Adaptive Conformal Predictions for Time Series

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, Aymeric Dieuleveut

Adaptive Conformal Predictions for Time Series

Margaux Zaffran [1 2 3] **Olivier Féron** [1 4] **Yannig Goude** [1 5] **Julie Josse** [2 6] **Aymeric Dieuleveut** [3]

Abstract

Uncertainty quantification of predictive models is crucial in decision-making problems. Conformal prediction is a general and theoretically sound answer. However, it requires exchangeable data, excluding time series. While recent works tackled this issue, we argue that Adaptive Conformal Inference (ACI, Gibbs & Candès, 2021), developed for distribution-shift time series, is a good procedure for time series with general dependency. We theoretically analyse the impact of the learning rate on its efficiency in the exchangeable and auto-regressive case. We propose a parameter-free method, AgACI, that adaptively builds upon ACI based on online expert aggregation. We lead extensive fair simulations against competing methods that advocate for ACI's use in time series. We conduct a real case study: electricity price forecasting. The proposed aggregation algorithm provides efficient prediction intervals for day-ahead forecasting. All the code and data to reproduce the experiments are made available on GitHub.

1. Introduction

The increasing use of renewable intermittent energy leads to more dependent and volatile energy markets. Therefore, an accurate electricity price forecasting is required to stabilize energy production planning, gathering loads of research work as evidenced by recent substantial reviews (Weron, 2014; Lago et al., 2018; 2021). Furthermore, probabilistic forecasts are needed to develop risk-based strategies (Gaillard et al., 2016; Maciejowska et al., 2016; Nowotarski & Weron, 2018; Uniejewski & Weron, 2021). On the one hand, the lack of uncertainty quantification of predictive models is a major barrier to the adoption of powerful machine learning methods. On the other hand, probabilistic forecasts are only valid asymptotically or upon strong assumptions on the data.

Conformal prediction (CP, Vovk et al., 1999; 2005; Papadopoulos et al., 2002) is a promising framework to overcome both issues. It is a general procedure to build predictive intervals for any (black box) predictive model, such as neural networks, which are *valid* (i.e. achieve nominal marginal coverage) in finite sample and without any distributional assumptions except that the data are exchangeable.

Thereby, CP has received increasing attention lately, favored by the development of *split conformal prediction* (SCP, Lei et al., 2018, reformulated from *inductive* CP, Papadopoulos et al., 2002). More formally, suppose we have n training samples $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i \in [\![1, n]\!]$, realizations of random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$, and that we aim at predicting a new observation y_{n+1} at x_{n+1}. Given a *miscoverage rate* $\alpha \in [0, 1]$ fixed by the user (typically 0.1 or 0.05) the aim is to build a predictive interval $\mathcal{C}_\alpha$ such that:

$$\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}_\alpha\left(X_{n+1}\right)\right\} \geq 1 - \alpha, \tag{1}$$

with $\mathcal{C}_\alpha$ as small as possible, in order to be informative. For the sequel, we call a *valid interval* an interval satisfying equation (1) and an *efficient interval* when it is as small as possible (Vovk et al., 2005; Shafer & Vovk, 2008).

To achieve this, SCP first splits the n points of the training set into two sets $\mathrm{Tr}, \mathrm{Cal} \subset [\![1, \mathrm{n}]\!]$, to create a *proper training set*, Tr, and a *calibration set*, Cal. On the proper training set a regression model $\hat{\mu}$ (chosen by the user) is fitted, and then used to predict on the calibration set. A *conformity score* is applied to assess the conformity between the calibration's response values and the predicted values, giving $S_{\mathrm{Cal}} = \{(s_i)_{i \in \mathrm{Cal}}\}$. In regression, usually the absolute value of the residuals is used, i.e. $s_i = |\hat{\mu}(x_i) - y_i|$. Finally, a corrected[1] $(1 - \hat{\alpha})$-th quantile of these scores $\widehat{Q}_{1-\hat{\alpha}}(S_{\mathrm{Cal}})$ is computed to define the size of the interval, which, in its simplest form, is centered on the predicted value: $\mathcal{C}_\alpha\left(x_{n+1}\right) = \widehat{C}_{\hat{\alpha}}(x_{n+1}) := [\hat{\mu}(x_{n+1}) \pm \widehat{Q}_{1-\hat{\alpha}}(S_{\mathrm{Cal}})]$. These steps are detailed in Appendix A, and illustrated in Figure 9. More details on CP, including beyond regres-

[1]Electricité de France R&D, Palaiseau, France [2]INRIA Sophia-Antipolis, Montpellier, France [3]CMAP, École Polytechnique, Institut Polytechnique de Paris, Palaiseau, France [4]FiME, Université Paris-Dauphine, France [5]LMO, Université Paris-Saclay, Orsay, France [6]IDESP, Montpellier, France. Correspondence to: Margaux Zaffran <margaux.zaffran@inria.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]The correction $\alpha \to \hat{\alpha}$ is needed because of the inflation of quantiles in finite sample (see Lemma 2 in Romano et al. (2019) or Section 2 in Lei et al. (2018)).

sion, are given in Vovk et al. (2005); Angelopoulos & Bates (2021).

The cornerstone of SCP *validity* results is the exchangeability assumption of the data (see Lei et al., 2018, and Appendix A.3). However, this assumption is not met in time series forecasting problems. Despite the lack of theoretical guarantees, several works have applied CP to time series. Dashevskiy & Luo (2008; 2011) apply original (*inductive*) CP (Papadopoulos et al., 2002) to both simulated (using Auto-Regressive Moving Average (ARMA) processes) and real network traffic data and obtain *valid* intervals. Wisniewski et al. (2020); Kath & Ziel (2021) apply SCP respectively to financial data (e.g. markets makers' net positions) and to electricity price forecasting on various markets. In order to account for the temporal aspect, they consider an online version of SCP. In both studies, the *validity* varied greatly depending on the markets and the underlying regression model, suggesting that further developments of CP and theoretical guarantees for time series are needed.

To this end, Chernozhukov et al. (2018) extend the CP theory to ergodic cases in order to include dependent data. Xu & Xie (2021a) improve on that theory and propose a new algorithm, Ensemble Prediction Interval (EnbPI), adapted to time series by adding a sequential aspect.

Another case that breaks the exchangeability assumption is *distribution shift*, which allows for example to deal with cases where the test data is shifted with respect to the training data. Tibshirani et al. (2019) consider covariate shift while Cauchois et al. (2020) tackle a joint distributional shift setting (that is, of (X, Y)). In both studies, a single shift in the distribution is considered, a major limitation for applying these methods to time series. In an adversarial setting, Gibbs & Candès (2021) propose Adaptive Conformal Inference (ACI), accounting for an undefined number of shifts on the joint distribution. It is based on refitting the predictive model, as well as updating online the quantile level used by a recursive scheme depending on an hyper-parameter γ (a learning rate). Furthermore, they prove an asymptotic *validity* result for any data distribution.

We argue in this work that the design and guarantees of ACI can be beneficial for dependent data without distribution shifts.

Contributions. We propose to analyse ACI (Gibbs & Candès, 2021) in the context of time series with general dependency and make the following contributions:

- Relying on an asymptotic analysis of ACI's behaviour for simple time series distribution, we prove that ACI deteriorates *efficiency* in an exchangeable case (closed-form expression) while improving it in an AR setting (numerical approximation) with a well-chosen γ (Section 3).
- We introduce AgACI, a parameter-free method using online expert aggregation, to avoid choosing γ, achieving good performances in terms of *validity* and *efficiency* (Section 4).
- We compare ACI to EnbPI and online SCP on extensive synthetic experiments and we propose an easy-to-interpret visualisation combining *validity* and *efficiency* (Section 5).
- We forecast and give predictive intervals on French electricity prices, an area where accurate predictions, but also controlled predictive intervals, are required (Section 6).

To allow for better benchmarking of existing and new methods, we provide (re-)implementations in Python of (all) the described methods and a complete pipeline of analysis on GitHub. As explained in Section 4, the code for AgACI is, for now, the only one available only in R.

Notations. In the sequel, the following notations are used: $[\![a, b]\!] := \{a, a+1, \dots, b\}$; $\mathbb{Q}$ refers to the set of rational numbers; $\mathcal{C}^4([0,1])$ refers to the set of 4-times continuously differentiable functions on $[0,1]$; $\overset{\text{not.}}{=}$ defines a notation; $\#A$ is the cardinal of the set A.

2. Setting: ACI for time series

In this section, we introduce ACI and our framework. We consider T_0 observations $(x_1, y_1), \dots, (x_{T_0}, y_{T_0})$ in $\mathbb{R}^d \times \mathbb{R}$. The aim is to predict the response values and give predictive intervals for T_1 subsequent observations $x_{T_0+1}, \dots, x_{T_0+T_1}$ sequentially: at any prediction step $t \in [\![T_0+1, T_0+T_1]\!]$, $y_{t-T_0}, \dots, y_{t-1}$ have been revealed. Thereby, the data $((x_{t-T_0}, y_{t-T_0}), \dots, (x_{t-1}, y_{t-1}))$ are used for the construction of the predicted interval.

Adaptive Conformal Inference. Proposed by Gibbs & Candès (2021), ACI is designed to adapt CP to temporal distribution shifts. The idea of ACI is twofold. First, one considers an online procedure with a random split[2], i.e., Tr_t and Cal_t are random subsets of the last T_0 points. Second, to improve adaptation when the data is highly shifted, an *effective miscoverage level* α_t, updated recursively, is used instead of the target level α. Set $\alpha_1 = \alpha$, and for $t \geq 1$

$$\begin{cases} \widehat{C}_{\alpha_t}(x_t) & = [\hat{\mu}(x_t) \pm \widehat{Q}_{1-\alpha_t}(S_{\mathrm{Cal}_t})] \\ \alpha_{t+1} & = \alpha_t + \gamma\left(\alpha - \mathbb{1}\{y_t \notin \widehat{C}_{\alpha_t}(x_t)\}\right), \end{cases} \tag{2}$$

for $\gamma \geq 0$[3]. If ACI does not cover at time t, then $\alpha_{t+1} \leq \alpha_t$, and the size of the predictive interval increases; conversely when it covers. Nothing prevents $\alpha_t \leq 0$ or $\alpha_t \geq 1$. While the later is rare (as α is small) and produces by convention $\widehat{C}_{\alpha_t}(\cdot) = \{\hat{\mu}(\cdot)\}$ (i.e. $\widehat{Q}_{1-\alpha_t} = 0$) , the former can happen frequently for some γ, giving $\widehat{C}_{\alpha_t} \equiv \mathbb{R}$ ($\widehat{Q}_{1-\alpha_t} = +\infty$).

How to deal with infinite intervals. A specificity of ACI's algorithm is thus to often produce infinite intervals. Defining

[2] Figure 5(a) with training and calibration part shuffled randomly.

[3] ACI actually wraps around *any* CP procedure, here the definition is given using mean regression SCP.

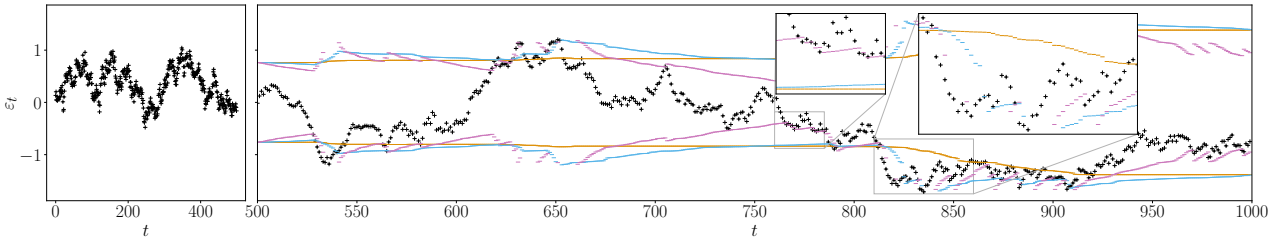


Figure 1. ACI on one simulated path ε_t, $t = 1, \dots, 1000$, from an AR(1) process (in black). The first 500 values form the initial calibration set (left subplot), and predicted interval bounds are computed on the last 500 points (right) for $\gamma = 0$, $\gamma = 0.01$ and $\gamma = 0.05$.

the *average* length of an interval is then impossible. In order to assess the *efficiency* in the following, we consider two solutions: (i) imputing the length of infinite intervals by (twice) the overall maximum of the residuals, or $Q(1)$ if the residual's quantile function is known and bounded[4]; (ii) focusing on the median instead.

ACI on time series with general dependency. As highlighted by Wisniewski et al. (2020); Kath & Ziel (2021), the first step to adapt a method for dependent time series is to work online which is the case for ACI. Moreover, the update of the quantile level according to the previous error implies that ACI could cope with a fitted model that has not correctly caught the temporal evolution, such as a trend, a seasonality pattern or a dependence on the past. Therefore, ACI is a perfect candidate for CP for time series with general dependency. To account for the temporal structure, we change the random split to a sequential split.[5]

To gain understanding on ACI in the context of dependent temporal data, we analyse a situation where a fitted regression model $\hat{\mu}$ produces AR(1) residuals, thus $y_t - \hat{\mu}(x_t) = \varepsilon_t$, where ε_t is an AR(1) process: $\varepsilon_{t+1} = 0.99\varepsilon_t + \xi_{t+1}$, with $\xi_t \sim \mathcal{N}(0, 0.01)$. We plot this toy example in Figure 1, for $T_0 = T_1 = 500$. Three versions of ACI are compared: $\gamma = 0$, the quantile level is not updated but the calibration set Cal_t is; $\gamma = 0.01$ and $\gamma = 0.05$. To obtain an insightful visualisation[6], we represent the interval $[\pm\widehat{Q}_{1-\alpha_t}(S_{\mathrm{Cal}_t})]$ instead of $\widehat{C}_{\alpha_t}(x_t)$. When no intervals are displayed, ACI is predicting $\mathbb{R}$. Here and in the sequel, we use $\alpha = 0.1$.

In this toy example, the coverage rate among many observations is *valid* for $\gamma \in \{0.01, 0.05\}$ (90% and 92% of points included) but not for $\gamma = 0$ (72.6%). Moreover, Figure 1 shows that the type of errors depends on γ. For $\gamma = 0$, ACI excludes consecutive observations (e.g. for $t \in [810, 860]$, zoomed-in plot). For $\gamma \in \{0.01, 0.05\}$, ACI manages to adapt to these observations, and the higher the γ, the less the adaptation is delayed. Furthermore, when the residuals are small and far from both interval bounds, ACI quickly reduces the interval's length and produces more *efficient* intervals. Consequently, ACI may also not cover on points for which the residuals have a relatively small values compared to the calibration's values (e.g. for $t \in [760, 785]$).

[4]This happens in practice when the response and prediction are bounded, e.g., thanks to physical/real constraints as for the spot prices presented in Section 6.1, that are bounded by market rules.

[5]As in Figure 5(a). This is also consistent with OSSCP (Sec. 5.3).

[6]We suggest focusing the visualisation on the scores to analyse the behaviour of CP methods, as they are at the core of the *validity* proof. A detailed discussion on this is given in App. A.5

3. Impact of γ on ACI efficiency

The choice of the parameter γ strongly impacts the behaviour of ACI: while the method always satisfies the *asymptotic validity* property, i.e. $\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{y_t \notin \widehat{C}_{\alpha_t}(x_t)\} \xrightarrow[T\to\infty]{a.s.} \alpha$ (Proposition 4.1 in Gibbs & Candès, 2021), this property does not give any insight on the length of resulting intervals. Besides, this guarantee directly stems from the fact that $\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{y_t \notin \widehat{C}_{\alpha_t}(x_t)\} - \alpha \leq 2/(\gamma T)$. This tends to suggest the use of larger γ values, that unfortunately generate frequent infinite intervals. Here, we thus analyse the impact of γ on ACI's *efficiency* in simple yet insightful cases: in Section 3.1, focusing on the exchangeable case, then in Section 3.2, with a simple AR process on the residuals.

Approach. Our focus is on the impact of the key parameter γ. Analysing simple theoretical distributions allows to build intuition on the behaviour of the algorithm for more complex data structure. In order to derive theoretical results, we thus make supplementary modelling assumptions on the residuals, and do not consider the impact of the calibration set: we introduce Q the quantile function of the scores and assume, for all $\hat{\alpha}$ and t, $\widehat{Q}_{1-\hat{\alpha}}(S_{\mathrm{Cal}_t}) = Q(1-\hat{\alpha})$. This corresponds to considering the limit as $\#\mathrm{Cal} \to \infty$. This allows to focus on the impact of recursive updates in (2) and describe their behaviour by relying on Markov Chain theory.

3.1. Exchangeable case

ACI is usually applied in an adversarial context. If the scores are actually exchangeable, ACI's *validity* would not

improve upon SCP (known to be quasi-exactly *valid*), thus assessing ACI's impact on *efficiency* is necessary. Define $L(\alpha_t) = 2Q(1-\alpha_t)$ the length of the interval predicted by the adaptive algorithm at time t, and $L_0 = 2Q(1-\alpha)$ the length of the interval predicted by the non-adaptive algorithm (or equivalently, $\gamma = 0$).

Theorem 3.1. *Assume that: (i) $\alpha \in \mathbb{Q}$; (ii) the scores are exchangeable with quantile function Q; (iii) the quantile function is perfectly estimated at each time (as defined above); (iv) the quantile function Q is bounded and $\mathcal{C}^4([0,1])$. Then, for all $\gamma > 0$, $(\alpha_t)_{t>0}$ forms a Markov Chain, that admits a stationary distribution π_γ, and*

$$\frac{1}{T}\sum_{t=1}^{T} L(\alpha_t) \xrightarrow[T\to+\infty]{a.s.} \mathbb{E}_{\pi_\gamma}[L] \overset{not.}{=} \mathbb{E}_{\tilde{\alpha}\sim\pi_\gamma}[L(\tilde{\alpha})].$$

Moreover, as $\gamma \to 0$,

$$\mathbb{E}_{\pi_\gamma}[L] = L_0 + Q''(1-\alpha)\frac{\gamma}{2}\alpha(1-\alpha) + O(\gamma^{3/2}).$$

Interpretation of assumptions. Assumption (i) is weak since a practitioner will always select $\alpha \in \mathbb{Q}$ while assumption (ii) describes the classical exchangeable setting. The main assumptions are (iii) and (iv): (iii) can be interpreted as considering an infinite calibration set while (iv) is necessary[7] in order to define $\mathbb{E}_{\pi_\gamma}[L]$: here, we extend $Q(1-\hat{\alpha})$ by $Q(1)$ for $\hat{\alpha} < 0$. When $\hat{Q} \equiv \hat{Q}_t$ is the empirical quantile function on a calibration set Cal, the convergence in Theorem 3.1 holds conditionally to Cal. Finally, the regularity assumption on Q is purely technical.

Interpretation of the result. For standard distributions, $Q''(1-\alpha) > 0$,[8] and Theorem 3.1 implies that ACI on exchangeable scores *degrades* the *efficiency* linearly with γ compared to CP. This is an important takeaway from the analysis, that underlines that such adaptive algorithms may actually hinder the performance if the data does not have any temporal dependency, and a small γ is preferable. For example, if the residuals are standard gaussians, for $\alpha = 0.01$, setting $\gamma = 0.03$ (resp. $\gamma = 0.05$) will increase the length by 1.59% (resp. by 3.38%) with respect to $\gamma = 0$.

3.2. AR(1) case

We now consider the case of (highly) correlated residuals, which happens in many practical time series applications.

