Transductive conformal inference with adaptive scores

Ulysse Gazin
LPSM, Université Paris cité, FR
ulysse.gazin@lpsm.paris

Gilles Blanchard
IMO, Université Paris-Saclay, FR
gilles.blanchard@universite-paris-saclay.fr

Etienne Roquain
LPSM, Sorbonne Université, FR
etienne.roquain@upmc.fr

Context

- Reliable machine learning: guarantees for blackboxes
- Large scale inference problem: simultaneous guarantees

Conformal Inference

- Distribution free: use only exchangeability
- Finite sample: true for any sample size

Two tasks:

- Prediction intervals with controlled coverage
- Novelty detection with controlled errors

Classical approach: Expectation guarantees of errors

Aim: Uniform in probability guarantees.

Split Conformal Prediction

Nonparametric regression.

- Data: X_i and Y_i resp. feature and response of unit i
- Regression function: $\mu(x) = \mathbb{E}[Y_i | X_i = x]$
- Good predictor: $\widehat{\mu}$

How close is $\widehat{\mu}(X_{n+1})$ to Y_{n+1} ?

- Training $\mathcal{D}_{\text{train}}$
- Non-Conformity scores $\widehat{S}(x,y) = \widehat{g}(x,y;\mathcal{D}_{\text{train}}) \in \mathbb{R}$
  For example, $\widehat{S}(x,y) = \|\widehat{\mu}(x) - y\|$.
- Calibration $\mathcal{D}_{\text{cal}} = (X_1, Y_1), \dots, (X_n, Y_n)$
- Test $\mathcal{D}_{\text{test}} = (X_{n+1}, Y_{n+1}), \cdots, (X_{n+m}, Y_{n+m})$, only the X are observed.

Split Conformal inference [Papadopoulos et al., 2002]

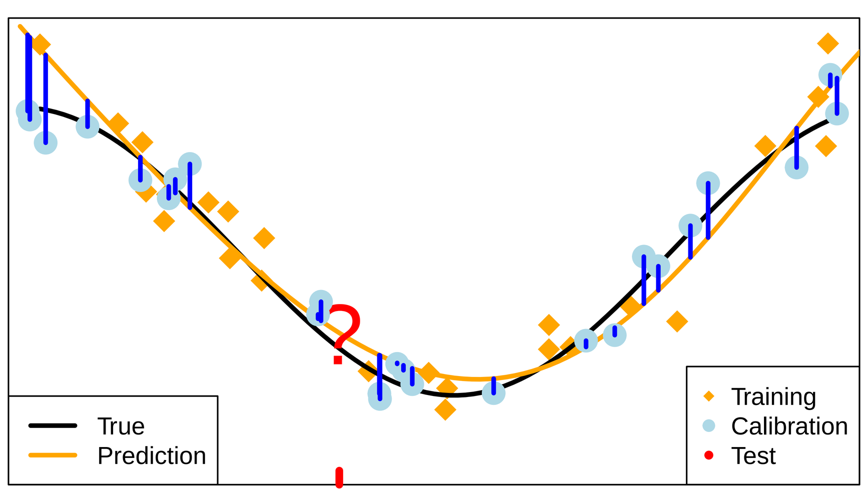
Conformal p-values

- Observed scores $S_k = \widehat{S}(X_k, Y_k)$, $1 \le k \le n$
- For all $y \in \mathbb{R}$ and $i \in [\![m]\!]$:

$$p_i^{(y)} = \frac{1 + \sum_{k \in [\![n]\!]} \mathbb{1}_{S_k \ge S(X_{n+i}, y)}}{n+1}$$

- Prediction interval $\mathcal{C}_i(\alpha) = \mathcal{C}_\alpha(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{cal}}, X_{n+i})$ for Y_{n+i}

$$\mathcal{C}_i(\alpha) = \left\{ y \in \mathbb{R};\ p_i^{(y)} > \alpha \right\}$$



Assumption (A)

$\left(\widehat{S}(X,Y)\right)_{(X,Y) \in \mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}}$ **exchangeable** and **no ties** a.s.

Under Assumption (A), $(n+1)p_1 \sim \text{Unif}([\![n+1]\!])$.

Theorem [Papadopoulos et al., 2002]

Under Assumption (A), $\mathbb{P}(Y_{n+i} \notin \mathcal{C}_i(\alpha)) \le \alpha$.

