Adaptive Conformal Predictions for Time Series

Margaux Zaffran[1,2,3] Olivier Féron[1] Yannig Goude[1] Julie Josse[2] Aymeric Dieuleveut[3]

[1]Electricité De France, Paris, France [2]INRIA, Montpellier, France [3]Ecole Polytechnique, Paris, France

Paper ⟶
5' talk ⟶
45' talk ⟶
Code ⟶

Control validity

Produce **predictive intervals** around forecasts, enjoying **theoretical guarantees** on their **coverage** with **few assumptions**.

Optimize efficiency

The intervals should be **as small as possible**.

Bad example: outputting $\begin{cases} \mathbb{R} \text{ 90\% of the time} \\ \emptyset \text{ 10\% of the time} \end{cases}$ is valid but **useless!**

Setting in time series

- **Data:** T_0 observations $(x_1, y_1), \dots, (x_{T_0}, y_{T_0})$ in $\mathbb{R}^d \times \mathbb{R}$.
- **Aim:** predict for T_1 subsequent observations $x_{T_0+1}, \dots, x_{T_0+T_1}$.

$\hookrightarrow$ Build the smallest interval $\hat{C}_\alpha^t$ such that:

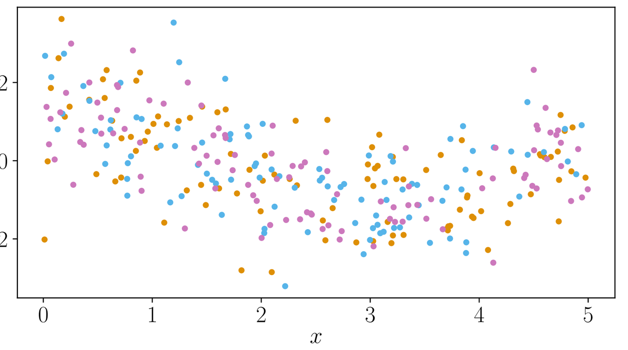
$$\mathbb{P}\left\{Y_t \in \hat{C}_\alpha^t(X_t)\right\} \geq 1-\alpha, \quad \text{for } t \in [\![T_0+1, T_0+T_1]\!].$$

Summary

Conformal prediction gives predictive intervals under exchangeability, not time series. **ACI can be used** but require a learning rate γ.

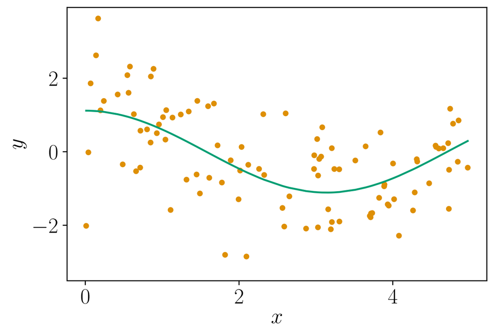
1. **Theory** on ACI's efficiency depending on the learning rate γ.
2. **Algorithm** based on expert aggregation, to avoid choosing γ.
3. **Numerical tests**: synthetic and French electricity prices.

Split Conformal Prediction (Vovk et al., 2005)



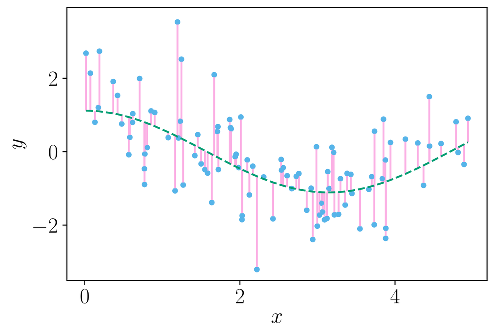
Randomly split the data to obtain a proper training set and a calibration set. Keep the test set.

Step 1



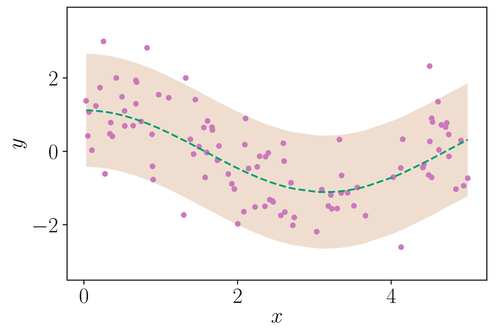
▶ Learn $\hat{\mu}$.

