Theorem 3.1.17) and the above elementary bound:

$$\begin{aligned} \mathbb{P}[S \geq t] &\leq \exp\left(-\frac{v}{b^2} \cdot h\left(\frac{bt}{v}\right)\right) \\ &\leq \exp\left(-\frac{v}{b^2} \cdot \frac{b^2t^2}{v^2} \cdot \frac{1}{2(1 + \frac{bt}{3v})}\right) \\ &= \exp\left(-\frac{t^2}{v} \cdot \frac{1}{2(1 + \frac{bt}{3v})}\right) \\ &= \exp\left(-\frac{t^2}{2(v + bt/3)}\right). \end{aligned}$$

□

3.2 Basic assumptions

We now present the general framework for split conformal prediction developed in Oliveira et al. (2022), starting with basic assumptions that will be needed. Regression setting will be the focus from now on and notation will follow Chapter 2 as close as possible.

The first assumption concerns data distribution. In contrast to usual requirements of exchangeable or iid data, the concentration of measure approach of Oliveira et al. (2022) is less stringent and allows for dependence.

Assumption 1 (Dependent data with stationary marginals)**.** *The sample $(X_i, Y_i)_{i=1}^n$ consists of n random covariate/response pairs with stationary marginals: $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are measurable spaces. An additional random pair $(X_*, Y_*) \in \mathcal{X} \times \mathcal{Y}$, 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 $i \in [n]$.*

Some financial time series can naturally be taken to satisfy Assumption 1. While prices are not stationary in general, simple transformations such as dividing a price by its immediate predecessor is enough to make the resulting series resemble a stationary one. Mandelbrot (1963) noticed that large cotton price changes tended to be followed by large changes and, similarly, small changes in the price were usually followed by small changes. This behaviour has since been observed in a myriad of assets and time frames, receiving the name of *volatility clustering*. Cont (2010) argues that volatility clustering points to nonlinear dependence in returns across time. Therefore, it is reasonable to assume asset returns satisfy Assumption 1.

The second assumption follows from usual conformal prediction, with the slight distinction of training, test and calibration sets being fixed beforehand, which is important for the theoretical treatment of dependent data but irrelevant to practical considerations. As customary, a nonconformity score, completely arbitrary, is trained on the training set.

Assumption 2 (Training, test, and calibration data; trained nonconformity score)**.** *We assume $n = n_{\text{train}} + n_{\text{cal}} + n_{\text{test}}$ is a sum of three positive integers. We partition the indices*

$$[n] = I_{\text{train}} \sqcup I_{\text{cal}} \sqcup I_{\text{test}},$$

where $I_{\text{train}} \coloneqq [n_{\text{train}}]$ 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. Consider any function $s : (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}+1} \to \mathbb{R}$. For $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we use the notation:

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

to denote the values of s when the first n_{train} pairs in the input correspond to the training data; $\widehat{s}_{\text{train}}$ is called a nonconformity score trained on the (training) data.

3.3 Split conformal prediction

Under the framework of Oliveira et al. (2022), marginal and conditional guarantees hold for standard split CP as described in Algorithm 2. One immaterial modification to be considered is that exact

quantiles will be taken, i.e., there will be no adjustment due to the calibration set size. This minute consideration can safely be ignored by practitioners when calibration data is abundant and accounted for otherwise. From now on, *split conformal prediction* is to be understood as Algorithm 2 without the quantile adjustment, as in Remark 2.1.7.

In this section, we wish to prove that prediction sets from split CP have adequate coverage over $i \in I_{\text{test}}$, for suitably small η and δ, in the following two senses:

$$\textbf{Marginal coverage:} \quad \mathbb{P}\left[Y_i \in C_{1-\alpha}(X_i)\right] \geq 1 - \alpha - \eta, \text{ for all } i \in I_{\text{test}}, \tag{3.5}$$

$$\textbf{Empirical coverage:} \quad \mathbb{P}\left[\frac{1}{n_{\text{test}}} \sum_{i \in I_{\text{test}}} \mathbb{1}\{Y_i \in C_{1-\alpha}(X_i)\} \geq 1 - \alpha - \eta\right] \geq 1 - \delta, \tag{3.6}$$

Note that η denotes a small penalty relative to Chapter 2, to be paid for the generality assumed by the data in Assumption 1.

3.3.1 Concentration and decoupling assumptions

The next assumptions will give us conditions on the data that suffice for (approximate) marginal and empirical coverage as in (3.5) and (3.6). To state the assumptions, we make the following definition.

Definition 3.3.1. *Given $q : (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}} \to \mathbb{R}$ (assumed measurable), define $q_{\text{train}} \coloneqq q((X_i, Y_i)_{i \in I_{\text{train}}})$ and*

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

The first assumption needed below is about the calibration data. Intuitively, it 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.

Assumption 3 (Concentration over calibration data)**.** *There exist $\varepsilon_{\text{cal}} \in (0, 1)$ and $\delta_{\text{cal}} \in (0, 1)$ such that the following holds: if $q : (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}} \to \mathbb{R}$ and q_{train}, $P_{q,\text{train}}$ are as in Definition 3.3.1, then*

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

The next two assumptions are about the test data. The first 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.

Assumption 4 (Marginal decoupling of test data)**.** *There exists $\varepsilon_{\text{test}}$ such that, if $q : (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}} \to \mathbb{R}$ and q_{train}, $P_{q,\text{train}}$ are as in Definition 3.3.1, then, for $i \in I_{\text{test}}$,*

$$|\mathbb{P}[\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}] - \mathbb{E}[P_{q,\text{train}}]| \leq \varepsilon_{\text{test}}.$$

Finally, we require concentration of the empirical c.d.f. over the test data.

Assumption 5 (Concentration over test data)**.** *There exist $\varepsilon_{\text{test}}, \delta_{\text{test}} \in (0, 1)$ such that, if $q : (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}} \to \mathbb{R}$ and q_{train}, $P_{q,\text{train}}$ are as in Definition 3.3.1, then*

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

3.3.2 Theoretical guarantees

We now combine the assumptions to obtain general coverage guarantees for split conformal prediction under dependent data. The first theorem achieves the goal of marginal coverage (3.5).

Theorem 3.3.2 (Marginal coverage over test data)**.** *Given $\alpha \in (0,1)$, $\delta_{\text{cal}} > 0$, if Assumptions 1, 2, 3 and 4 hold, then, for all $i \in I_{\text{test}}$ and $\alpha \in (0,1)$:*

$$\mathbb{P}[Y_i \in C_{1-\alpha}(X_i)] \geq 1 - \alpha - \varepsilon_{\text{cal}} - \delta_{\text{cal}} - \varepsilon_{\text{test}}.$$

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

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

The second general theorem gives empirical coverage over the test data.

Theorem 3.3.3 (Empirical coverage over test data)**.** *Given $\alpha \in (0,1)$, $\delta_{\text{cal}} > 0$ and $\delta_{\text{test}} > 0$, if Assumptions 1, 2, 3 and 5 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:

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

3.4 Split conformal prediction with conditional guarantees

The results from the previous section extend naturally to the conditional setting given suitably adapted assumptions.

Consider the problem proposed by Barber et al. (2020), where one wants good coverage for $C_{1-\alpha}(X_i)$ conditionally on X_i belonging to a subset $K \subset \mathcal{X}$. As explained in Barber et al. (2020), it is not possible to obtain guarantees for completely general (measurable) sets K. We will thus restrict K to a family $\mathcal{K}$ of subsets of $\mathcal{X}$. For the sake of notation, given a measurable set $K \subset \mathcal{X}$, let:

$$I_{\text{cal}}(K) \coloneqq \{i \in I_{\text{cal}} \,:\, X_i \in K\}; \tag{3.7}$$

$$n_{\text{cal}}(K) \coloneqq \#I_{\text{cal}}(K); \tag{3.8}$$

$$I_{\text{test}}(K) \coloneqq \{i \in I_{\text{test}} \,:\, X_i \in K\}; \tag{3.9}$$

$$n_{\text{test}}(K) \coloneqq \#I_{\text{test}}(K). \tag{3.10}$$

We now introduce the corresponding empirical quantiles and prediction sets.

Definition 3.4.1 (Empirical conditional quantiles and prediction sets)**.** *Let $\mathcal{K}$ denote a family of measurable subsets of $\mathcal{X}$. Given $\phi \in [0,1)$, $K \in \mathcal{K}$, $I_{\text{cal}}(K)$ as in (3.7) and $n_{\text{cal}}(K)$ as in (3.8), denote*

the empirical ϕ-quantile of $\widehat{s}_{\text{train}}(X_i, Y_i)$ over $i \in I_{\text{cal}}$:

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

For $x \in K$, and assuming again $n_{\text{cal}}(K) > 0$, define the prediction set:

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

Fix a coverage level $1 - \alpha$ and a parameter $\gamma > 0$. We wish to prove that prediction sets have adequate coverage over $i \in I_{\text{test}}$. Here, however, we must restrict ourselves to sets that have sufficiently large measure. Thus, we require that $\mathbb{P}[X_* \in K] \ge \gamma$ for each $K \in \mathcal{K}$ and we want to show that for any $i \in I_{\text{test}}$:

$$\textbf{Marginal coverage:} \quad \mathbb{P}\left[Y_i \in C_{1-\alpha}(X_i; K) \mid X_i \in K\right] \ge 1 - \alpha - \eta, \tag{3.11}$$

$$\textbf{Empirical coverage:} \quad \mathbb{P}\left[\inf_{K \in \mathcal{K}} \frac{1}{n_{\text{test}}(K)} \sum_{i \in I_{\text{test}}(K)} \mathbb{1}\{Y_i \in C_{1-\alpha}(X_i; K)\} \ge 1 - \alpha - \eta\right] \ge 1 - \delta, \tag{3.12}$$

with suitably small η and δ and a quantile $\widehat{q}_{1-\alpha,\text{cal}}(K)$ depending on the set K.

3.4.1 Conditional concentration and decoupling assumptions

The assumptions below are the analogues for the conditional coverage setting to those of §3.3.1. To state them, we require the following analogue of Definition 3.3.1.

Definition 3.4.2. *Given $q : (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}} \to \mathbb{R}$ (assumed measurable) and $K \in \mathcal{K}$ with $\mathbb{P}[X \in K] > 0$, define $q_{\text{train}} \coloneqq q((X_i, Y_i)_{i \in I_{\text{train}}})$ as in Definition 3.3.1 and*

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

The following assumptions are analogous to Assumptions 3, 4 and 5. Recall the notation for $I_{\text{cal}}(K)$ and $n_{\text{cal}}(K)$ introduced in (3.7) and (3.8), respectively.

Assumption 6 (Concentration over calibration data)**.** *There exist $\delta_{\text{cal}}, \varepsilon_{\text{cal}} \in (0, 1)$ such that for all $q : (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}} \to \mathbb{R}$, letting q_{train} and $P_{q,\text{train}}(K)$ be as in Definition 3.4.2,*

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

Assumption 7 (Marginal decoupling from test data)**.** *There exists $\varepsilon_{\text{test}} \in (0, 1)$ such that, for all $q : (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}} \to \mathbb{R}$, letting q_{train} and $P_{q,\text{train}}(K)$ be as in Definition 3.4.2,*

$$|\mathbb{P}[\widehat{s}_{\text{train}}(X_k, Y_k) \le q_{\text{train}} \mid X_k \in K] - \mathbb{E}[P_{p,\text{train}}(K)]| \le \varepsilon_{\text{test}}.$$

Assumption 8 (Concentration over test data)**.** *There exist $\delta_{\text{test}}, \varepsilon_{\text{test}} \in (0, 1)$ such that for all*

$q : (\mathcal{X} \times \mathcal{Y})^{n_{\text{train}}} \to \mathbb{R}$, letting q_{train} and $P_{q,\text{train}}(K)$ be as in Definition 3.4.2,

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

3.4.2 Theoretical guarantees under conditioning

We now combine the assumptions to obtain general conditional coverage guarantees for split conformal prediction under dependent data. The first theorem achieves the goal of marginal coverage (3.11).

Theorem 3.4.3 (Conditional coverage over test data)**.** *Given $\alpha \in (0,1)$ and $\delta_{\text{cal}} > 0$, if Assumptions 1, 2, 6 and 7 hold, then, for each $K \in \mathcal{K}$ and any $i \in I_{\text{test}}$:*

$$\mathbb{P}[Y_i \in C_{1-\alpha}(X_i; K) \mid X_i \in K] \ge 1 - \alpha - \varepsilon_{\text{cal}} - \delta_{\text{cal}} - \varepsilon_{\text{test}}.$$

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

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

The second general theorem gives empirical conditional coverage over the test data.

Theorem 3.4.4 (Empirical conditional coverage over test data)**.** *Given $\alpha \in (0,1)$, $\delta_{\text{cal}} > 0$ and $\delta_{\text{test}} > 0$, if Assumptions 1, 2, 6 and 8 hold, then for each $K \in \mathcal{K}$:*

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

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:

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

3.5 Application to the iid case

As an example, we sketch how the framework above applies to iid data. For the marginal coverage of Theorem 3.3.2 and empirical coverage of Theorem 3.3.3, Assumptions 3, 4 and 5 need to be checked. First, note that, in the iid case, when $i \in I_{\text{test}}$,

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

showing that Assumption 4 holds with $\varepsilon_{\text{test}} = 0$.

Moreover, using the fact that $(\mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \le q_{\text{train}}\})_{i=1}^n$ is an iid sample of bounded random variables, by Hoeffding's inequality (Theorem 3.1.8), with probability at least $1 - \delta$,

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

Therefore, taking

$$\varepsilon_{\text{cal}} = \sqrt{\frac{1}{2n_{\text{cal}}} \log\left(\frac{2}{\delta_{\text{cal}}}\right)} \quad \text{and} \quad \varepsilon_{\text{test}} = \sqrt{\frac{1}{2n_{\text{test}}} \log\left(\frac{2}{\delta_{\text{test}}}\right)} \tag{3.13}$$

proves Assumptions 3 and 5.

For conditional guarantees, note that, as in the marginal case, when $i \in I_{\text{test}}(K)$,

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

proving Assumption 7.

Next, suppose the family $\mathcal{K}$ has finite VC dimension $\text{VC}(\mathcal{K}) = d$. Oliveira et al. (2022) show that, if for some $\gamma > 0$, $\mathbb{P}[K] > \gamma$ for all $K \in \mathcal{K}$,

$$\sup_{K \in \mathcal{K}} \left| P_{q,\text{train}}(K) - \frac{1}{n(K)} \sum_{i \in I(K)} \mathbb{1}\{s(X_i, Y_i) \leq q_{\text{train}}\} \right| \leq \varepsilon,$$

where

$$\varepsilon = \frac{1}{\gamma}\left(4\sqrt{\frac{\log(2(n+1)^d)}{n}} + 2\sqrt{\frac{1}{2n}\log\left(\frac{4}{\delta}\right)}\right)$$

Thus, it is possible to pick n and δ to guarantee Assumptions 6 and 8.

Chapter 4

Stochastic Processes

Sequences of random variables, also known as stochastic processes, are the backbone of a myriad of applications. Scientists, statisticians or engineers who want to model a random phenomenon most likely do so via stochastic processes. We refer the reader to Parzen (1962) for a brief account on how stochastic processes are used in statistical physics, population growth models, communication and control theory, management science and operations research, and time series analysis. In finance, Bachelier (1900) used Brownian motion — one of the most ubiquitous stochastic processes — to model prices in the Paris stock market.

With applications in astronomy, biology, ecology, economics, epidemiology, finance, geology, medicine, meteorology, oceanography, physics, psychology, and seismology, stochastic processes' versatility makes them extremely valuable. We now formalize the notion of a stochastic process and its processes, then give illustrative examples. The main motivation for this chapter is that the concentration of measure framework for conformal prediction presented in Chapter 3 can be applied to a general class of stochastic processes, known as β-mixing, that go beyond exchangeability.

4.1 Basic definitions

Definition 4.1.1 (Stochastic process)**.** *Let $\mathbb{T}$ be an arbitrary, countable or uncountable, index set. For each $t \in \mathbb{T}$, let Z_t be a random variable defined on a probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$ and taking values in some measurable space $(E, \mathcal{E})$. The sequence of random variables $\{Z_t\}_{t\in\mathbb{T}}$ is called a stochastic process.*

The index set $\mathbb{T}$ may be the natural numbers $\mathbb{N}$, the integers $\mathbb{Z}$, the reals $\mathbb{R}$ or any other set. One could think of $\mathbb{T}$ as time, with random variables observed one after the other, but it could just as well be the Cartesian plane or a high-dimensional Euclidean space. Working with the interpretation that $\mathbb{T}$ is a time-index should serve us well for financial applications. If $\mathbb{T}$ is continuous (respectively, discrete), our process is deemed a continuous-time (discrete-time) stochastic process. From now on, we will take $\mathbb{T} \equiv \mathbb{Z}$ for simplicity, that is, we will focus on discrete-time stochastic processes.

Definition 4.1.2 (Stationarity)**.** *A stochastic process $\{Z_t\}_{t\in\mathbb{Z}}$ is stationary if, for any $t \in \mathbb{Z}$ and $m, k \in \mathbb{N}$,*

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

That is, the finite-dimensional distributions of a stationary process are time-invariant.

The discussion carried out in Section 2.1 related to the concepts of exchangeability and independence naturally apply to stochastic processes. Indeed, although not yet formally identified as stochastic

processes by then, we were dealing with sequences of random variables, which is precisely Definition 4.1.1. For completeness, we now restate those important concepts.

Definition 4.1.3 (Exchangeability)**.** *A finite stochastic process $\{Z_t\}_{t\in\mathbb{T}}$ is exchangeable if, for any permutation function $\pi\colon [n] \to [n]$,*

$$(Z_1, \ldots, Z_n) \overset{d}{=} (Z_{\pi(1)}, \ldots, Z_{\pi(n)}).$$

An infinite stochastic process $\{Z_t\}_{t\in\mathbb{Z}}$ is exchangeable if every finite subsequence is exchangeable.

Definition 4.1.4 (Independence within a stochastic process)**.** *A stochastic process is independent $\{Z_t\}_{t\in\mathbb{T}}$ if and only if for all $n \in \mathbb{N}$ and for all $t_1, \ldots, t_n \in \mathbb{T}$*

$$F_{Z_{t_1},\ldots,Z_{t_n}}(z_1, \ldots, z_n) = \prod_{i=1}^{n} F_{Z_{t_i}}(z_i),$$

for all $z_1, \ldots, z_n$. In words, a stochastic process is independent if the joint cdf equals the product of the individual marginals for any of its subsequences.

Definition 4.1.5 (Independence between two stochastic processes)**.** *Let $\{U_t\}_{t\in\mathbb{T}}$ and $\{V_t\}_{t\in\mathbb{T}}$ taking values in $(E, \mathcal{E})$ and $(E', \mathcal{E}\prime)$, respectively, be stochastic processes defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If the σ-fields $\mathcal{F}^U \coloneqq \sigma(U_t : t \in \mathbb{T})$ and $\mathcal{F}^V \coloneqq \sigma(V_t : t \in \mathbb{T})$ are independent, we say that the stochastic processes $\{U_t\}_{t\in\mathbb{T}}$ and $\{V_t\}_{t\in\mathbb{T}}$ are independent. Alternatively, this holds if for all $n \in \mathbb{N}$ and for all $t_1, \ldots, t_n \in \mathbb{T}$, the random vectors $\{U_t\}_{t=t_1}^{t_n}$ and $\{V_t\}_{t=t_1}^{t_n}$ are independent, that is,*

$$F_{U_{t_1},\ldots,U_{t_n},V_{t_1},\ldots,V_{t_n}}(u_1, \ldots, u_n, v_1, \ldots, v_n) = F_{U_{t_1},\ldots,U_{t_n}}(u_1, \ldots, u_n) \cdot F_{V_{t_1},\ldots,V_{t_n}}(v_1, \ldots, v_n).$$

We define next a natural condition for dependent data.

Definition 4.1.6 (β-mixing (absolute regularity))**.** *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*

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

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

Intuitively, the β-mixing coefficient measures how close, in total variation distance, the law of two blocks of random variables a units apart is from being independent. Therefore, the β-mixing condition may be thought of as asymptotic independence. Many natural classes of stochastic processes satisfy this property, including ARMA and GARCH models (Carrasco and Chen 2002; Mokkadem 1988) and more general Markov processes (Doukhan 2012). In particular, the β-mixing coefficients decay exponentially fast for ARMA and GARCH models, and likewise for stationary geometrically ergodic Markov chains.

4.2 Examples

4.2.1 Markov chains

Let $(W_0, W_1, W_2, \ldots)$ be a homogeneous recurrent Markov chain with state space $\mathcal{W}$, transition matrix P and stationary distribution π. Then, for $r \in \mathbb{N}$, its β-mixing coefficient is given by

$$\beta(r) = \mathbb{E}_\pi[\|P^r(X, \cdot) - \pi(\cdot)\|_{\mathrm{TV}}] = \int_{\mathcal{X}} \pi(dx)\|P^r(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}},$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation norm.