Joint law of conformal p-values ? Try with i.i.d. scores

Focus on $P_{n,m}$

$P_{n,m}$ is the distribution of the colors of m draws in a Pòlya urn model with $n+1$ colors labelled $\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\}$.
Let $(U_1, \cdots, U_n)$ i.i.d. Unif(0,1).
Sort $0 = U_{(0)} < U_1 < \cdots < U_{(n)} < U_{(n+1)} = 1$ and define the discrete distribution P^U on the set $\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\}$, with

$$P^U(\{\ell/(n+1)\}) = U_{(\ell)} - U_{(\ell-1)}.$$

Draw $(q_1, \dots, q_m | U) \overset{\text{i.i.d.}}{\sim} P^U$ then

$$P_{n,m} = \mathcal{D}\big(q_i, i \in [\![m]\!]\big).$$

Transductive Prediction

Construct $\mathcal{C}_\alpha = (\mathcal{C}_i(\alpha))_{i \in [\![m]\!]}$ for m test points.

False Coverage Proportion :

$$\begin{aligned}\text{FCP}(\mathcal{C}_\alpha) &= m^{-1} \sum_{i \in [\![m]\!]} \mathbb{1}_{Y_{n+i} \notin \mathcal{C}_i(\alpha)} \\ &= m^{-1} \sum_{i \in [\![m]\!]} \mathbb{1}_{p_i^{(Y_{n+i})} \le \alpha} = \widehat{F}_m(\alpha)\end{aligned}$$

Known results under Assumption (A):

- [Marginal Control] For all α, $\mathbb{E}\left(\widehat{F}_m(\alpha)\right) \le \alpha$
- [Vovk, 2013] FWER control: $\mathbb{P}\big(\text{FCP}(\mathcal{C}_{\alpha/m}) > 0\big) \le \alpha$,
- [Marques F., 2023] : for all t, $\widehat{F}_m(t)$ follows a Pòlya urn distribution.

Contribution

- **Joint law of conformal p-values,**
- **Precise (in probability) control of the FCP for simultaneous inference,**
- **Uniform error bounds for data-driven threshold,**
- **Use of adaptive scores.**

Joint Law and Concentration

General results on conformal p-values.

Theorem Joint Law

Under Assumption (A), $\left(p_1^{(Y_{n+1})}, \cdots, p_m^{(Y_{n+m})}\right) \sim P_{n,m}$.
p-values law free of scores underlying distribution.

Theorem DKW Inequality

Under Assumption (A), for all $\lambda > 0$, $n, m \ge 1$,

$$\mathbb{P}\left(\sup_{t \in [0,1]} \big(\widehat{F}_m(t) - I_n(t)\big) > \lambda\right) \le \left[1 + \frac{2\sqrt{2\pi}\lambda\tau_{n,m}}{(n+m)^{1/2}}\right] e^{-2\tau_{n,m}\lambda^2}$$

where $\tau_{n,m} := nm(n+m)^{-1}$ "effective sample size", and $I_n(t) = \frac{\lfloor (n+1)t \rfloor}{n+1}$.

Since Theorems holds for any exchangeable scores, one can use $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$ in an exchangeable way to create adaptive scores, as in [Marandon et al., 2022] for novelty detection.

Application 1: Adaptive Scores

Adaptive and Exchangeable Scores

- Semi-supervised:

$$\widehat{S}(x,y) = \widehat{S}\left(x, y; \mathcal{D}_{\text{train}}; \mathcal{D}^X_{\text{cal+test}}\right);$$