Definition 3.2 (AR(1) clipped)**.** $\varepsilon_{t+1} = \varphi\varepsilon_t + \xi_{t+1}$ with $(\xi_t)_t$ i.i.d. random variables admitting a continuous density with respect to Lebesgue measure, of support $\mathcal{S}$ clipped at a large value R, and $[-R, R] \subset \mathcal{S}$

Theorem 3.3. *Assume that: (i) $\alpha \in \mathbb{Q}$; (ii) the residuals follow an AR(1) process clipped at R of parameter φ (Definition 3.2); (iii) the quantile function Q of the stationary distribution of $(\varepsilon_t)_t$ is known. Then $(\alpha_t, \varepsilon_{t-1})$ is a homogeneous Markov Chain in $\mathbb{R}^2$ that admits a unique stationary distribution $\pi_{\gamma,\varphi}$. Moreover,*

$$\frac{1}{T}\sum_{t=1}^{T} L(\alpha_t) \xrightarrow[T\to+\infty]{a.s.} \mathbb{E}_{\pi_{\gamma,\varphi}}[L].$$

We numerically estimate $\gamma^*_\varphi = \mathrm{argmin}_\gamma \mathbb{E}_{\pi_{\gamma,\varphi}}[L]$ in Figure 2. To do so, AR(1) processes of length $T = 10^6$ are simulated for various φ and asymptotic variance 1. ACI is applied on each of them, with 100 different $\gamma \in [0, 0.2]$. Figure 2 (left) represents the average length depending on γ for each φ, and (right) the values of γ minimizing this average length for each φ (for 25 repetitions of the experiment). The average length is computed after imputing all the infinite intervals' length by the maximum of the process, as explained in Section 2. A similar study using instead the median length is provided after the proofs in Appendix B.

Interpretation. We make the following observations:

1. For high φ, ACI indeed improves for a strictly positive γ upon $\gamma = 0$. This proves that ACI can be used to produce smaller intervals for time series CP. The function $\gamma \mapsto \mathbb{E}_{\pi_{\gamma,\varphi}}[L]$ decreases until γ^*_φ, then increases again, as expected because very large γ cause the algorithm to be less stable and produce numerous infinite intervals.
2. In Figure 2 (left), zoomed-in plot, the black line represents asymptotic result of Theorem 3.1. We retrieve here that the expected length is minimal for $\gamma = 0$ and grows linearly with γ around 0. This behaviour is very similar for $\varphi = 0.6$.
3. For any γ, the function $\varphi \mapsto \mathbb{E}_{\pi_{\gamma,\varphi}}[L]$ is decreasing (Figure 2, left). Indeed, stronger correlation between residuals (i.e., a higher φ), allows to build smaller intervals. This confirms that ACI's impact strengthens with the strength of

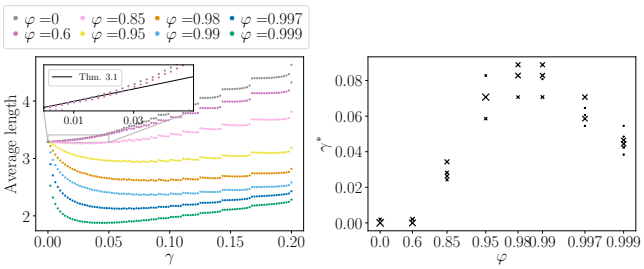


Figure 2. Left: evolution of the mean length depending on γ for various φ. Right: γ^* minimizing the average length for each φ (each cross has a size proportional to the number of runs for which γ^* was the minimizer).

[7] $\forall\gamma > 0$, $\mathbb{P}_{\pi_\gamma}(\tilde{\alpha} \leq 0) > 0$: we need $|Q(1)| < \infty$ to define $\mathbb{E}_{\pi_\gamma}[L]$.
[8] as $Q'(x) = \frac{1}{f(Q(x))}$ with f the scores' probability density function, $Q'(x)$ increases locally around x if and only if f decreases locally around $Q(x)$ (Q is increasing). Thus, $Q''(x) > 0$ if and only if f decreases locally around $Q(x)$. Thereby, for $x = 1-\alpha$ high (usually the case), $Q''(1-\alpha) > 0$ for standard distributions.

the temporal dependence.

4. Surprisingly, the function $\varphi \mapsto \gamma^*_\varphi$, that corresponds to the optimal learning rate for a given signal, *is non-monotonic*, (Figure 2, right). As $\gamma = 0$ is optimal for $\varphi = 0$, the function first increases. However, the optimal learning rate then diminishes as φ increases. This sheds light on the complex intrinsic tradeoffs of the method: for small values of φ, using $\gamma > 0$ simply degrades the *efficiency*; for "moderate" values of φ using a larger γ is necessary to quickly benefit from the short-term dependency between residuals; finally, for larger values of φ, the process exhibits a longer memory, thus it is crucial to find a smaller learning rate that produces more stable intervals, even if it means that the algorithm won't adapt as quickly.

What if $Q \neq \hat{Q}$? While our analysis provides a first step by comparing ACI to CP in the ideal case where the quantile distribution is known (for both methods), the impact of the finite-Cal is of interest. Indeed, if Cal is small, ACI can help to attain coverage **conditionally to a given** Cal even in the i.i.d. case. Yet intuitively, **marginally**, the randomness induced by ACI in the i.i.d. case would negatively impact efficiency w.r.t. $\gamma = 0$, even in the finite-Cal case. Finite sample trade-offs and general analysis of the case $Q \neq \hat{Q}$ is an important open direction.

Overall, these results highlight the importance of the choice of γ, as not choosing γ^* can lead to significantly larger intervals. In addition, they provide insights on the corresponding dynamics. Yet the choice of γ in more complex practical settings remains difficult: this calls for adaptive strategies.

4. Adaptive strategies based on ACI

To prevent the critical choice of γ an ideal solution is an adaptive strategy with a time dependent γ. We propose two strategies based on running ACI for $K \in \mathbb{N}$ values $\{(\gamma_k)_{k\leq K}\}$ of γ, chosen by the user. Overall, this does not increase the computational cost because $\mathrm{Tr_t}$ and $\mathrm{Cal_t}$ are shared between all ACI; thus the only additional cost is the computation of the K different quantiles. For any x_t, denote $\widehat{C}_{\alpha_{t,k}}(x_t)$ the interval at time t built by ACI using γ_k.

Naive strategy. A simple strategy is to use at each step the γ that achieved in the past the best *efficiency* while ensuring *validity*. For stability purposes, consider a warm-up period $T_w \leq T_1 - 1$. For each $t \geq T_0 + T_w$, we select $k^*_{t+1} \in \operatorname{argmin}_{k\in\mathcal{A}_t} \left\{ t^{-1} \sum_{s=1}^t \mathrm{length}(\widehat{C}_{\alpha_{s,k}}(x_s)) \right\}$ with $\mathcal{A}_t = \{k \in [\![1, K]\!] \mid t^{-1} \sum_{s=1}^t \mathbb{1}_{y_s \in \widehat{C}_{\alpha_{s,k}}(x_s)} \geq 1 - \alpha\}$ or $k^*_{t+1} \in \operatorname{argmin}_{k\in[\![1,K]\!]} \{|1-\alpha-t^{-1} \sum_{s=1}^t \mathbb{1}_{y_s \in \widehat{C}_{\alpha_{s,k}}(x_s)}|\}$ if $\mathcal{A}_t = \emptyset$. For the first T_w steps, an arbitrary strategy is applied (in simulations, $\gamma = 0$ for $t \leq T_w = 50$).

Online Expert Aggregation on ACI (AgACI). Instead of picking one γ in the grid, we introduce an adaptive aggregation of *experts* (Cesa-Bianchi & Lugosi, 2006), with expert k being ACI with parameter γ_k. This strategy is detailed in Algorithm 1, and schematised in Figure 3. At each step t, it performs two independent aggregations of the K-ACI intervals $\widehat{C}_{\alpha_{t,k}}(\cdot) \overset{\text{not.}}{=} [\hat{b}^{(\ell)}_{t,k}(\cdot), \hat{b}^{(u)}_{t,k}(\cdot)]$, one for each bound, and outputs $\widetilde{C}_t(\cdot) \overset{\text{not.}}{=} [\tilde{b}^{(\ell)}_t(\cdot), \tilde{b}^{(u)}_t(\cdot)]$. Aggregation computes an optimal weighted mean of the experts (Line 11), where the weights $\omega^{(\ell)}_{t,k}, \omega^{(u)}_{t,k}$ assigned to expert k depend on all experts performances (suffered *losses*) at time steps $1, \cdots, t$ (Line 9). We use the pinball loss ρ_β, as it is frequent in quantile regression, where the pinball parameter β is chosen to $\alpha/2$ (resp. $1 - \alpha/2$) for the lower (resp. upper) bound. These losses are plugged in the *aggregation rule* Φ. Finally, the aggregation rule can include the computation of the gradients of the loss (*gradient trick*, see Cesa-Bianchi & Lugosi, 2006, for more details). As aggregation rules require bounded experts, a thresholding step is added (Line 5).

Note that the pinball loss helps to avoid large intervals (e.g. it strongly penalizes infinite or very large intervals).

We chose Φ to be the Bernstein Online Aggregation (BOA, Wintenberger, 2017, see Appendix C.1 for a brief description), that was successfully applied for financial data (Berrisch & Ziel, 2021; Remlinger et al., 2021). We rely on R package OPERA (Gaillard & Goude, 2021), which allows the user to easily select among many other aggregation rules (EWA (Vovk, 1990), ML-Poly (Gaillard et al., 2014) or FTRL (Shalev-Shwartz & Singer, 2007; Hazan, 2019), etc.) that give similar results in our experiments. We use the gradient trick in the simulations. In the sequel, AgACI refers to AgACI using BOA and gradient trick.

Algorithm 1 Online Expert Aggregation on ACI (AgACI)

Input: Miscoverage rate α, grid $\{\gamma_k, k \in [\![1, K]\!]\}$, aggregation rule Φ, threshold values $M^{(\ell)}, M^{(u)}$.

1: Let $\beta^{(\ell)} = \alpha/2$ and $\beta^{(u)} = 1 - \alpha/2$
2: **for** $t \in [\![T_0 + 1, T_0 + T_1]\!]$ **do**
3: **for** $k \in [\![1, K]\!]$ **do**
4: Compute $\hat{b}^{(\cdot)}_{t,k}(x_t)$ using ACI with γ_k.
5: **if** $\hat{b}^{(\cdot)}_{t,k}(x_t) \notin \mathbb{R}$ **then** set $\hat{b}^{(\cdot)}_{t,k}(x_t) = M^{(\cdot)}$
6: **end for**
7: Set $\widetilde{C}_t(x_t) = [\tilde{b}^{(\ell)}_t(x_t), \tilde{b}^{(u)}_t(x_t)]$
8: **for** $k \in [\![1, K]\!]$ **do**
9: $\omega^{(\cdot)}_{t,k} = \Phi\left(\{ \rho_{\beta^{(\cdot)}}(y_s, \hat{b}^{(\cdot)}_{s,l}(x_s)), s \in [\![T_0 + 1, t]\!], l \in [\![1, K]\!]\}\right)$
10: **end for**
11: Define $\tilde{b}^{(\cdot)}_{t+1}(x) = \frac{\sum_{k=1}^K \omega^{(\cdot)}_{t,k} \hat{b}^{(\cdot)}_{t+1,k}(x)}{\sum_{l=1}^K \omega^{(\cdot)}_{t,l}}$ for any $x \in \mathbb{R}^d$
12: **end for**

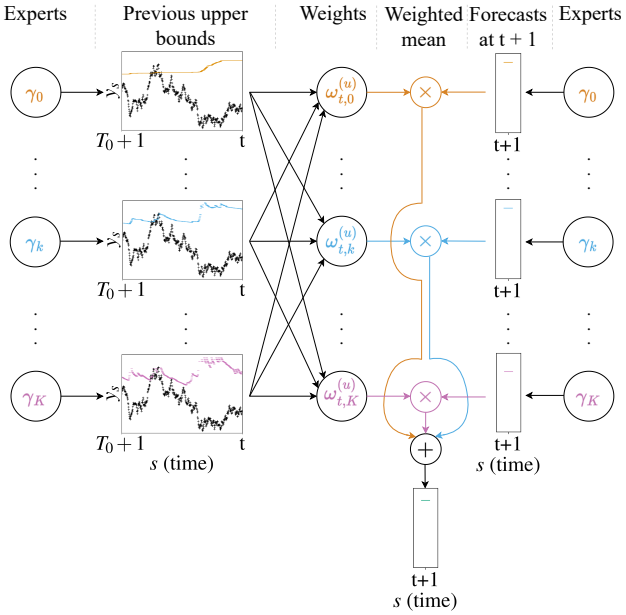


Figure 3. Scheme of AgACI algorithm, upper bound u only, for a forecast at time $t+1$. A similar procedure is performed independently for the lower bound ℓ in parallel.

5. Numerical evaluation on synthetic data sets

In this section we conduct synthetic experiments on a wide range of data sets presented in Section 5.1. The goal of this section is twofold. First, in Section 5.2, comparing our proposed adaptive strategies to ACI with a wide range of γ values. Second, in Section 5.4, comparing performances of AgACI and ACI to that of competitors – namely EnbPI and online sequential SCP, described in Section 5.3.

5.1. Data generation process and settings

We generate data according to:

$$\begin{aligned} Y_t =& 10 \sin\left(\pi X_{t,1} X_{t,2}\right) + 20\left(X_{t,3} - 0.5\right)^2 \\ &+ 10 X_{t,4} + 5 X_{t,5} + 0 X_{t,6} + \varepsilon_t, \end{aligned} \tag{3}$$

where the X_t are multivariate uniformly distributed on $[0,1]$, and $X_{t,6}$ represents an uninformative variable. The noise ε_t is generated from an ARMA(1,1) process of parameters φ and θ, i.e. $\varepsilon_{t+1} = \varphi\varepsilon_t + \xi_{t+1} + \theta\xi_t$, with ξ_t a white noise called the *innovation* (see Appendix C.2 for details). When the noise is i.i.d., one retrieves the simulations from Friedman et al. (1983). The temporal dependence is present only in the noise in order to control its strength and its impact on the algorithms' performance.

Given the non-linear structure of the data generating process, we use a random forest (RF) as predictive model, with the same hyper-parameters through all the experiments (specified in Appendix C.3).

To assess the impact of the temporal structure, we vary φ and θ in $\{0.1, 0.8, 0.9, 0.95, 0.99\}$. To focus on the impact of the dependence structure, the value of the innovation's variance is selected so that the asymptotic variance of ε_t is independent of φ, θ: here we choose $\lim_{t\to\infty} \mathrm{Var}(\varepsilon_t) = 10$. For each set of parameters, we generate $n = 500$ samples $(\varepsilon_t)_{t\in[\![1,T_0+T_1]\!]}$ with $T_0 = 200$. In the sequel we display the results on an ARMA(1,1) which are representative of all the results obtained. For the sake of simplicity, we consider $\varphi = \theta$. Complementary results (i) for an asymptotic variance of 1 (corresponding to a higher *signal to noise* ratio), (ii) for AR(1) and MA(1) models are available in Appendix D.

Joint visualisation of validity & efficiency. In order to simultaneously assess *validity* and *efficiency*, in Figures 4, 6 and 8, we represent on the same graph the empirical coverage and average median length (used for *efficiency* as imputing the infinite bounds by the maximum of the whole sequence is not always feasible in practice). In those three figures, the vertical dotted line represents the target miscoverage rate, $\alpha = 0.1$. Consequently, a method is *valid* when it lies at the right of this line, and the lower the better.

5.2. Impact of γ, performance of AgACI

Figure 4 illustrates the behaviour of ACI (with multiple values of γ), the naive strategy (empty triangles) and AgACI (black stars) for increasing (from left to right) values of φ, θ, with $T_1 = 200$. In particular, the top row shows the joint *validity* & *efficiency* and, for this figure only, we add in the bottom row the same graph using the average length after imputation (see details in Appendix D) to assess *efficiency* in another way.

When γ is small, one observes an undercoverage, which increases when the temporal dependency of ε increases. Increasing γ enables ACI to increase the interval's size faster when we do not cover, and thus to improve *validity*, which is achieved for high values of γ; however this also increases the frequency of uninformative (infinite) intervals, as deduced from the bottom row of Figure 4, where the average length after imputation grows with γ. Remark that these results do not contradict the validity result recalled at the beginning of Section 3, which is only asymptotic while we predict on 200 points. For φ, θ small, we observe that similarly to Theorem 3.1, the *efficiency* does not improve with γ. For moderate values of $\varphi, \theta \in \{0.8, 0.9, 0.95\}$, we observe that the average median length is decreasing with γ for $\gamma \geq 0.01$. This effect is observable on average but not present in all the 500 experiments. One possible explanation is that the shrinking effect of ACI on the predicted interval enables to significantly reduce the predicted interval when γ is large, and this effect is, on average, more important than the number of large intervals.

Moreover, the naive strategy is clearly not *valid*: indeed it results in greedily choosing a γ that achieved good results in the past, and is consequently slightly more likely to fail

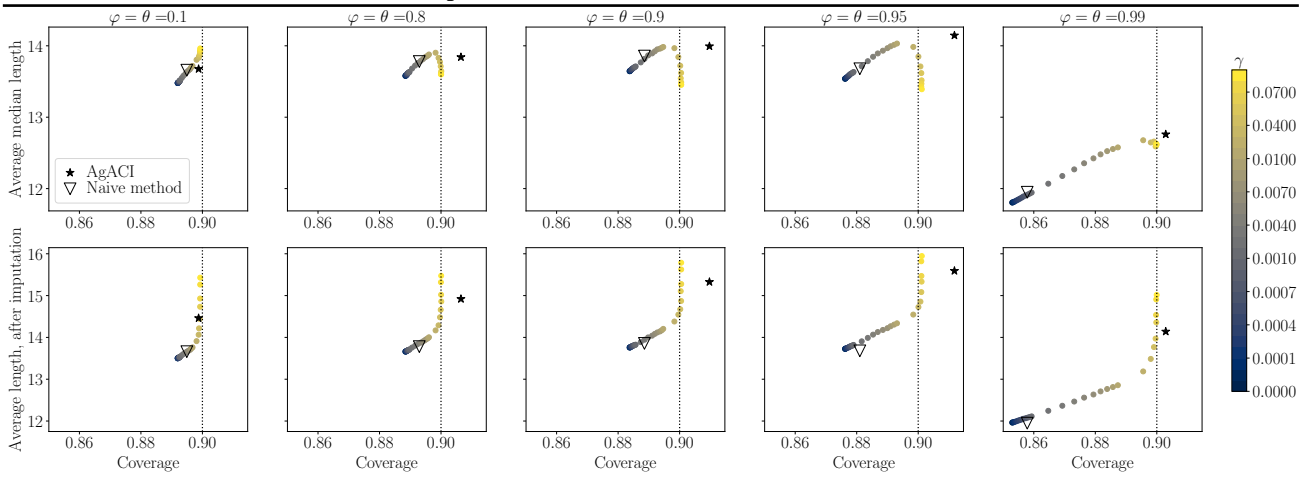


Figure 4. ACI performance with various θ, φ and γ on data simulated according to equation (3) with a Gaussian ARMA(1,1) noise of asymptotic variance 10 (see Appendix C.2). Top row: average median length w.r.t. the coverage. Bottom row: average length after imputation w.r.t. the coverage. Stars correspond to AgACI, and empty triangles to the naive choice.

to cover in future steps. Thereby, we do not consider it anymore. Finally, AgACI achieves *valid* coverage without increasing the median length with respect to each expert, and even improves the coverage. Overall, it appears to be a good candidate as a parameter-free method.