In general, every homogeneous recurrent Markov chain is absolutely regular and its β-mixing coefficients can be calculated, given the transition matrix P and stationary distribution π (Davydov 1974; McDonald, Shalizi, and Schervish 2015; Vidyasagar and Karandikar 2016).

For a concrete example, consider the Markov chain with state space $\mathcal{W} = \{0, 1\}$ and transition matrix $P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$, as depicted below

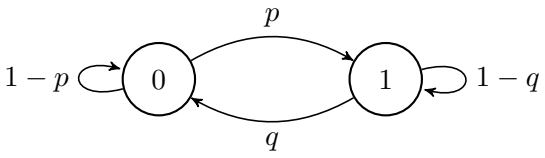

Figure 4.1: Markov chain with two states.

For this two-state Markov chain, the $\beta(r)$ coefficient can be calculated explicitly. Let $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_{|\mathcal{W}|}$ be the ordered eigenvalues of the transition matrix P. From the characterization of the stationary distribution π, i.e., $\pi P = \pi$, it is clear that π is an eigenvector associated with eigenvalue 1. Moreover, $|\lambda_j| \leq 1$ for all $j \in \{1, 2, \ldots, |\mathcal{W}|\}$ (Levin, Peres, and Wilmer 2017, Lemma 12.1), which yields $\lambda_1 = 1$. Now, it follows that

$$\begin{aligned}
\beta(r) &= \mathbb{E}_\pi[\|P^r(X, \cdot) - \pi(\cdot)\|_{\mathrm{TV}}] \\
&= \sum_{x \in \mathcal{W}} \pi(x)\|P^r(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \\
&= \sum_{x \in \mathcal{W}} \pi(x) \cdot \tfrac{1}{2} \sum_{y \in \mathcal{W}} |P^r(x, y) - \pi(y)| && \text{(Levin, Peres, and Wilmer 2017, Proposition 4.2)} \\
&= \sum_{x \in \mathcal{W}} \pi(x) \cdot \tfrac{1}{2} \sum_{y \in \mathcal{W}} |\lambda_2^{r-1}(P(x, y) - \pi(y))| && \text{(Levin, Peres, and Wilmer 2017, Remark 1.2)} \\
&= \sum_{x \in \mathcal{W}} \pi(x) \cdot \tfrac{1}{2}|\lambda_2^{r-1}| \sum_{y \in \mathcal{W}} |P(x, y) - \pi(y)| \\
&= \tfrac{1}{2}|\lambda_2^{r-1}| \sum_{x \in \mathcal{W}} \pi(x) \sum_{y \in \mathcal{W}} |P(x, y) - \pi(y)|.
\end{aligned}$$

The stationary distribution of this Markov chain is $\pi = \begin{bmatrix} \frac{q}{p+q} & \frac{p}{p+q} \end{bmatrix}$ and the eigenvalues of the transition matrix are the roots of the characteristic polynomial

$$\det(P - \lambda I) = \det\left(\begin{bmatrix} 1-p-\lambda & p \\ q & 1-q-\lambda \end{bmatrix}\right) = (1-p-\lambda)(1-q-\lambda) - pq = (\lambda - 1)(\lambda - 1 + p + q).$$

Therefore, $\lambda_1 = 1$, as previously noted, and the second-largest eigenvalue of the transition matrix, λ_2, equals $1 - p - q$.

The β-mixing coefficient for $r \in \mathbb{Z}_{>0}$ is thus given by

$$\begin{aligned}
\beta(r) &= \tfrac{1}{2}|1-p-q|^{r-1}(\tfrac{q}{p+q}(|1-p-\tfrac{q}{p+q}| + |p-\tfrac{p}{p+q}|) + \tfrac{p}{p+q}(|q-\tfrac{q}{p+q}| + |1-q-\tfrac{p}{p+q}|)) \\
&= \tfrac{1}{2}|1-p-q|^{r-1}(\tfrac{q}{p+q}(2 \cdot |1-p-\tfrac{q}{p+q}|) + \tfrac{p}{p+q}(2 \cdot |q-\tfrac{q}{p+q}|)) \\
&= \tfrac{1}{2}|1-p-q|^{r-1}(\tfrac{2}{p+q}(q \cdot |1-p-\tfrac{q}{p+q}| + p \cdot |q-\tfrac{q}{p+q}|) \\
&= \tfrac{1}{2}|1-p-q|^{r-1} \cdot \tfrac{4q}{p+q} \cdot |1-p-\tfrac{q}{p+q}| \\
&= |1-p-q|^{r-1} \cdot |1-p-\tfrac{q}{p+q}| \cdot \tfrac{2q}{p+q} \\
&= |1-p-q|^{r-1} \cdot \tfrac{1}{p+q} \cdot |(p+q)(1-p)-q| \cdot \tfrac{2q}{p+q} \\
&= |1-p-q|^{r-1} \cdot \tfrac{1}{p+q} \cdot |p(1-p-q)| \cdot \tfrac{2q}{p+q} \\
&= |1-p-q|^{r-1} \cdot \tfrac{p}{p+q} \cdot |1-p-q| \cdot \tfrac{2q}{p+q} \\
&= |1-p-q|^{r} \cdot \tfrac{p}{p+q} \cdot \tfrac{2q}{p+q} \\
&= |1-p-q|^{r} \cdot \tfrac{2pq}{(p+q)^2}
\end{aligned}$$

This shows that the two-state Markov chain is geometrically β-mixing in the sense that $\beta(r) = O(c^r)$ for $c \coloneqq |1-p-q| \in (0,1)$. In other words, the β-mixing coefficient not only converges to zero, but it does so exponentially fast.

Note that $p = q = 0.5$ implies $P^r(x,y) = \pi(y)$ for all $x, y \in \mathcal{W}$ and $r \in \mathbb{N}_{>0}$, so $\beta(r)$ will always be zero, indicating total lack of dependence. This is consistent with the observation that, with such probabilities, the Markov chain reduces to a Bernoulli process, i.e., a sequence of iid Bernoulli trials. On the other hand, as p and q tend towards zero, dependence increases in the sense that $\beta(r)$ becomes larger for every r and the Markov chain is more likely to be stuck in the current state.

Another Markov chain of interest is described by the random walk on a cycle graph of v vertices, portrayed in Figure 4.2, whose state space is the set of integers modulo v, i.e., $\mathbb{Z}/v\mathbb{Z} = \{0, \ldots, v-1\}$. On any given state, there is a probability b of moving backward, f of moving forward and s of staying in the current state, such that $b + f + s = 1$ with $b, s, f \in [0,1)$. Thus, for $j, k \in \mathbb{Z}/v\mathbb{Z}$, the process' transition matrix is defined by

$$P(j,k) = \begin{cases} b, & \text{if } k \equiv j-1 \pmod{v} \\ f, & \text{if } k \equiv j+1 \pmod{v} \\ s, & \text{if } k \equiv j \\ 0, & \text{otherwise} \end{cases}.$$

Moreover, as all vertices have the same degree, the chain's stationary distribution equals the uniform distribution.

Since the random walk on the cycle under consideration has a fixed transition matrix over time, it is homogeneous; moreover, as the set of vertices is finite and we required a strictly positive probability of changing state, it is also recurrent. Therefore, β-mixing coefficients can be calculated as showcased previously,

$$\beta(r) = \sum_{x \in \mathcal{W}} \pi(x) \cdot \tfrac{1}{2} \sum_{y \in \mathcal{W}} |P^r(x,y) - \pi(y)|,$$

simply requiring the transition matrix and stationary distribution, both of which we have available.

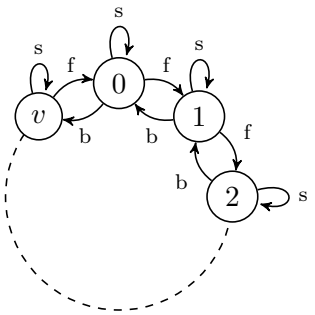

Figure 4.2: Random walk on the cycle graph of v vertices.

As in the case of the two-state Markov chain, β-mixing coefficients are sensitive to the probability of moving, b and f for the random walk on the cycle graph. If the probability s of staying on the current state is high, and consequently b and f are small, dependence should be larger due to more frequent repeated states. In the same vein, it is likely that the cycle will not be properly explored and some states may not even manifest on small samples. Intuitively, the number of vertices will also play an important role, as it should take a larger sample to account for the 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}.

4.2.2 Autoregressive processes

For a different kind of Markovian process, consider the autoregressive process of order one (AR(1)), defined by the recurrence $W_t = \lambda W_{t-1} + \varepsilon_t$, with $t \in \mathbb{N}_{>0}$, $\lambda \in \mathbb{R}$, and ε_t independent normally distributed random variables with mean zero and variance one. The sequence is stationary as long as $|\lambda| < 1$, and iid for $\lambda = 0$. Although the β-mixing coefficients cannot be calculated explicitly, it is possible to numerically integrate $\beta(r) = \int \|P^r(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \pi(dx)$ to approximate it, as in McDonald, Shalizi, and Schervish (2015).

4.2.3 Renewal processes

Lastly, we turn our attention to renewal processes, which generalize Poisson processes for arbitrary iid inter-arrival times. Formally, we follow Berbee (1987) to introduce renewal sequences $(W_t)_{t \geq 0}$. Let $T_1, T_2, \ldots$ be independent random variables with law according to an aperiodic probability distribution F on $\mathbb{N}_{>0}$. In order to ensure stationarity, the first state T_0 — independent of $T_1, T_2, \ldots$ — is assumed to follow $\mathbb{P}[T_0 = i] = \frac{1}{\mathbb{E}[T_1]} \mathbb{P}[T_1 > i]$ for all $i \geq 0$. Finally, the renewal sequence is given by

$$W_t = \begin{cases} 1, & \text{if } t = T_0 + T_1 + \ldots T_k \text{ for some } k \\ 0, & \text{otherwise} \end{cases}.$$

Stationary renewal processes are known to be β-mixing and although their coefficients cannot be directly calculated, they can be upper bounded (Heinrich 1992).

Chapter 5

Conformal Prediction for β-mixing Processes

The alternative approach to split CP developed in Oliveira et al. (2022) and described in Chapter 3, employing concentration of measure and decoupling inequalities, can be readily applied to the concrete case of β-mixing processes. As in that article, this chapter's focus will be on stationary series in particular, although the results can be extended to nonstationary settings.

The β-mixing condition allows us to replace independence with asymptotic independence and still retain some important concentration results. Due to the so-called blocking technique (Yu 1994; Mohri and Rostamizadeh 2010; Kuznetsov and Mohri 2017), it is possible 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 (Theorem 3.1.19).

Remark 5.0.1. *Theoretical bounds that follow are in terms of "optimal block sizes". However, this is a purely mathematical device: while they appear in the* performance bounds *below, the split CP method is not dependent on this optimization. In fact, the method does not require a choice of block sizes (unlike, e.g., Chernozhukov, Wüthrich, and Zhu (2018)).*

The proofs for all following results can be found in Oliveira et al. (2022).

5.1 Standard coverage guarantees

We now argue that Assumptions 3, 4 and 5 hold for stationary β-mixing processes. As is standard with the blocking technique, the error bounds obtained will depend on an optimization of block sizes (cf. Remark 5.0.1).

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

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

and

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

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:

$$\varepsilon_{\text{cal}} \coloneqq \inf_{(a,m,r)\in F_{\text{cal}}} \left\{ \widetilde{\sigma}(a)\sqrt{\frac{4}{n_{\text{cal}}-r+1}\log\left(\frac{4}{\delta_{\text{cal}}-4(m-1)\beta(a)-\beta(r)}\right)} + \frac{1}{3m}\log\left(\frac{4}{\delta_{\text{cal}}-4(m-1)\beta(a)-\beta(r)}\right) + \frac{r-1}{n_{\text{cal}}} \right\}, \tag{5.1}$$

where

$$\widetilde{\sigma}(a) = \sqrt{\frac{1}{4} + \frac{2}{a}\sum_{j=1}^{a-1}(a-j)\beta(j)}. \tag{5.2}$$

Similarly, we define the test error correction factor for a stationary β-mixing process as

$$\varepsilon_{\text{test}} = \inf_{(a,m,s)\in F_{\text{test}}} \left\{ \widetilde{\sigma}(a)\sqrt{\frac{4}{n_{\text{test}}}\log\left(\frac{4}{\delta_{\text{test}}-4(m-1)\beta(a)-\beta(n_{\text{cal}})}\right)} + \frac{1}{3m}\log\left(\frac{4}{\delta_{\text{test}}-4(m-1)\beta(a)-\beta(n_{\text{cal}})}\right) + \frac{s}{n_{\text{test}}} \right\}. \tag{5.3}$$

With ε_{cal} defined as above, Theorem 3.3.2 yields the following result for stationary β-mixing processes:

Theorem 5.1.1 (Marginal coverage: stationary β-mixing processes)**.** *Suppose that $(X_i, Y_i)_{i=1}^n$ 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 (5.1) and $\varepsilon_{\text{train}} = \beta(k - n_{\text{train}})$. Additionally, if $\widehat{s}_{\text{train}}(X_*, Y_*)$ almost surely has a continuous distribution conditionally on the training data, then:

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

Remark 5.1.2. *Under certain assumptions over the dependence of the processes, the stationary β-mixing bounds given by (5.1) are of the same asymptotic order as the iid bounds (3.13) from Section 3.5. 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*

$$m = o(n_{\text{cal}}^{(b-c)/(b+1)}) \quad \text{and} \quad \sqrt{n_{\text{cal}}\log(n_{\text{cal}})} = o\left(m\right),$$

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 3.3.3 yields the following:

Theorem 5.1.3 (Empirical coverage: stationary β-mixing processes)**.** *Suppose that $(X_i, Y_i)_{i=1}^n$ 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 (5.1) and (5.3). Additionally, if $\widehat{s}_{\text{train}}(X_*, Y_*)$ almost surely has a continuous distribution conditionally on the training data, then:

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

Remark 5.1.4. *The expression in (5.1) follows from a stationary β-mixing version of Bernstein's inequality, proved in Oliveira et al. (2022), which might be of independent interest. The factor of $1/4$ that appears in the variance term $\widetilde{\sigma}$ (cf. Equation (5.2)) is due to the fact that for any q_{train}, we have*

$$\text{Var}\left[\mathbb{1}\{\widehat{s}_{\text{train}}(X_*, Y_*) \leq q_{\text{train}}\}\right] \leq 1/4.$$

However, given a coverage level $\alpha \in (0,1)$, it is possible to improve on this bound by considering $q_{\text{train}} = \tilde{q}_{\alpha,\text{train}}$, where $\tilde{q}_{\alpha,\text{train}}$ is a slight adaptation of the $(1-\alpha)$-quantile, such that

$$\text{Var}\left[\mathbb{1}\left\{\widehat{s}_{\text{train}}(X_*, Y_*) \leq \tilde{q}_{\alpha,\text{train}}\right\}\right] \leq (1-\alpha)\alpha,$$

provided, for example, that $\widehat{s}_{\text{train}}(X_*, Y_*)$ almost surely has a continuous distribution conditionally on the training data. Therefore, the calibration adjustment becomes

$$\begin{aligned}\varepsilon_{\text{cal}} = \inf_{(a,m,r)\in F_{\text{cal}}} \Bigg\{ & \widetilde{\sigma}(a,\alpha)\sqrt{\frac{4}{n_{\text{cal}}-r+1}\log\left(\frac{4}{\delta_{\text{cal}} - 4(m-1)\beta(a) - \beta(r)}\right)} \\ & + \frac{1}{3m}\log\left(\frac{4}{\delta_{\text{cal}} - 4(m-1)\beta(a) - \beta(r)}\right) + \frac{r-1}{n_{\text{cal}}}\Bigg\},\end{aligned} \tag{5.4}$$

and the test adjustment becomes,

$$\begin{aligned}\varepsilon_{\text{test}} = \inf_{(a,m,s)\in F_{\text{test}}} \Bigg\{ & \widetilde{\sigma}(a,\alpha)\sqrt{\frac{4}{n_{\text{test}}}\log\left(\frac{4}{\delta_{\text{test}} - 4(m-1)\beta(a) - \beta(n_{\text{cal}})}\right)} \\ & + \frac{1}{3m}\log\left(\frac{4}{\delta_{\text{test}} - 4(m-1)\beta(a) - \beta(n_{\text{cal}})}\right) + \frac{s}{n_{\text{test}}}\Bigg\},\end{aligned} \tag{5.5}$$

where

$$\widetilde{\sigma}(a,\alpha) = \sqrt{(1-\alpha)\alpha + \frac{2}{a}\sum_{j=1}^{a-1}(a-j)\beta(j)}, \tag{5.6}$$

which is never worse than (5.1) since $\alpha \in (0,1)$. Nonetheless, as shown in Remark 5.1.2, our original bound (5.1) is enough to recover the same asymptotic order of the iid case.

Finally, note that in the iid case, with $KL(\cdot\|\cdot)$ denoting the Kullback-Leibler divergence between two probability distributions, using the fact that for any q_{train}

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\widehat{s}_{\text{train}}(X_i, Y_i) \leq q_{\text{train}}\} - P_{q,\text{train}}\right| \geq \varepsilon\right] \leq 2e^{-n\text{KL}(P_{q,\text{train}}+\varepsilon\|P_{q,\text{train}})},$$

and applying a similar argument as before, it is possible to show that, provided that $\widehat{s}_{\text{train}}(X_*, Y_*)$ almost

surely has a continuous distribution conditionally on the training data and $\alpha \in (0, 1/2)$, we can take

$$\varepsilon_{\text{cal}} = \sqrt{\frac{2\alpha(1-\alpha)}{n_{\text{cal}}} \log\left(\frac{2}{\delta_{\text{cal}}}\right)}, \quad \varepsilon_{\text{test}} = \sqrt{\frac{2\alpha(1-\alpha)}{n_{\text{test}}} \log\left(\frac{2}{\delta_{\text{test}}}\right)}. \tag{5.7}$$

Remark 5.1.5. *Computationally, ε_{cal} from Equation (5.4) and $\varepsilon_{\text{test}}$ from Equation (5.5) can be calculated in an efficient manner via Algorithm 4. Given a calibration set of size n_{cal} and a probability of $1-\delta$, one must find the amount $2m$ of consecutive blocks as well as their size a and the gap r between training and calibration that minimize $(\frac{n_{\text{cal}}}{n_{\text{cal}}-r} - 1)\alpha + \varepsilon(a, m, \delta) + \beta(r)$ in order to calculate η. From the constraint $2ma + r = n_{\text{cal}}$ and the fact that $r > 0$, it follows that $2ma < n_{\text{cal}}$. Since both m and a are integers, the inequality is equivalent to $ma \leq \lfloor \frac{n_{\text{cal}}}{2} \rfloor$. For ease of notation, let $h \coloneqq \lfloor \frac{n_{\text{cal}}}{2} \rfloor$ such that $ha \leq m$. Note that when h (respectively, a) assumes a value $i \in \{1, \ldots, h\}$, then a (respectively, h) must be no larger than $\lfloor \frac{h}{i} \rfloor$ for the inequality to hold.*

Algorithm 4: Calculate extra miscoverage cost due to data dependence.