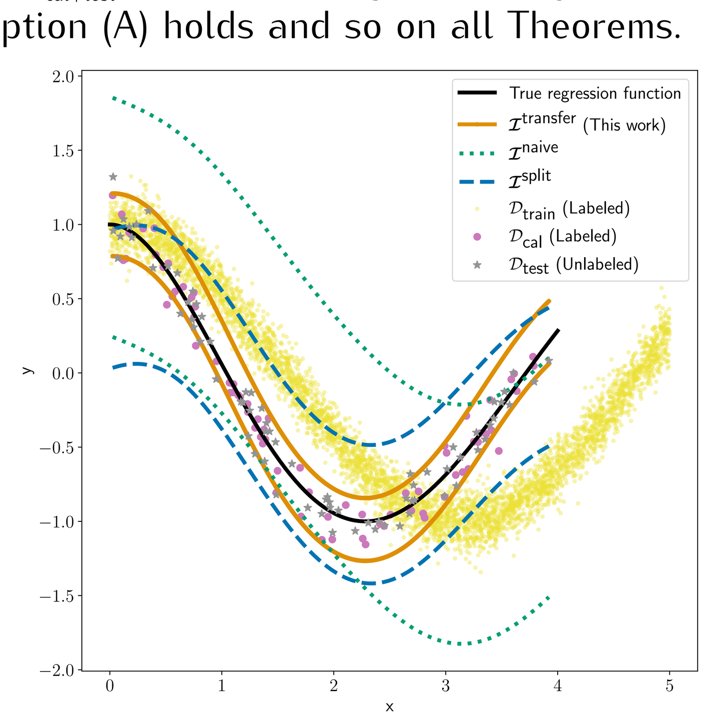
- $\widehat{S}$ permutation invariant w.r.t. $(X_1, \dots, X_{n+m})$;
- Assumption (A) holds if $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$ i.i.d.

only exchangeable even with i.i.d. data

Transfer Learning, Domain Adaptation

- $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$ i.i.d. with distribution P,
- $\mathcal{D}_{\text{train}}$ i.i.d. with distribution $Q \neq P$.

Use of $\mathcal{D}^X_{\text{cal+test}}$ in an exchangeable way for "transfer" and Assumption (A) holds and so on all Theorems.



Application 2: Data Driven Level

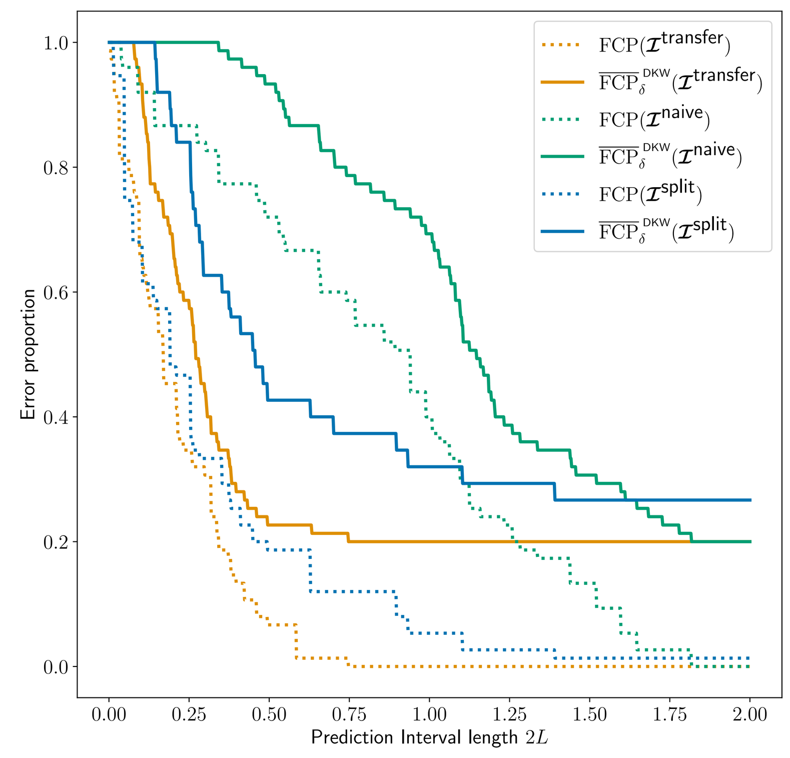
Corollary GBR

Let $\widehat{\alpha} \in (0,1)$ a random prediction level.
Under Assumption (A), with probability at least $1-\delta$,

$$\text{FCP}(\mathcal{C}_{\widehat{\alpha}}) \le \widehat{\alpha} + \lambda^{\text{DKW}}_{\delta,n,m} = \overline{\text{FCP}}^{\text{DKW}}_{\delta}$$

With $\lambda^{\text{DKW}}_{\delta,n,r}$ inverse function of the RHS of above DKW-type inequality.