5.3. Description of baseline methods

We consider as baseline *online sequential split conformal prediction* (OSSCP), a generalisation of SCP[9]. The other competitor is EnbPI (Xu & Xie, 2021a), specifically designed for time series. Pseudo-codes and details are given in Appendix C.4. Offline SCP (for which $\mathrm{Tr}_t \equiv \mathrm{Tr}_0$ and $\mathrm{Cal}_t \equiv \mathrm{Cal}_0$) is not considered as a competitor because it is unfair to compare an *offline* algorithm to one that uses more recent data points. This corresponds to comparing a prediction at horizon T_{large} to one at horizon T_{small}. This is a limitation of the comparison in Xu & Xie (2021a).

OSSCP. We consider an online version of SCP by refitting the underlying regression model and recalibrating using the newest points. Moreover, to appropriately account for the temporal structure of the data, we use a *sequential split* as in Wisniewski et al. (2020): at any t, the time indices in Tr_t are smaller than those of Cal_t. Not randomizing aims at excluding future observations from Tr_t, which may lead to an under-estimation of the errors on Cal_t, thus eventually to smaller intervals with under-coverage. We compare both splitting strategies on simulations in Appendix D.4. OSSCP procedure is schematised in Figure 5(a).

Original EnbPI. EnbPI, Ensemble Prediction Interval (Xu & Xie, 2021a), works by updating the list of *conformity scores* with the most recent ones so that the intervals adapt to latest performances, without refitting the underlying regression model. Thereby, the predicted intervals can adapt to seasonality and trend. In EnbPI, B bootstrap samples of the training set are generated and the regression algorithm is fitted on each bootstrap sample producing B predictors. Finally, the predictors are aggregated in two ways: first, for each training point of index $t \leq T_0$, EnbPI aggregates only the subset of predictors trained on bootstrap sample *excluding* (x_t, y_t). This way, EnbPI constructs a set of hold-out calibration scores. Second, for test points of index $t > T_0$ EnbPI aggregates all the B predictors. A sketch of EnbPI is presented in Figure 5(b). Note that in Xu & Xie (2021a) they use a classical bootstrap procedure, not dedicated to time series.

They show empirically that it leads to *valid* coverage on real world time series, such as hourly wind power production and solar irradiation, while offline SCP fails to attain *valid* coverage.

EnbPI V2. Xu & Xie (2021a) used the mean aggregation during the training phase and the $(1-\alpha)$-th quantile of the predictors for the prediction. We consider using the mean aggregation all along the procedure as mixing both aggregations may hurt the performance of the algorithm (as shown in the following simulations). Note that simultaneously to

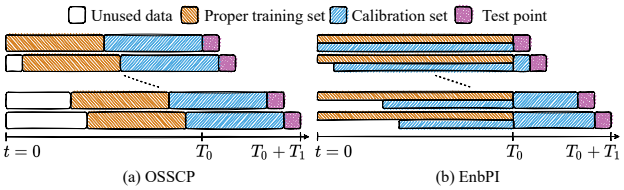


Figure 5. Scheme of the two baselines: OSSCP and EnbPI. In (a), Tr and Cal have equal size, but it can be changed.

[9]Recall here that inductive CP and SCP are equivalent methods.

our work, authors released an updated version on ArXiv (Xu & Xie, 2021b), incorporating a similar change.

5.4. Experimental results: impact of φ, θ

Figure 6 presents the results for data generated as in Section 5.1, for various (φ, θ). Each sample contains 300 observations, with $T_0 = 200$ and $T_1 = 100$. We compare AgACI (with $K = 30$ experts), ACI (with $\gamma \in \{0.01, 0.05\}$), OSSCP, EnbPI and EnbPI V2 (with mean aggregation). To assess the impact and interest of an online procedure, we also add offline SCP. Finally, to ensure the robustness of our conclusions each experiment is repeated $n = 500$ times, and Figure 6 includes the standard errors (given by $\frac{\hat{\sigma}_n}{\sqrt{n}}$, where $\hat{\sigma}_n$ is the empirical standard deviation).

Each color is associated to a set (φ, θ), each marker to an algorithm. To improve readability, we often link markers of the same method. There are thus two ways of analysing Figure 6: for a given method, the lines highlight the evolution of its performance with (φ, θ); for a given data distribution, the set of markers of its color allows to compare the methods. Figure 6, and results on AR(1) in Appendix D.2.1, highlight that in an AR(1) or ARMA(1,1) process:

- Refitting the method (OSSCP vs Offline SCP) brings a significant improvement, that increases with higher dependence (higher values for φ and θ).
- All methods produce smaller intervals for $\varphi = \theta = 0.99$.
- EnbPI loses coverage while producing shorter intervals when the dependence increases. The performance of EnbPI depends significantly on the type and strength of dependence.
- EnbPI V2 is closer to the target coverage than EnbPI.
- OSSCP loses *validity* & coverage as φ and θ increase.
- While ACI with $\gamma = 0.01$ also struggles for high values of φ and θ such as 0.99, we observe that it still attains *valid* coverage with a well chosen γ. Most importantly, ACI performances are robust to the increase of the dependence strength: except for the $\varphi = \theta = 0.99$, its markers are really close to each other.
- AgACI always nearly attains *validity* (coverage is over 89.8% for all φ), and achieves the best *efficiency* performance among *valid* methods.

Note that ACI's *valid* coverage with some γ comes at the price of predicting more infinite intervals. A more detailed analysis on this phenomenon is conducted in Appendix D.3. This can also be observed in graphs obtained with the average length after imputation, which are similar to Figure 6 and Appendix D.2.1. In these graphs, the *validity* remains unchanged as expected, but the *efficiency* is more degraded for ACI with $\gamma = 0.05$ and for AgACI, since they produce more often uninformative intervals, as observed in Figure 4.

Summary. We highlight the following takeaways:

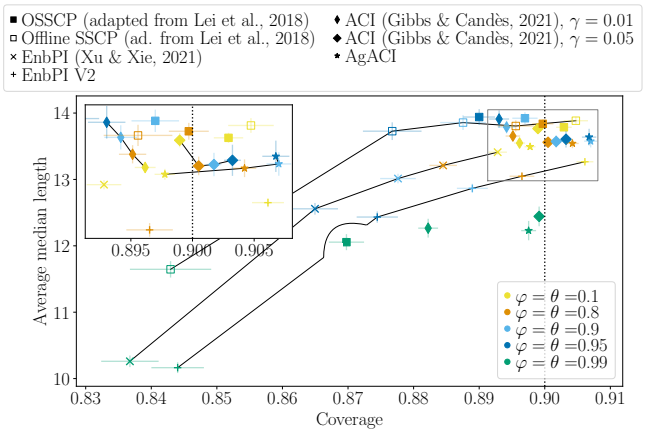


Figure 6. Performance of various CP methods on data simulated according to equation (3) with a Gaussian ARMA(1,1) noise of asymptotic variance 10 (see Appendix C.2). Results aggregated from 500 independent runs. Empirical standard errors displayed.

1. The temporal dependence impacts the *validity*.
2. Online is significantly better than offline.
3. **OSSCP.** Achieves *valid* coverage for φ and θ smaller than 0.9, but is not robust to the increasing dependence.
4. **EnbPI.** Its *validity* strongly depends on the data distribution (it is *valid* on a MA(1) noise, not in AR(1) and ARMA(1,1) noise). When the method is *valid*, it produces the smallest intervals. EnbPI V2 method should be preferred.
5. **ACI.** Achieves *valid* coverage for every simulation settings with a well chosen γ, or for dependence such that $\varphi < 0.95$. It is robust to the strength of the dependence.
6. **AgACI.** Achieves *valid* coverage for every simulation setting, with good *efficiency*.

6. Forecasting French electricity spot prices

In this last section, the task of forecasting French electricity spot prices with predictive intervals is considered in order to assess the methods on a real time series, and most importantly to show the relevance of ACI and AgACI in practice for time series without distribution shifts.

6.1. Presentation of the price data

The data set contains the French electricity spot prices, set by an auction market, from 2016 to 2019. Each day D before 12 AM, any producer (resp. supplier) submit their orders for the 24 hours of day $D + 1$. An order consists of an electricity volume in MWh offered for sale (resp. required to be purchased) and a price in €/MWh, at which they accept to sell (resp. buy) this volume. At 12 AM, the algorithm "Euphemia" (EUPHEMIA) fixes the 24 hourly prices of day $D + 1$ according to these offers and additional constraints. Thereby, it is an hourly data set, containing

$(3 \times 365 + 366) \times 24 = 35064$ observations. Our aim is to predict at day D (before 12 AM) the 24 prices of day $D+1$. Given the prices' construction, we consider the following explanatory variables: day-ahead forecast consumption, day-of-the-week, 24 prices of the day $D-1$ and 24 prices of the day $D-7$. An extract of the considered data set is presented in Appendix E.1.

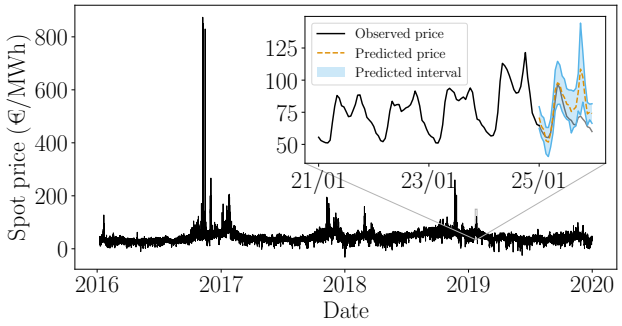


Figure 7. French electricity spot prices, from 2016 to 2019. Predicted intervals on the 25th of January 2019, using AgACI.

These prices exhibits medium to high peaks, as illustrated in Figure 7 where the French prices had reached 800 €/MWh in fall 2016, compared to an average price of approximately 40 €/MWh in 2019. These extreme events are mainly due to the non-storability of electricity and the inelasticity of the demand: when the demand is high compared to the available production, production units with expensive production costs must be called, leading to a huge market price.

6.2. Price prediction with predictive intervals in 2019

Since the 24 hours have very distinct patterns, we fit one model per hour, using again RF. We predict for the year 2019, using a sliding window of 3 years, as described in Figure 5(a), using one year and 6 months as proper training set and the most recent year and a half for calibration. The results are represented in Figure 8.

OSSCP over-covers but to a lesser extent than the offline version. This can be explained by a low presence of peaks during the test period. Indeed, by updating the whole procedure, the high peaks are "forgotten" which leads to small intervals while it is not the case for the offline version which leads to too large intervals. Thereby, online versions can help to improve *efficiency*, in addition to *validity*. **EnbPI** attains a *valid* coverage by over-covering. The under-coverage observed in the simulation study is not systematic, as in Xu & Xie (2021a). **ACI** gives the smallest intervals with a correct coverage, for $\gamma = 0.01$ and $\gamma = 0.05$. The update of the quantile level enables to shrink the intervals. While the simulation in Section 5.4 study outlines that ACI improves *validity*, this application illustrates that it can provide *efficient* interval. **AgACI** is more *efficient* than $\gamma = 0$ while maintaining *validity*. Yet it slightly over-covers, and is slightly less *efficient* than ACI with a well chosen γ.

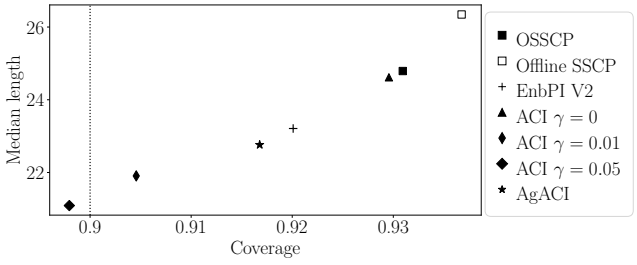


Figure 8. Performance of different CP methods on hourly spot electricity prices in France, trained from 2016 to 2018 and forecasted on 2019. Median length with respect to empirical coverage.

An illustration of the predicted intervals is given in the inset graphic of Figure 7, for AgACI, to highlight the practical relevance of such an approach on the spot prices.

However, as expected, these intervals only enjoy a *marginally* valid frequency. They do not have *conditional* guarantees. Especially, in this forecasting task, the predicted intervals cover the true prices around 88% of the time on week ends and Mondays, and 93% of the time on Tuesdays to Fridays (see Appendix E.2). Further developments are needed to improve this unbalanced coverage.

7. Conclusion

This article shows why and how ACI can be used for interval prediction in the context of time series with general dependencies. We prove that ACI deteriorates *efficiency* compared to CP in the exchangeable case and analyse the dependency on γ in the AR case with the support of numerical simulations. We propose an algorithm, AgACI, based on online expert aggregation, that wraps around ACI to avoid the choice of γ. We conduct extensive experiments on synthetic time series for various strengths and structures of time dependence, demonstrating ACI's robustness and better performances than baselines, with well chosen γ or using AgACI. Finally we perform a detailed application study on the high-stakes electricity price forecasting problem in the energy transition era. Future work includes theoretical study of the proposed aggregation algorithm, including whether it preserves the asymptotic *validity* observed experimentally or to quantify its *efficiency* with respect to the performances of each expert.

Acknowledgements

We thank Maximilien Germain, Pablo Jiménez and Constantin Philippenko for interesting discussions. We thank the anonymous reviewers for their useful comments on an earlier draft. The work of A. Dieuleveut is partially supported by ANR-19-CHIA-0002-01/chaire SCAI and Hi! Paris.

References

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

Berrisch, J. and Ziel, F. CRPS learning. *Journal of Econometrics*, 2021.

Cai, Y. and Davies, N. A simple bootstrap method for time series. *Communications in Statistics-Simulation and Computation*, 41(5):621–631, 2012.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. Robust Validation: Confident Predictions Even When Distributions Shift. *arXiv preprint arXiv:2008.04267*, 2020.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge University Press, 2006.

Chernozhukov, V., Wüthrich, K., and Yinchu, Z. Exact and Robust Conformal Inference Methods for Predictive Machine Learning with Dependent Data. In *Conference On Learning Theory*, pp. 732–749. PMLR, 2018.

Dashevskiy, M. and Luo, Z. Network traffic demand prediction with confidence. In *IEEE Global Telecommunications Conference*. IEEE, 2008.

Dashevskiy, M. and Luo, Z. Time series prediction with performance guarantee. *IET communications*, 5(8):1044–1051, 2011.

EUPHEMIA. Euphemia public description, single price coupling algorithm, April 2019. URL https://www.nemo-committee.eu/assets/files/190410_Euphemia%20Public%20Description%20version%20NEMO%20Committee.pdf.

Friedman, J. H., Grosse, E., and Stuetzle, W. Multidimensional additive spline approximation. *SIAM J. Sci. Stat. Comput.*, 1983.

Gaillard, P. and Goude, Y. *OPERA*, 2021. URL https://cran.r-project.org/package=opera. R package version 1.2.0.

Gaillard, P., Stoltz, G., and Van Erven, T. A second-order bound with excess losses. In *Conference on Learning Theory*, pp. 176–196. PMLR, 2014.

Gaillard, P., Goude, Y., and Nedellec, R. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. *International Journal of Forecasting*, 32(3):1038–1050, 2016.

Gibbs, I. and Candès, E. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, 2021.

Goehry, B. Random forests for time-dependent processes. *ESAIM: Probability and Statistics*, 24:801–826, 2020.

Goehry, B., Yan, H., Goude, Y., Massart, P., and Poggi, J.-M. Random forests for time series. *HAL hal-03129751*, 2021.

Härdle, W., Horowitz, J., and Kreiss, J.-P. Bootstrap methods for time series. *International Statistical Review*, 71 (2):435–459, 2003.

Hazan, E. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Kath, C. and Ziel, F. Conformal prediction interval estimation and applications to day-ahead and intraday power markets. *International Journal of Forecasting*, 37(2): 777–799, 2021.

Kreiss, J.-P. and Paparoditis, E. The hybrid wild bootstrap for time series. *Journal of the American Statistical Association*, 107(499):1073–1084, 2012.

Lago, J., De Ridder, F., and De Schutter, B. Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. *Applied Energy*, 221:386–405, 2018.

Lago, J., Marcjasz, G., De Schutter, B., and Weron, R. Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, 293:116983, 2021.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Maciejowska, K., Nowotarski, J., and Weron, R. Probabilistic forecasting of electricity spot prices using Factor Quantile Regression Averaging. *International Journal of Forecasting*, 32(3):957–965, 2016.

Meyn, S. P. and Tweedie, R. L. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Nowotarski, J. and Weron, R. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81:1548–1568, 2018.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive Confidence Machines for Regression. In *Machine Learning: ECML 2002*, pp. 345–356. Springer, 2002.

Remlinger, C., Brière, M., Alasseur, C., and Mikael, J. Expert aggregation for financial forecasting. *arXiv preprint arXiv:2111.15365*, 2021.

Romano, Y., Patterson, E., and Candès, E. Conformalized Quantile Regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Saha, A., Basu, S., and Datta, A. Random forests for spatially dependent data. *Journal of the American Statistical Association*, 0(0):1–19, 2021.

Shafer, G. and Vovk, V. A Tutorial on Conformal Prediction. *JMLR*, 9:51, 2008.

Shalev-Shwartz, S. and Singer, Y. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69(2-3): 115–142, 2007.

Tibshirani, R. J., Barber, R. F., Candès, E., and Ramdas, A. Conformal Prediction Under Covariate Shift. *Advances in Neural Information Processing Systems*, 32:11, 2019.

Uniejewski, B. and Weron, R. Regularized quantile regression averaging for probabilistic electricity price forecasting. *Energy Economics*, pp. 105121, 2021.

Vovk, V., Gammerman, A., and Saunders, C. Machine-Learning Applications of Algorithmic Randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 444–453. Morgan Kaufmann Publishers Inc., 1999.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer US, 2005.

Vovk, V. G. Aggregating strategies. *Proc. of Computational Learning Theory*, 1990.

Weron, R. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30(4):1030–1081, 2014.

Wintenberger, O. Optimal learning with bernstein online aggregation. *Machine Learning*, 106(1):119–141, 2017.

Wisniewski, W., Lindsay, D., and Lindsay, S. Application of conformal prediction interval estimations to market makers' net positions. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pp. 285–301. PMLR, 2020.

Xu, C. and Xie, Y. Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11559–11569. PMLR, 2021a.

Xu, C. and Xie, Y. Conformal prediction for dynamic time-series. *arXiv preprint arXiv:2010.09107*, 2021b.

Appendices

The appendices are organized as follows. First, Appendix A provides details about the Split Conformal Prediction procedure. Second, Appendix B proves the results of Section 3 and conducts the numerical analysis of Section 3.2 in the case where the *efficiency* is computed using the median length. Then, Appendix C contains details on the experimental setup (brief description of BOA, hyper-parameters, settings, pseudo-codes of competing algorithms). Finally, Appendices D and E contain complementary numerical results, respectively on synthetic data sets and on the French electricity spot prices data set.

A. Details on Split Conformal Prediction

In this section, we introduce and review the simplest theoretical properties of Split Conformal Prediction (SCP). More specifically, we present the whole algorithm, the theoretical guarantees and discuss the visualisation challenges arising when visualising a CP procedure.

A.1. Split Conformal Prediction Algorithm

Algorithm 2 Split Conformal Algorithm, with absolute value residuals scores

Input: Regression algorithm $\mathcal{A}$, significance level α, examples $z_1, \ldots, z_T$ with $z_t = (x_t, y_t)$.
Output: Prediction interval $\hat{\mathcal{C}}_\alpha(x)$ for any $x \in \mathbb{R}^d$.