function `minimize`$(n_{\text{cal}}, \alpha, \delta)$**:**
- $\ell \leftarrow \lfloor \frac{n_{\text{cal}}}{2} \rfloor$
- $K \leftarrow \sum_{i=1}^{\ell} \lfloor \frac{\ell}{i} \rfloor$
- $\boldsymbol{a}[1, \ldots, K] \leftarrow 0$
- $\boldsymbol{m}[1, \ldots, K] \leftarrow 0$
- $k \leftarrow 1$
- **for** $i \in \{1, \ldots, \ell\}$ **do**
  - **for** $j \in \{1, \ldots, \lfloor \frac{\ell}{i} \rfloor\}$ **do**
    - $\boldsymbol{m}[k] \leftarrow i$
    - $\boldsymbol{a}[k] \leftarrow j$
    - $k \leftarrow k + 1$
  - **end**
- **end**
- $\boldsymbol{r} \leftarrow n_{\text{cal}} - 2\boldsymbol{m}\boldsymbol{a} + 1$
- $\mathcal{T} \leftarrow \{k \in \{1, \ldots, K\} : \delta > 4(\boldsymbol{m}[k] - 1)\beta(\boldsymbol{a}[k]) + \beta(\boldsymbol{r}[k]) \wedge \boldsymbol{r}[k] \geq 1\}$
- $\boldsymbol{m} \leftarrow \boldsymbol{m}[\mathcal{T}]; \boldsymbol{a} \leftarrow \boldsymbol{a}[\mathcal{T}]; \boldsymbol{r} \leftarrow \boldsymbol{r}[\mathcal{T}]$
- $\boldsymbol{B}[1, \ldots, |\mathcal{T}|] \leftarrow 0$
- **for** $i \in \{1, \ldots, |\mathcal{T}|\}$ **do**
  - **for** $j \in \{1, \ldots, \max(\boldsymbol{a})\}$ **do**
    - $\boldsymbol{B}[i] \leftarrow \boldsymbol{B}[i] + \beta(j) \cdot \max(0, \boldsymbol{a}[i] - 1 - j)$
  - **end**
- **end**
- $\boldsymbol{L} \leftarrow \log\left(\frac{4}{\delta - 4(\boldsymbol{m}-1)\beta(\boldsymbol{a}) - \beta(\boldsymbol{r})}\right)$
- $\boldsymbol{\sigma} \leftarrow (1-\alpha)\alpha + (2/\boldsymbol{a}) * \boldsymbol{B}$
- $\boldsymbol{\varepsilon}_{\textbf{cal}} \leftarrow \sqrt{4\boldsymbol{\sigma}\boldsymbol{L}/(n_{\text{cal}} - \boldsymbol{r} + 1)} + \frac{1}{3\boldsymbol{m}}\boldsymbol{L} + \frac{\boldsymbol{r}-1}{n_{\text{cal}}}$
- $\boldsymbol{\varepsilon}_{\textbf{train}} \leftarrow \beta(n_{\text{cal}} + 1)$
- $\boldsymbol{\eta} \leftarrow \boldsymbol{\varepsilon}_{\textbf{cal}} + \boldsymbol{\varepsilon}_{\textbf{train}} + \delta$
- $\star \leftarrow \text{argmax}_{t \in \{1, \ldots, |\mathcal{T}|\}}(1 - \alpha - \boldsymbol{\eta}[t])(1 - \delta)$
- $\eta_\star \leftarrow \boldsymbol{\eta}[\star]$
- $m_\star \leftarrow \boldsymbol{m}[\star]$
- $a_\star \leftarrow \boldsymbol{a}[\star]$
- $r_\star \leftarrow \boldsymbol{r}[\star]$

return $\eta_\star, m_\star, a_\star, r_\star$

5.2 Conditional guarantees

To apply Theorems 3.4.3 and 3.4.4 for stationary β-mixing processes, we need to specify a family $\mathcal{K}$ of Borel measurable sets in $\mathcal{X}$ satisfying certain conditions that allow us to verify Assumptions 6, 7 and 8. In the remaining of this section we assume the following:

Assumption 9 (Family complexity)**.** *For a fixed value $\gamma > 0$, the family $\mathcal{K}$ of Borel measurable sets in $\mathcal{X}$ has finite VC dimension $\mathrm{VC}(\mathcal{K}) = d$ and $\mathbb{P}[X_* \in K] > \gamma$ for all $K \in \mathcal{K}$.*

The assumption that $\mathcal{K}$ has finite VC dimension allows us to obtain concentration bounds for the empirical processes in Assumptions 6 and 8. Moreover, the condition $\mathbb{P}[X_* \in K] > \gamma$ is important to ensure the conditioned empirical quantile is well defined.

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

$$\varepsilon_{\mathrm{cal}} = \inf_{(a,m,r)\in G_{\mathrm{cal}}} \left\{ \frac{1}{\gamma} \left(4\sqrt{\frac{\log(2(m+1)^d)}{m}} + \frac{2(r-1)}{n_{\mathrm{cal}}} + 2\sqrt{\frac{1}{2m}\log\left(\frac{16}{\delta_{\mathrm{cal}} - 16(m-1)\beta(a) - \beta(r)}\right)}\right)\right\} \tag{5.8}$$

where

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

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 $\mathbb{P}[X_* \in K]$ can be close to zero, and thus a larger sample is necessary to estimate the empirical quantile well.

Similarly, we define the test error correction factor for a stationary β-mixing process conditioned to the family $\mathcal{K}$ as

$$\varepsilon_{\mathrm{test}} = \inf_{(a,m,s)\in G_{\mathrm{test}}} \left\{ \frac{1}{\gamma} \left(4\sqrt{\frac{\log(2(m+1)^d)}{m}} + \frac{2s}{n_{\mathrm{test}}} + 2\sqrt{\frac{1}{2m}\log\left(\frac{8}{\delta_{\mathrm{test}} - 8(m-1)\beta(a) - \beta(n_{\mathrm{cal}})}\right)}\right)\right\}, \tag{5.9}$$

where

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

Finally, Theorem 3.4.3 yields the following result.

Theorem 5.2.1 (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 $K \in \mathcal{K}$ and any $i \in I_{\mathrm{test}}$*

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

with $\eta = \varepsilon_{\mathrm{cal}} + \varepsilon_{\mathrm{test}}$, where $\varepsilon_{\mathrm{cal}}$ is as in (5.8) and $\varepsilon_{\mathrm{test}} = \beta(k - n_{\mathrm{train}})$.

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

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

And Theorem 3.4.4 yields the following:

Theorem 5.2.2 (Empirical conditional coverage: stationary β-mixing processes)**.** *Suppose that* $(X_i, Y_i)_{i=1}^n$ *is stationary* β*-mixing, then given* $\alpha \in (0,1)$*,* $\gamma > 0$*,* $\delta_{\text{cal}} > 0$ *and* $\delta_{\text{test}} > 0$*, for each* $K \in \mathcal{K}$*:*

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

where $\eta = \varepsilon_{\text{cal}} + \varepsilon_{\text{test}}$*, for* ε_{cal} *as in (5.8) and* $\varepsilon_{\text{test}}$ *as in (5.9).*

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

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

We conclude this chapter by noting that collective results from traditional conformal prediction discussed in Chapter 2 could be naturally extended to the general class of stationary β-mixing processes through the concentration of measure approach from Chapter 3 with the small addition of a coverage penalty due to the assumption of data being dependent (Assumption 1). Once again, we refer the reader to Oliveira et al. (2022) for all technical proofs. Finally, it remains to be seen how split conformal prediction behaves in practice for dependent data. That will be the focus of the next chapter.

Chapter 6

Experiments and Applications

In this chapter, synthetic and real-world experiments are conducted in order to evaluate the claim that split conformal prediction, traditionally used for exchangeable data, performs well beyond such setting. Synthetic experiments are tailored to the theoretical setup in Chapter 5: the data-generating processes are stationary and β-mixing. For real-world experiments, the datasets are all nonexchangeable due to time-dependence, but can be assumed stationary after suitable transformations. A display of how the coverage guarantees fare in all experiments corroborate that split CP works well for non-exchangeable data, unless the dependence is extreme. Code to reproduce these results is available at `https://github.com/jv-rv/split-conformal-nonexchangeable`.

6.1 Data

6.1.1 Synthetic

Stochastic processes introduced in Chapter 4 will be the base of our synthetic experiments, given their β-mixing nature and stationarity under adequate parameters and initialization. On the one hand, autoregressive processes are continuous and can be readily used for regression tasks. On the other hand, Markov chains and renewal processes take values in a discrete set and are unfit for regression. However, given a β-mixing sequence $\{X_i\}_{i=1}^n$ and a measurable function f, a new sequence $\{f(X_i)\}_{i=1}^n$ is guaranteed to be β-mixing with coefficients upper bounded by those of the original sequence (Yu 1994). Moreover, stationarity is still ensured by sampling the initial state from π. Therefore, the discrete sequences can be made continuous by adding a Gaussian noise with small variance. In the experiments below, except for the autoregressive process which is already continuous, a Gaussian noise of zero mean and variance of 10^{-6} is considered. Such transformation pertain to a more general class usually referred to as *hidden models* in the literature.

6.1.2 Financial time series

Financial time series will also be considered: euro spot exchange rate (`eurusd`), Brent crude oil future (`bcousd`) and S&P 500 stock index future (`spxusd`). To ensure reproducibility, data is retrieved from an open provider: HistData (2022). Minute-by-minute bid prices are available, so the series should be understood as being from a buyer's perspective. As we have selected three of the most liquid contracts, the spread between bid and ask should not be too large in general, although it may be the case in periods of high volatility, for example. Real-world data can be noisy and checking assumptions may often be non-trivial. As series of prices are highly nonstationary in general, we compute linear

returns by dividing a price at time t by the preceding price, at $t-1$, and subtracting 1 from the result. Stationarity of returns vary from asset class and data frequency, but augmented Dickey-Fuller tests show it is reasonable to assume that all series we consider are stationary.

The markets considered usually operate fully from Monday through Thursday, partially on Friday and Sunday and do not open on Saturdays. Figure 6.1 shows the histograms of data points per day for all three financial datasets with $0.1, \ldots, 0.9$ quantiles shown as vertical red lines. On Sunday, markets open late, around 17:00 – 20:00, so there is only 20% – 30% of usual data points available. Likewise, market closes early on Fridays, around 16:00 – 17:00, so we have only 70% – 80% of the usual amount of observations. Holidays apart, the first mass on the histograms represent Sundays, the second mass represents Friday and the largest mass on the far right consists of days from Monday to Thursday. In order to deal only with days of similar number of observations, Fridays and Sundays were discarded for the entire period.

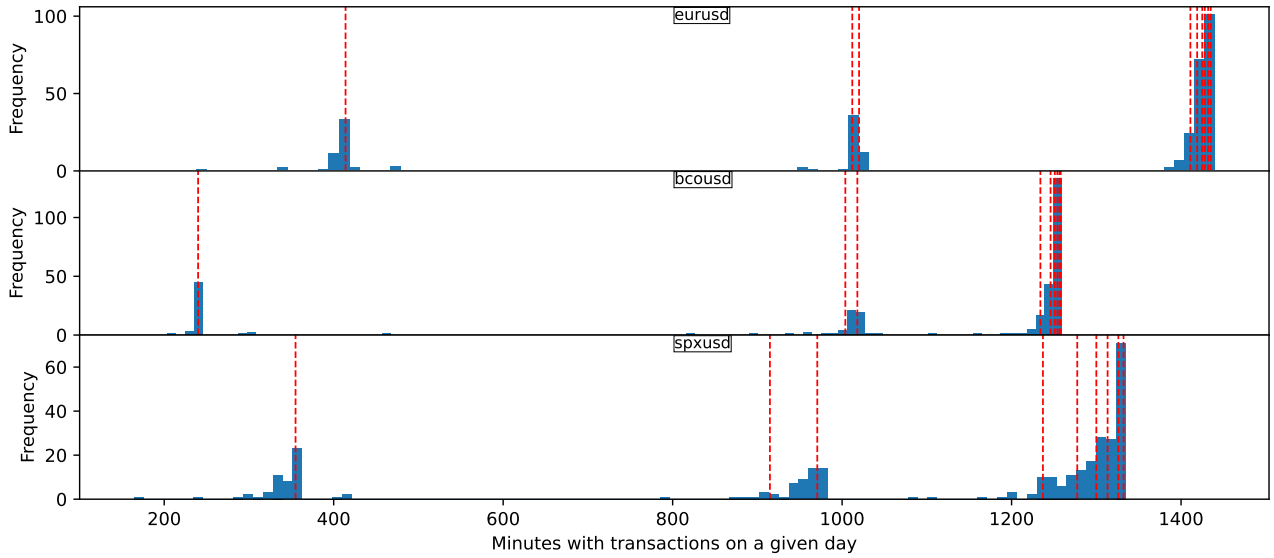

Figure 6.1: Data availability representation via histograms of observations per day for all financial datasets. Vertical dashed lines show the $0.1, \ldots, 0.9$ quantiles. First mass on far left is mostly composed of Sundays, middle mass of Fridays and larger mass on the far right of Mondays through Thursdays. Market is closed on Saturdays.

6.1.3 Setting

In all experiments that follow, data may be thought of as a time series, and the feature set to predict each data point comprises 11 lagged observations. Quantile regression models are trained with fixed hyperparameters on training sets (refer to Appendix A for a detailed list of hyperparameters). Nonconformity scores are calculated over the calibration sets and conformalized following split conformal quantile regression (Romano, Patterson, and Candès 2019), so valid prediction intervals can be generated for new data points. Nominal coverage is set to $1-\alpha = 0.9$ and quantiles are calculated in accordance to the procedure outlined in Chapter 3, i.e., no correction factor is considered when calculating quantiles. Since models trained with the pinball loss (linear regression, gradient boosting and neural networks) have no monotonicity guarantee of estimated quantiles — a phenomenon known as quantile crossing (Bassett Jr and Koenker 1982) pointed out in Chapter 3 —, in the rare cases of crossing, we swap lower and upper predictions in accordance to the methodology described in Chernozhukov, Fernández-Val, and Galichon (2010). As outlined in Section 2.4, quantile regression forests and quantile k-nearest

neighbors do not suffer from quantile crossing, so they do not need post-processing in this regard.

Assessing β-mixing is difficult and literature on the topic is scarce. McDonald, Shalizi, and Schervish (2015) proposed the first estimator for β-mixing coefficients; Khaleghi and Lugosi (2021) introduced estimators and goodness-of-fit tests. To the best of our knowledge, there has been no further work on the area and, unfortunately, the two aforementioned approaches suffer from drawbacks: while the former's estimator is complex and provides no convergence rate beyond Markov processes, the latter's implementation can be impractical due to the fine partitions needed to be taken over the entire state space. However, Cont (2010) argues that for short timeframes, microstructure effects induce autocorrelation of asset returns, illustrating this point with the autocorrelation function of log-returns for the exchange rate between dollar and yen: the correlation of the series with itself lagged in 5 minutes is negatively intense and vanishes for further lags. While vanishing autocorrelation does not imply β-mixing data, it does imply lack of independence in the short term and leads to independence in the long term, which we will take as a reasonable indicator of β-mixing.

6.2 Marginal coverage

Recall that the marginal coverage guarantee for a coverage level $1-\alpha \in (0,1)$ is stated as

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

for every $i \in I_{\text{test}}$ and suitable η. In the case of β-mixing sequences, η is as in Theorem 5.1.1.

Our goal in this section is to show that split conformal prediction works well not only under exchangeability assumptions, but also for stationary β-mixing data. Theorem 5.1.1 gives a marginal coverage guarantee that translates to results below.

We consider five base models (gradient boosting, k-nearest neighbors, linear regression, neural network and random forest) which are used as quantile estimators[1], four synthetic datasets (two-state hidden Markov model, hidden random walk on the cycle graph, autoregressive process and hidden renewal model) and three real-world datasets (EUR/USD spot exchange rate, Brent crude oil futures and S&P 500 futures). For synthetic experiments, 10000 simulations were performed, each comprising 1000 training points and 500 calibration points, with 1 single prediction performed for each previously unseen covariate, that is, $I_{\text{train}} = \{1,\dots,1000\}$, $I_{\text{cal}} = \{1001,\dots,1500\}$ and $I_{\text{test}} = \{1501\}$ for 10000 randomly generated sequences of 1501 points each. Nominally prescribed iid level was set to $1-\alpha = 0.9$. All experiments yield the same conclusion: split CP's marginal coverage is close to nominal iid levels, even for moderately dependent data, and the method undercovers only when dependence is extremely high. A natural question is whether the behavior is the same for different coverage levels. In Appendix C, we answer in the positive for $1-\alpha = 0.95$ and $1-\alpha = 0.85$.

6.2.1 Two-state hidden markov model

The level of dependence on a two-state hidden Markov model is dictated by the probabilities $1-p$ and $1-q$ of repeating the previous state: the closer to one, the more dependent is the generated sequence. When perfectly balanced with $p = q = 0.5$, the model was shown to be iid. For simplicity, we will consider $1-p = 1-q$. Figure 6.2 shows that for all prediction intervals, marginal coverage is maintained

[1]As quantile estimators, the models receive the names of gradient boosting quantile regressor, quantile k-nearest neighbors, linear quantile regression, neural network quantile regressor and quantile regression forest, as outlined in Section 2.4. For the sake of presentation, we will refer to the base models' names in all experiments.

unless the level of dependence is extreme, close to the maximum value of 1. Coverage remains above 89% even for large values of dependence, and falls below 88% only after $1 - p = 1 - q = 0.999$. It is possible to use the guarantees provided in Chapter 5 to adjust the quantile according to the desired nominal levels.

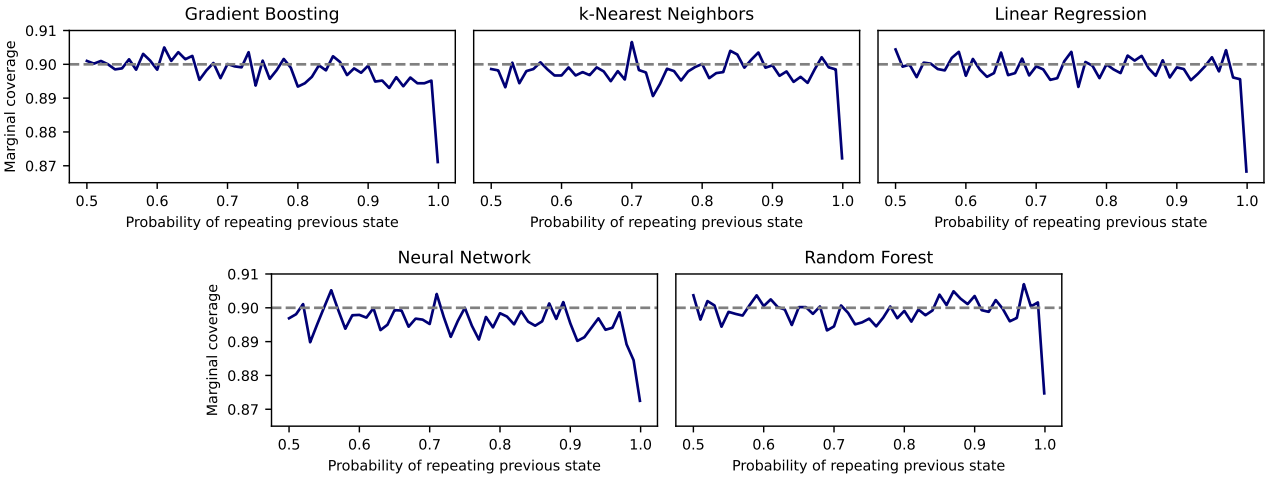

Figure 6.2: Marginal coverage for hidden Markov model with two underlying states (solid) and nominally prescribed iid level of $1 - \alpha = 0.9$ (dashed). Split CP guarantees hold well even under moderate dependence. Significant undercoverage only happens at extreme levels of dependence.

Figure 6.3 shows how the correction η in Theorem 5.1.1 depends on the calibration set sizes for the two-state hidden Markov model, quickly converging to the iid limit, even for moderately dependent data.

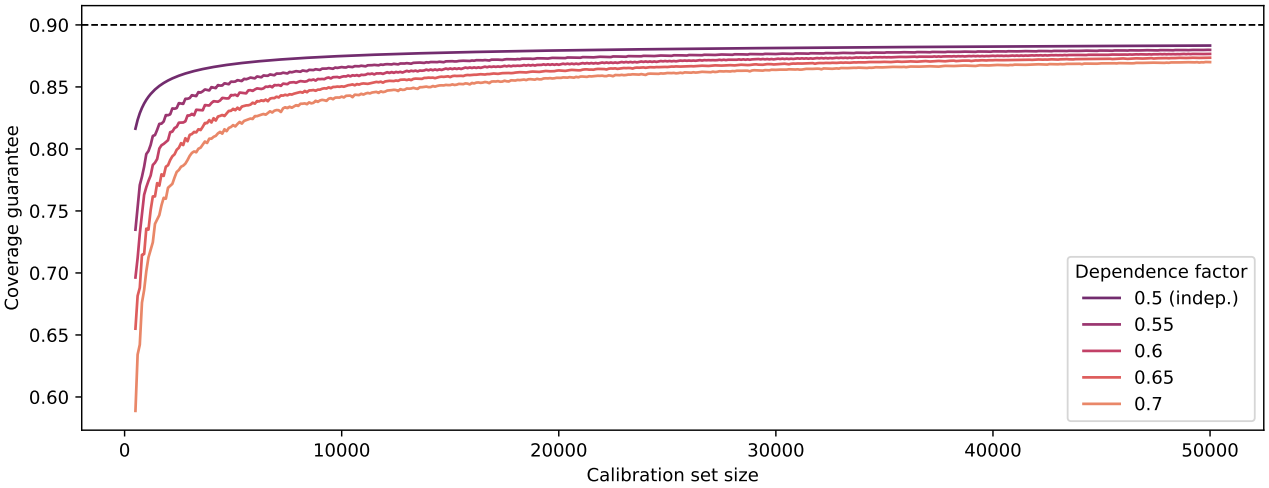

Figure 6.3: Marginal coverage guarantees (Theorem 5.1.1) for varying calibration set sizes, dependence levels, fixed $\delta = 0.01$ and fixed $\alpha = 0.1$. As the calibration sets increase in size, the guarantee under dependence converges to the iid case (cf. Section 3.5).