Step 2



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

Step 3



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

- Given any regression function $\hat{\mu}$
- For any sample size n (finite-sample)
- If the (X_i, Y_i) are exchangeable

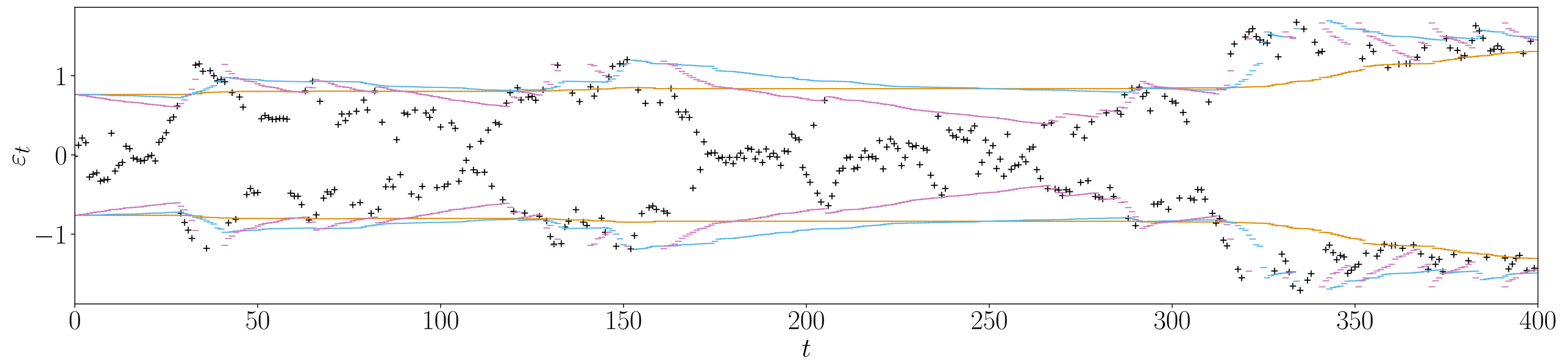
$$\mathbb{P}\left(Y \in \hat{C}_{\hat{\alpha}}(X)\right) \geq 1-\alpha$$

$\hookrightarrow$ what is essential is that the **scores** $\{s_i\}_i$ are exchangeable.

Adaptive Conformal Inference (ACI, Gibbs and Candès, 2021)

Use an **effective quantile level** based on a recursive equation and a **learning rate** γ: $\alpha_{t+1} := \alpha_t + \underbrace{\gamma}_{\geq 0}\left(\alpha - \mathbb{1}\left\{y_t \notin \hat{C}_{\alpha_t}(x_t)\right\}\right)$.

Illustration: ACI with $\gamma = 0$, $\gamma = 0.01$ and $\gamma = 0.05$.



Theory

For any distribution:

$$\left|\underbrace{\frac{1}{T_1}\sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{y_t \in \hat{C}_{\alpha_t}(x_t)\right\}}_{\text{Average coverage}} - (1-\alpha)\right| \leq \frac{2}{\gamma T_1}$$

❶ Impact of the learning rate γ

Exchangeable case

Theorem 1 (informal)

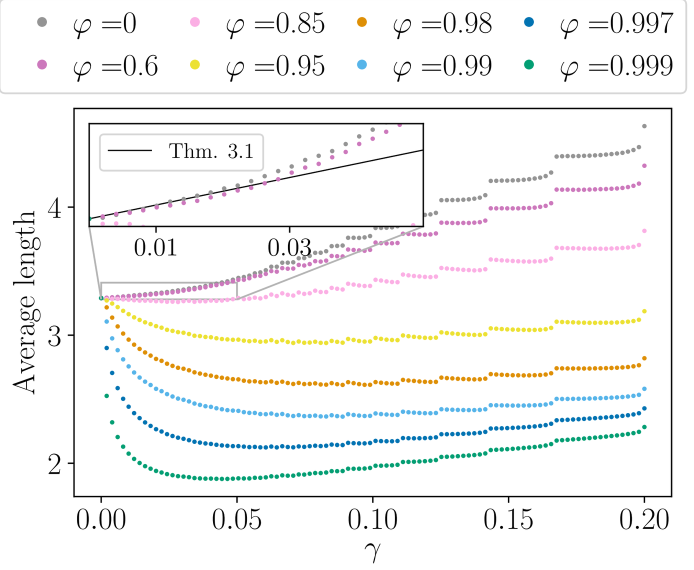
Assume exchangeable scores and perfect calibration. As $\gamma \to 0$:

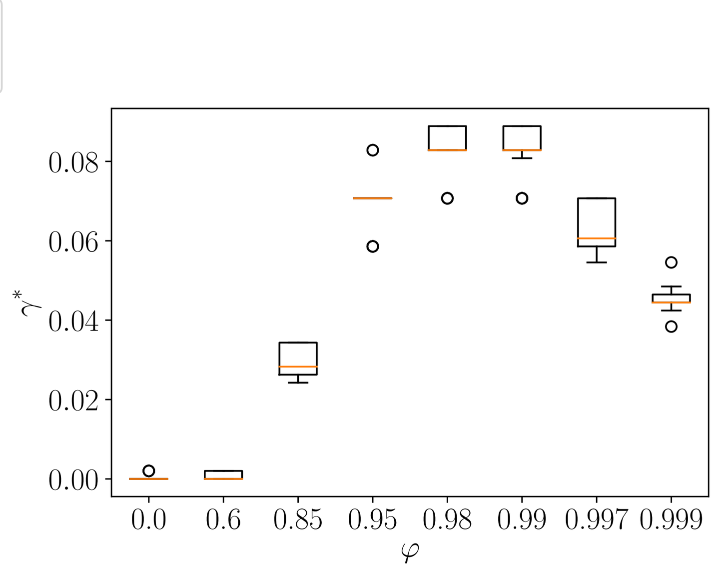
Average length of intervals from ACI using γ
= Average length of intervals from Split Conformal Prediction
$+ \gamma \times \underbrace{\mathcal{C}(\alpha, \text{distribution of the data})}_{>0 \text{ in general}}$.

Auto-regressive case: $\varepsilon_{t+1} = \varphi\varepsilon_t + \xi_{t+1}$.

Theorem 2 (informal)

Assume auto-regressive residuals and perfect calibration. There exists an optimal $\gamma^* > 0$ minimizing the average length for high φ.

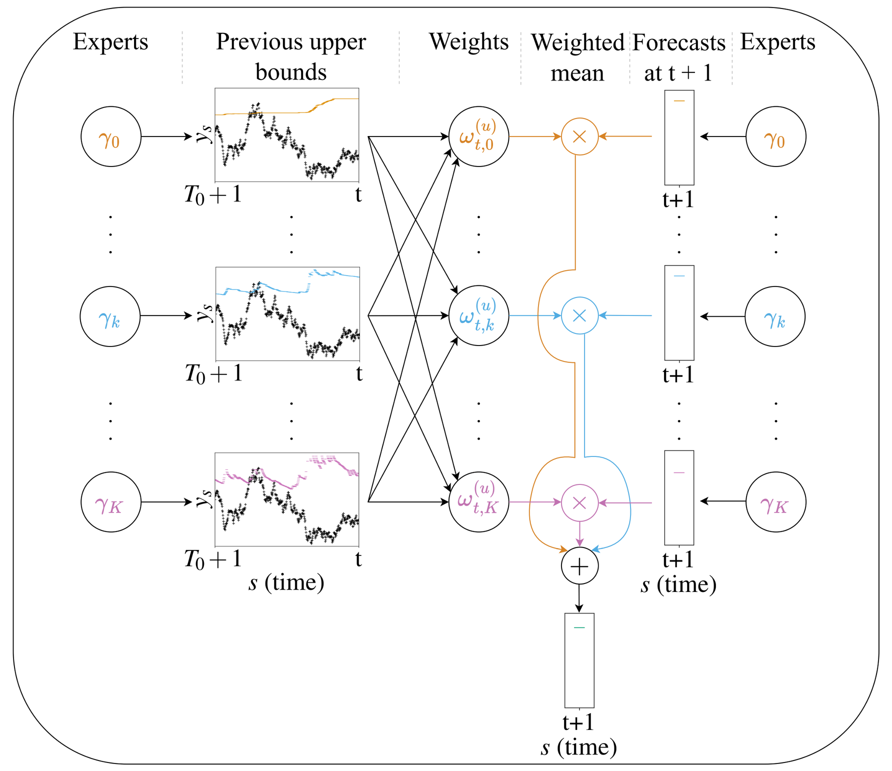




Conclusion: choosing γ is crucial but difficult.

❷ AgACI

- Experts (ACI with many γ) aggregation (Cesa-Bianchi and Lugosi, 2006).
- One algorithm for the upper bound, another for the lower bound.
- Based on the pinball loss of level $1 - \frac{\alpha}{2}$, or of level $\frac{\alpha}{2}$.