1: Randomly split $\{1, \ldots, T\}$ into two disjoint sets Tr and Cal.
2: Fit a mean regression function: $\hat{\mu}(\cdot) \leftarrow \mathcal{A}\left(\{z_t, t \in \mathrm{Tr}\}\right)$
3: **for** $j \in \mathrm{Cal}$ **do**
4: Set $s_j = |y_j - \hat{\mu}(x_j)|$, the *conformity scores*
5: **end for**
6: Set $S_{\mathrm{Cal}} = \{s_j, j \in \mathrm{Cal}\}$
7: Compute $\widehat{Q}_{1-\alpha^{\mathrm{SCP}}}\left(S_{\mathrm{Cal}}\right)$, the $1-\alpha^{\mathrm{SCP}}$-th empirical quantile of S_{Cal}, with $1-\alpha^{\mathrm{SCP}} := (1-\alpha)\left(1 + 1/\#\mathrm{Cal}\right)$.
8: Set $\hat{\mathcal{C}}_\alpha(x) = \left[\hat{\mu}(x) \pm \widehat{Q}_{1-\alpha^{\mathrm{SCP}}}\left(S_{\mathrm{Cal}}\right)\right]$, for any $x \in \mathbb{R}^d$.

A.2. Illustration of the SCP procedure

Figure 9 provides a visualisation of the SCP procedure in a regression task. The conformity scores are taken to be the absolute value of the residuals.

A.3. Theoretical guarantees of Split Conformal Prediction

Conformal prediction relies on the assumption that the data is exchangeable.

Definition A.1 (Exchangeability)**.** $(X_t, Y_t)_{t=1}^T$ are exchangeable if for any permutation σ of $[\![1, T]\!]$ we have:

$$\mathcal{L}\left((X_1, Y_1), \ldots, (X_T, Y_T)\right) = \mathcal{L}\left(\left(X_{\sigma(1)}, Y_{\sigma(1)}\right), \ldots, \left(X_{\sigma(T)}, Y_{\sigma(T)}\right)\right),$$

where $\mathcal{L}$ designates the joint distribution.

Lei et al. (2018) proves the following Theorem A.2 about SCP quasi-exact *validity*.

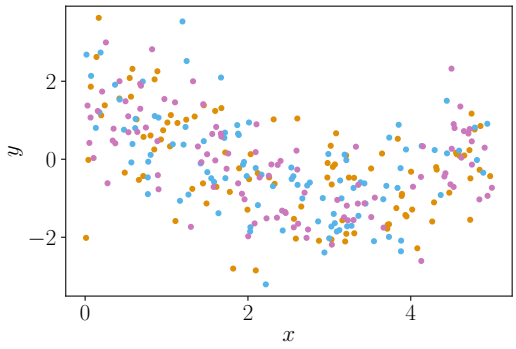
Theorem A.2. *Suppose $(X_t, Y_t)_{t=1}^{T+1}$ are exchangeable, and we apply algorithm 2 on $(X_t, Y_t)_{t=1}^T$ to predict an interval on X_{T+1}, $\hat{\mathcal{C}}_\alpha\left(X_{T+1}\right)$. Then we have:*

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

If, in addition, the scores S_{Cal} have a continuous joint distribution, we also have an upper bound:

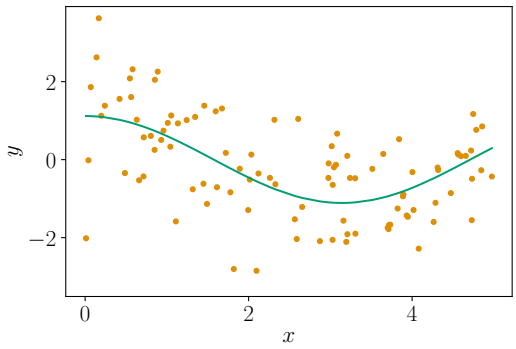
$$\mathbb{P}\left\{Y_{T+1} \in \hat{\mathcal{C}}_\alpha\left(X_{T+1}\right)\right\} \leq 1 - \alpha + \frac{1}{\#\mathrm{Cal} + 1}.$$

Step 1



- Create a proper training set, a calibration set, and keep your test set, by randomly splitting your data set.

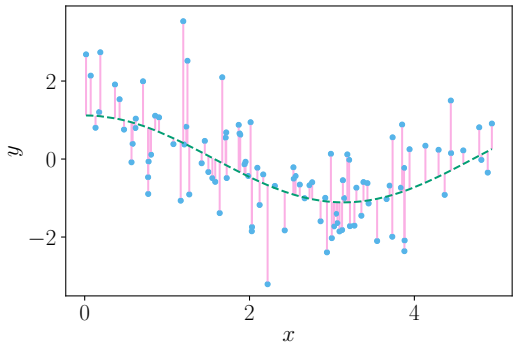
Step 2



On the proper training set:

- Learn $\hat{\mu}$

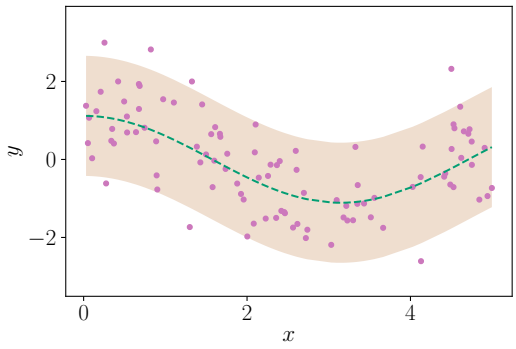
Step 3



On the calibration set:

- Predict with $\hat{\mu}$
- Get the residuals $\hat{\varepsilon}_i$ and form the scores $s_i = |\hat{\varepsilon}_i|$
- Compute the $(1-\alpha) \times (1 + \frac{1}{\#\text{Cal}})$ empirical quantile of the s_i, noted $Q_{1-\hat{\alpha}}(s_i)$

Step 4



On the test set:

- Predict with $\hat{\mu}$
- Build $\hat{C}_{\hat{\alpha}}(x)$: $[\hat{\mu}(x) \pm Q_{1-\hat{\alpha}}(s_i)]$

Figure 9. Schematic illustration of the Split Conformal Prediction procedure. Special case of a regression task, where the conformity scores are the absolute value of the residuals, as it is standard.

A.4. Examples of dependent scores when data noise is exchangeable

In this subsection, we provide two examples that highlight the importance of adapting CP to time series. In these examples, the scores are non exchangeable while the true noise of the data is exchangeable.

Example A.3 (Endogenous and not perfectly estimated)**.** Assume $X_t = Y_{t-1} \in \mathbb{R}$ and that

$$Y_t = aY_{t-1} + \varepsilon_t,$$

where ε_t is a white noise. This corresponds to an order-1 Auto-Regressive (i.e. AR(1)).

Assume that the fitted model is $\hat{f}_t(x) = \hat{a}x$, with $\hat{a} \neq a$. Then, for any t, we have that:

$$\begin{aligned}\hat{\varepsilon}_t &= Y_t - \hat{Y}_t = (a - \hat{a})\, Y_{t-1} + \varepsilon_t \\ \hat{\varepsilon}_t &= a\hat{\varepsilon}_{t-1} + \xi_t\end{aligned}$$

with $\xi_t = \varepsilon_t - \hat{a}\varepsilon_{t-1}$.

The residual process $(\hat{\varepsilon}_t)_{t\geq 0}$ is an ARMA(1,1) (Auto-Regressive Moving-Average, see section C.2) of parameters $\varphi = a$ and $\theta = -\hat{a}$.

Thus, we have generated dependent residuals (ARMA residuals) even though the underlying model only had white noise. □

Example A.4 (Exogenous and misspecified)**.** Assume $X_t \in \mathbb{R}^2$ and that:

$$Y_t = aX_{1,t} + bX_{2,t} + \varepsilon_t,$$

with $\varepsilon_t \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $X_{2,t+1} = \varphi X_{2,t} + \xi_t$, $\xi_t \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $X_{1,t}$ can be any random variable.

Assume that we misspecify the model so that the fitted model is $\hat{f}_t(x) = ax_1$ for any $t \geq 0$. Then, for any $t \geq 0$, we have that

$$\hat{\varepsilon}_t = Y_t - \hat{Y}_t = bX_{2,t} + \varepsilon_t.$$

Thus, we have generated dependent residuals (Auto-Regressive residuals) even if the underlying model only had i.i.d. Gaussian noise. □

A.5. How should we visualise CP predicted intervals?

We propose to have a closer look at how are constructed the prediction of this method. In this aim, we introduce model A.5.

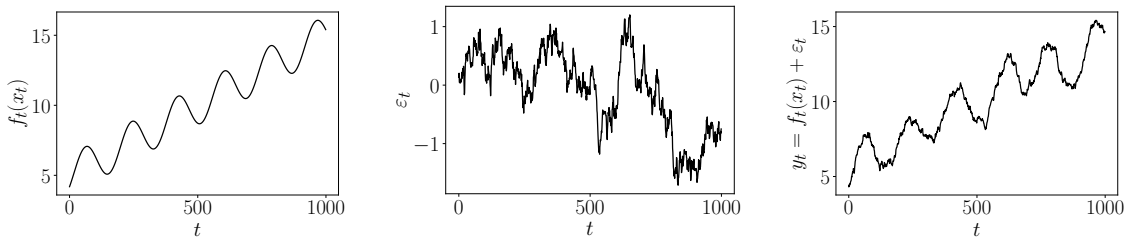
Model A.5.

$$\begin{aligned}x_t &= \cos\left(\frac{2\pi}{180}t\right) + \sin\left(\frac{2\pi}{180}t\right) + \frac{t}{100} \\ \varepsilon_{t+1} &= 0.99\varepsilon_t + \xi_{t+1}, \quad \xi_t \sim \mathcal{N}(0, 0.01) \\ Y_t &= f_t(x_t) + \varepsilon_t = x_t + \varepsilon_t\end{aligned} \quad .$$

In this model A.5, the explanatory variables are deterministic. A generation from this model is represented in Figure 10. The first subplot, Figure 10(a), represents x_t across time. The second subplot, Figure 10(b), represents the noise ε_t across time. Finally, the last subplot, Figure 10(c), represents the whole process Y_t across time.

The aim is to predict intervals of coverage 0.9 for values of Y_t, at $t > 500$, that is to say $T_0 = 500$ here. For simplicity, we assume $\hat{f}_t = f_t$ at each time step t and we do not represent the points used to obtain this perfect regression model. There are two ways of visualizing the predictions, that are represented in each row of Figure 11. If the focus of the analysis is on a specific application with the aim of analysing the whole prediction, it is relevant to represent the response y_t itself and the associated intervals. This is represented in the first row of Figure 11. Nevertheless, to better understand a CP method, it is relevant to represent the scores and the corresponding intervals, rescaled. This is represented in the second row of Figure 11 (even if the residuals are displayed and not their absolute value, i.e. the scores).

To better understand the difference between the two visualizations, let's look specifically at some observations. In the first line of the Figure 11, we can see that the intervals widen for $t \in [801; 900]$, while struggling to include the observations.



(a) $f_t(x_t)$ (b) ε_t (c) y_t

Figure 10. Representation of data simulated according to model A.5.

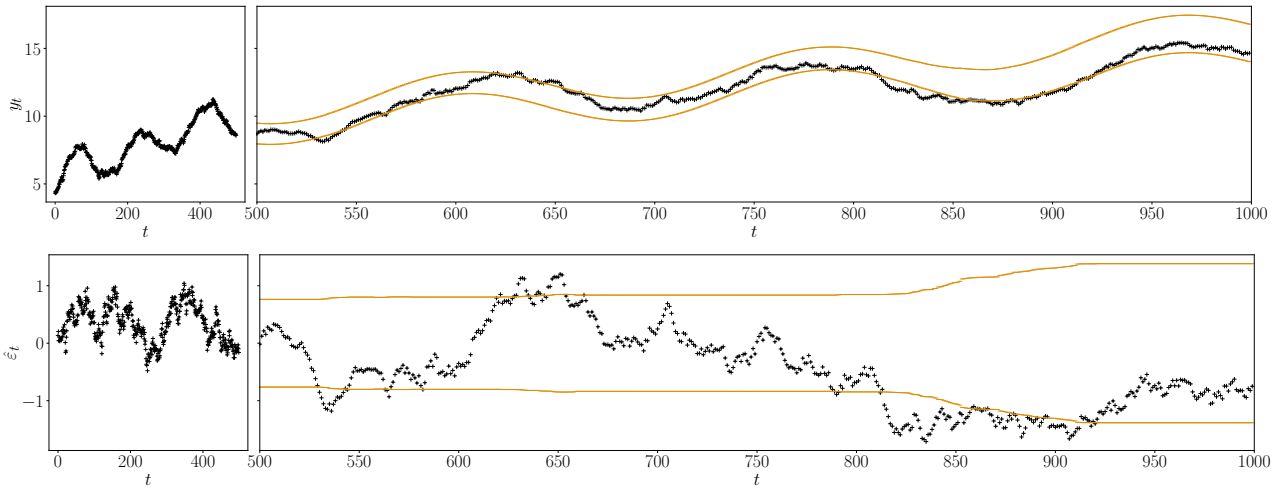


Figure 11. Visualisation of OSSCP on simulated data, from model model A.5. 1000 data points are generated. The 500 first ones form the initial calibration set, displayed on the first subplot of each row. The 500 last ones are the ones the algorithm tries to predict. They are displayed on the right subplot of each row. Observed values are in black, predicted intervals bounds are displayed in orange

Nevertheless, it is difficult to understand the underlying phenomenon on such a plot. Indeed, the points seem very similar to those for $t \in [660; 720]$. What considerably influences the CP are the scores and not the observed values. Thus, in the second line, at times $t \in [801; 900]$, we observe more clearly that the values go out of the previous range of values, being around 1.5 in absolute value. This explains why the intervals widen: the calibration set contains more and more high values, which increases the value of the quantile and, therefore, the length of the interval. To conclude, to analyse and assess the performances of CP procedures, we recommend representing the intervals around the *conformity scores* (or the residuals, depending on the score function) rather than the observed values. This is because the scores are what truly determine the conformal behaviour.

B. Proof of the results presented in Section 3 and additional numerical experiments

B.1. Proof of Theorem 3.1

We recall here Theorem 3.1.

Theorem 3.1. *Assume that: (i) $\alpha \in \mathbb{Q}$; (ii) the scores are exchangeable with quantile function Q; (iii) the quantile function is perfectly estimated at each time (as defined above); (iv) the quantile function Q is bounded and $\mathcal{C}^4([0,1])$. Then, for all $\gamma > 0$, $(\alpha_t)_{t>0}$ forms a Markov Chain, that admits a stationary distribution π_γ, and*

$$\frac{1}{T}\sum_{t=1}^{T} L(\alpha_t) \underset{T\to+\infty}{\overset{a.s.}{\longrightarrow}} \mathbb{E}_{\pi_\gamma}[L] \overset{not.}{=} \mathbb{E}_{\tilde{\alpha}\sim\pi_\gamma}[L(\tilde{\alpha})].$$

Moreover, as $\gamma \to 0$*,*

$$\mathbb{E}_{\pi_\gamma}[L] = L_0 + Q''(1-\alpha)\frac{\gamma}{2}\alpha(1-\alpha) + O(\gamma^{3/2}).$$

To prove Theorem 3.1, we rely on the following lemmas, that will be proved after the theorem. We denote B_β a Bernoulli random variable of parameter β and $P(x)$ designates the projection of x onto $[0,1]$. Finally, for $\gamma > 0$, define the following Markov Chain:

$$\alpha_{t+1} = \alpha_t + \gamma\left(\alpha - B_{P(\alpha_t)}\right) \text{ for } t > 0, \tag{4}$$

We introduce $(p,q) \in \mathbb{N} \times \mathbb{N}^*$, $p < q$, s.t. $\alpha = \frac{p}{q}$, and:

$$\mathcal{A} = \left\{\alpha + \gamma\frac{\gcd(q-p,p)}{q}\mathbb{Z}\right\} \cap\ ]\gamma(\alpha-1), 1+\gamma\alpha[. \tag{5}$$

Lemma B.1 (Finite state space)**.** *Assume that* $\alpha \in \mathbb{Q}$*. Then, for any* $\gamma > 0$*, the Markov Chain defined by* $\alpha_1 \in \mathcal{A}$ *and* $\alpha_{t+1} = \alpha_t + \gamma\left(\alpha - B_{P(\alpha_t)}\right)$*, for* $t > 0$ *has a finite state space* $\mathcal{A}$*.*

Lemma B.2 (Irreducibility)**.** *Assume that* $\alpha \in \mathbb{Q}$*. Then, for any* $\gamma > 0$*, the Markov Chain defined by Equation* (4)*, for* $t > 0$ *and* $\alpha_1 \in \mathcal{A}$*, is irreducible.*

Thereby we will prove that the chain admits a unique stationary distribution π_γ, we now compute the first four moments of the stationary distribution in Lemmas B.3 to B.6. The final proof relies on a Taylor expansion, that requires to control these four moments.

Lemma B.3 (Expectation)**.** *Let* $\gamma > 0$ *and consider again the Markov Chain defined in equation* (4)*. We have:*

$$\mathbb{E}_{\pi_\gamma}\left[(P(\tilde{\alpha}) - \alpha)\right] = 0.$$

Lemma B.4 (Second order moment)**.** *Let* $\gamma > 0$ *and consider again the Markov Chain defined in equation* (4)*. As* $\gamma \to 0$*, we have:*

$$\mathbb{E}_{\pi_\gamma}\left[(P(\tilde{\alpha}) - \alpha)^2\right] = \frac{\gamma}{2}\alpha(1-\alpha) + O(\gamma^2).$$

Lemma B.5 (Third order moment)**.** *Let* $\gamma > 0$ *and consider again the Markov Chain defined in equation* (4)*. As* $\gamma \to 0$*, we have:*

$$\mathbb{E}_{\pi_\gamma}\left[(P(\tilde{\alpha}) - \alpha)^3\right] = O(\gamma^{3/2}).$$

Lemma B.6 (Fourth order moment)**.** *Let* $\gamma > 0$ *and consider again the Markov Chain defined in equation* (4)*. As* $\gamma \to 0$*, we have:*

$$\mathbb{E}_{\pi_\gamma}\left[(P(\tilde{\alpha}) - \alpha)^4\right] = O(\gamma^2).$$

The proofs of these Lemmas are postponed to Appendices B.2 and B.3. Here, we first give the proof of the main theorem.

Proof of Theorem 3.1. Let $\gamma > 0$. For any $t > 0$ we have, for the recursion introduced in Equation (2), that

$$\alpha_{t+1} := \alpha_t + \gamma\left(\alpha - \mathbb{1}_{y_t \notin \widehat{C}_{\alpha_t}(x_t)}\right) = \alpha_t + \gamma\left(\alpha - \mathbb{1}_{S_t > \hat{Q}_{1-P(\alpha_t)}}\right),$$

where S_t is the conformity score at time t. Noting that $\mathbb{1}_{S_t > \hat{Q}_t(1-P(\alpha_t))} \overset{d}{=} B_{\mathbb{P}\left(S_t > \hat{Q}_t(1-P(\alpha_t))\right)}$, we obtain:

$$\begin{aligned}
\alpha_{t+1} &\overset{d}{=} \alpha_t + \gamma\left(\alpha - B_{\mathbb{P}\left(S_t > \hat{Q}_t(1-P(\alpha_t))\right)}\right) \\
&\overset{d}{=} \alpha_t + \gamma\left(\alpha - B_{\mathbb{P}(S_t > Q(1-P(\alpha_t)))}\right) \\
&\overset{d}{=} \alpha_t + \gamma\left(\alpha - B_{P(\alpha_t)}\right),
\end{aligned}$$

where the second line results from assumption (ii) and (iii), and the last equation from assumption (iii) only. Consequently, by induction, the chain defined by Equation (2) and

$$\alpha_{t+1} = \alpha_t + \gamma \left(\alpha - B_{P(\alpha_t)}\right), \tag{6}$$

with $\alpha_1 = \alpha$, have the same distribution.