6.2.2 Hidden random walk on the cycle graph

The second stochastic process we evaluate is the hidden random walk on the cycle graph, whose dependence increases as the probability of not moving in the cycle increases. Marginal coverage is well behaved in general once again and undercoverage is observed only when the dependence is extreme (Figure 6.4).

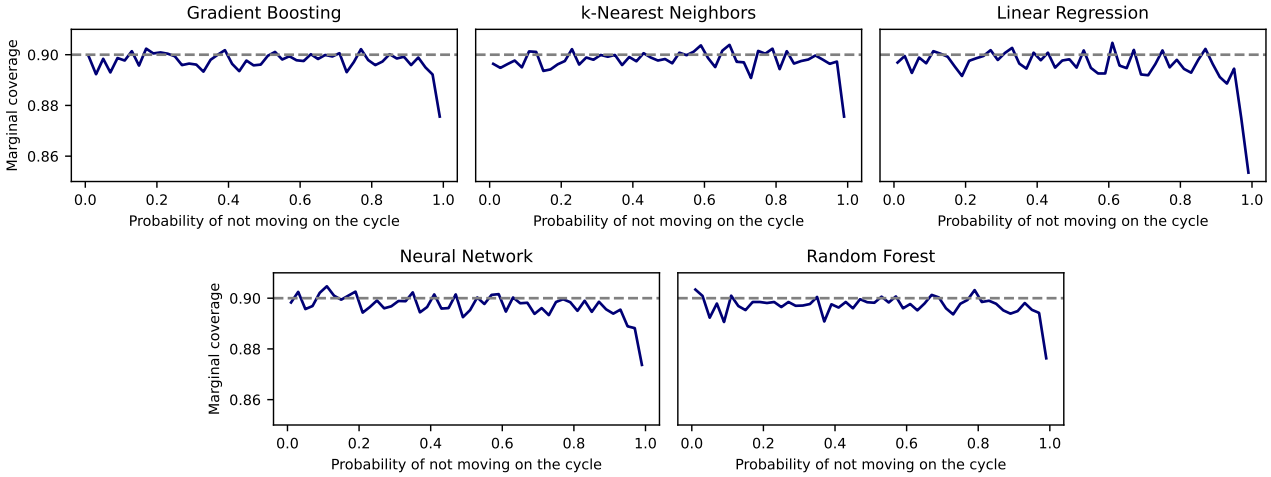

Figure 6.4: Marginal coverage for the hidden random walk on the cycle graph of 5 vertices (solid) and nominally prescribed iid level of $1 - \alpha = 0.9$ (dashed). Split CP guarantees hold well even under moderate dependence. Significant undercoverage only happens at extreme levels of dependence.

6.2.3 Autoregressive process

The autoregressive process is iid when its coefficient equals zero, but Figure 6.5 strikingly shows that marginal coverage remains close to prescribed nominal levels even when the coefficient is far from zero. Dependence increases from zero towards one and the gap between marginal and iid nominal coverages is extremely tight even for highly dependent data. Autoregressive coefficients up to $\lambda = 0.99$ achieve coverage higher than 89%. In particular, a significant loss of coverage only occurs when $\lambda = 0.999$.

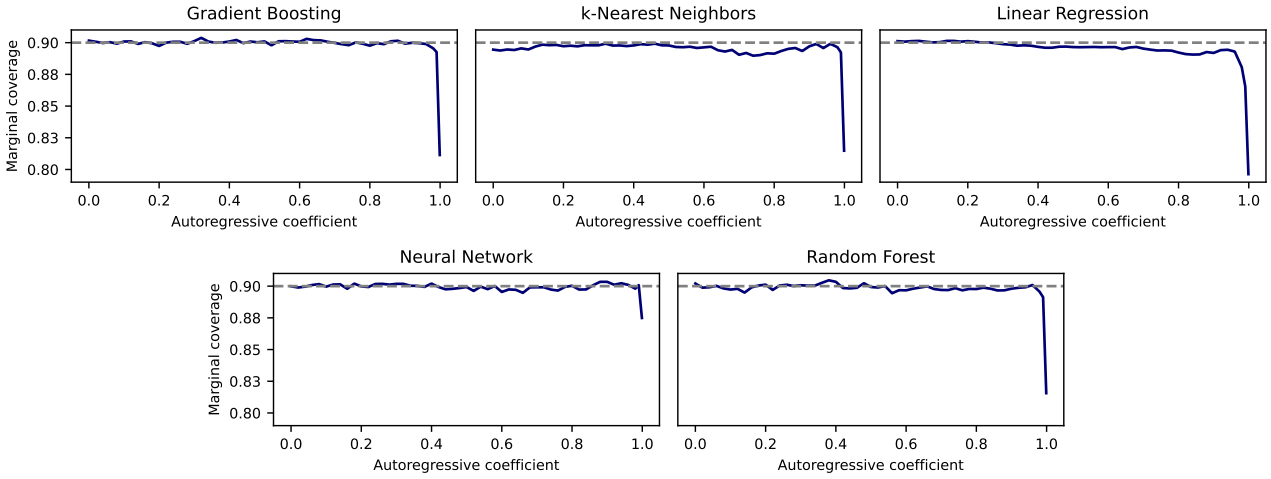

Figure 6.5: Marginal coverage for autoregressive process of order 1 (solid) and nominally prescribed iid level of $1 - \alpha = 0.9$ (dashed). Split CP guarantees hold well even under moderate dependence. Significant undercoverage only happens at extreme levels of dependence.

6.2.4 Hidden renewal model

Recall that the hidden renewal model previously presented was defined in terms of a base distribution F. Let $F(i) = 1 - \frac{n!}{\prod_{j=1}^{n} i+j}$, where $n \in \mathbb{N}_{\geq 0}$ is a parameter we can vary. No matter the n, the hidden renewal model is not independent. However, dependence does not increase or decrease monotonically as in the previous experiments. Nevertheless, Figure 6.6 shows that marginal coverage is generally between

0.890 and 0.905 for the prescribed nominal level of 0.90, for a number of parameters n and machine learning models. Therefore, once again, split conformal prediction behaves well under dependence.

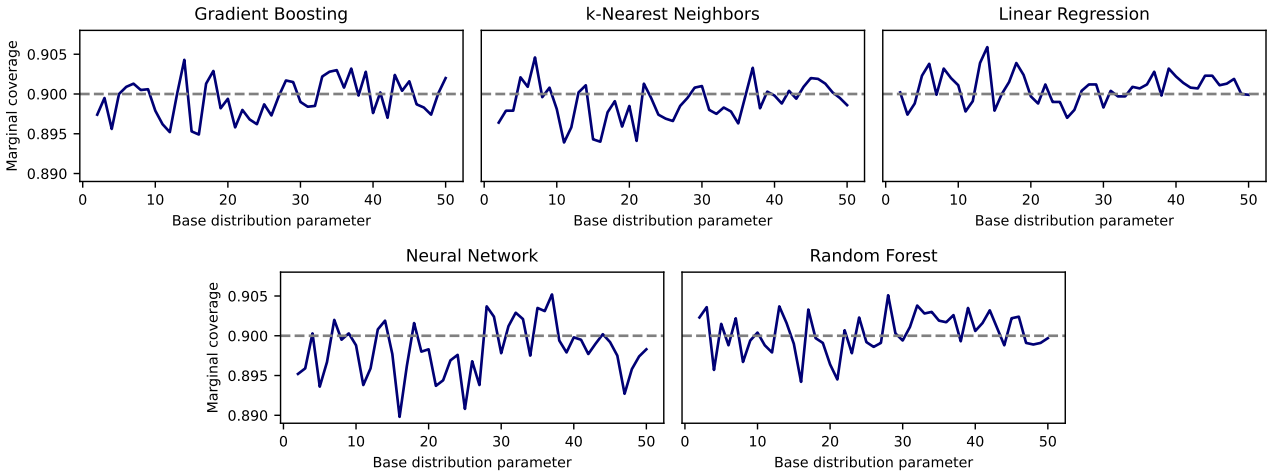

Figure 6.6: Marginal coverage for hidden renewal model (solid) and nominally prescribed iid level of $1 - \alpha = 0.9$ (dashed). Split CP guarantees hold well even under moderate dependence. Significant undercoverage only happens at extreme levels of dependence.

6.2.5 EUR/USD spot exchange rate

A foreign exchange (FX) spot transaction consists of two counterparties exchanging currencies at an agreed price *on the spot*, i.e., as soon as possible, usually within two business days. A currency pair gives the quotation of one currency (base currency) against another (quote currency). For example, consider the EUR/USD currency pair. Conventionally, the base currency comes first, so 1 EUR/USD indicates how many US dollars must be disbursed in order to acquire 1 euro. In an EUR/USD transaction, the buyer is understood as the one buying euro (selling dollar) and the seller is understood as the one selling euro (buying dollar), so the trading action can be thought as being applied to the base currency.

The average daily trading volume of the global spot FX market is about $2 trillion, with the FX market as a whole — essential for international transactions in goods, services and financial assets — surpassing $6 trillion, making it the largest financial market in the world (Chaboud, Rime, and Sushko 2022).

Foreign exchange is mostly traded in over-the-counter (OTC) markets, without centralized trading mechanisms. Trades are usually done privately between market participants or intermediated by dealers and brokers, so information is less readily available in comparison to centralized exchanges (Duffie 2012).

We performed 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, each point corresponding to a minute. Figure 6.7 shows the daily marginal coverage (Equation (3.5)) of the method. 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.

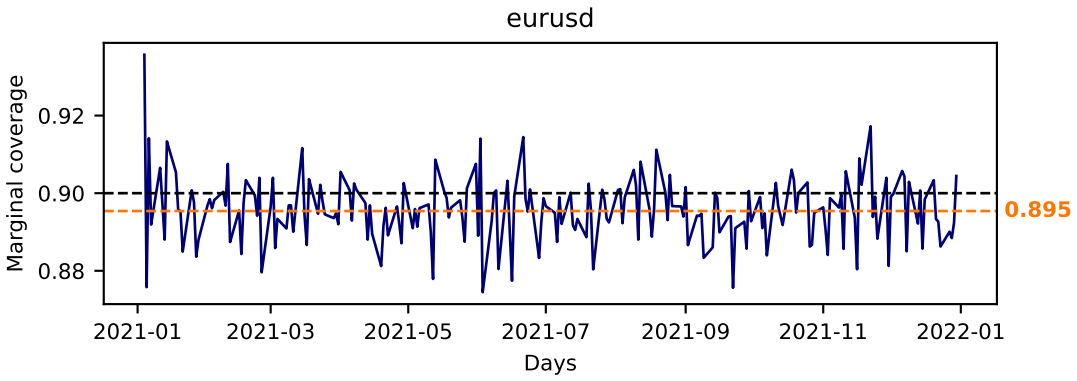

Figure 6.7: Daily marginal coverage of minute-by-minute online prediction for EUR/USD spot exchange rate (solid blue), nominally prescribed iid level of $1 - \alpha = 0.9$ (dashed black) and marginal coverage over the entire year (dashed orange).

6.2.6 Brent crude oil futures

Futures contracts are standardized financial instruments that allow one to buy or sell an underlying asset at a predetermined price and prespecified delivery time in the future. Delivery may be physical or cash-settled depending on the asset. Cash settlement stipulates that expiring contracts are settled by the transfer of cash from the seller to the buyer if the final settlement price is higher than the fixed trade price and from the buyer to the seller otherwise. Physical settlement means that sellers have the obligation of delivering the underlying asset once the contract expires and buyers have the obligation of receiving it, with incurring costs due to transportation, insurance and storage, for example. However, trading futures is highly facilitated by their standardized nature and it should be possible to transfer ownership of a contract ahead of expiration if one desires and there is enough liquidity. In contrast to spot exchange rates, futures are usually traded on centralized exchanges, such as the Chicago Mercantile Exchange (CME), the London Metal Exchange (LME) or the Intercontinental Exchange (ICE), which guarantee the specification of contracts. Besides speculation, futures contracts can be used for hedging a position, i.e., reducing downside (and upside) risk. A farmer may sell a futures contract to guarantee a specific price for their crop even before harvest. Likewise, an airplane company may buy crude oil or jet fuel futures to lock in the price and mitigate the risk of prices skyrocketing.

We will focus on a cash-settled commodity future of worldwide importance: Brent crude oil[2]. Following the same methodology from EUR/USD spot exchange rate experiment, online conformal prediction was performed for the year of 2021 and both the daily marginal coverage (solid blue line) and overall marginal coverage (dashed orange line) were calculated. The nominally prescribed iid level $1 - \alpha = 0.9$ is presented in as the dashed orange line, which allows us to conclude to split conformal prediction worked well also for Brent crude oil futures.

[2]The West Texas Intermediate (WTI) crude oil future is another contract of global importance. However, its delivery is physical, which can cause market turmoil and operational complications. On April 2020, for example, WTI futures plunged into negative territory near delivery date since associated buyer costs, such as storage of received barrels, were extremely high.

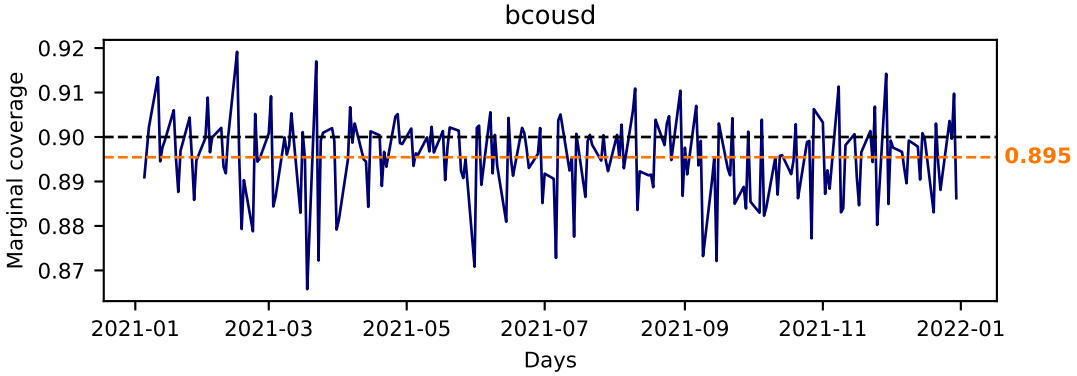

Figure 6.8: Daily marginal coverage of minute-by-minute online prediction for Brent crude oil futures (solid blue), nominally prescribed iid level of $1 - \alpha = 0.9$ (dashed black) and marginal coverage over the entire year (dashed orange).

6.2.7 S&P 500 futures

Finally, we consider another cash-settled future, the Standard and Poor's (S&P) 500 stock index future. The S&P 500 tracks the performance of 500 large companies traded on the United States, covering about 80% of available market capitalization. Its future contract allows one to buy or sell the index at a future date for a price determined beforehand. As in the EUR/USD and Brent experiments, we performed online conformal prediction with $1 - \alpha = 0.9$ in an online fashion over a sliding window with $n_{\text{train}} = 1000$, $n_{\text{cal}} = 500$ and $n_{\text{test}} = 1$, for test points comprising the entire year of 2021. Figure 6.9 shows once again that marginal coverage is close to 90%, as one would expect from the theory outlined in Chapter 3.

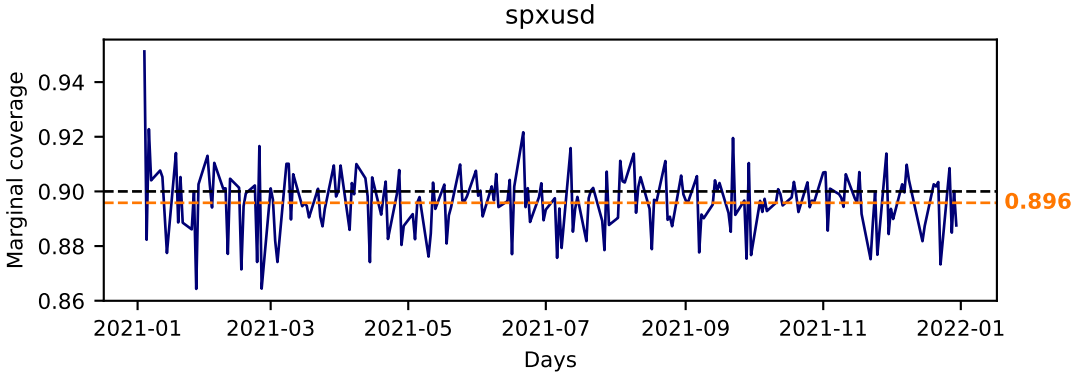

Figure 6.9: Daily marginal coverage of minute-by-minute online prediction for S&P 500 futures (solid blue), nominally prescribed iid level of $1 - \alpha = 0.9$ (dashed black) and marginal coverage over the entire year (dashed orange).

6.3 Empirical coverage

Recall the empirical coverage guarantee from Theorem 5.1.3:

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

for any $\delta_{\text{cal}}, \delta_{\text{test}} \in (0,1)$ and η as outlined in the theorem.

Consider a gradient boosting model trained with 1000 points and calibrated over a set of 15000 points with prescribed coverage level $1-\alpha = 0.9$. Prediction intervals are generated for 15000 test points. Figure 6.10 illustrates how empirical coverage $\frac{1}{n_{\text{test}}} \sum_{i \in I_{\text{test}}} \mathbb{1}\{Y_i \in C_{1-\alpha}(X_i)\}$ is always above the theoretical bound $1-\alpha-\eta$ with $\delta_{\text{cal}} = \delta_{\text{test}} = 0.005$ for the two-state hidden Markov model over a wide range of dependence levels, where the empirical coverage region represents 1000 simulations conducted for each individual dependence factor. Theorem 5.1.3 indicated that the empirical coverage should be above the theoretical bound of $1-\alpha-\eta$ at least 99% $(1-\delta_{\text{cal}}-\delta_{\text{test}})$ of the time. We observe that this indeed happens, but that the theoretical bound is conservative and empirical coverage is actually above it 100% of the time. Moreover, it is clear that the theoretical bound quickly decreases when dependence increases, but empirical coverage remains consistently high, indicating that moderate dependence is even less problematic than theory seems to suggest. Nevertheless, it provides a conservative lower bound of practical value, especially for mildly dependent processes.

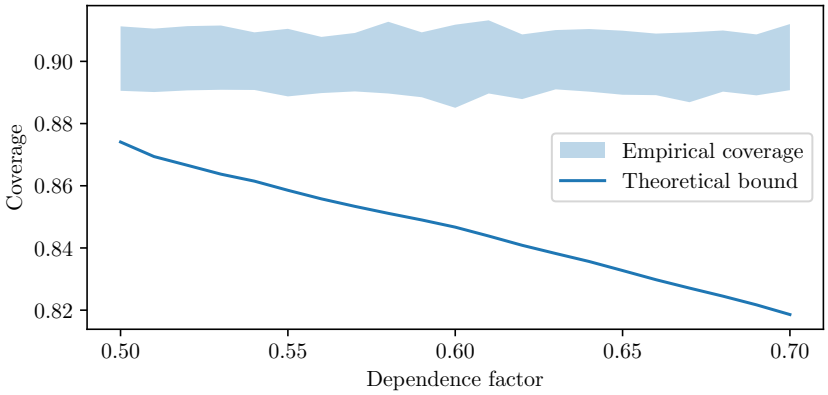

Figure 6.10: Empirical coverage and theoretical guarantee (Theorem 5.1.3) for $\delta_{\text{cal}} = \delta_{\text{test}} = 0.005$ and $1-\alpha = 0.9$. Empirical coverage is always above the theoretical bound.

6.4 Conditional coverage

Table 6.1 presents the conditional coverage (Equation (3.11)) on four distinct events of interest for all three financial datasets. 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 datasets. Following the results in Chapter 5, larger calibration sets have an important effect in improving coverage. In Appendix C, we report conditional coverage results for $1-\alpha = 0.95$ and $1-\alpha = 0.85$, strengthening the conclusion that coverage is retained after conditioning and that results generally improve when more calibration data is available.

Dataset	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 6.1: 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, as expected.

As expected, marginal, conditional and empirical coverage guarantees for split conformal prediction, traditionally known to hold for iid data, carry over to the β-mixing case, according to results presented in Chapter 5. Bounds can be conservative in practice, but were shown to still be useful, especially in a setting of low to moderate dependence.

6.5 Conformalized algorithmic trading

In this section, we argue that conformal prediction can aid algorithmic trading strategies in the quest for higher risk-adjusted returns. Intuitively, prediction intervals should be more informative than point predictions and quantifying uncertainty should give an edge when making investment decisions. Following the theoretical developments in Chapters 3, 4 and 5 and experiments above in this chapter, data dependence is not as disconcerting as once thought for split conformal prediction. Indeed, as proved and empirically observed, dependence becomes problematic only at extreme levels. Split CP will thus be used without further ado or concerns, even if the financial time series considered present temporal dependence.