Example: $\widehat{\alpha} = \inf\{\alpha \in (0,1), \forall i \in [\![m]\!], \text{Length}[\mathcal{C}_i(\alpha)] \le 2L\}$.



Application 3: Novelty Detection

Novelty detection setting [Marandon et al., 2022]

- $\mathcal{D}_{\text{train}}$ i.i.d. law P_0,
- $\mathcal{D}_{\text{cal}}$ i.i.d. law P_0,
- $\mathcal{D}_{\text{test}}$ independent either distributed as P_0 or not,
- $\text{FDP}(\mathcal{R})$ ="proportion of incorrect detection in $\mathcal{R}$".

Compute novelty scores $\widehat{S}(Z_i)$ for $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$, compute p-values and consider thresholding procedures:

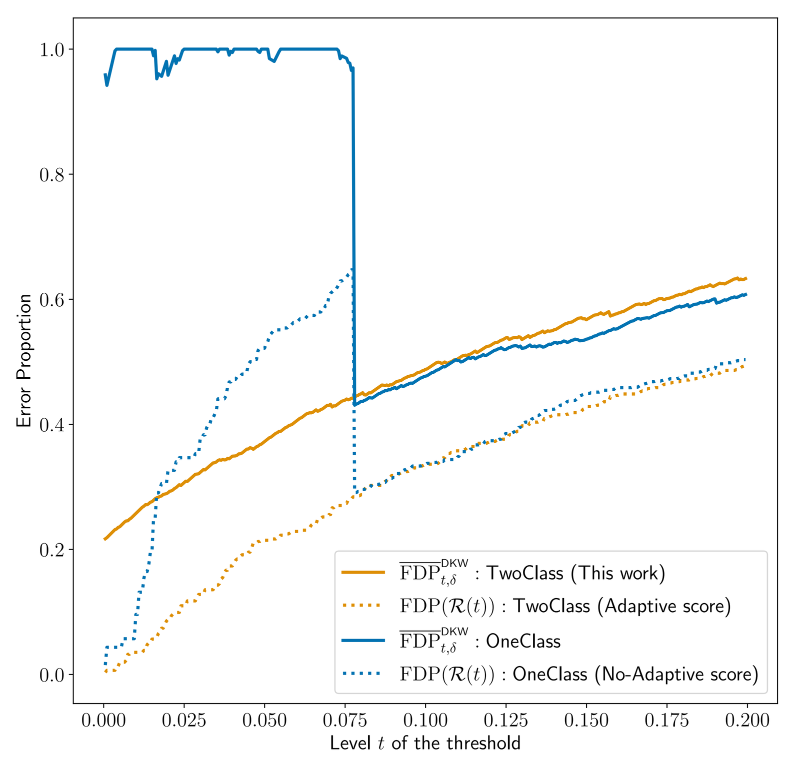
$$\mathcal{R}(t) = \{i \in [\![m]\!], p_i \le t\},\ t \in (0,1)$$

Corollary Uniform Bound

Under Assumption (A) for $\mathcal{D}_{\text{cal}} \cup \{Z_{n+i}, i \in \mathcal{H}_0\}$, with probability at least $1-\delta$ we have for all $t \in (0,1)$,

$$\text{FDP}(\mathcal{R}(t)) \le \frac{\widehat{m}_0 I_n(t) + \max_{r \in [\![\widehat{m}_0]\!]}\{r\lambda^{\text{DKW}}_{\delta,n,r}\}}{1 \vee |\mathcal{R}(t)|} = \overline{\text{FDP}}^{\text{DKW}}_{t,\delta}$$

With $\widehat{m}_0$ a specific estimator of $|\mathcal{H}_0|$ and $\lambda^{\text{DKW}}_{\delta,n,r}$ inverse function of the RHS of above DKW-type inequality.



References

[Courty et al., 2017] Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. *NIPS 2018.*

[Marandon et al., 2022] Marandon, A., Lei, L., Mary, D., and Roquain, E. (2022). Machine learning meets false discovery rate. *Annals of Statistics.*

[Marques F., 2023] Marques F., P. C. (2023). On the universal distribution of the coverage in split conformal prediction. *arXiv preprint.*

[Papadopoulos et al., 2002] Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. *ECML 2002.*

[Vovk, 2013] Vovk, V. (2013). Transductive conformal predictors. *AIAI 2013.*