Using assumption (i), Lemma B.1 ensures that the state space $\mathcal{A}$ of the Markov Chain defined in equation (6) is finite. Furthermore, Lemma B.2 also ensures that the chain is irreducible. Therefore, the chain is irreducible on a finite state space, thus it admits a unique stationary distribution, noted π_γ and for any positive function f such that $\int f \mathrm{d}\pi_\gamma < \infty$, we have (Meyn & Tweedie, 2012, Theorem 17.1.7):

$$\frac{1}{T}\sum_{t=1}^{T} f(\alpha_t) \xrightarrow[T\to\infty]{a.s.} \int f \mathrm{d}\pi_\gamma.$$

Remark that $L(\beta) = 2Q(1 - P(\beta))$ for any β. Therefore, combined with previous result we get the first result of Theorem 3.1:

$$\frac{1}{T}\sum_{t=1}^{T} L(\alpha_t) \xrightarrow[T\to+\infty]{a.s.} \mathbb{E}_{\tilde{\alpha}\sim\pi_\gamma}\left[L(\tilde{\alpha})\right].$$

We now need to characterize $\mathbb{E}_{\tilde{\alpha}\sim\pi_\gamma}\left[L(\tilde{\alpha})\right] = 2\mathbb{E}_{\tilde{\alpha}\sim\pi_\gamma}\left[Q(1 - P(\tilde{\alpha}))\right]$ as $\gamma \to 0$. Assume that $Q \in \mathcal{C}^4([0,1])$. Using Taylor series expansion, for any $\tilde{\alpha} \in \mathcal{A}$, there exists $\beta(\tilde{\alpha}) \in [0,1]$:

$$\begin{aligned} Q(1 - P(\tilde{\alpha})) =& Q(1-\alpha) + Q'(1-\alpha)(\alpha - P(\tilde{\alpha})) + \frac{Q''(1-\alpha)}{2}(\alpha - P(\tilde{\alpha}))^2 \\ &+ \frac{Q'''(1-\alpha)}{6}(\alpha - P(\tilde{\alpha}))^3 + \frac{Q''''(1-\beta(\tilde{\alpha}))}{24}(\alpha - P(\tilde{\alpha}))^4. \end{aligned} \tag{7}$$

To conclude, we take the expectation under π_γ of equation (7), which gives:

$$\begin{aligned} \mathbb{E}_{\pi_\gamma}\left[Q(1 - P(\tilde{\alpha}))\right] =& Q(1-\alpha) + Q'(1-\alpha)\mathbb{E}_{\pi_\gamma}\left[(\alpha - P(\tilde{\alpha}))\right] + \frac{Q''(1-\alpha)}{2}\mathbb{E}_{\pi_\gamma}\left[(\alpha - P(\tilde{\alpha}))^2\right] \\ &+ \frac{Q'''(1-\alpha)}{6}\mathbb{E}_{\pi_\gamma}\left[(\alpha - P(\tilde{\alpha}))^3\right] + \mathbb{E}_{\pi_\gamma}\left[\frac{Q''''(1-\beta(\tilde{\alpha}))}{24}(\alpha - P(\tilde{\alpha}))^4\right]. \end{aligned} \tag{8}$$

Injecting results of Lemmas B.3 to B.5 in equation (8), we obtain:

$$\mathbb{E}_{\pi_\gamma}\left[Q(1 - P(\tilde{\alpha}))\right] = Q(1-\alpha) + \frac{Q''(1-\alpha)}{4}\gamma\alpha(1-\alpha) + O(\gamma^{3/2}) + \mathbb{E}_{\pi_\gamma}\left[\frac{Q''''(1-\beta(\tilde{\alpha}))}{24}(\alpha - P(\tilde{\alpha}))^4\right]. \tag{9}$$

Finally, we can control the last term since $Q \in \mathcal{C}^4([0,1])$ by assumption, thus there exists $M > 0$ such that for any $x \in [0,1]$, $|Q''''(1-x)| < M$. Hence, using Lemma B.6 we obtain:

$$\begin{aligned} \left|\mathbb{E}_{\pi_\gamma}\left[Q''''(1-\beta(\tilde{\alpha}))(\alpha - P(\tilde{\alpha}))^4\right]\right| &\leq \mathbb{E}_{\pi_\gamma}\left[\left|Q''''(1-\beta(\tilde{\alpha}))\right|(\alpha - P(\tilde{\alpha}))^4\right] \\ &\leq M\mathbb{E}_{\pi_\gamma}\left[(\alpha - P(\tilde{\alpha}))^4\right] \\ &\leq MO(\gamma^{3/2}) \\ \mathbb{E}_{\pi_\gamma}\left[Q''''(1-\beta(\tilde{\alpha}))(\alpha - P(\tilde{\alpha}))^4\right] &= O(\gamma^{3/2}). \end{aligned} \tag{10}$$

Finally, combining equations (10) and (9) to conclude the proof by obtaining:

$$\mathbb{E}_{\pi_\gamma}\left[Q(1 - P(\tilde{\alpha}))\right] = Q(1-\alpha) + \frac{Q''(1-\alpha)}{4}\gamma\alpha(1-\alpha) + O(\gamma^{3/2}). \tag{11}$$

□

This concludes the proof of Theorem 3.1.

Remark: is it possible to use only 3 moments? The proof here relies on the control of the first four moments. It is not clear that the same result could be obtained using only a third order Taylor expansion, as we would then require a bound on $\mathbb{E}[|P(\tilde{\alpha}) - \alpha|^3]$, which is *not* guaranteed to be $O(\gamma^{3/2})$, contrary to $\mathbb{E}[(P(\tilde{\alpha}) - \alpha)^3]$.

B.2. Proof of Lemmas B.1 and B.2

Proof of Lemma B.1. Let $\gamma > 0$ and denote $\alpha = \frac{p}{q}$ with $0 < p < q$ and $(p, q) \in \mathbb{N}^2$. We denote E the state space of the Markov Chain defined by equation (6), starting from $a \in \mathcal{A}$. We show that $E = \mathcal{A}$.

First, (α_t) is stritcly bounded by $\gamma(\alpha - 1)$ and $1 + \gamma\alpha$. Thus $E \subset]\gamma(\alpha - 1), \gamma\alpha[$. Secondly, for any starting point $\alpha_1 \in \mathcal{A}$, we can observe that:

$$\begin{aligned}
\{\alpha_t, t \geq 1\} &\overset{a.s.}{\subset} \alpha_1 + \{k\gamma(\alpha - 1) + n\gamma\alpha, (k, n) \in \mathbb{N}^2\} \\
&\subset \alpha_1 + \{k\gamma(\alpha - 1) + n\gamma\alpha, (k, n) \in \mathbb{Z}^2\} \\
&= \alpha_1 + \{k\gamma\frac{p-q}{q} + n\gamma\frac{p}{q}, (k, n) \in \mathbb{Z}^2\} \\
&= \alpha_1 + \frac{\gamma}{q}\{(q-p)\mathbb{Z} + p\mathbb{Z}\} \\
&= \alpha_1 + \frac{\gamma}{q}\text{gcd}(q-p, p)\mathbb{Z} \\
&= \alpha + \frac{\gamma}{q}\text{gcd}(q-p, p)\mathbb{Z}
\end{aligned}$$

where $\text{gcd}(a, b)$ is the greatest common divisor of a and b. We have used at the last line that $\alpha_1 \in \mathcal{A}$ writes as $\alpha + \frac{\gamma}{q}\text{gcd}(q - p, p)k$, for some $k \in \mathbb{Z}$. Combining both results, we get that:

$$E \subset \left\{\alpha + \frac{\gamma}{q}\text{gcd}(q-p, p)\mathbb{Z}\right\} \cap \ ]\gamma(\alpha - 1), \gamma\alpha[.$$

This shows that the state space is finite and a subset of $\mathcal{A}$. The reciprocal implication is proved in the following Lemma, together with irreducibility. □

Proof of Lemma B.2. Our objective is to show that there is a path of positive probability going from any point of the state space $\mathcal{A}$ to any point of the same state space $\mathcal{A}$. Note that the chain always has at most two options when on a state x: make a step $\gamma\alpha$, with probability $1 - P(x)$, or a step $\gamma(\alpha - 1)$, with probability $P(x)$.

Let $(x, y) \in \mathcal{A}^2$. Thereby, there exist $(k, n), (l, m) \in \mathbb{N}^2$ such that:

$$\begin{aligned}
x &= \alpha + k\gamma\alpha + n\gamma(\alpha - 1) \\
y &= \alpha + l\gamma\alpha + m\gamma(\alpha - 1).
\end{aligned}$$

Thus, starting from x, to attain y, the chain has to make the path $y - x = (l - k)\gamma\alpha + (m - n)\gamma(\alpha - 1)$.

Noting that for any $h \in \mathbb{N}$ we have $\gamma\alpha(q - p)h + \gamma(\alpha - 1)hp = 0$, we can equivalently write that:

$$y - x = u\gamma\alpha + v\gamma(\alpha - 1), \tag{12}$$

with $(u, v) \in \mathbb{N}^2 \setminus \{(0, 0)\}$.

Thus, for any $(x, y) \in \mathcal{A}^2$ there exists $(u, v) \in \mathbb{N}^2 \setminus \{(0, 0)\}$ such that $y - x = u\gamma\alpha + v\gamma(\alpha - 1)$.

Let's show by induction on $u + v$ that for any $(u, v) \in \mathbb{N}^2$, and $(x, y) \in \mathcal{A}^2$ satisfying Equation (12) there exists a path of strictly positive probability between x and y.

Initialization. Suppose first that $u + v = 1$. Then, there are two options: $u = 1$ and $v = 0$ or the reverse. Assume the former: Equation (12) gives $y = x + \gamma\alpha$ and necessarily $x < 1$ since $y < 1 + \gamma\alpha$ because $y \in \mathcal{A}$. Thereby the step $\gamma\alpha$ has a probability $1 - P(x) > 0$ to occur. Thus the chain can attain y starting from x, i.e., $\mathbb{P}(\alpha_2 = y | \alpha_1 = x) > 0$. The second case works similarly, by observing that necessarily $x > 0$.

Heredity. Let $m \in \mathbb{N}$. We assume that for any $(u, v) \in \mathbb{N}^2$ such that $u + v = m$, and $(x, y) \in \mathcal{A}^2$ satisfying Equation (12) there exists a path of strictly positive probability between x and y, or formally there exists $t \in \mathbb{N}$ such that $\mathbb{P}(\alpha_t = y | \alpha_1 = x) > 0$.

Suppose now that $u + v = m + 1$ with $m \in \mathbb{N}^*$. If $v = 0$, then $y = x + u\gamma\alpha$ and similarly than for $v = 0$ and $u = 1$, the step $\gamma\alpha$ is probable. Let $z = x + \gamma\alpha$. We have:

- $\mathbb{P}(\alpha_2 = z|\alpha_1 = x) = 1 - P(x) > 0$.
- By our induction hypothesis, (y, z) satisfy Eq. 12 with $u + v = m$, thus there exists t such that $\mathbb{P}(\alpha_t = y|\alpha_2 = y) > 0$.

Overall, $\mathbb{P}(\alpha_t = y|\alpha_1 = x) > 0$.

If instead $u = 0$, then $y = x + v\gamma(\alpha - 1)$ and as for $u = 0$ and $v = 1$, the step $\gamma\alpha$ is of strictly positive probability and we conclude similarly.

Finally, if both u and v are non-null, then we can make the step $\gamma(\alpha - 1)$ if $x > 0$ and the step $\gamma\alpha$ otherwise, before using our induction hypothesis.

This shows that we can build a path of strictly positive probability for any $(x, y) \in \mathcal{A}^2$, and thereby that the chain is irreducible. □

B.3. Control of the first four moments: Lemmas B.3 to B.6

In the following Lemmas, to compute the first order moments of π_γ, we consider the chain $\alpha_{t+1} = \alpha_t + \gamma\left(\alpha - B_{P(\alpha_t)}\right)$ for $t > 0$, launched from the stationary distribution $\alpha_1 \sim \pi_\gamma$. Thanks to the stationarity property, for all $t \geq 1$, $\alpha_t \sim \pi_\gamma$.

Proof of Lemma B.3. Let $\gamma > 0$. To derive $\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)\right]$ we start by equation (6) with $t = 1$:

$$\begin{aligned}
& & \alpha_2 &= \alpha_1 + \gamma\left(\alpha - B_{P(\alpha_1)}\right) \\
&\text{taking expectation} & \mathbb{E}\left[\alpha_2\right] &= \mathbb{E}\left[\alpha_1\right] + \gamma\left(\alpha - \mathbb{E}\left[B_{P(\alpha_1)}\right]\right) \\
&\text{using} \quad \mathbb{E}\left[\alpha_1\right] = \mathbb{E}\left[\alpha_2\right] = \mathbb{E}_{\pi_\gamma}[\alpha], & 0 &= \gamma\left(\alpha - \mathbb{E}_{\pi_\gamma}\left[B_{P(\alpha_1)}\right]\right) \\
& & \mathbb{E}_{\pi_\gamma}\left[\mathbb{E}\left[B_{P(\alpha_1)}|\alpha_1\right]\right] &= \alpha \\
& & \mathbb{E}_{\pi_\gamma}\left[P(\alpha_1)\right] &= \alpha.
\end{aligned}$$

□

Proof of Lemma B.4. Let $\gamma > 0$. To derive $\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)^2\right]$ we start by equation (6) with $t = 1$:

$$\begin{aligned}
(\alpha_2 - \alpha)^2 &= (\alpha_1 - \alpha)^2 + \gamma^2(\alpha - B_{P(\alpha_1)})^2 + 2\gamma(\alpha - B_{P(\alpha_1)})(\alpha_1 - \alpha) \\
\mathbb{E}_{\pi_\gamma}\left[(\alpha_2 - \alpha)^2\right] &= \mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - \alpha)^2\right] + \gamma^2\mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)})^2\right] + 2\gamma\mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)})(\alpha_1 - \alpha))\right] \\
0 &= \gamma^2\mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)})^2\right] + 2\gamma\mathbb{E}_{\pi_\gamma}\left[(\alpha - P(\alpha_1))(\alpha_1 - \alpha)\right]
\end{aligned}$$

Consequently,

$$\begin{aligned}
2\gamma\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)(\alpha_1 - P(\alpha_1) + P(\alpha_1) - \alpha)\right] &= \gamma^2\mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)} + P(\alpha_1) - P(\alpha_1))^2\right] \\
2\gamma\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)^2\right] - 2\gamma\mathbb{E}_{\pi_\gamma}\left[(\alpha - P(\alpha_1))(\alpha_1 - P(\alpha_1))\right] &= \gamma^2\mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)} + P(\alpha_1) - P(\alpha_1))^2\right] \\
(2 - \gamma)\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)^2\right] &= \gamma\mathbb{E}_{\pi_\gamma}\left[P(\alpha_1)(1 - P(\alpha_1))\right] \\
&\quad + 2\mathbb{E}_{\pi_\gamma}\left[(\alpha - P(\alpha_1))(\alpha_1 - P(\alpha_1))\right]. \qquad (13)
\end{aligned}$$

We can compute $\mathbb{E}_{\pi_\gamma}\left[P(\alpha_1)(1 - P(\alpha_1))\right]$:

$$\begin{aligned}
\mathbb{E}_{\pi_\gamma}\left[P(\alpha_1)(1 - P(\alpha_1)) - \alpha(1 - \alpha)\right] &= \mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)(1 - P(\alpha_1)) + \alpha(1 - P(\alpha_1)) - \alpha(1 - \alpha)\right] \\
&= \mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)(1 - P(\alpha_1)) + \alpha(\alpha - P(\alpha_1))\right] \\
&= \mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)(1 - P(\alpha_1) - \alpha)\right] \\
&= \mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)(\alpha - P(\alpha_1) + 1 - 2\alpha)\right] \\
&= -\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)^2\right] + \mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)(1 - 2\alpha)\right] \\
&= -\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)^2\right] \\
\Rightarrow \mathbb{E}_{\pi_\gamma}\left[P(\alpha_1)(1 - P(\alpha_1))\right] &= \alpha(1 - \alpha) - \mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)^2\right] \qquad (14)
\end{aligned}$$

Reinjecting equation (14) in equation (13):

$$\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2\right] = \frac{\gamma}{2}\alpha(1-\alpha) + \mathbb{E}_{\pi_\gamma}\left[(\alpha - P(\alpha_1))(\alpha_1 - P(\alpha_1))\right] \tag{15}$$

We are now going to derive an upper and lower bound of $\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2\right]$. Note that $\text{sign}(\alpha - P(\alpha_1)) = -\text{sign}(\alpha_1 - P(\alpha_1))$, thus $\mathbb{E}_{\pi_\gamma}\left[(\alpha - P(\alpha_1))(\alpha_1 - P(\alpha_1))\right] \leq 0$. Hence we obtain the following upper bound:

$$\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2\right] \leq \frac{\gamma}{2}\alpha(1-\alpha). \tag{16}$$

Furthermore, using again this observation, and additionally that $|\alpha - P(\alpha_1)| \leq 1$ and $|\alpha_1 - P(\alpha_1)| \leq \gamma$ and from equation (15), we can obtain:

$$\begin{aligned}
\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2\right] \geq& \frac{\gamma}{2}\alpha(1-\alpha) - \gamma\mathbb{P}_{\pi_\gamma}(\alpha_1 \notin]0,1[) \\
\geq& \frac{\gamma}{2}\alpha(1-\alpha) - \gamma C_\alpha^{-1}\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2\right] \\
\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2\right] \geq& \frac{1}{1+\gamma C_\alpha^{-1}}\frac{\gamma}{2}\alpha(1-\alpha),
\end{aligned} \tag{17}$$

where the second inequality holds by observing that:

$$\begin{aligned}
&\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2\right] \geq (1-\alpha)^2\mathbb{P}_{\pi_\gamma}(\alpha_1 > 1) + \alpha^2\mathbb{P}_{\pi_\gamma}(\alpha_1 < 0) \\
&\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2\right] \geq C_\alpha\mathbb{P}_{\pi_\gamma}(\alpha_1 \notin [0,1]) \\
&\Rightarrow \mathbb{P}_{\pi_\gamma}(\alpha_1 \notin [0,1]) \leq C_\alpha^{-1}\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2\right]
\end{aligned}$$

with $C_\alpha = \min(\alpha^2, (1-\alpha)^2)$.