6.5.1 Setting

We will consider in the experiments that follow seven highly liquid currencies quoted against the US dollar: Australian dollar (`audusd`), euro (`eurusd`), British pound (`gbpusd`), Japanese yen (`jpyusd`), New Zealand dollar (`nzdusd`), Canadian dollar (`usdcad`) and Swiss franc (`usdchf`). In contrast with the high frequency, minute-by-minute, data of previous experiments, we now consider daily prices, from the end of each business day, to calculate linear returns.

Gradient boosting regressors were trained to estimate the conditional median, which is equivalent to using the pinball loss (Equation (A.1)) with $\tau = 0.5$, which results in $L(y, \widehat{y}) = |y - \widehat{y}|$. Training occurred over 252 days (close to one year in business days), using five lagged observations of each one of the seven currency pair, totalling 35 features. The target variables were set as the subsequent observations of the seven currency pairs, so seven predictions were made at each time step. Unconventionally, but in agreement with the theory, the calibration set was chosen as the 252 days preceding training. The idea is that calibrating with past data is still valid and allows for the training procedure to use more

recent, possibly more relevant, observations. Prediction intervals were then generated for the 505th day following both calibration and training for different coverage levels. In an online fashion, the model was retrained and recalibrated for every new daily return.

6.5.2 Results

Following the procedure outlined above, prediction intervals for each of the seven currency pairs over different coverage levels were obtained for test points comprising the years of 2009 through 2021. Figure 6.11 shows, as expected, that higher coverage translates to larger intervals in all cases and that sufficiently small coverage levels converge to point estimates.

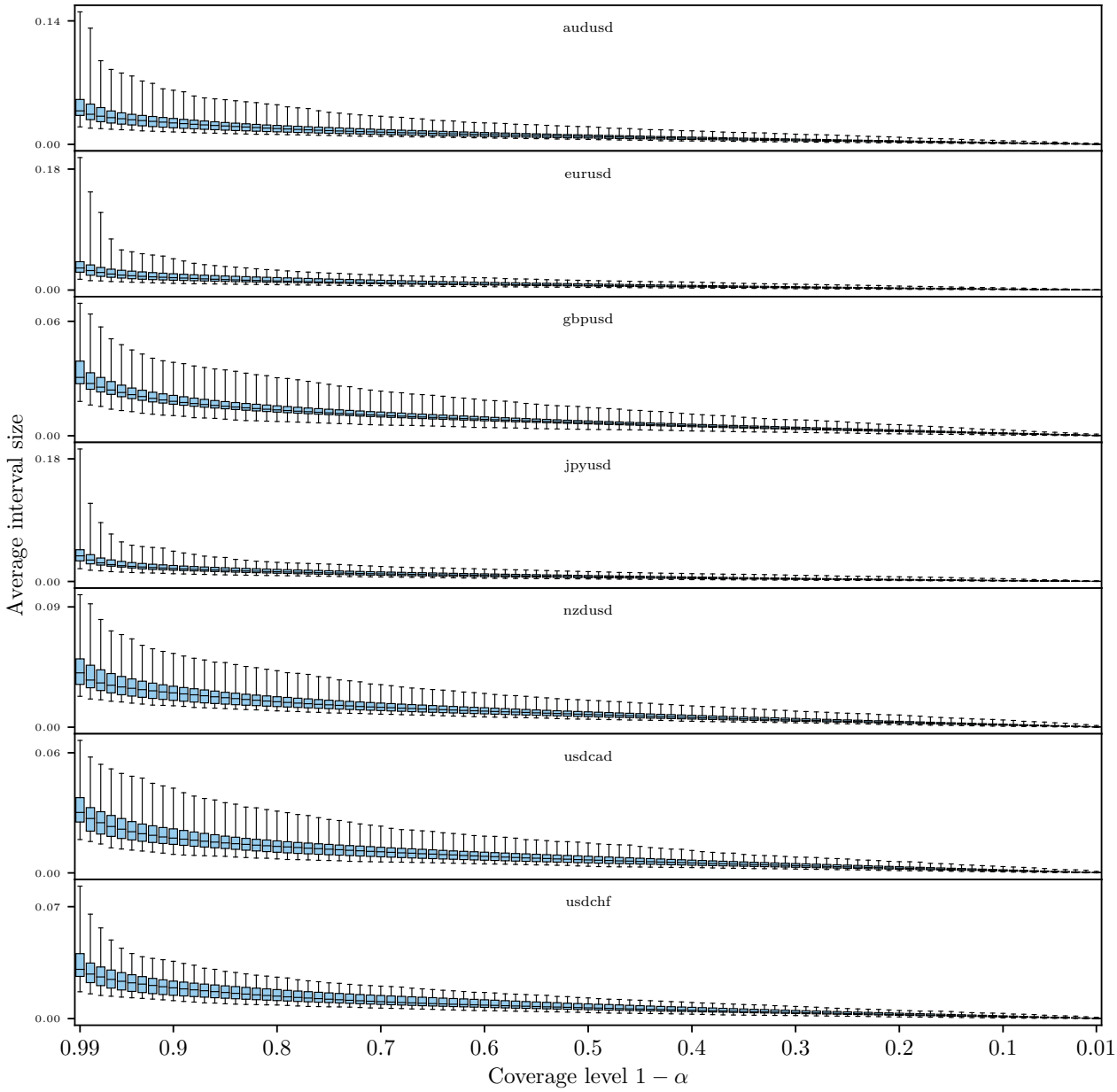

Figure 6.11: Average interval size given by split conformal prediction for different coverage levels $1 - \alpha$ on seven of the most liquid currency pairs. Prediction intervals generated on daily data from 2009 to 2021.

Given prediction intervals, a trading strategy can be developed to build portfolios as follows. Let $w_{i,j}$ represent the weight in the portfolio of a currency pair j on day i. Moreover, let $\widehat{y}_{i,j}^{\text{lo}}$ and $\widehat{y}_{i,j}^{\text{hi}}$ denote

the lower and upper bounds of the prediction interval, respectively, and $\mathcal{J}$ the set of currency pairs. We highlight in Figure 6.12 the result of the following strategy:

$$w'_{i,j} = \frac{\widehat{y}^{\text{hi}}_{i,j} + \widehat{y}^{\text{lo}}_{i,j}}{\widehat{y}^{\text{hi}}_{i,j} - \widehat{y}^{\text{lo}}_{i,j}} \cdot \mathbb{1}\{\text{sgn}(\widehat{y}^{\text{lo}}_{i,j}) = \text{sgn}(\widehat{y}^{\text{hi}}_{i,j})\} \cdot \mathbb{1}\{\min\{|\widehat{y}^{\text{lo}}_{i,j}|, |\widehat{y}^{\text{hi}}_{i,j}|\} > 0.0005\}, \tag{6.1}$$

$$w_{i,j} = \frac{w'_{i,j}}{\sum_{j \in \mathcal{J}} |w'_{i,j}|}. \tag{6.2}$$

Intuitively, weights are proportional to the prediction interval's midpoint, but normalized by the uncertainty as measured by the interval size. The indicator functions act as a filter: no investment is made unless the entire interval is above 0.0005 or below −0.0005, indicating that we expect, for the prescribed nominal coverage level, a profit that at least compensates transaction costs. Lastly, the transformation of $w'_{i,j}$ into $w_{i,j}$ is to ensure that, on any given day, weights sum up to one in absolute value, ensuring all capital is invested, without leverage or deleverage. The result of the strategy on 2021 for k-nearest neighbors and neural networks as base models, showcased in Figure 6.12, is measured in terms of risk-adjusted return, defined here as the average return of the portfolio normalized by the standard deviation of returns. Risk-adjusted return metrics are widely used to compare strategies, taking into account both financial gains and incurred risk. Note that a transaction cost of 0.05% was considered for all operations. It is evident that a good choice of coverage level $1 - \alpha$ can be useful for the strategy.

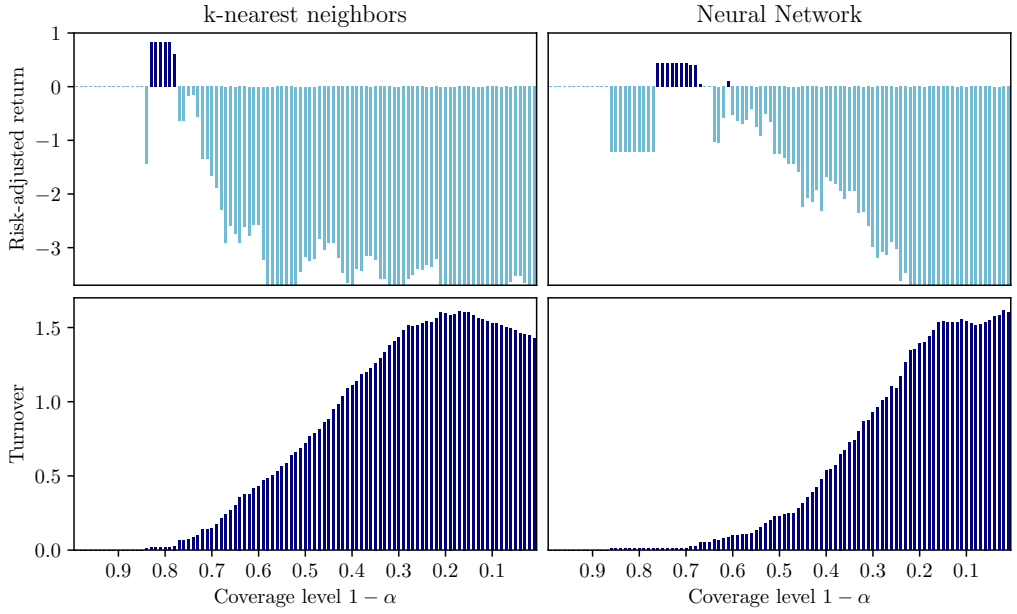

Figure 6.12: Risk-adjusted return (top) and turnover (bottom) of the portfolio for different coverage levels $1 - \alpha$, showing that there are gains to be extracted from prediction intervals generated with k-nearest neighbors (left) and neural networks (right) as base quantile regression models for CQR.

The detrimental result can be partially explained by transaction costs. Turnover peaks around $1 - \alpha = 0.2$ and is greatly reduced for higher coverage levels, an expected behavior, given that more weights will be equal to zero in this scenario due to the filter considered in Equation (6.1).

Although it is possible to select a coverage level that ensures profit in hindsight for many years, that could be a difficulty in practice. Moreover, there are years in which no profit could be made with this strategy, no matter the coverage level. Indeed, our intention here is merely to show a simple and

modest way in which prediction intervals could be of value, not to claim a robust and profitable strategy. With proper feature engineering, an ensemble of models and a method for selecting coverage levels on the fly, as well as a thoughtful allocation strategy, this idea could perhaps lead to more consistent results, but it remains as a proof of concept as of right now.

Chapter 7

Conclusion

We have given a thorough overview of standard conformal prediction, starting with the general full CP method and showing how split CP follows as a particular case. Different notions of coverage were presented as well as a discussion on nonconformity scores. Motivated by the nonexchangeable nature of financial time series, a recent concentration of measure approach to split conformal prediction was presented. Under this framework, split CP was shown to retain marginal and conditional coverage guarantees, with the addition of a small penalty term that empirically had a minute effect. We then defined stochastic processes and some important properties such as stationarity and β-mixing (absolute regularity). Markov chains, autoregressive processes and renewal processes were given as examples of stochastic processes that are stationary and absolutely regular under certain conditions and we showed how to exactly calculate the β-mixing coefficient of Markov chains in particular. Next, we showcased that split CP could be used for β-mixing processes by applying the concentration of measure framework, which suddenly enabled the method to reach further than previously thought, much beyond exchangeable data. Indeed, experiments on synthetic dependent data corroborated theoretical results. More importantly, split CP was well justified to be applied to financial time series that exhibited vanishing autocorrelation and stationarity. Real data experiments with three financial datasets further cemented the argument that split conformal prediction excels for uncertainty quantification even for nonexchangeable data: in all experiments, we observed marginal and conditional coverage close to nominally prescribed iid levels. Lastly, we displayed how conformal prediction could be useful in algorithmic trading strategies by providing more information than point predictions. Still, applications of CP to finance remain a largely unexplored area of research.

Statistical arbitrage is a general class of quantitative trading strategies that employ statistical methods to identify temporal price differences between assets and profit from those deviations (Guijarro-Ordonez, Pelger, and Zanotti 2021). Although many parametric and nonparametric models have been considered for signal extraction and allocation decisions, uncertainty quantification is usually a minor or nonexistent concern. Avellaneda and Lee (2010) build a mean-reversion strategy based on an Ornstein-Uhlenbeck stochastic process and filter signals whose estimated speed of mean reversion is too low, as the short-term behavior is more uncertain and poses higher risk. Yeo and Papanicolaou (2017) go one step further by rejecting trading signals generated from an Ornstein-Uhlenbeck estimation whose coefficient of determination R^2 is below a given threshold[1]. Such minor uncertainty filters (estimated mean-reversion speed and R^2) were shown to improve the strategy and reaffirm its importance in finance. Conformal prediction could likely be used as a more fundamental technique for quantifying

[1]The Ornstein-Uhlenbeck process is discretized as an AR(1) process, so the coefficient of determination refers to this regression.

uncertainty, in the form of intervals for some estimations done throughout statistical arbitrage pipelines. Hopefully, the richer information would translate to more profitable strategies.

Wisniewski, Lindsay, and Lindsay (2020) applied split CP to generate prediction intervals of market makers' net positions, as discussed in Chapter 1. Since the field of distribution-free uncertainty quantification is rapidly evolving, many novelties could be employed in the same market making problem. We highlight Feldman, Bates, and Romano (2022) and Bastani et al. (2022) as recent approaches that do not require a calibration set but are computationally efficient. Financial time series in general might benefit from those new approaches, as making use of all data, especially more recent data, could improve predictions. One advantage of building upon the work of Wisniewski, Lindsay, and Lindsay (2020) is that the dataset was made public and makes for a good benchmark.

Lastly, we point out that Sun and Boyd (2018) developed a strategy for practical betting with uncertainty which is optimal in the long run. Classic Kelly gambling suffers from the optimizer's curse: empirical in-sample distributions usually significantly differ from out-of-sample distributions in investment and decisions based on empirical nominal distributions can lead to unsatisfactory results when deployed. The optimizer's curse is overcome by embracing the fact that the distribution is not known and considering a given set of possible distributions. The authors briefly mention that conformal prediction could likely be used as an alternative to the construction of the uncertainty sets from investment data.

In summary, financial problems are abundant and inherently surrounded by uncertainty. Conformal prediction has been enjoying flourishing theoretical and practical developments recently, but applications to finance remain scarce. By reviewing recent results that prove that split CP can be used for dependent data, we were able to overcome the nonexchangeable nature of financial data and scratch the surface of possible applications. We believe that conformal prediction in finance will flourish as a field and continue to provide insights to many real-world problems.

Appendix A

Hyperparameters

All quantile regression models used in Chapter 6 were trained with fixed hyperparameters. The implementation of linear quantile regression was from `Scikit-learn` (Pedregosa et al. 2011, version 1.0.2); quantile regression forests and quantile k-nearest neighbors from `sklearn-quantile` (Roebroek 2022, version 0.0.18); gradient boosting from `LightGBM` (Ke et al. 2017, version 3.3.2) and the neural network from `PyTorch` (Paszke et al. 2019, version 1.11.0). The pinball loss, used to optimize some of the models, is defined in terms of a true label y, a prediction $\widehat{y}$ and a quantile τ to be estimated:

$$L_\tau(y, \widehat{y}) = \tau(y - \widehat{y})\,\mathbb{1}\{y \geq \widehat{y}\} + (1 - \tau)(\widehat{y} - y)\,\mathbb{1}\{y < \widehat{y}\}. \tag{A.1}$$

Unless otherwise noted, quantile regressors making use of the pinball loss had τ set to $\alpha/2$ and $1 - \alpha/2$, where α is the acceptable miscoverage level.

Linear Quantile Regression The model was fit with an intercept; pinball loss was used; L1 regularization was added to the loss function; HiGHS (Huangfu and Hall 2018) was used to solve the linear programming formulation of the problem.

Gradient Boosting The model was set to boost 100 trees with a learning rate of 0.1 and pinball loss function; trees of any depth were allowed; the minimal number of data in one leaf was set to 20; the minimal sum hessian in one leaf was set to 0.001; no minimal gain to perform a split was required; no more than 31 leaves were allowed per tree; no L1 or L2 regularization was set.

Quantile k-Nearest Neighbors Five neighbors were considered; the weight function chosen was uniform, i.e., all points were weighted equally; the distance metric used for the tree was the Euclidean distance (Minkowski metric with power 2); the algorithm used to compute the nearest neighbors is automatically selected by the package between BallTree, KDTree and brute force; a leaf size of 30 is passed to BallTree or KDTree in case they are selected.

Quantile Regression Forest The model was trained with 10 estimators; mean squared error was used to measure the quality of a split; no maximum tree depth was set, so nodes are expanded until all leaves contain less than 2 samples; all features are considered when looking for the best split.

Neural Network The neural network consisted of three fully connected layers, with rectified linear unit (ReLU) activations between them; 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 (Loshchilov and Hutter 2019) was used as the stochastic optimization algorithm, with learning rate of 10^{-3} and weight decay of 10^{-6}; training occurred over 100 epochs with batches of size 64; pinball loss was used.

Appendix B

Technical Results

In Section 3.1, we proved McDiarmid's inequality (Theorem 3.1.14) based on Azuma's inequality (Theorem 3.1.12) as done in Mohri, Rostamizadeh, and Talwalkar (2018). For completeness and given its widespread use in machine learning theory, we provide an alternative proof of McDiarmid's inequality based on the notion of entropy, following Boucheron, Lugosi, and Massart (2013) in most part.