❸ Numerical results

$$Y_t = 10\sin(\pi X_{t,1}X_{t,2}) + 20(X_{t,3} - 0.5)^2 + 10X_{t,4} + 5X_{t,5} + \varepsilon_t$$

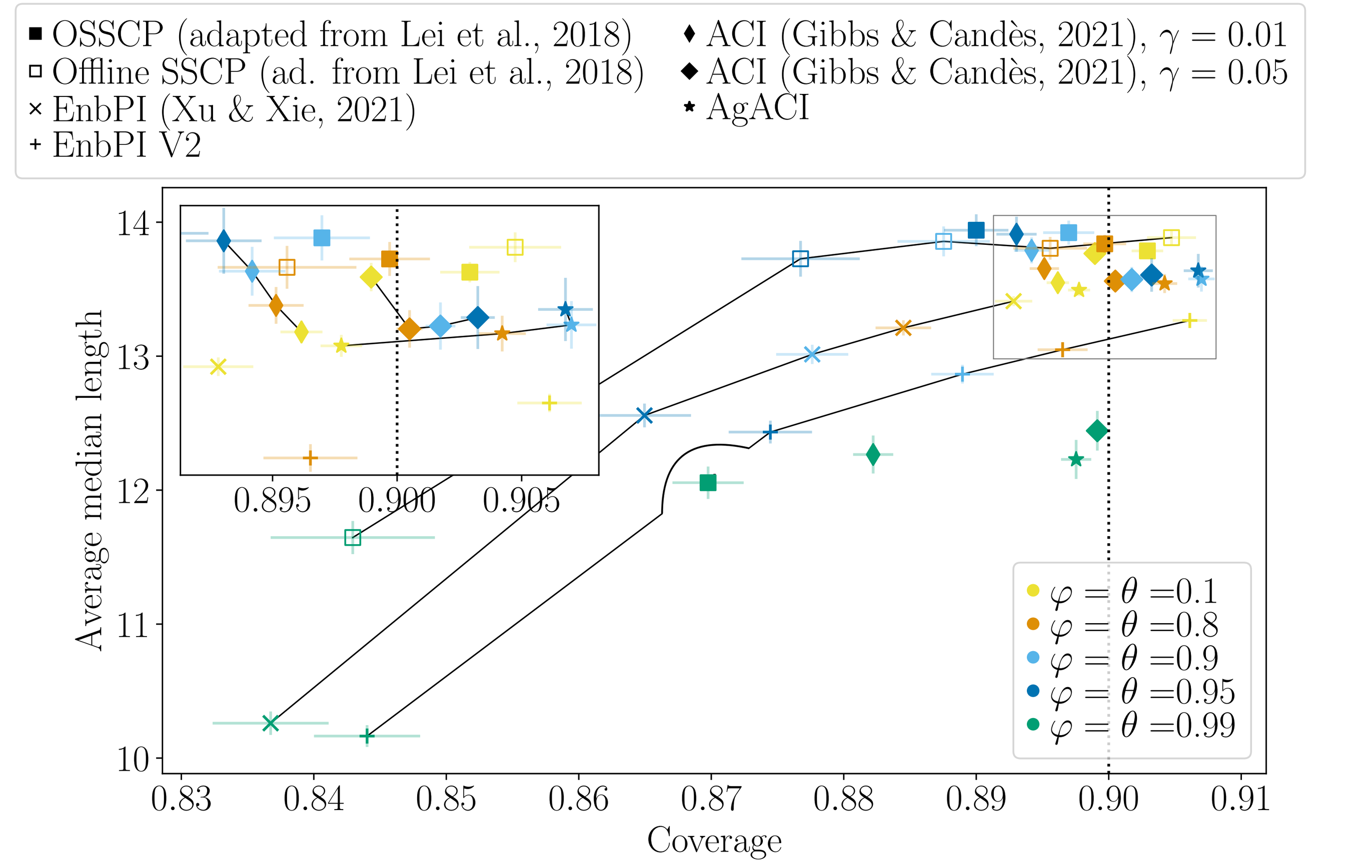
with $X_{t,\cdot} \sim \mathcal{U}([0,1])$ and ε_t an ARMA(1,1) process:

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

with ξ_t is a white noise of variance σ^2.

- $\varphi = \theta$ range in $[0.1, 0.8, 0.9, 0.95, 0.99]$.
- σ is fixed to keep the variance $\text{Var}(\varepsilon_t)$ constant to 10.
- Random forest are used as regressor.
- For each setting (pair variance and φ,θ):
  - 300 points, the last 100 kept for prediction and evaluation,
  - 500 repetitions,

$\Rightarrow$ in total, $100 \times 500 = 50000$ predictions are evaluated.



- Increasing the temporal dependence impacts benchmarks validity.
- ACI is robust and maintains validity for some well-chosen γ.
- AgACI is robust and maintains validity without choosing γ.

Open directions

Theory on AgACI: is it asymptotically valid? Efficient?

Main references

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games.*
Gibbs, I. and Candès, E. (2021). Adaptive Conformal Inference Under Distribution Shift. *NeurIPS.*
Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World.*