Gathering equations (17) and (16), we obtain:

$$\begin{aligned}
\frac{1}{(1+\gamma C_\alpha^{-1})}\frac{\gamma}{2}\alpha(1-\alpha) \leq \mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2\right] \leq \frac{\gamma}{2}\alpha(1-\alpha) \\
\left(\frac{1}{(1+\gamma C_\alpha^{-1})} - 1\right)\frac{\gamma}{2}\alpha(1-\alpha) \leq \mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2\right] - \frac{\gamma}{2}\alpha(1-\alpha) \leq 0 \\
\left|\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2\right] - \frac{\gamma}{2}\alpha(1-\alpha)\right| \leq \frac{\gamma^2 C_\alpha^{-1}}{2(1+\gamma C_\alpha^{-1})}\alpha(1-\alpha) \\
\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2\right] - \frac{\gamma}{2}\alpha(1-\alpha) = O(\gamma^2).
\end{aligned} \tag{18}$$

□

Proof of Lemma B.5. Let $\gamma > 0$. We start again by using equation (6) and removing the first terms as $\mathbb{E}_{\pi_\gamma}\left[(\alpha_2-\alpha)^3\right] = \mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^3\right]$. Then we will isolate $\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^3\right]$ and finally we will dominate each term obtained.

$$
\begin{aligned}
0 =& 3\gamma \mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2(\alpha - B_{P(\alpha_1)})\right] + 3\gamma^2 \mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha - B_{P(\alpha_1)})^2\right] \\
&+ \gamma^3 \mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)})^3\right] \\
0 =& 3\gamma \mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2(\alpha - P(\alpha_1))\right] + 3\gamma^2 \mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha - P(\alpha_1))^2)\right] \\
&+ 6\gamma^2 \mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha - P(\alpha_1))(P(\alpha_1) - B_{P(\alpha_1)})\right] \\
&+ 3\gamma^2 \mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(P(\alpha_1) - B_{P(\alpha_1)})^2\right] + \gamma^3 \mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)})^3\right] \\
3\gamma \mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^3\right] =& 3\gamma \mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^2(\alpha - P(\alpha_1))\right] + 6\gamma \mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)(\alpha - P(\alpha_1))\right] \\
&+ 3\gamma^2 \mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha - P(\alpha_1))^2)\right] + 3\gamma^2 \mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)P(\alpha_1)(1-P(\alpha_1))\right] \\
&+ \gamma^3 \mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)})^3\right] \\
3\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^3\right] =& 3\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^2(\alpha - P(\alpha_1))\right] - 6\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)^2\right] \\
&+ 3\gamma \mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha - P(\alpha_1))^2)\right] + 3\gamma \mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)P(\alpha_1)(1-P(\alpha_1))\right] \\
&+ \gamma^2 \mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)})^3\right] \\
3\left|\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^3\right]\right| \leq& 3\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^2(\alpha - P(\alpha_1))\right]\right| + 6\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)^2\right]\right| \\
&+ 3\gamma\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha - P(\alpha_1))^2)\right]\right| + 3\gamma\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)P(\alpha_1)(1-P(\alpha_1))\right]\right| \\
&+ \gamma^2\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)})^3\right]\right|. \qquad (19)
\end{aligned}
$$

To conclude, we can bound each term of the right hand side of equation (19). In order of appearance we obtain:

$$
\begin{aligned}
\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^2(\alpha - P(\alpha_1))\right]\right| &\leq \mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^2\left|\alpha - P(\alpha_1)\right|\right] \\
\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^2(\alpha - P(\alpha_1))\right]\right| &\leq \gamma^2. \qquad (20)
\end{aligned}
$$

$$
\begin{aligned}
\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)^2\right]\right| &\leq \mathbb{E}_{\pi_\gamma}\left[\left|\alpha_1 - P(\alpha_1)\right|(P(\alpha_1) - \alpha)^2\right] \\
&\leq \gamma \mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)^2\right] \\
\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))(P(\alpha_1) - \alpha)^2\right]\right| &\leq \frac{\gamma^2}{2}\alpha(1-\alpha) + O(\gamma^3), \qquad (21)
\end{aligned}
$$

where the last equality is obtained by using Lemma B.4.

$$
\begin{aligned}
\gamma\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha - P(\alpha_1))^2)\right]\right| &\leq \gamma \mathbb{E}_{\pi_\gamma}\left[\left|\alpha_1 - \alpha\right|(\alpha - P(\alpha_1))^2)\right] \\
&\leq \gamma D_{\gamma,\alpha}\mathbb{E}_{\pi_\gamma}\left[(\alpha - P(\alpha_1))^2)\right] \\
\gamma\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha - P(\alpha_1))^2)\right]\right| &\leq D_{\gamma,\alpha}\frac{\gamma^2}{2}\alpha(1-\alpha) + O(\gamma^3), \qquad (22)
\end{aligned}
$$

again using Lemma B.4, and with $D_{\gamma,\alpha} = \max(1+\gamma\alpha, \gamma(1-\alpha)) - \alpha = O(1)$.

$$
\begin{aligned}
\gamma\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)P(\alpha_1)(1-P(\alpha_1))\right]\right| \leq& \gamma\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))P(\alpha_1)(1-P(\alpha_1))\right]\right| \\
&+ \gamma\left|\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)P(\alpha_1)(1-P(\alpha_1))\right]\right| \\
\leq& \gamma\frac{1}{4}\mathbb{E}_{\pi_\gamma}\left[\left|\alpha_1 - P(\alpha_1)\right|\right] + \gamma\frac{1}{4}\mathbb{E}_{\pi_\gamma}\left[\left|P(\alpha_1) - \alpha\right|\right] \\
\leq& \frac{\gamma^2}{4} + \frac{\gamma}{4}\sqrt{\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1) - \alpha)^2\right]} \\
\leq& \frac{\gamma^2}{4} + \frac{\gamma}{4}\sqrt{\frac{\gamma}{2}\alpha(1-\alpha) + O(\gamma^2)} \\
\gamma\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)P(\alpha_1)(1-P(\alpha_1))\right]\right| \leq& O(\gamma^{3/2}), \qquad (23)
\end{aligned}
$$

where the last inequality comes from Lemma B.4 a third time.

$$\gamma^2 \left|\mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)})^3\right]\right| \leq \gamma^2 \max(\alpha^3, (1-\alpha)^3). \tag{24}$$

Gathering equations (20) to (24) together with equation (19), we obtain the following upper bound:

$$3\left|\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^3\right]\right| \leq 3\gamma^2 + 3\gamma^2\alpha(1-\alpha) + O(\gamma^3) + 3D_{\gamma,\alpha}\frac{\gamma^2}{2}\alpha(1-\alpha) + O(\gamma^3) + O(\gamma^{3/2}) + \gamma^2\max(\alpha^3,(1-\alpha)^3),$$

which leads to:

$$\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^3\right] = O(\gamma^{3/2}). \tag{25}$$

□

Proof of Lemma B.6. Let $\gamma > 0$. For the fourth order moment, the proof works in the same way for the third order moment, Lemma B.5.

$$\begin{aligned}
0 =& 4\gamma\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^3(\alpha - B_{P(\alpha_1)})\right] + 6\gamma^2\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2(\alpha - B_{P(\alpha_1)})^2\right] \\
&+ 4\gamma^3\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha - B_{P(\alpha_1)})^3\right] + \gamma^4\mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)})^4\right] \\
0 =& 4\gamma\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1) + P(\alpha_1) - \alpha)^3(\alpha - P(\alpha_1))\right] \\
&+ 6\gamma^2\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2(\alpha - P(\alpha_1) + P(\alpha_1) - B_{P(\alpha_1)})^2\right] \\
&+ 4\gamma^3\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha - B_{P(\alpha_1)})^3\right] + \gamma^4\mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)})^4\right] \\
4\gamma\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^4\right] =& 4\gamma\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^3(\alpha - P(\alpha_1))\right] + 12\gamma\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^2(P(\alpha_1)-\alpha)(\alpha - P(\alpha_1))\right] \\
&+ 12\gamma\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))(P(\alpha_1)-\alpha)^2(\alpha - P(\alpha_1))\right] \\
&+ 6\gamma^2\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2(\alpha - P(\alpha_1))^2\right] + 0 + 6\gamma^2\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2(P(\alpha_1) - B_{P(\alpha_1)})^2\right] \\
&+ 4\gamma^3\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha - B_{P(\alpha_1)})^3\right] + \gamma^4\mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)})^4\right] \\
4\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^4\right] =& 4\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^3(\alpha - P(\alpha_1))\right] - 12\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^2(P(\alpha_1)-\alpha)^2\right] \\
&- 12\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))(P(\alpha_1)-\alpha)^3\right] \\
&+ 6\gamma\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2(\alpha - P(\alpha_1))^2\right] + 6\gamma\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2P(\alpha_1)(1-P(\alpha_1))\right] \\
&+ 4\gamma^2\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha - B_{P(\alpha_1)})^3\right] + \gamma^3\mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)})^4\right] \\
4\left|\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^4\right]\right| \leq& 4\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^3(\alpha - P(\alpha_1))\right]\right| + 12\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^2(P(\alpha_1)-\alpha)^2\right]\right| \\
&+ 12\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))(P(\alpha_1)-\alpha)^3\right]\right| \\
&+ 6\gamma\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2(\alpha - P(\alpha_1))^2\right]\right| + 6\gamma\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2P(\alpha_1)(1-P(\alpha_1))\right]\right| \\
&+ 4\gamma^2\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha - B_{P(\alpha_1)})^3\right]\right| + \gamma^3\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha - B_{P(\alpha_1)})^4\right]\right|.
\end{aligned} \tag{26}$$

We are now going to dominate each term of the right hand side of equation (26) in order of appearance.

$$\begin{aligned}
\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^3(\alpha - P(\alpha_1))\right]\right| &\leq \mathbb{E}_{\pi_\gamma}\left[\left|\alpha_1 - P(\alpha_1)\right|^3\left|\alpha - P(\alpha_1)\right|\right] \\
\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^3(\alpha - P(\alpha_1))\right]\right| &\leq \gamma^3
\end{aligned} \tag{27}$$

$$\begin{aligned}
\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^2(P(\alpha_1)-\alpha)^2\right]\right| &= \mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^2(P(\alpha_1)-\alpha)^2\right] \\
\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))^2(P(\alpha_1)-\alpha)^2\right]\right| &\leq \gamma^2.
\end{aligned} \tag{28}$$

$$\begin{aligned}
\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))(P(\alpha_1)-\alpha)^3\right]\right| &\leq \mathbb{E}_{\pi_\gamma}\left[\left|\alpha_1 - P(\alpha_1)\right|\left|P(\alpha_1)-\alpha\right|^3\right] \\
&\leq \gamma\mathbb{E}_{\pi_\gamma}\left[\left|P(\alpha_1)-\alpha\right|^3\right] \\
\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1 - P(\alpha_1))(P(\alpha_1)-\alpha)^3\right]\right| &\leq O(\gamma^{5/2}).
\end{aligned} \tag{29}$$

where the last inequality holds using Lemma B.5.

$$\begin{aligned}\gamma\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2(\alpha-P(\alpha_1))^2\right]\right| &= \gamma\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2(\alpha-P(\alpha_1))^2\right]\\ &\leq \gamma D_{\gamma,\alpha}^2(\frac{\gamma}{2}\alpha(1-\alpha)+O(\gamma^2))\\ \gamma\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2(\alpha-P(\alpha_1))^2\right]\right| &\leq D_{\gamma,\alpha}^2\frac{\gamma^2}{2}\alpha(1-\alpha)+O(\gamma^3). \end{aligned} \tag{30}$$

again where we've used Lemma B.5, and re-used its notation $D_{\gamma,\alpha}=\max(1+\gamma\alpha,\gamma(1-\alpha))-\alpha=O(1)$.

$$\begin{aligned}\gamma\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2P(\alpha_1)(1-P(\alpha_1))\right]\right| =&\gamma\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-P(\alpha_1))^2P(\alpha_1)(1-P(\alpha_1))\right]\right.\\ &+2\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-P(\alpha_1))(P(\alpha_1)-\alpha)P(\alpha_1)(1-P(\alpha_1))\right]\\ &\left.+\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2P(\alpha_1)(1-P(\alpha_1))\right]\right|\\ \leq&\frac{\gamma}{4}\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-P(\alpha_1))^2\right]+\frac{\gamma}{2}\mathbb{E}_{\pi_\gamma}\left[|\alpha_1-P(\alpha_1)|\,|P(\alpha_1)-\alpha|\right]\\ &+\frac{\gamma}{4}\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^2\right]\\ \gamma\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)^2P(\alpha_1)(1-P(\alpha_1))\right]\right| \leq&\frac{\gamma^3}{4}+\frac{\gamma^2}{2}+\frac{\gamma^2}{8}\alpha(1-\alpha)+O(\gamma^3). \end{aligned} \tag{31}$$

again where we've used Lemma B.5.

$$\begin{aligned}\gamma^2\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha-B_{P(\alpha_1)})^3\right]\right| &\leq \gamma^2\mathbb{E}_{\pi_\gamma}\left[|\alpha_1-\alpha|\left|\alpha-B_{P(\alpha_1)}\right|^3\right]\\ &\leq \gamma^2 D_{\gamma,\alpha}\mathbb{E}_{\pi_\gamma}\left[\left|\alpha-B_{P(\alpha_1)}\right|^3\right]\\ \gamma^2\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha_1-\alpha)(\alpha-B_{P(\alpha_1)})^3\right]\right| &\leq \gamma^2 D_{\gamma,\alpha}\max(\alpha^3,(1-\alpha)^3). \end{aligned} \tag{32}$$

$$\gamma^3\left|\mathbb{E}_{\pi_\gamma}\left[(\alpha-B_{P(\alpha_1)})^4\right]\right| \leq \gamma^3\max(\alpha^4,(1-\alpha)^4). \tag{33}$$

Gathering equations (27) to (33) together with equation (26), we obtain finally:

$$\mathbb{E}_{\pi_\gamma}\left[(P(\alpha_1)-\alpha)^4\right]=O(\gamma^2). \tag{34}$$

□

B.4. Proof of Theorem 3.3

In this section, we prove Theorem 3.3. Recall the theorem:

Theorem 3.3. *Assume that: (i) $\alpha\in\mathbb{Q}$; (ii) the residuals follow an AR(1) process (i.e., $\varepsilon_{t+1}=\varphi\varepsilon_t+\xi_{t+1}$ with $(\xi_t)_t$ i.i.d. random variables admitting a continuous density with respect to Lebesgue measure, of support $\mathcal{S}$) clipped at a large value R, and $[-R,R]\subset\mathcal{S}$; (iii) the quantile function Q of the stationary distribution of $(\varepsilon_t)_t$ is known. Then $(\alpha_t,\varepsilon_{t-1})$ is a homogeneous Markov Chain in $\mathbb{R}^2$ that admits a unique stationary distribution $\pi_{\gamma,\varphi}$. Moreover,*

$$\frac{1}{T}\sum_{t=1}^{T}L(\alpha_t)\underset{T\to+\infty}{\overset{a.s.}{\longrightarrow}}\mathbb{E}_{\pi_{\gamma,\varphi}}[L].$$

We consider $Z_t=(\alpha_t,\varepsilon_{t-1})$ defined in the state-space $\mathcal{Z}=\mathcal{A}\times[-R,R]$ by

$$\begin{cases}\alpha_{t+1} &= \alpha_t+\gamma\left(\alpha-\mathbb{1}\{|\varepsilon_t|>Q_{1-P(\alpha_t)}\}\right),\\ \varepsilon_t &= -R\vee(\varphi\varepsilon_{t-1}+\xi_t)\wedge R\end{cases}$$

That is, $(\alpha_t)_{t\geq 0}$ is the recurrence defined by Equation (2), and $(\varepsilon_t)_{t\geq 0}$ is an AR(1) process with parameters φ clipped at some large value R. Finally, $(\xi_t)_t$ is a sequence of i.i.d. r.v. admitting a continuous density with respect to the Lebesgue measure, of support $\mathcal{S} \supset [-R, R]$.

This chain is defined for parameters α, R considered as fixed, and we focus on the influence of γ, φ. The main difference w.r.t. the previous section is that the state space is not countable anymore. More precisely, the state space is a product of a finite discrete set and an interval of $\mathbb{R}$.

The state-space $\mathcal{Z}$ is $\mathcal{A} \times [-R, R]$, where $\mathcal{A}$ is defined in the previous Appendix B.1, equation (5). We equip $\mathcal{Z}$ with the σ-algebra $\mathcal{F} = \mathcal{P}(\mathcal{A}) \times \mathcal{B}(\mathbb{R})$, where $\mathcal{P}(\mathcal{A})$ is the power-set of the finite set $\mathcal{A}$ and $\mathcal{B}(\mathbb{R})$ is the borel set of $\mathbb{R}$.

Lemma B.7. *The sequence $(Z_t)_{t\geq 0}$ is a Markov chain. Moreover, the chain is Harris-recurrent, and admits a stationary distribution $\pi_{\gamma,\varphi}$.*

Proof. We observe that

$$Z_t = \begin{pmatrix} \alpha_{t+1} \\ \varepsilon_t \end{pmatrix} = \begin{pmatrix} \alpha_t + \gamma\left(\alpha - \mathbb{1}\{|\varphi\varepsilon_{t-1} + \xi_t| > Q_{1-P(\alpha_t)}\}\right) \\ -R \vee (\varphi\varepsilon_{t-1} + \xi_t) \wedge R \end{pmatrix} =: F_{\gamma,\varphi}(Z_{t-1}, \xi_t). \tag{35}$$

For a function $F_{\gamma,\varphi} : \mathbb{R}^2 \times \mathbb{R}$. Consequently, Z_t follows a *Non-Linear State Space* model (Meyn & Tweedie, 2012, Section 2.2.2 and Chapter 7). We denote $P_{\gamma,\varphi}$ the probability kernel or Markov transition function, that is, for any $z = (a, e) \in \mathcal{Z}$, and $F \in \mathcal{F}$:

$$P_{\gamma,\varphi}(z, F) = \mathbb{P}(Z_1 \in F | Z_0 = z).$$

Remark that relying on Equation (35), we have an explicit formula for $P_{\gamma,\varphi}$. Defining the sequence of functions $(F_t)_{t\geq 1}$ such that

$$F_{t+1}\left(z_0, \xi_1, \ldots \xi_{t+1}\right) = F_{\gamma,\varphi}\left(F_t\left(z_0, \xi_1, \ldots \xi_t\right), \xi_{t+1}\right)$$

where z_0 and (ξ_i) are arbitrary real numbers. By induction we have that for any initial condition $Z_0 = z_0 \in \mathcal{Z}$ and any $t \in \mathbb{N}$,

$$Z_t = F_t\left(z_0, \xi_1, \ldots, \xi_t\right),$$

which immediately implies that the t-step transition function may be expressed as

$$P^t_{\gamma,\varphi}(z, F) = \mathbb{P}\left(F_t\left(z, \xi_1, \ldots, \xi_t\right) \in F\right) = \int \cdots \int \mathbb{1}\left\{F_t\left(z, \xi_1, \ldots, \xi_t\right) \in F\right\} p\left(d\xi_1\right) \ldots p\left(d\xi_t\right)$$

where p is the distribution of ξ.