Definition B.0.1 (Φ-entropy)**.** *Let Φ be a convex function defined on an interval and Z an integrable random variable taking values in the same interval. The Φ-entropy of Z is defined as*

$$\mathrm{Ent}_\Phi[Z] \coloneqq \mathbb{E}[\Phi(Z)] - \Phi(\mathbb{E}[Z]).$$

Remark B.0.2. *The usual notion of variance can be retrieved as a special case of Φ-entropy by taking $\Phi(x) = x^2$:*

$$\mathrm{Ent}_{x\mapsto x^2}[Z] = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 = \mathrm{Var}[Z].$$

Definition B.0.3 (Entropy)**.** *The entropy of a random variable Z is defined as its Φ-entropy with $\Phi(x) = x\log(x)$:*

$$\begin{aligned}\mathrm{Ent}[Z] &\coloneqq \mathrm{Ent}_{x\mapsto x\log(x)}[Z] \\ &= \mathbb{E}[Z\log Z] - \mathbb{E}[Z]\log\mathbb{E}[Z]\end{aligned}$$

Remark B.0.4. *Shannon entropy, defined for a discrete random variable Z with probability mass function $p(Z)$ by $\mathbb{E}[-\log p(Z)]$, may also be referred to simply as entropy in the literature. However, the concept of entropy for us will always be that of Definition B.0.3.*

Definition B.0.5 (Conditional entropy)**.** *Let $W_1, \ldots, W_n$ be independent random variables. Define $W^{(i)} \coloneqq \{W_j\}_{j\in[n]\setminus i}$ the same set of random variables but with W_i excluded. Moreover, define $\mathbb{E}^{(i)} \coloneqq \mathbb{E}[\cdot|W^{(i)}]$ the conditional expectation with respect to $W^{(i)}$ for every $i \in [n]$. Finally, let $Z = f(W_1, \ldots, W_n)$ be a nonnegative measurable function of $W_1, \ldots, W_n$ such that $Z\log Z$ is integrable. Then, the conditional entropy of Z given $W^{(i)}$ is defined as*

$$\mathrm{Ent}^{(i)}[Z] \coloneqq \mathbb{E}^{(i)}[Z\log Z] - \mathbb{E}^{(i)}[Z]\log\mathbb{E}^{(i)}[Z].$$

Theorem B.0.6 (Duality formula of entropy)**.** *Let Z be a nonnegative random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}[\Phi(Z)]$ is bounded and let $\mathcal{U}$ be the set of all random variables*

$U : \Omega \to \mathbb{R} \cup \{-\infty, +\infty\}$ *with* $\mathbb{E}[e^U] = 1$. *Then, the duality formula of entropy is given by*

$$\mathrm{Ent}[Z] = \sup_{U \in \mathcal{U}} \mathbb{E}[UZ].$$

Proof. Define $\mathrm{Ent}_{e^U\mathbb{P}}$ as Ent but with expectations taken with respect to the $e^U\mathbb{P}$ measure instead of the usual $\mathbb{P}$ measure. Then, for any U with $\mathbb{E}[e^U] = 1$,

$$\mathrm{Ent}[Z] - \mathbb{E}[UZ] = \mathrm{Ent}_{e^U\mathbb{P}}[Ze^{-U}],$$

which implies

$$\mathrm{Ent}[Z] - \mathbb{E}[UZ] \geq 0.$$

Therefore, as $\mathrm{Ent}[Z]$ will always be greater than or equal to $\mathbb{E}[UZ]$, no matter the U, with equality achieved for $e^U = Z/\mathbb{E}[Z]$. □

Theorem B.0.7 (Subadditivity of entropy)**.** *Let $W_1, \ldots, W_n$ be independent random variables. For $Z = f(W_1, \ldots, W_n)$ a nonnegative measurable function such that $Z \log Z$ is integrable,*

$$\mathrm{Ent}[Z] \leq \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{Ent}^{(i)}[Z]\right].$$

Proof. Let $\mathbb{E}_i[\cdot]$ be the expectation operator conditioned on $W_1, \ldots, W_i$ for $i = 1, \ldots, n$, that is, $\mathbb{E}_i[\cdot] \coloneqq \mathbb{E}[\cdot|W_1, \ldots, W_i]$. Convention that E_0 is the usual expectation operator $\mathbb{E}$. Note that $\mathbb{E}_n$ is the identity when restricted to $(W_1, \ldots, W_n)$-measurable and integrable random variables, so

$$Z(\log Z - \log \mathbb{E}[Z]) = \sum_{i=1}^{n} Z(\log(\mathbb{E}_i[Z]) - \log(\mathbb{E}_{i-1}[Z])).$$

The duality formula from Theorem B.0.6 applied to $U \coloneqq \log T - \log \mathbb{E}[T]$for nonnegative and integrable random variables T gives

$$\mathbb{E}^{(i)}\left[Z\left(\log(\mathbb{E}_i[Z]) - \log(\mathbb{E}^{(i)}[E_i[Z]])\right)\right] \leq \mathrm{Ent}^{(i)}[Z].$$

Noting that $\mathbb{E}^{(i)}[\mathbb{E}_i[Z]] = \mathbb{E}_{i-1}[Z]$ due to independence of $W_1, \ldots, W_n$, taking expectations yields

$$\begin{aligned}\mathbb{E}[Z(\log Z - \log \mathbb{E}[Z])] &= \sum_{i=1}^{n} \mathbb{E}[\mathbb{E}^{(i)}[Z(\log(\mathbb{E}_i[Z]) - \log(\mathbb{E}^{(i)}[\mathbb{E}_i[Z]]))]] \\ &\leq \sum_{i=1}^{n} \mathbb{E}[\mathrm{Ent}^{(i)}[Z]].\end{aligned}$$

□

Proposition B.0.8 (Herbst's argument)**.** *Let Z be an integrable random variable such that for some $v > 0$, we have*

$$\forall \lambda > 0 \colon \quad \frac{\mathrm{Ent}[e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} \leq \frac{\lambda^2 v}{2}.$$

Then,

$$\forall \lambda > 0 \colon \quad \log \mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] \leq \frac{\lambda^2 v}{2}.$$

Proof. It follows from the definition of entropy and algebraic manipulation that

$$\begin{aligned}\frac{\mathrm{Ent}[e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} &= \frac{\mathbb{E}[e^{\lambda Z}\lambda Z] - \mathbb{E}[e^{\lambda Z}]\log\mathbb{E}[e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} \\ &= \lambda\frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \log\mathbb{E}[e^{\lambda Z}] \\ &= \lambda\psi_Z'(\lambda) - \psi_Z(\lambda),\end{aligned} \tag{B.1}$$

Recall that $\psi_Z(\lambda)$ was first defined in Lemma 3.1.7 and its derivative also calculated therein. Now, let $\widetilde{Z} \coloneqq Z - \mathbb{E}[Z]$ and observe that

$$\begin{aligned}\lambda\psi_{\widetilde{Z}}'(\lambda) &= \lambda\frac{\mathbb{E}[(Z-\mathbb{E}[Z])e^{\lambda(Z-\mathbb{E}[Z])}]}{\mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}]} \\ &= \frac{\lambda\mathbb{E}[Ze^{\lambda(Z-\mathbb{E}[Z])}]}{\mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}]} - \frac{\lambda\mathbb{E}[\mathbb{E}[Z]e^{\lambda(Z-\mathbb{E}[Z])}]}{\mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}]} \\ &= \frac{\lambda\mathbb{E}[Ze^{\lambda Z}e^{-\lambda\mathbb{E}[Z]}]}{\mathbb{E}[e^{\lambda Z}e^{-\lambda\mathbb{E}[Z]}]} - \frac{\lambda\mathbb{E}[Z]\mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}]}{\mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}]} \\ &= \frac{\lambda\mathbb{E}[Ze^{\lambda Z}]e^{-\lambda\mathbb{E}[Z]}}{\mathbb{E}[e^{\lambda Z}]e^{-\lambda\mathbb{E}[Z]}} - \lambda\mathbb{E}[Z] \\ &= \frac{\lambda\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \lambda\mathbb{E}[Z],\end{aligned}$$

and that

$$\begin{aligned}\psi_{\widetilde{Z}}(\lambda) &= \log\mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] \\ &= \log\mathbb{E}[e^{\lambda Z}e^{-\lambda\mathbb{E}[Z]}] \\ &= \log(\mathbb{E}[e^{\lambda Z}]e^{-\lambda\mathbb{E}[Z]}) \\ &= \log\mathbb{E}[e^{\lambda Z}] - \log e^{\lambda\mathbb{E}[Z]} \\ &= \log(\mathbb{E}[e^{\lambda Z}]) - \lambda\mathbb{E}[Z],\end{aligned}$$

which gives

$$\begin{aligned}\lambda\psi_{\widetilde{Z}}'(\lambda) - \psi_{\widetilde{Z}}(\lambda) &= \frac{\lambda\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \lambda\mathbb{E}[Z] - \log(\mathbb{E}[e^{\lambda Z}]) + \lambda\mathbb{E}[Z] \\ &= \frac{\lambda\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \log\mathbb{E}[e^{\lambda Z}] \\ &= \lambda\psi_Z'(\lambda) - \psi_Z(\lambda).\end{aligned}$$

Returning to Equation (B.1), we have

$$\begin{aligned}\frac{\mathrm{Ent}[e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} &= \lambda\psi_Z'(\lambda) - \psi_Z(\lambda) \\ &= \lambda\psi_{\widetilde{Z}}'(\lambda) - \psi_{\widetilde{Z}}(\lambda) \\ &\leq \frac{\lambda^2 v}{2},\end{aligned}$$

which implies

$$\frac{1}{\lambda}\psi_{\widetilde{Z}}'(\lambda) - \frac{1}{\lambda^2}\psi_{\widetilde{Z}}(\lambda) \leq \frac{v}{2}.$$

Defining $G(\lambda) \coloneqq \frac{1}{\lambda}\psi_{\widetilde{Z}}(\lambda)$ so that

$$G'(\lambda) = \frac{1}{\lambda}\psi'_{\widetilde{Z}}(\lambda) - \frac{1}{\lambda^2}\psi_{\widetilde{Z}}(\lambda) \leq \frac{v}{2},$$

it follows that

$$\int_0^\lambda G'(\theta)d\theta \leq \int_0^\lambda \frac{v}{2}d\theta \implies G(\lambda) - \lim_{\theta\to 0} G(\theta) \leq \frac{\lambda v}{2} \implies G(\lambda) \leq \frac{\lambda v}{2} \implies \lambda G(\lambda) \leq \frac{\lambda^2 v}{2},$$

since

$$\lim_{\theta\to 0} G(\theta) = \lim_{\theta\to 0} \frac{\log \mathbb{E}[e^{\theta(Z-\mathbb{E}[Z])}]}{\theta} = \lim_{\theta\to 0} \frac{\log(1)}{\theta} = 0.$$

Rewriting $\lambda G(\lambda)$ in the desired form, we conclude:

$$\lambda G(\lambda) = \psi_{\widetilde{Z}}(\lambda) = \log \mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] \leq \frac{\lambda^2 v}{2}. \qquad \square$$

Theorem B.0.9 (McDiarmid's inequality (same as Theorem 3.1.14; alternative proof))**.** *Let f be a function that satisfies the bounded differences condition (Definition 3.1.13) with constants $c_1, \ldots, c_n > 0$ and define*

$$v \coloneqq \frac{1}{4}\sum_{i=1}^n c_i^2.$$

For independent random variables $W_1, \ldots, W_n$, set $Z = f(W_1, \ldots, W_n)$. Then, for any $t > 0$,

$$\mathbb{P}[Z - \mathbb{E}[Z] > t] \leq \exp\left(\tfrac{-t^2}{2v}\right),$$

and, by symmetry of the bounded differences assumption,

$$\mathbb{P}[Z - \mathbb{E}[Z] < -t] \leq \exp\left(\tfrac{-t^2}{2v}\right).$$

Proof. Let us consider the zero-mean random variable $\widetilde{W} \coloneqq W - \mathbb{E}[W]$ defined on the interval $[a, b]$. Applying Hoeffding's lemma (Lemma 3.1.7) yields, for every $\lambda \in \mathbb{R}$,

$$\psi''_{\widetilde{W}}(\lambda) \leq \frac{(b-a)^2}{4},$$

which implies

$$\int_0^\lambda \theta\psi''_{\widetilde{W}}(\theta)d\theta \leq \int_0^\lambda \theta\frac{(b-a)^2}{4}d\theta = \frac{(b-a)^2\lambda^2}{8}.$$

The left-hand side of the inequality can be solved by integration by parts, giving us

$$\begin{aligned}
\int_0^\lambda \theta\psi''_{\widetilde{W}}(\theta)d\theta &= \theta\psi'(\theta)\Big|_0^\lambda - \int_0^\lambda \psi'(\theta)d\theta \\
&= \lambda\psi'(\lambda) - \psi(\lambda) + \psi(0) \\
&= \lambda\psi'(\lambda) - \psi(\lambda) \\
&= \frac{\mathrm{Ent}[e^{\lambda\widetilde{W}}]}{\mathbb{E}[e^{\lambda\widetilde{W}}]},
\end{aligned}$$

where the last equality follows from Equation (B.1). The resulting bound,

$$\frac{\mathrm{Ent}[e^{\lambda \widetilde{W}}]}{\mathbb{E}[e^{\lambda \widetilde{W}}]} \leq \frac{(b-a)^2\lambda^2}{8}, \tag{B.2}$$

implies Hoeffding's inequality (Theorem 3.1.8) after applying Herbst's argument (Proposition B.0.8) and can be seen as a stronger version of it.

Now, note that conditioned on $W^{(i)}$ for any $i \in [n]$, the random variable Z takes values in an interval whose length does not exceed c_i due to the bounded differences assumption. Therefore, Equation (B.2) can be applied to yield

$$\frac{\mathrm{Ent}^{(i)}[e^{\lambda Z}]}{\mathbb{E}^{(i)}[e^{\lambda Z}]} \leq \frac{c_i^2\lambda^2}{8} \implies \mathrm{Ent}^{(i)}[e^{\lambda Z}] \leq \frac{c_i^2\lambda^2}{8} \cdot \mathbb{E}^{(i)}[e^{\lambda Z}].$$

By the subadditivity of entropy (Theorem B.0.7),

$$\begin{aligned}
\mathrm{Ent}[e^{\lambda Z}] &\leq \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{Ent}^{(i)}[e^{\lambda Z}]\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^{n} \frac{c_i^2\lambda^2}{8} \cdot \mathbb{E}^{(i)}[e^{\lambda Z}\right] \\
&= \sum_{i=1}^{n} \frac{c_i^2\lambda^2}{8} \cdot \mathbb{E}[e^{\lambda Z}],
\end{aligned}$$

where the last equality holds due to the law of iterated expectations. We may equivalently state

$$\begin{aligned}
\frac{\mathrm{Ent}[e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} &\leq \frac{\lambda^2 \sum_{i=1}^{n} c_i^2}{8} \\
&= \frac{\lambda^2 v}{2},
\end{aligned}$$

for v defined as in the theorem's statement. Herbst's argument for $\widetilde{Z} \coloneqq Z - \mathbb{E}[Z]$ now gives

$$\psi_{\widetilde{Z}}(\lambda) = \log \mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] \leq \frac{\lambda^2 v}{2},$$

and the generalized Markov's inequality (Theorem 3.1.2) applied to the random variable $Z - \mathbb{E}[Z]$ with $\phi(x) = e^{\lambda x}$ lets us conclude

$$\begin{aligned}
\mathbb{P}[Z - \mathbb{E}[Z] > t] &\leq \frac{\mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}]}{e^{\lambda t}} \\
&= \exp(\log \mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}]) \cdot \exp(-\lambda t) \\
&= \exp(\psi_{\widetilde{Z}}(\lambda) - \lambda t) \\
&\leq \exp\left(\tfrac{\lambda^2 v}{2} - \lambda t\right).
\end{aligned}$$

Finally, taking $\lambda = \frac{t}{v}$:

$$\begin{aligned}
\mathbb{P}[Z - \mathbb{E}[Z] > t] &\leq \exp\left(\tfrac{\lambda^2 v}{2} - \lambda t\right) \\
&= \exp\left(\tfrac{t^2 v}{2v^2} - \tfrac{t^2}{v}\right) \\
&= \exp\left(\tfrac{t^2}{2v} - \tfrac{2t^2}{2v}\right) \\
&= \exp\left(\tfrac{-t^2}{2v}\right).
\end{aligned}$$

The complementary tail bound, $\mathbb{P}[Z - \mathbb{E}[Z] < -t] \leq \exp\left(\frac{-t^2}{2v}\right)$, is a direct consequence of the symmetry due to the bounded differences assumption, as noted in the theorem's statement. □

Appendix C

Further Experiments

All experiments presented in Chapter 6 were conducted for a nominally prescribed iid level of $1-\alpha = 0.9$. It is natural to wonder if the same conclusions would be reached for other coverage levels. As anticipated in Chapter 6 and as expected from the theoretical results, that is indeed the case. Below, experiments are replicated for $1 - \alpha = 0.95$ and $1 - \alpha = 0.85$, showcasing coherent results.

Nominally prescribed iid level: 95%.

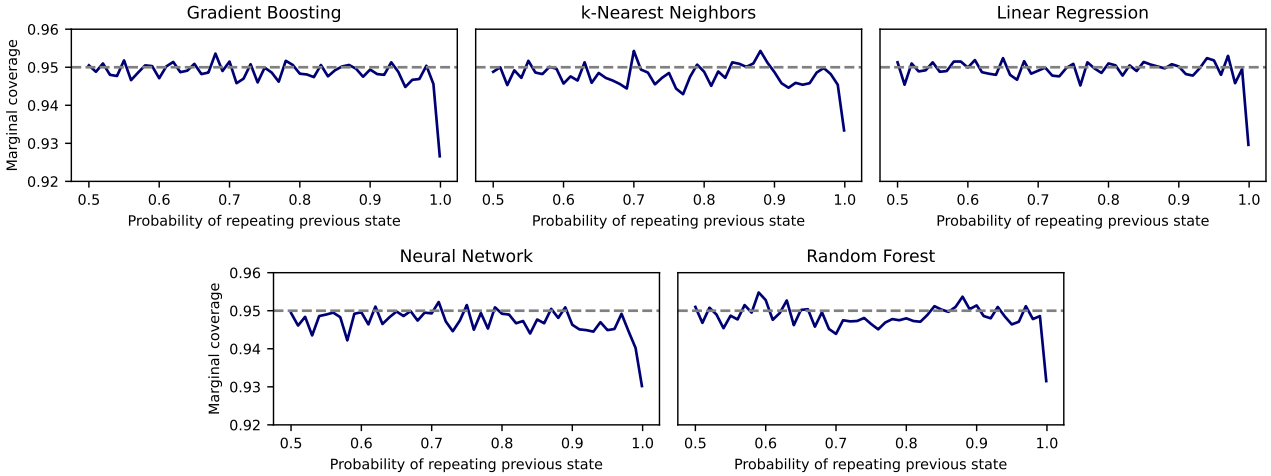

Figure C.1: Marginal coverage for hidden Markov model with two underlying states (solid) and nominally prescribed iid level of $1 - \alpha = 0.95$ (dashed). Split CP guarantees hold well even under moderate dependence. Significant undercoverage only happens at extreme levels of dependence.

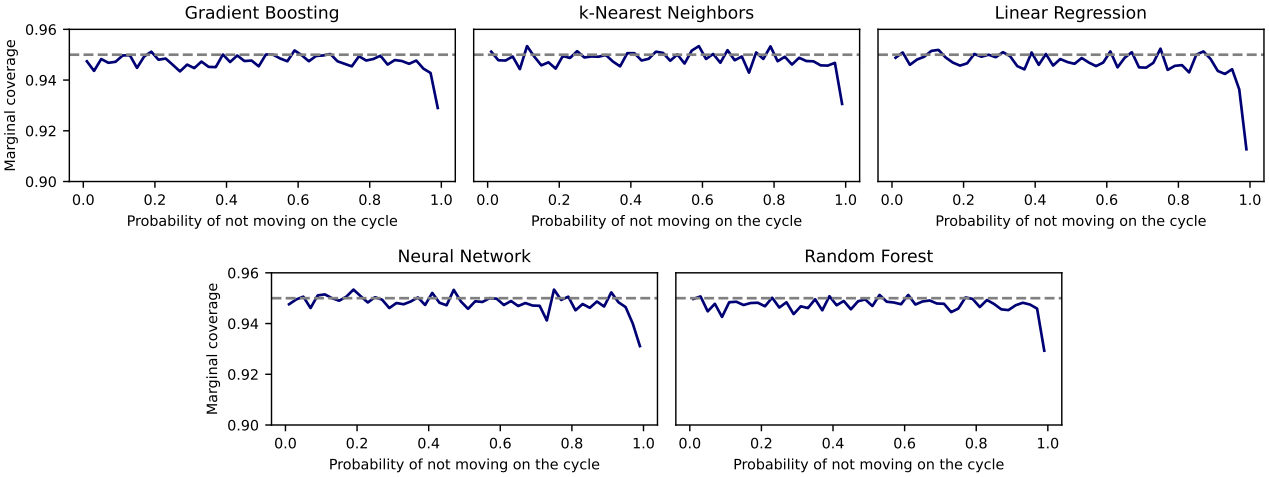

Figure C.2: Marginal coverage for the hidden random walk on the cycle graph of 5 vertices (solid) and nominally prescribed iid level of $1 - \alpha = 0.95$ (dashed). Split CP guarantees hold well even under moderate dependence. Significant undercoverage only happens at extreme levels of dependence.

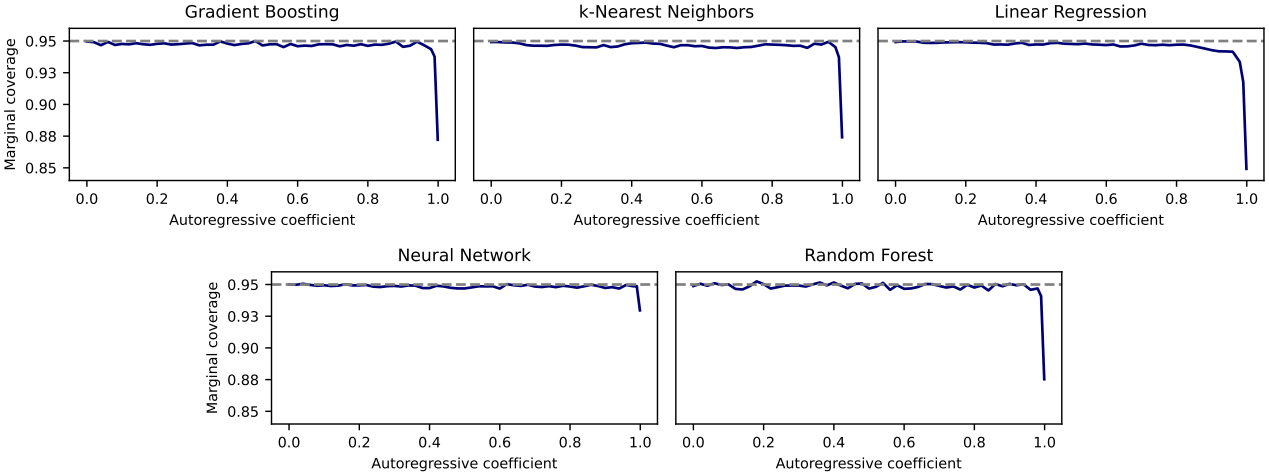

Figure C.3: Marginal coverage for autoregressive process of order 1 (solid) and nominally prescribed iid level of $1 - \alpha = 0.95$ (dashed). Split CP guarantees hold well even under moderate dependence. Significant undercoverage only happens at extreme levels of dependence.