We first prove that the chain is ψ-irreductible, for $\psi = \mu \otimes \lambda_{\text{Leb}}$, with μ the uniform probability measure on $\mathcal{A}$ and λ_{Leb} the Lebesgue measure on $[-R; R]$. [10]

For any $z_0 = (a_0, e_0) \in \mathcal{Z}$ and $F = \{a'\} \times \mathcal{O}$, with $\mathcal{O}$ open set, such that $\psi(F) \neq 0$ we have that, for some t large enough

$$\mathbb{P}(Z_t \in F | Z_0 = z_0) > 0.$$

Indeed,

1. There exists a path $(a_0, \ldots, a_t = a')$ in $\mathcal{A}$ from a_0 to a' such that for all $s \in \{1, \ldots, t-1\}$, $0 < a_s < 1$; and $a_{s+1} - a_s \in \{\gamma(\alpha - 1), \gamma\alpha\}$ similarly to the proof of Lemma B.2 since $\alpha \in \mathbb{Q}$.
2. Let E_{s+1} be the event such that we obtain a_{s+1} from a_s. Technically, if
   a. if $a_{s+1} - a_s = \gamma(\alpha - 1)$, $E_{s+1} = \{\xi_s$ such that $|\varphi\varepsilon_{s-1} + \xi_s| > Q_{1-a_s}\}$
   b. conversely, if $a_{s+1} - a_s = \gamma\alpha$, $E_{s+1} = \{\xi_s$ such that $|\varphi\varepsilon_{s-1} + \xi_s| \leq Q_{1-a_s}\}$.

[10] Moreover ψ is transformed to remove mass from the sets that cannot be reached by the chain $(Z_t)_t$, i.e., if B is such that $\mathbb{P}(Z_t \in B) = 0$ for all t. This only concerns extremely marginal points, possible only $\alpha > 1$ or $\alpha = \min \mathcal{A}$, for which we assign a zero mass to $\alpha \times \mathcal{U}$ for some $\mathcal{U}$.

3. Then if $0 < a' < 1$, we can directly conclude, as we have that for all $s \in \{1, \dots, t\}$,

$$\mathbb{P}\big(Z_{s+1} \in \{a_{s+1}\} \times E_{s+1} | Z_s = (a_s, z_s)\big) = \mathbb{P}(E_{s+1}) > \delta > 0.$$

Indeed (for case a.):

$$\begin{aligned}\mathbb{P}(E_{s+1}) &= \mathbb{P}\{\xi_s \text{ such that } |\varphi\varepsilon_{s-1} + \xi_s| > Q_{1-a_s}\} \\ &> \min\Big(\mathbb{P}\{\xi_s \text{ such that } \xi_s > Q_{1-\min \mathcal{A}\cup\mathbb{R}_+^*}\}, \mathbb{P}\{\xi_s \text{ such that } \xi_s < -Q_{1-\min \mathcal{A}\cup\mathbb{R}_+^*}\}\Big) =: \delta,\end{aligned}$$

with $\delta > 0$ by the assumption (ii) (esp. the fact that the support $\mathcal{S}$ of ξ_s includes $[-R, R]$).
The proof is similar for case b.
Consequently, $\mathbb{P}(Z_t \in F | Z_0 = z_0) > \delta^t > 0$.

4. The argument extends to the case where $a' < (0, 1)$, relying on the fact that $\psi(F) > 0$.

Moreover, the argument can be extended to show that for any $a', \mathcal{O}$, there exists δ' such that for all a_0, e_0, there exists $t \leq C_{\alpha,\gamma}$ (e.g., $C_{\alpha,\gamma} = \frac{2}{\alpha\gamma}$) such that

$$\mathbb{P}(Z_t \in F | Z_0 = z_0) > \delta'.$$

Which proves that the chain will visit infinitely many times any Borel set F with probability 1, and is consequently Harris-recurrent (Meyn & Tweedie, 2012, Chapter 9). Using Theorem 10.0.1 in Meyn & Tweedie (2012), we conclude that the chain admits a unique stationary distribution $\pi_{\gamma,\varphi}$.

Finally, applying (Theorem 17.1.7 Meyn & Tweedie, 2012) to the later result gives:

$$\frac{1}{T}\sum_{t=1}^{T} L(\alpha_t) \underset{T\to+\infty}{\overset{a.s.}{\longrightarrow}} \mathbb{E}_{\pi_{\gamma,\varphi}}[L].$$

□

B.5. Numerical study of ACI efficiency with AR(1) residuals, with respect to the median length

We here reproduce the same experiment as in Section 3.2, but focus on the *efficiency* as the median of the intervals' lengths instead of the average (after imputation). Results are given in Figure 12.

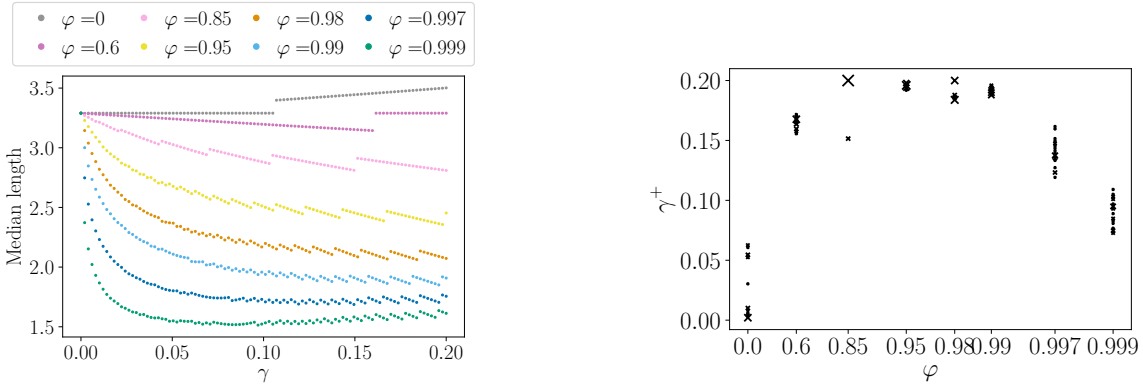


Figure 12. Left: evolution of the median length depending on γ for various φ. Right: γ^+ minimizing the median length for each φ.

Observations are very similar to the average length case, especially regarding (i) the monotonicity of the median interval length w.r.t. φ, (ii) the existence of a minimum γ_φ^+ to the function $\gamma \mapsto \mathrm{Med}_{\pi_{\gamma,\varphi}}[\tilde{\alpha}] := \mathrm{argmin}_{\mathrm{m}} \mathbb{E}_{\pi_{\gamma,\varphi}}[|\tilde{\alpha} - \mathrm{m}|]$ (iii) the non-monotonicity of $\varphi \mapsto \gamma_\varphi^+$.

C. Experimental details.

C.1. Details on the BOA procedure

The Bernstein Online Aggregation (BOA) procedure (Wintenberger, 2017) is a type of aggregation rule Φ. The weights outputted by BOA have an exponential form. In the exponent is plugged the difference between the loss suffered by the last aggregated forecast and the current squared loss suffered by the expert, instead of plugging the losses suffered by the experts (this would be Exponential Weighted Aggregation, Vovk, 1990). As stated in Wintenberger (2017), "this procedure favors online learners that predicted accurately and which past predictions losses are close to the loss of the last aggregative online learner, ensuring the stability in time and a small quadratic variation". For more details, we refer the reader to the original paper Wintenberger (2017).

C.2. Details ARMA(1,1) processes

Definition C.1 (ARMA(1,1) process)**.** We say that ε_t is an ARMA(1,1) process if for any t:

$$\varepsilon_{t+1} = \varphi\varepsilon_t + \xi_{t+1} + \theta\xi_t,$$

with:

- $\theta + \varphi \neq 0$, $|\varphi| < 1$ and $|\theta| < 1$;
- ξ_t is a white noise of variance σ^2, called the *innovation*.

The asymptotic variance of this process is:

$$\text{Var}(\varepsilon_t) = \sigma^2 \frac{1 - 2\varphi\theta + \theta^2}{1 - \varphi^2}. \tag{36}$$

An ARMA(1,1) is thus characterised by three parameters: the coefficients φ and θ and the innovation's variance σ^2. The larger the coefficients, in absolute value, the greater the time dependence and variance. Note that when $\varphi = 0$, the ARMA(0,1) process corresponds to a MA(1) and when $\theta = 0$, the ARMA(1,0) process corresponds to an AR(1).

To fix the asymptotic variance of an ARMA(1,1) of parameters φ and θ to v, we fix $\sigma^2 = v\frac{1-\varphi^2}{1-2\varphi\theta+\theta^2}$.

C.3. Random forest parameters

All the random forests model have the same parameters, that are the following:

- Number of trees: 1000
- Minimum sample per leaf: 1 (default)
- Maximum number of features: d (default)

Furthermore, for EnbPI, as there is already an individual bootstrap in the algorithm, the random forest regressors do not bootstrap them again.

C.4. Details about the baselines and comparison

C.4.1. EnbPI full algorithm

In order to be self-contained and precise the modifications done in EnbPI V2, the EnbPI algorithm from Xu & Xie (2021a) is recalled in the following. In purple we precise the difference in EnbPI V2.

Remark on the bootstrap approach. The bootstrap scheme is not adapted to time series, even if such strategies have been developed (Härdle et al., 2003; Kreiss & Paparoditis, 2012; Cai & Davies, 2012), and could be used to improve the adequation of EnbPI with the time series framework. Furthermore, recent works have proposed modifications of RF in the dependent setting (Goehry, 2020; Goehry et al., 2021; Saha et al., 2021). Generalizing these improvements to any ensemble method and use it for EnbPI could also enhance its performance, but is out of the scope of this paper.

Algorithm 3 Sequential Distribution-free Ensemble Batch Prediction Intervals (EnbPI)

Input: Training data $\{(x_i, y_i)\}_{i=1}^{T}$, regression algorithm $\mathcal{A}$, decision threshold α, aggregation function φ, number of bootstrap models B, the batch size s, and test data $\{(x_t, y_t)\}_{t=T+1}^{T+T_1}$, with y_t revealed only after the batch of s prediction intervals with t in the batch are constructed.

Output: Ensemble prediction intervals $\{C_\alpha(x_t)\}_{t=T+1}^{T+T_1}$

for $b = 1, \ldots, B$ **do**
  Sample with replacement an index set $S_b = (i_1, \ldots, i_T)$ from indices $(1, \ldots, T)$
  Compute $\hat{f}^b = \mathcal{A}\left(\{(x_i, y_i) \mid i \in S_b\}\right)$
end for
Initialise $\varepsilon = \{\}$
for $i = 1, \ldots, T$ **do**
  $\hat{f}^{\varphi}_{-i}(x_i) = \varphi\left(\left\{\hat{f}^b(x_i) \mid i \notin S_b\right\}\right)$
  Compute $\hat{\varepsilon}^{\varphi}_i = \left|y_i - \hat{f}^{\varphi}_{-i}(x_i)\right|$
  $\varepsilon = \varepsilon \cup \{\hat{\varepsilon}^{\varphi}_i\}$
end for
for $t = T+1, \ldots, T+T_1$ **do**
  Let $\hat{f}^{\varphi}_{-t}(x_t) = (1-\alpha)$ quantile of $\left\{\hat{f}^{\varphi}_{-i}(x_t)\right\}_{i=1}^{T}$ **EnbPI V2:** this is replaced by $\hat{f}^{\varphi}_{-t}(x_t) = \varphi\left(\left\{\hat{f}^{\varphi}_{-i}(x_t)\right\}_{i=1}^{T}\right)$.
  Let $w^{\varphi}_t = (1-\alpha)$ quantile of ε
  Return $C^{\varphi,\alpha}_{T,t}(x_t) = \left[\hat{f}^{\varphi}_{-t}(x_t) \pm w^{\varphi}_t\right]$
  if $t - T = 0 \mod s$ **then**
    for $j = t-1, \ldots, t-1$ **do**
      Compute $\hat{\varepsilon}^{\varphi}_j = \left|y_j - \hat{f}^{\varphi}_{-j}(x_t)\right|$
      $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon} - \{\hat{\varepsilon}^{\varphi}_1\}) \cup \{\hat{\varepsilon}^{\varphi}_i\}$ and reset index of $\boldsymbol{\varepsilon}$
    end for
  end if
end for

C.4.2. DETAILS ON THE IMPLEMENTATION

We conclude this section by summarizing computational aspects of the methods. One of the contributions is to provide a unified experimental framework. Therefore, in Table 1, we display the current available code for these methods, and what is available in the proposed repository.

Table 1. Summary of available code online for each method and the proposed code in the repository. The programming language is specified, and, when relevant, the nature of the code.

	Currently available		Contribution	
Methods	Language	Details	Language	Options
CP	R		Python	
OSCP	not available		Python	randomised split
EnbPI	Python		Python	same aggregation function
ACI	R script	no general function	Python	randomised split

D. Additional experiments on synthetic data sets

In this section, we provide supplemental results on the synthetic data sets presented in Section 5.1.

First, in Appendix D.1 the sensitivity analysis of ACI γ as well as the comparison to the naive strategy and AgACI is extended to AR(1) and MA(1) processes of asymptotic variance 10.

Then, in Appendix D.2, the comparison of all the CP methods for time series (initiated in Section 5.4) is also extended to these noises, that is AR(1) and MA(1) processes of asymptotic variance 10 (Appendix D.2.1), and to ARMA(1,1), AR(1) and MA(1) processes of asymptotic variance 1 (Appendix D.2.2).

Next, we discuss in Section 5.4 that the improved *validity* for $\gamma = 0.05$ in comparison to $\gamma = 0.01$ comes at the cost of more infinite intervals. This analysis is detailed in Appendix D.3.

Finally, we compare randomized and sequential split in Appendix D.4.

Imputation. The rationale to impute the infinite intervals is the following. We take the maximum of the absolute values of the residuals on the test set, noted $|\varepsilon|_{\max}$. Then, for any $t \in [\![T_0 + 1, T_0 + T_1]\!]$, if the predicted upper (resp. lower) bound $\hat{b}_t^{(u)}(x_t)$ is such that $\hat{b}_t(x_t) > \hat{\mu}_t(x_t) + |\varepsilon|_{\max}$ (resp. $\hat{b}_t^{(\ell)}(x_t) < \hat{\mu}_t(x_t) - |\varepsilon|_{\max}$) we impute it by $\hat{\mu}_t(x_t) + |\varepsilon|_{\max}$ (resp. $\hat{\mu}_t(x_t) - |\varepsilon|_{\max}$).

D.1. Additional experimental results of ACI sensitivity to γ, presented in Section 5.2

In this subsection, we provide similar results to those of Section 5.2, for different models on the noise. Especially, we consider AR(1) and MA(1) processes.

Observations. The behaviour of the AR(1) process is very similar to the one of ARMA(1,1). On the other hand, for the MA case, the dependence structure is too weak to observe a significant effect of γ. All ACI methods produce nearly valid intervals, with coverage above 89.25%.

Results are given in Figures 13 and 14.

D.2. Comparison to baselines, extension of Section 5.4

D.2.1. ASYMPTOTIC VARIANCE FIXED TO 10.

Figure 15 displays the results on data generated according to Section 5.1, for an asymptotic variance of the noise of 10 (as in Figure 6), when this noise is an AR(1) or MA(1) process.

Observations. As in the previous section, the methods' performances are greatly impacted by the type and strength of dependence structure. Figure 15 shows that while ARMA(1,1) and AR(1) noises lead to similar patterns, it is not the case for an MA(1) noise. In the latter, θ has little influence: the five performances (one for each θ) are similar within each method.

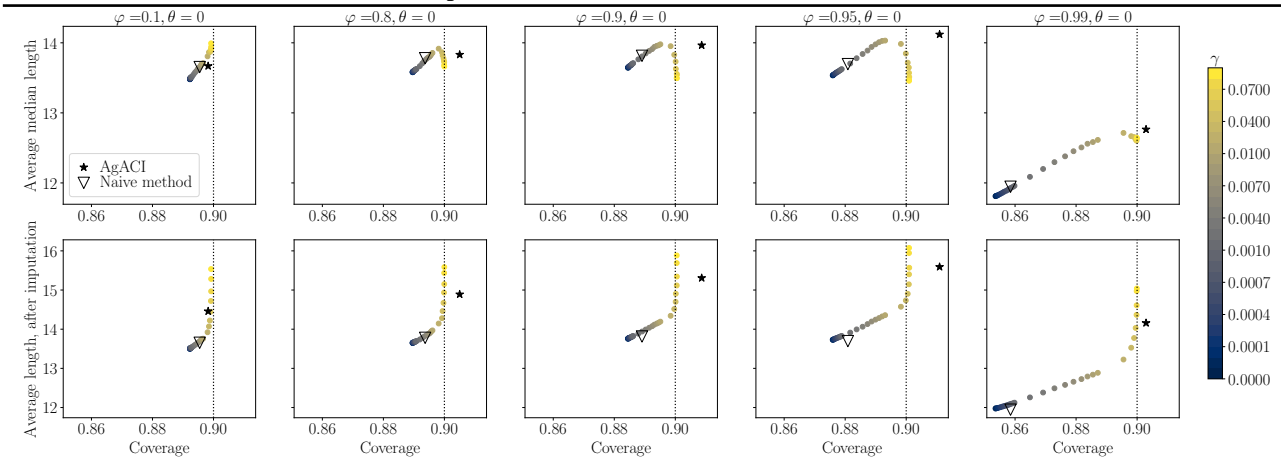


Figure 13. ACI performance with various θ, φ and γ on data simulated according to equation (3) with a Gaussian AR(1) noise of asymptotic variance 10 (see Appendix C.2). Top row: average median length with respect to the coverage. Bottom row: percentage of infinite intervals. Stars correspond to the proposed online expert aggregation strategy, and empty triangles to the naive choice.

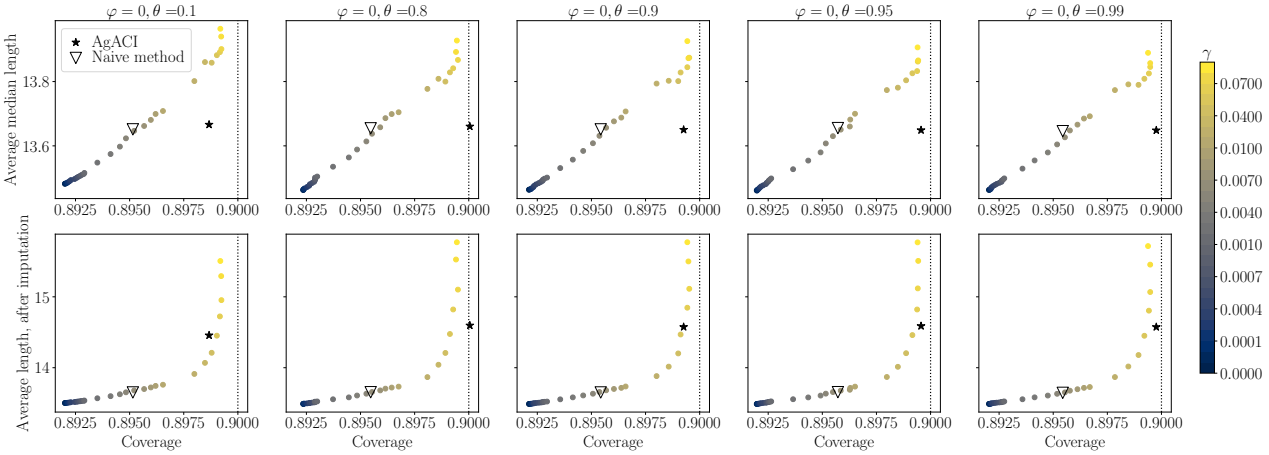


Figure 14. ACI performance with various θ, φ and γ on data simulated according to equation (3) with a Gaussian MA(1) noise of asymptotic variance 10 (see Appendix C.2). Top row: average median length with respect to the coverage. Bottom row: percentage of infinite intervals. Stars correspond to the proposed online expert aggregation strategy, and empty triangles to the naive choice.