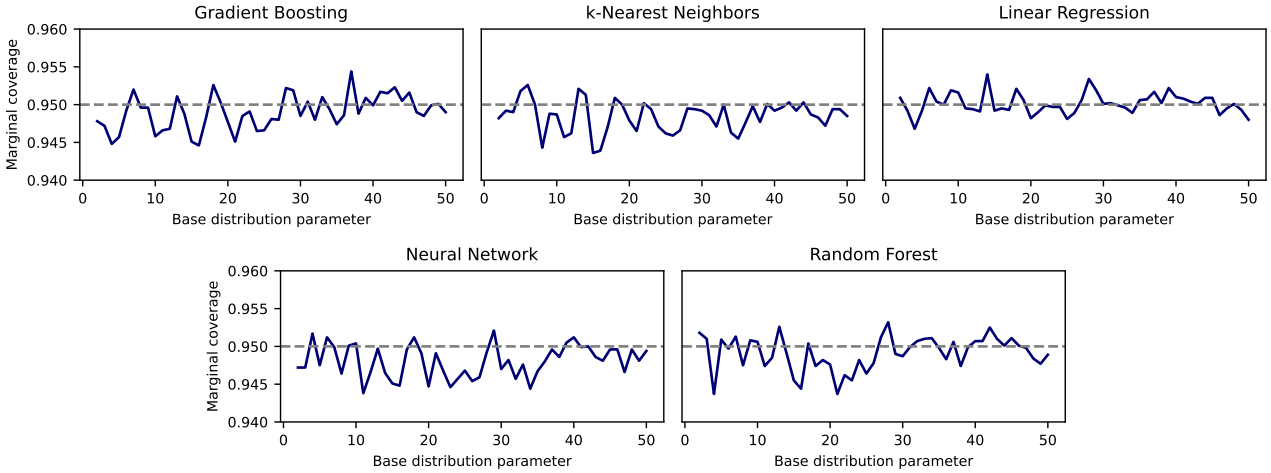

Figure C.4: Marginal coverage for hidden renewal model (solid) and nominally prescribed iid level of $1 - \alpha = 0.95$ (dashed). Split CP guarantees hold well even under moderate dependence. Significant undercoverage only happens at extreme levels of dependence.

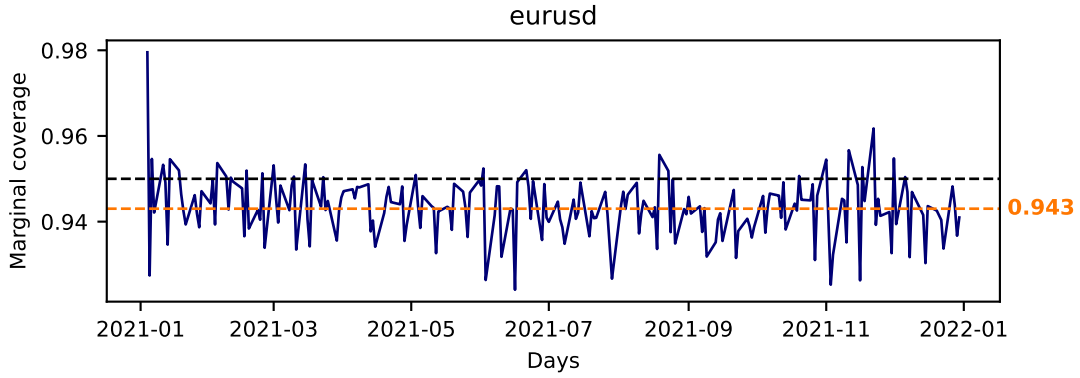

Figure C.5: Daily marginal coverage of minute-by-minute online prediction for EUR/USD spot exchange rate (solid blue), nominally prescribed iid level of $1 - \alpha = 0.95$ (dashed black) and marginal coverage over the entire year (dashed orange).

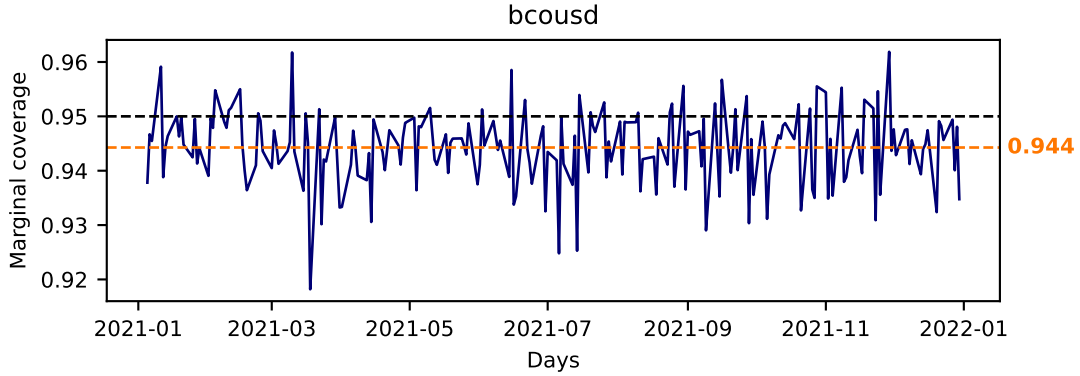

Figure C.6: Daily marginal coverage of minute-by-minute online prediction for Brent crude oil futures (solid blue), nominally prescribed iid level of $1 - \alpha = 0.95$ (dashed black) and marginal coverage over the entire year (dashed orange).

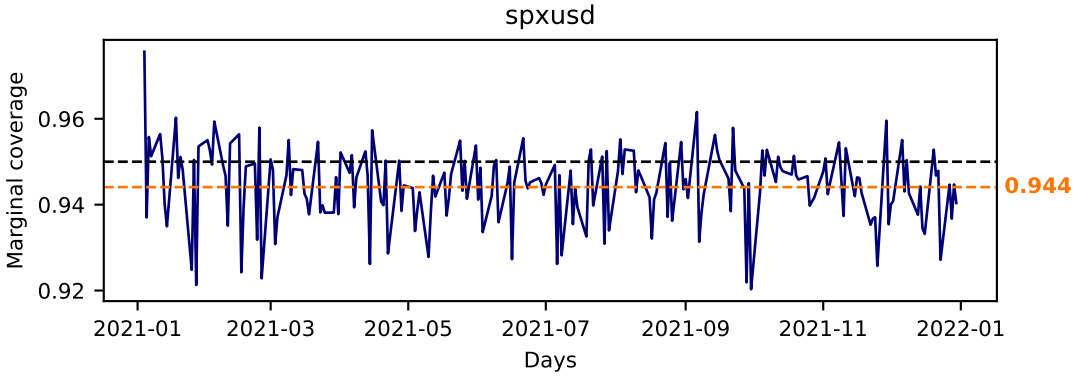

Figure C.7: Daily marginal coverage of minute-by-minute online prediction for S&P 500 futures (solid blue), nominally prescribed iid level of $1 - \alpha = 0.95$ (dashed black) and marginal coverage over the entire year (dashed orange).

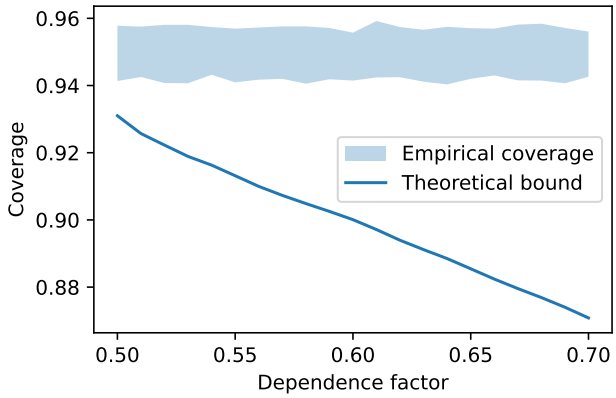

Figure C.8: Empirical coverage and theoretical guarantee (Theorem 5.1.3) for $\delta_{\text{cal}} = \delta_{\text{test}} = 0.005$ and $1 - \alpha = 0.95$. Empirical coverage is always above the theoretical bound.

Dataset	Cal. set size	Conditional coverage			
		Uptrend	Downtrend	High vol.	Low vol.
`eurusd`	500	93.50%	93.47%	92.18%	94.73%
	1000	93.80%	93.89%	92.95%	94.93%
	5000	94.84%	94.88%	94.66%	94.99%
`bcousd`	500	93.66%	93.45%	91.30%	94.38%
	1000	94.23%	94.18%	92.18%	94.84%
	5000	94.78%	94.83%	94.54%	95.05%
`spxusd`	500	93.86%	93.82%	92.86%	94.57%
	1000	94.32%	94.24%	93.38%	94.89%
	5000	94.82%	94.69%	94.39%	95.20%

Table C.1: 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 $1 - \alpha = 0.95$ and results improve given more calibration points, as expected.

Nominally prescribed iid level: 85%.

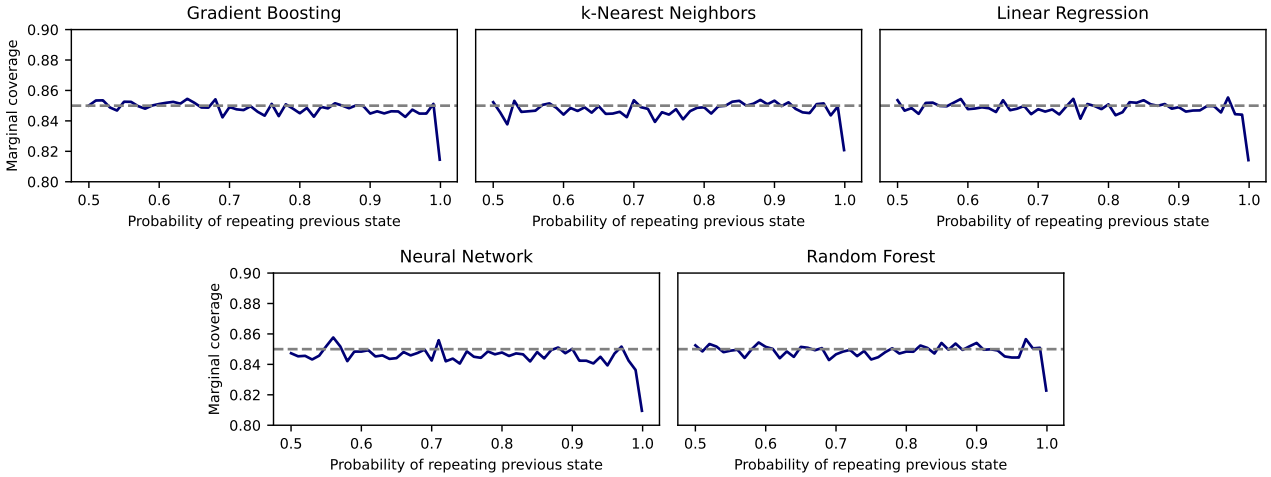

Figure C.9: Marginal coverage for hidden Markov model with two underlying states (solid) and nominally prescribed iid level of $1 - \alpha = 0.85$ (dashed). Split CP guarantees hold well even under moderate dependence. Significant undercoverage only happens at extreme levels of dependence.

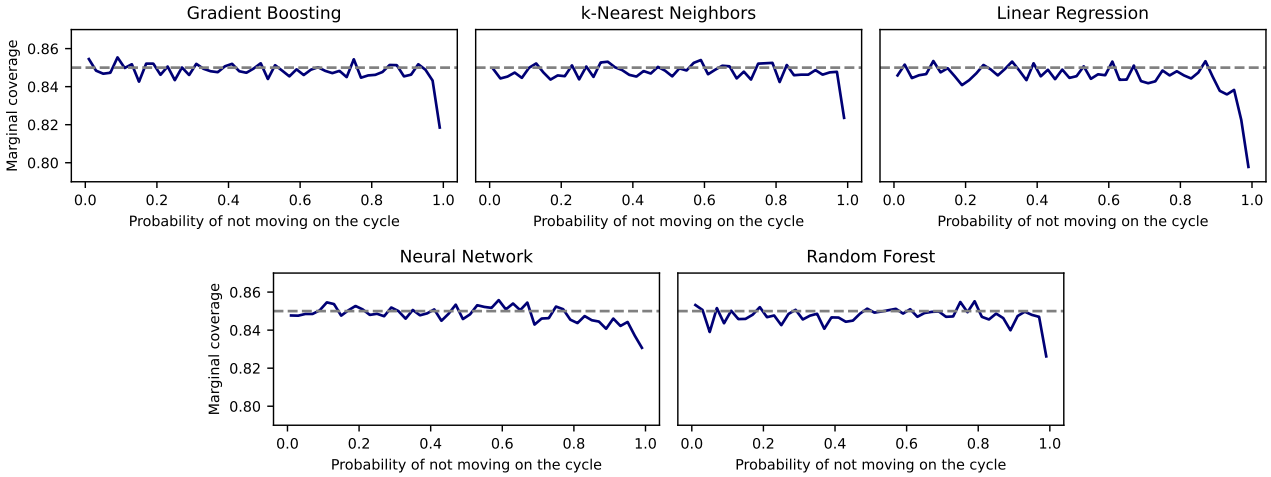

Figure C.10: Marginal coverage for the hidden random walk on the cycle graph of 5 vertices (solid) and nominally prescribed iid level of $1 - \alpha = 0.85$ (dashed). Split CP guarantees hold well even under moderate dependence. Significant undercoverage only happens at extreme levels of dependence.

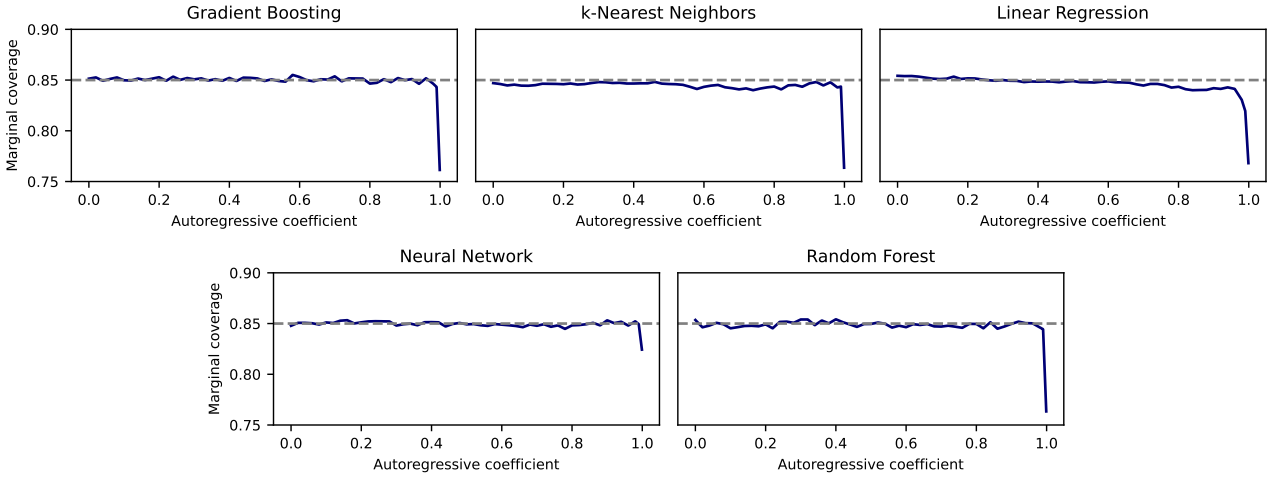

Figure C.11: Marginal coverage for autoregressive process of order 1 (solid) and nominally prescribed iid level of $1 - \alpha = 0.85$ (dashed). Split CP guarantees hold well even under moderate dependence. Significant undercoverage only happens at extreme levels of dependence.

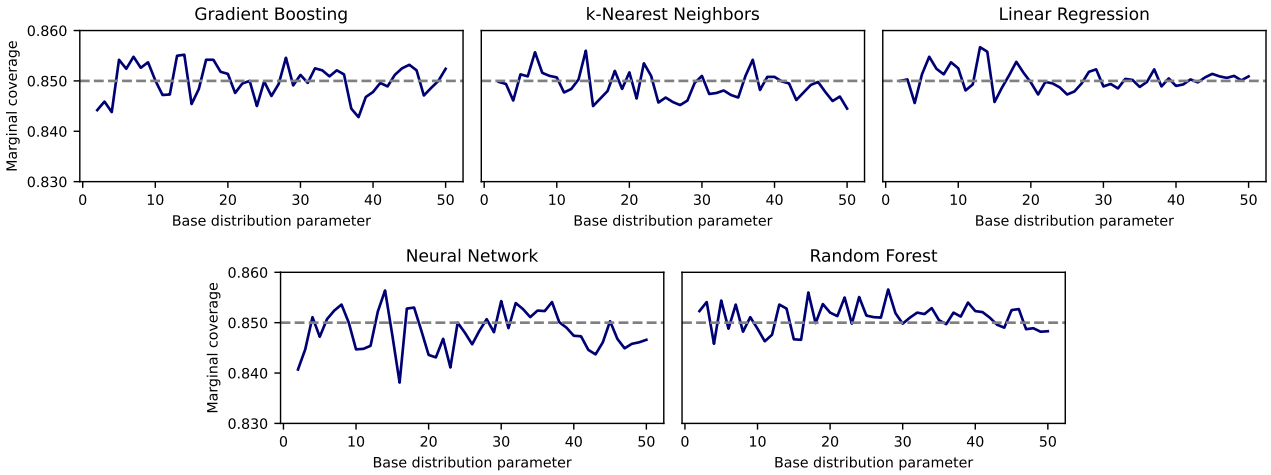

Figure C.12: Marginal coverage for hidden renewal model (solid) and nominally prescribed iid level of $1 - \alpha = 0.85$ (dashed). Split CP guarantees hold well even under moderate dependence. Significant undercoverage only happens at extreme levels of dependence.

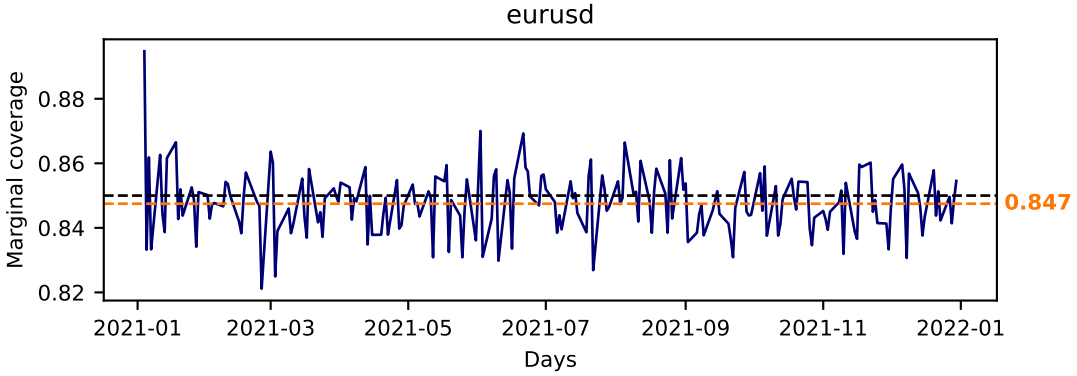

Figure C.13: Daily marginal coverage of minute-by-minute online prediction for EUR/USD spot exchange rate (solid blue), nominally prescribed iid level of $1 - \alpha = 0.85$ (dashed black) and marginal coverage over the entire year (dashed orange).

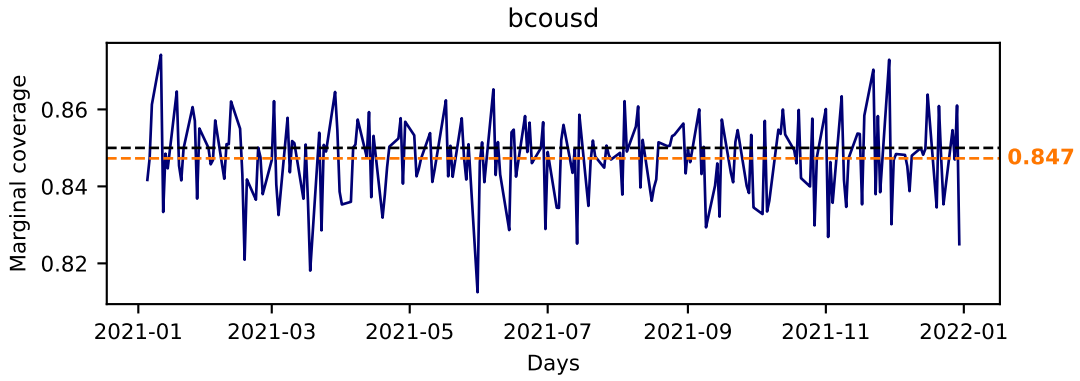

Figure C.14: Daily marginal coverage of minute-by-minute online prediction for Brent crude oil futures (solid blue), nominally prescribed iid level of $1 - \alpha = 0.85$ (dashed black) and marginal coverage over the entire year (dashed orange).