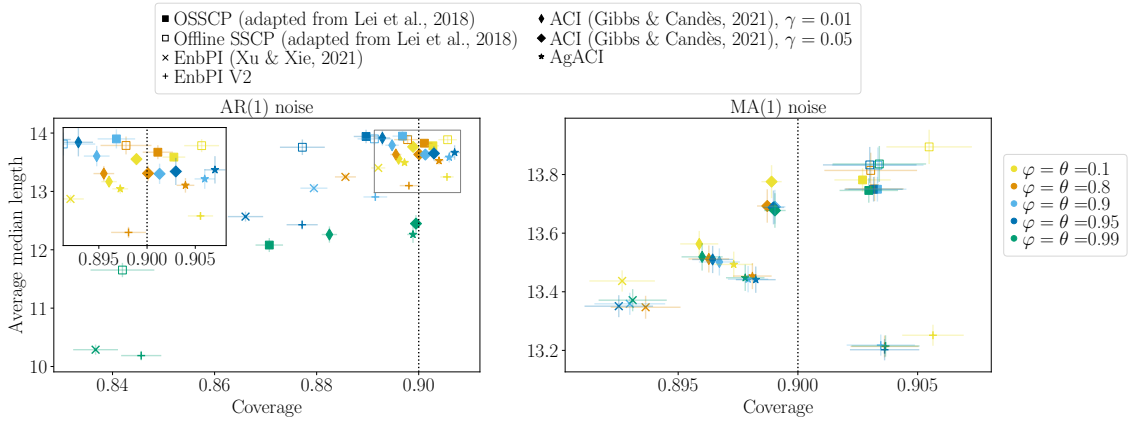


Figure 15. Performance of various interval prediction methods on data simulated according to equation (3) with a Gaussian AR(1) (left) and MA(1) (right) noise of asymptotic variance 10 (see Appendix C.2). Results aggregated from 500 independent runs. Empirical standard errors are displayed.

In addition, offline sequential SCP is very close to OSSCP. This is expected as a MA(1) process has very short memory, and the temporal dependence is thus small even for $\theta = 0.99$.

D.2.2. Asymptotic variance fixed to 1.

We now fix the asymptotic variance of the noise to 1. The results are plotted in Figure 16. Note that this is an easier setting than previously, as the signal to noise ratio is higher for this asymptotic variance.

Observations. Similarly to Figure 15, θ has little influence when the noise is a MA(1). On AR(1) and ARMA(1,1) noises (left and middle subplots), the patterns are similar. First, we observe again the improvement thanks to the online mode (empty squares versus solid ones), which increases when the dependence increases. Second, all the methods achieve *validity* or are significantly closer to achieving it than when the asymptotic variance is set to 10 (this is related to the high signal to noise ratio mentioned at the beginning of this section). Third, EnbPI V2 is *valid* for $\varphi = \theta \leq 0.95$ and provides the most *efficient* intervals for theses values. Nevertheless, its performances, as well as those of EnbPI, follow a clear trend (similar to that of Figure 6): when the dependence increases, the coverage decreases, as well as the length. EnbPI does not seem to be robust to the increasing temporal dependence in these experiments.

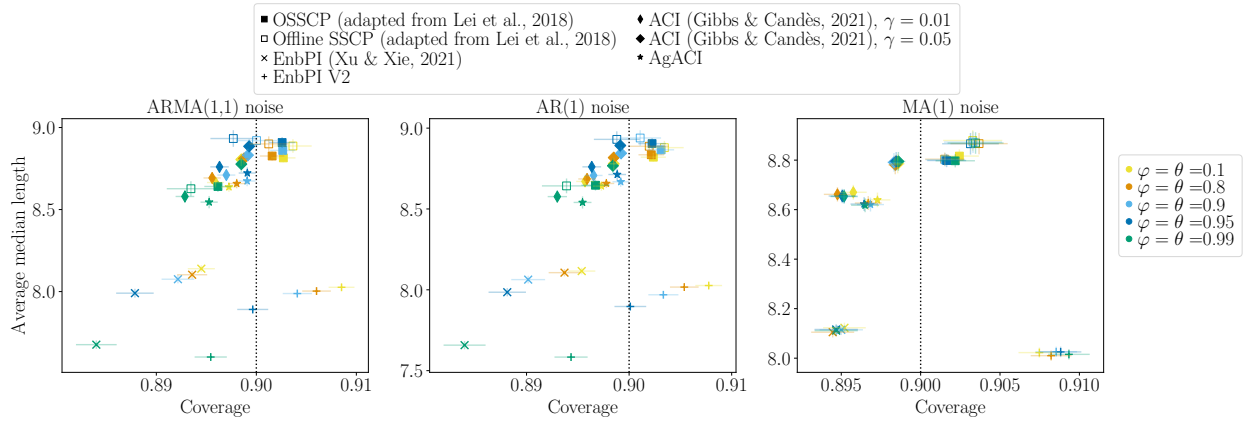


Figure 16. Performance of interval prediction methods on data simulated according to equation (3) with an ARMA(1,1) (left), AR(1) (center) and MA(1) (right) noise with a $\mathcal{N}(0, 1\frac{1-\varphi^2}{1-2\varphi\theta+\theta^2})$ *innovation*. Results aggregated from 500 independent runs. Empirical standard errors are displayed.

D.3. Closer look at infinite intervals

In this subsection, we investigate further the infite intervals generated by ACI for ARMA(1,1), AR(1) and MA(1) noise models. We report the results in Table 2. The central two columns present the percentage of infinite intervals, for $\gamma = 0.01$ and $\gamma = 0.05$. A first obvious observation is that the number of infinite intervals is orders of magnitude smaller for $\gamma = 0.01$ than for $\gamma = 0.05$. The last column represents the proportion of points for which $\gamma = 0.05$ predicts $\mathbb{R}$ and that are *not* covered for $\gamma = 0.01$. This suggests that for those intervals, predicting an infinite interval was somehow justified in the sense that the point was seemingly challenging to cover (as $\gamma = 0.01$ failed to cover). For example, in the first line ($\varphi = \theta = 0.1$) we read that there are 562 points that result in infinite intervals for $\gamma = 0.05$, among which 53 lead to finite predictions for $\gamma = 0.01$ failing to cover on that point. This means only 9.43 % of 562 infinite intervals that can be considered as "somehow justified". This analysis highlights that $\gamma = 0.05$ seem to predict more infinite

Table 2. Percentage of infinite intervals for ACI, on an ARMA(1,1) noise (first five rows), on an AR(1) noise ($\theta = 0$, next five rows) and a MA(1) noise ($\varphi = 0$, last five rows). The central two columns present the percentage of infinite intervals, for $\gamma = 0.01$ and $\gamma = 0.05$. The last column represents the proportion of points for which $\gamma = 0.05$ predicts $\mathbb{R}$ and that are *not* covered for $\gamma = 0.01$.

Noise parameters	$\gamma = 0.01$	$\gamma = 0.05$	Intersection	
$\varphi = \theta = 0.1$	0	1.12	53 out of 562	(9.43%)
$\varphi = \theta = 0.8$	0	2.76	263 out of 1381	(19.04%)
$\varphi = \theta = 0.9$	0	3.72	425 out of 1862	(22.83%)
$\varphi = \theta = 0.95$	0.03	4.45	514 out of 2224	(23.11%)
$\varphi = \theta = 0.99$	0.04	6.22	554 out of 3109	(17.82%)
$\varphi = 0.1$	0	1	37 out of 500	(7.40%)
$\varphi = 0.8$	0	2.75	212 out of 1373	(15.44%)
$\varphi = 0.9$	0	3.24	359 out of 1622	(22.13%)
$\varphi = 0.95$	0.03	4.32	488 out of 2160	(22.59%)
$\varphi = 0.99$	0.06	6.15	560 out of 3073	(18.22%)
$\theta = 0.1$	0	1.03	38 out of 516	(7.36%)
$\theta = 0.8$	0	1.42	49 out of 710	(6.90%)
$\theta = 0.9$	0	1.54	47 out of 772	(6.09%)
$\theta = 0.95$	0	1.54	45 out of 770	(5.84%)
$\theta = 0.99$	0	1.56	53 out of 781	(6.79%)

intervals than necessary, to compensate for easy errors as explained in Section 2.

D.4. Randomised, sequential and other splits.

In Figure 17, we compare the sequential split strategy (dark markers) used in our experiments to the randomised version (clear markers), on online SCP. We observe that the intervals produced by the randomised version are significantly smaller than the sequential one, while covering slightly less.

Another splitting strategy would consist in calibrating on the first points and training on the last ones. Up to our knowledge, this has not been used in practice. This way, we could hope to obtain a better model for the point prediction task. Nevertheless, we would be calibrating on really different data than the test ones. Thereby, the impact of this scheme regarding the interval prediction task performance is not straightforward. This is why we focus here on the sequential split, which is the most intuitive approach. Analysing further all of these effects theoretically or with extensive numerical experiments would be beneficial to the time series conformal prediction domain.

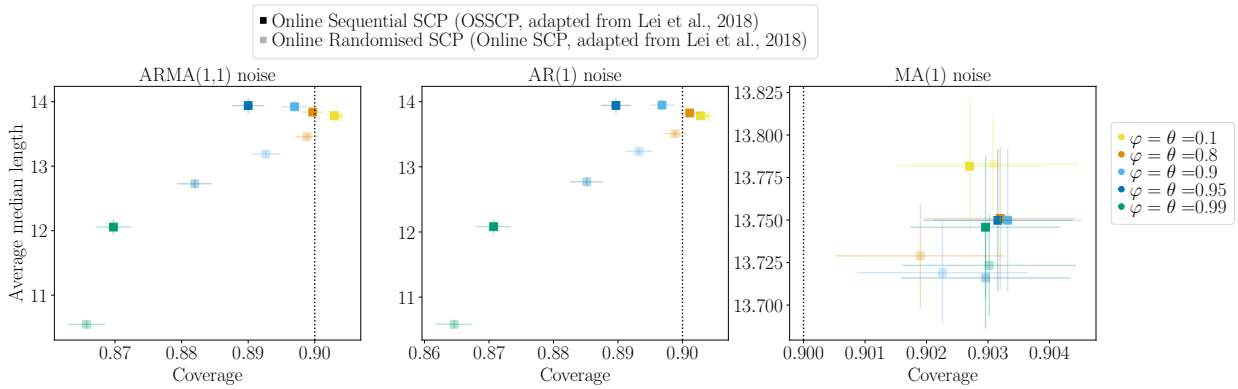


Figure 17. Performance of interval prediction methods on data simulated according to equation (3) with a Gaussian ARMA(1,1) (left), AR(1) (middle) and MA(1) (right) noise of asymptotic variance 10 (see Appendix C.2). Randomised methods are displayed. Results aggregated from 500 independent runs. Empirical standard errors are displayed.

E. Forecasting French electricity spot prices

E.1. Details about the data set

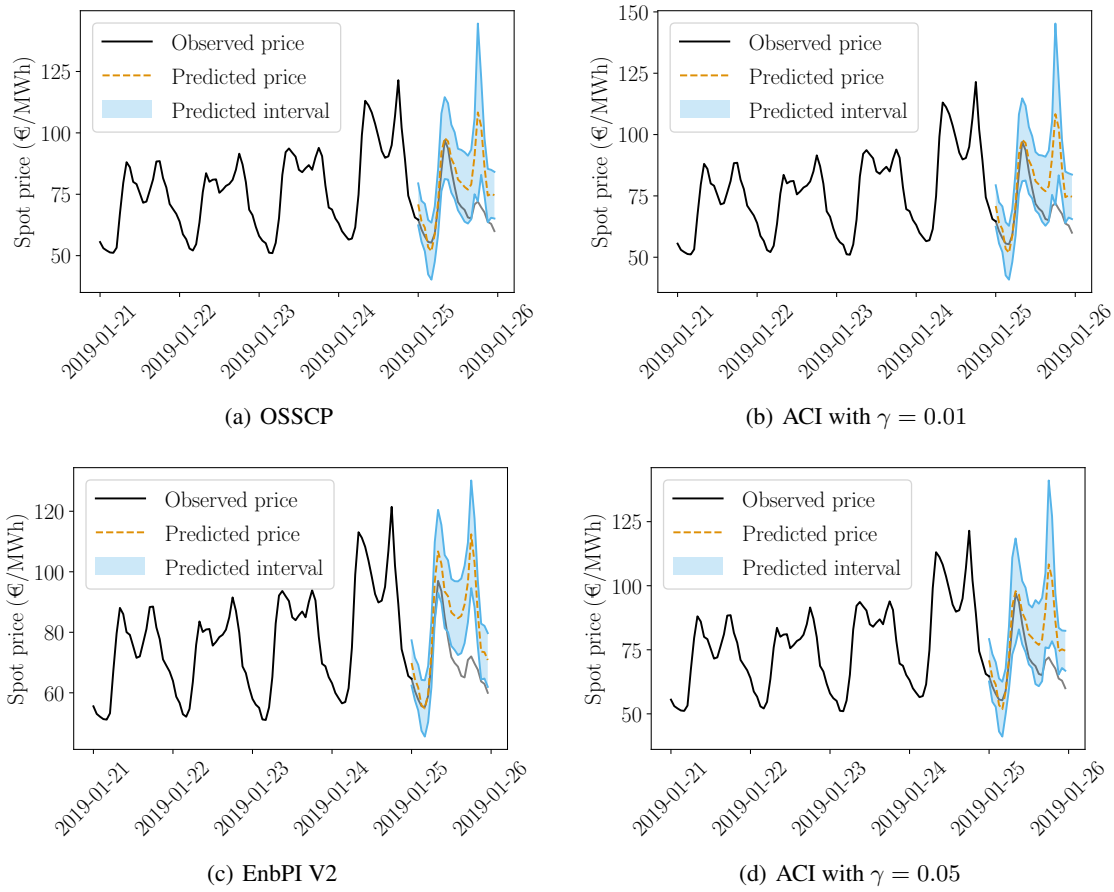
Table 3 presents an extract of the French electricity spot prices data set used in Section 6. In this table, 2×23 columns are hidden for clarity and space: the 24 prices of $D-7$ and the 24 prices of $D-7$ are used as variables.

Table 3. Extract of the built data set, for French electricity spot price forecasting.

Date and time	Price	Price D-1	Price D-7	For. cons.	DOW
11/01/16 0PM	21.95	15.58	13.78	58800	Monday
11/01/16 1PM	20.04	19.05	13.44	57600	Monday
⋮	⋮	⋮	⋮	⋮	⋮
12/01/16 0PM	21.51	21.95	25.03	61600	Tuesday
12/01/16 1PM	19.81	20.04	24.42	59800	Tuesday
⋮	⋮	⋮	⋮	⋮	⋮
18/01/16 0PM	38.14	37.86	21.95	70400	Monday
18/01/16 1PM	35.66	34.60	20.04	69500	Monday
⋮	⋮	⋮	⋮	⋮	⋮

E.2. Forecasting year 2019

In Figure 18 we observe that on January 25, 2019, the forecasts are very different from the actual values. Nevertheless, the prediction intervals manage to include these observations for almost all hours (except after 5 pm) and almost all methods (EnbPI does not include points earlier, starting at 11 am).

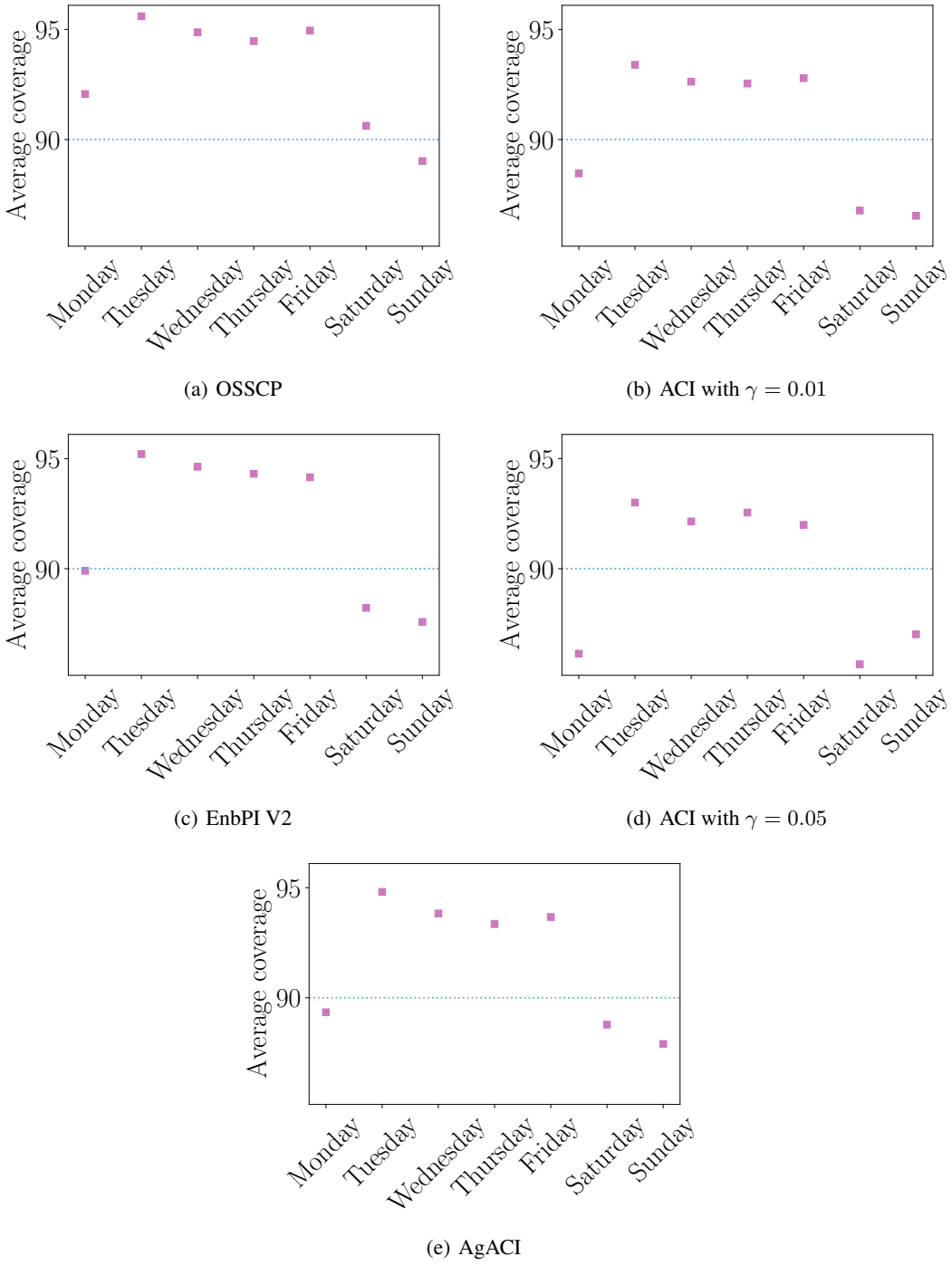


(a) OSSCP (b) ACI with $\gamma = 0.01$

(c) EnbPI V2 (d) ACI with $\gamma = 0.05$

Figure 18. Representation of predicted intervals around point forecasts on the 25th of January of 2019.

In Figure 19 we observe that the four algorithms suffer from an unbalanced coverage depending on the day-of-the-week (each algorithm in a different extent). That is, they cover more than 90% of the observations on Tuesdays to Fridays, but less than 90% on Mondays and week-ends (Saturdays and Sundays).



(a) OSSCP

(b) ACI with $\gamma = 0.01$

(c) EnbPI V2

(d) ACI with $\gamma = 0.05$

(e) AgACI

Figure 19. Coverage proportion during 2019 depending on the day-of-the-week.