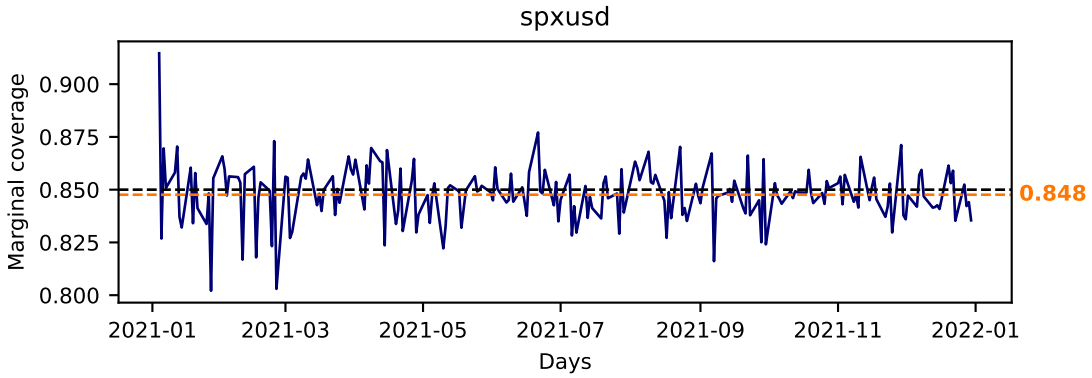

Figure C.15: Daily marginal coverage of minute-by-minute online prediction for S&P 500 futures (solid blue), nominally prescribed iid level of $1 - \alpha = 0.85$ (dashed black) and marginal coverage over the entire year (dashed orange).

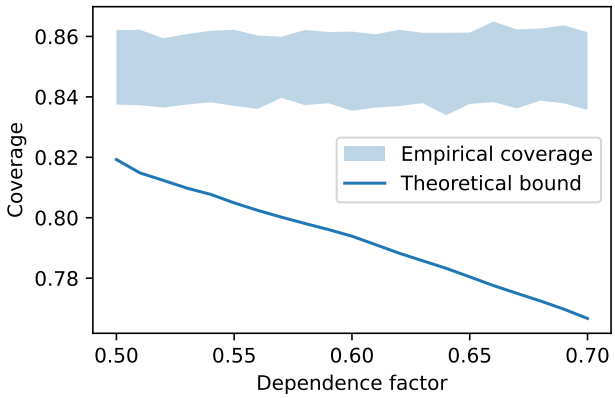

Figure C.16: Empirical coverage and theoretical guarantee (Theorem 5.1.3) for $\delta_{\text{cal}} = \delta_{\text{test}} = 0.005$ and $1 - \alpha = 0.85$. Empirical coverage is always above the theoretical bound.

Dataset	Cal. set size	Conditional coverage			
		Uptrend	Downtrend	High vol.	Low vol.
eurusd	500	84.06%	84.04%	83.09%	85.25%
	1000	84.57%	84.42%	83.66%	85.36%
	5000	85.02%	85.17%	84.74%	85.12%
bcousd	500	84.00%	84.04%	82.72%	84.60%
	1000	84.45%	84.39%	83.21%	85.03%
	5000	84.82%	84.73%	84.55%	84.98%
spxusd	500	84.28%	84.21%	84.80%	84.90%
	1000	84.65%	84.66%	84.22%	85.21%
	5000	85.12%	84.62%	84.71%	85.38%

Table C.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 $1 - \alpha = 0.85$ and results improve given more calibration points, as expected.

Bibliography

Abad, Javier, Umang Bhatt, Adrian Weller, and Giovanni Cherubin (2022). "Approximating Full Conformal Prediction at Scale via Influence Functions". In: *arXiv preprint arXiv:2202.01315*.

Aldous, David J (1985). "Exchangeability and related topics". In: *École d'Été de Probabilités de Saint-Flour XIII—1983*. Springer, pp. 1–198.

Anderson, George (1740). "Letter from Anderson in Leyden to Jones." In: *Correspondence of Scientific Men of the Seventeenth Century*. Ed. by S.P. Rigaud, S.J. Rigaud, I. Barrow, J. Flamsteed, J. Wallis, I. Newton, and A. De Morgan. University Press, pp. 360–366.

Angelopoulos, Anastasios Nikolas and Stephen Bates (2021). "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification". In: *arXiv preprint arXiv:2107.07511*.

Angelopoulos, Anastasios Nikolas, Stephen Bates, Michael Jordan, and Jitendra Malik (2021). "Uncertainty Sets for Image Classifiers using Conformal Prediction". In: *International Conference on Learning Representations*.

Avellaneda, Marco and Jeong-Hyun Lee (2010). "Statistical arbitrage in the US equities market". In: *Quantitative Finance* 10.7, pp. 761–782.

Bachelier, Louis (1900). "Théorie de la spéculation". In: vol. 17. Societe Mathematique de France, pp. 21–86. DOI: 10.24033/asens.476.

Baker, George A. and Peter Graves-Morris (1996). *Padé Approximants*. 2nd ed. Encyclopedia of Mathematics and its Applications. Cambridge University Press. DOI: 10.1017/CBO9780511530074.

Barber, Rina Foygel, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani (2020). "The limits of distribution-free conditional predictive inference". In: *Information and Inference: A Journal of the IMA* 10.2, pp. 455–482. DOI: 10.1093/imaiai/iaaa017.

Barber, Rina Foygel, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani (2021). "Predictive inference with the jackknife+". In: *The Annals of Statistics* 49.1, pp. 486–507.

Barber, Rina Foygel, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani (2022). "Conformal prediction beyond exchangeability". In: *arXiv preprint arXiv:2202.13415*.

Bassett Jr, Gilbert and Roger Koenker (1982). "An empirical quantile function for linear models with iid errors". In: *Journal of the American Statistical Association* 77.378, pp. 407–415.

Bastani, Osbert, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth (2022). "Practical Adversarial Multivalid Conformal Prediction". In: *arXiv preprint arXiv:2206.01067*.

Berbee, Henry (1987). "Convergence rates in the strong law for bounded mixing sequences". In: *Probability Theory and Related Fields* 74 (2), pp. 255–270. DOI: 10.1007/bf00569992.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford. ISBN: 9780199535255.

Breiman, Leo (2001). "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". In: *Statistical Science* 16.3, pp. 199–231. DOI: 10.1214/ss/1009213726.

Burnaev, Evgeny and Vladimir Vovk (2014). "Efficiency of conformalized ridge regression". In: *Proceedings of The 27th Conference on Learning Theory*. Ed. by Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári. Vol. 35. Proceedings of Machine Learning Research. Barcelona, Spain: PMLR, pp. 605–622.

Carrasco, Marine and Xiaohong Chen (2002). "Mixing and moment properties of various GARCH and stochastic volatility models". In: *Econometric Theory* 18.1, pp. 17–39.

Chaboud, Alain, Dagfinn Rime, and Vladyslav Sushko (2022). "The Foreign Exchange Market". In: *The Research Handbook of Financial Markets*. Ed. by Refet Gürkaynak and Jonathan Wright. Available at SSRN 4063213. Forthcoming.

Chen, Wenyu, Kelli-Jean Chun, and Rina Foygel Barber (2018). "Discretized conformal prediction for efficient distribution-free inference". In: *Stat* 7.1, e173.

Chernozhukov, Victor, Iván Fernández-Val, and Alfred Galichon (2010). "Quantile and probability curves without crossing". In: *Econometrica* 78.3, pp. 1093–1125.

Chernozhukov, Victor, Kaspar Wüthrich, and Yinchu Zhu (2018). "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*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, pp. 732–749.

Chernozhukov, Victor, Kaspar Wüthrich, and Yinchu Zhu (2021). "Distributional conformal prediction". In: *Proceedings of the National Academy of Sciences* 118.48. DOI: 10.1073/pnas.2107794118.

Cont, Rama (2010). "Stylized properties of asset returns". In: *Encyclopedia of Quantitative Finance*. Wiley.

Cousin, Areski, Hassan Maatouk, and Didier Rullière (2016). "Kriging of financial term-structures". In: *European Journal of Operational Research* 255.2, pp. 631–648.

Cox, D. R. (2001). "[Statistical Modeling: The Two Cultures]: Comment". In: *Statistical Science* 16 (3), pp. 216–218. DOI: 10.2307/2676682.

Cramér, H. (1938). "Sur un nouveau théorème limite dans la théorie des probabilités". In: *Colloque consacré à la théorie des probabilités*. Vol. 736. Paris: Wiley, pp. 2–23.

Davydov, Yu A (1974). "Mixing conditions for Markov chains". In: *Theory of Probability and Its Applications* 18.2, pp. 312–328.

de Finetti, Bruno (1931). "Funzione caratteristica di un fenomeno aleatorio". In: *Atti della R. Accademia Nazionale dei Lincei, Ser. 6. Memorie, Classe di Scienze Fisiche, Matematiche e Naturali 4*, pp. 251–299.

de Finetti, Bruno (1937). "La prévision: ses lois logiques, ses sources subjectives". In: *Annales de l'institut Henri Poincaré* 7.1, pp. 1–68.

Dewolf, Nicolas, Bernard De Baets, and Willem Waegeman (2022). "Valid prediction intervals for regression problems". In: *Artificial Intelligence Review*. DOI: 10.1007/s10462-022-10178-5.

Dixon, Matthew F, Igor Halperin, and Paul Bilokon (2020). *Machine learning in Finance*. Springer.

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

Dua, Dheeru and Casey Graff (2017). *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml.

Duffie, Darrell (2012). *Dark Markets: Asset Pricing and Information Transmission in Over-the-Counter Markets*. Illustrated. Princeton Lectures in Finance, 6. Princeton University Press. ISBN: 9780691138961.

Efron, Brad (2001). "[Statistical Modeling: The Two Cultures]: Comment". In: *Statistical Science* 16 (3), pp. 218–219. DOI: 10.2307/2676683.

Fedorova, Valentina, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk (2012). "Plug-in Martingales for Testing Exchangeability on-Line". In: *Proceedings of the 29th International Coference on International Conference on Machine Learning*. ICML'12. Edinburgh, Scotland: Omnipress, pp. 923–930.

Feldman, Shai, Stephen Bates, and Yaniv Romano (2022). "Conformalized Online Learning: Online Calibration Without a Holdout Set". In: *arXiv preprint arXiv:2205.09095*.

Frobenius, Georg (1881). "Ueber Eelationen zwischen den Näherungsbrüchen von Potenzreihen." In: *Journal für die reine und angewandte Mathematik* (90), pp. 1–17.

Gammerman, A., V. Vovk, and V. Vapnik (1998). "Learning by Transduction". In: *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. UAI'98. Madison, Wisconsin: Morgan Kaufmann Publishers Inc., pp. 148–155.

Gibbs, Isaac and Emmanuel Candès (2021). "Adaptive Conformal Inference Under Distribution Shift". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan.

Guijarro-Ordonez, Jorge, Markus Pelger, and Greg Zanotti (2021). "Deep Learning Statistical Arbitrage". In: *Available at SSRN 3862004*.

Hastie, Trevor, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani (2022). "Surprises in high-dimensional ridgeless least squares interpolation". In: *The Annals of Statistics* 50.2, pp. 949–986.

Hechtlinger, Yotam, Barnabás Póczos, and Larry Wasserman (2019). "Cautious Deep Learning". In: *arXiv preprint arXiv:1805.09460*.

Heinrich, Lothar (1992). "Bounds for the absolute regularity coefficient of a stationary renewal process". In: *Yokohama Mathematical Journal* 40.1, pp. 25–33. URL: http://hdl.handle.net/10131/5619.

Hewitt, Edwin and Leonard J Savage (1955). "Symmetric measures on Cartesian products". In: *Transactions of the American Mathematical Society* 80.2, pp. 470–501.

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

Hoadley, Bruce (2001). "[Statistical Modeling: The Two Cultures]: Comment". In: *Statistical Science* 16 (3), pp. 220–224. DOI: 10.2307/2676684.

Huangfu, Qi and JA Julian Hall (2018). "Parallelizing the dual revised simplex method". In: *Mathematical Programming Computation* 10.1, pp. 119–142.

Hull, J. J. (1994). "A database for handwritten text recognition research". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16.5, pp. 550–554. DOI: 10.1109/34.291440.

Jacobi, Carl Gustav Jakob (1846). "Über die Darstellung einer Reihe gegebner Werthe durch eine gebrochne rationale Function." In: *Journal für die reine und angewandte Mathematik* (30), pp. 127–156.

Kath, Christopher and Florian Ziel (2021). "Conformal prediction interval estimation and applications to day-ahead and intraday power markets". In: *International Journal of Forecasting* 37.2, pp. 777–799.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu (2017). "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc.

Khaleghi, Azadeh and Gábor Lugosi (2021). "Inferring the mixing properties of an ergodic process". In: *arXiv preprint arXiv:2106.07054*.

Koenker, Roger and Gilbert Bassett (1978). "Regression Quantiles". In: *Econometrica* 46.1, pp. 33–50.

Kuchibhotla, Arun Kumar (2020). "Exchangeability, conformal prediction, and rank tests". In: *arXiv preprint arXiv:2005.06095*.

Kuznetsov, Vitaly and Mehryar Mohri (2017). "Generalization bounds for non-stationary mixing processes". In: *Machine Learning* 106.1, pp. 93–117. DOI: 10.1007/s10994-016-5588-2.

Larson, S. C. (1931). "The shrinkage of the coefficient of multiple correlation." In: *Journal of Educational Psychology* 22.1, p. 45.

Lei, J (2019). "Fast exact conformalization of the lasso using piecewise linear homotopy". In: *Biometrika* 106.4, pp. 749–764. DOI: 10.1093/biomet/asz046.

Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman (2018). "Distribution-Free Predictive Inference for Regression". In: *Journal of the American Statistical Association* 113.523, pp. 1094–1111. DOI: 10.1080/01621459.2017.1307116.

Lei, Jing, Alessandro Rinaldo, and Larry Wasserman (2015). "A conformal prediction approach to explore functional data". In: *Annals of Mathematics and Artificial Intelligence* 74 (1-2), pp. 29–43. DOI: 10.1007/s10472-013-9366-6.

Levin, David A., Yuval Peres, and Elizabeth L. Wilmer (2017). *Markov Chains and Mixing Times*. 2nd ed. Vol. 107. American Mathematical Society.

Liu, Bingqing, Ivan Kiskin, and Stephen Roberts (2020). "An overview of Gaussian process regression for volatility forecasting". In: *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, pp. 681–686.

Loshchilov, Ilya and Frank Hutter (2019). "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations*.

Mandelbrot, Benoit (1963). "The Variation of Certain Speculative Prices". In: *The Journal of Business* 36 (4), pp. 394–419. DOI: 10.2307/2350970.

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

Meinshausen, Nicolai (2006). "Quantile Regression Forests". In: *Journal of Machine Learning Research* 7.35, pp. 983–999.

Mohri, Mehryar and Afshin Rostamizadeh (2010). "Stability Bounds for Stationary φ-mixing and β-mixing Processes." In: *Journal of Machine Learning Research* 11.2.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar (2018). *Foundations of Machine Learning*. 2nd ed. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press. 504 pp. ISBN: 978-0-262-03940-6.

Mokkadem, Abdelkader (1988). "Mixing properties of ARMA processes". In: *Stochastic processes and their applications* 29.2, pp. 309–315.

Moran, P. A. P. (1973). "Dividing a Sample into Two Parts a Statistical Dilemma". In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 35.3, pp. 329–333. ISSN: 0581572X. (Visited on 04/17/2022).

Ndiaye, Eugene and Ichiro Takeuchi (2019). "Computing Full Conformal Prediction Set with Approximate Homotopy". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach,

H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc.

Oliveira, Roberto I., Paulo Orenstein, Thiago Ramos, and João Vitor Romano (2022). "Split Conformal Prediction for Dependent Data". In: *arXiv preprint arXiv:2203.15885*. URL: https://arxiv.org/abs/2203.15885.

Padé, Henri (1892). "Sur la représentation approchée d'une fonction par des fractions rationnelles". In: *Annales scientifiques de l'école Normale Supérieure*. Vol. 9, pp. 3–93.

Papadopoulos, Harris, Kostas Proedrou, Volodya Vovk, and Alexander Gammerman (2002). "Inductive Confidence Machines for Regression". In: *Proceedings of the 13th European Conference on Machine Learning*. ECML '02. Berlin, Heidelberg: Springer-Verlag, pp. 345–356.

Parzen, Emanuel (1962). *Stochastic Processes*. Holden-Day series in probability and statistics. Holden-Day. ISBN: 9780816266647.

Parzen, Emanuel (2001). "[Statistical Modeling: The Two Cultures]: Comment". In: *Statistical Science* 16 (3), pp. 224–226. DOI: 10.2307/2676685.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., pp. 8024–8035.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Roebroek, Jasper (2022). *sklearn-quantile*. URL: https://github.com/jasperroebroek/sklearn-quantile. Version 0.0.18.

Romano, Joseph P. and Andrew F. Siegel (1986). *Counterexamples in probability and statistics*. Wadsworth & Brooks/Cole statistics/probability series. Wadsworth & Brooks/Cole Advanced Books & Software. ISBN: 9780534055684.

Romano, Yaniv, Evan Patterson, and Emmanuel Candès (2019). "Conformalized Quantile Regression". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc.

Saunders, C., A. Gammerman, and V. Vovk (1999). "Transduction with Confidence and Credibility". In: *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'99. Stockholm, Sweden: Morgan Kaufmann Publishers Inc., pp. 722–726.

Srinivasan, Ashwin (1993). *Statlog (Landsat Satellite) Data Set*. URL: https://archive.ics.uci.edu/ml/datasets/Statlog+(Landsat+Satellite).

Sun, Qingyun and Stephen Boyd (2018). "Distributional Robust Kelly Gambling: Optimal Strategy under Uncertainty in the Long-Run". In: *arXiv preprint arXiv:1812.10371*.

Tegner, Martin and Stephen Roberts (2021). "A Bayesian take on option pricing with Gaussian processes". In: *arXiv preprint arXiv:2112.03718*.

Tibshirani, Rob and Trevor Hastie (2021). "A Melting Pot". In: *Observational Studies* 7.1, pp. 213–215.

Tibshirani, Ryan J, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas (2019). "Conformal Prediction Under Covariate Shift". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc.

Vidyasagar, Mathukumalli and Rajeeva L. Karandikar (2016). "A Learning Theory Approach to System Identification and Stochastic Adaptive Control". In: *Probabilistic and Randomized Methods for Design under Uncertainty*. Springer-Verlag, pp. 265–302. DOI: 10.1007/1-84628-095-8_10.

Vovk, Vladimir (2015). "Cross-conformal predictors". In: *Ann. Math. Artif. Intell.* 74.1-2, pp. 9–28. DOI: 10.1007/s10472-013-9368-4.

Vovk, Vladimir (2021). "Testing Randomness Online". In: *Statistical Science* 36.4, pp. 595–611. DOI: 10.1214/20-STS817.

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

Vovk, Vladimir, Ilia Nouretdinov, and Alexander Gammerman (2003). "Testing exchangeability on-line". In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 768–775.

Vovk, Vladimir, Ilia Nouretdinov, Valery Manokhin, and Alexander Gammerman (Nov. 2018). "Cross-conformal predictive distributions". In: *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*. Ed. by Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov, and Ralf Peeters. Vol. 91. Proceedings of Machine Learning Research. PMLR, pp. 37–51.

Vovk, Volodya, Alexander Gammerman, and Craig Saunders (1999). "Machine-Learning Applications of Algorithmic Randomness". In: *Proceedings of the Sixteenth International Conference on Machine Learning*. ICML '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 444–453. ISBN: 1558606122.

Wisniewski, Wojciech, David Lindsay, and Sian Lindsay (2020). "Application of conformal prediction interval estimations to market makers' net positions". In: *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*. Ed. by Alexander Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov, and Giovanni Cherubin. Vol. 128. Proceedings of Machine Learning Research. PMLR, pp. 285–301.

Xu, Chen and Yao Xie (2021). "Conformal prediction interval for dynamic time-series". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 11559–11569.

Yeo, Joongyeub and George Papanicolaou (2017). "Risk control of mean-reversion time in statistical arbitrage". In: *Risk and Decision Analysis* 6.4, pp. 263–290.

Yu, Bin (1994). "Rates of Convergence for Empirical Processes of Stationary Mixing Sequences". In: *The Annals of Probability* 22.1, pp. 94–116. DOI: 10.1214/aop/1176988849.