Scalable and adaptive prediction bands with kernel sum-of-squares

ETICS 2025

Louis Allain

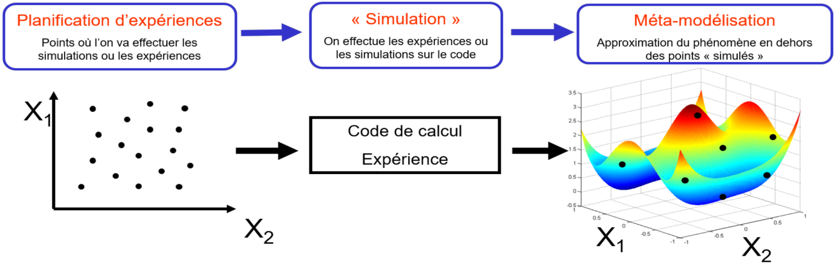
October 9, 2025

Thesis advisor : Sébastien Da Veiga
Supervisor : Brian Staber

Table of Contents

Industrial challenges

- Widespread use of supervised learning for computer experiments, where expensive simulation outputs are approximated with a ML model from a DoE dataset



In practice

- Strategy adopted across multiple industries, using various ML models:
  - Linear/logistic regression
  - Random forests
  - Gaussian processes
  - Neural networks...

- Critical applications require **confidence intervals around predictions**, with guaranteed coverage:
  - Denote $\widehat{C}(X)$ a confidence interval for a prediction at X, estimated from training data
  - The guarantee of **marginal coverage** at level α writes

$$\mathbb{P}(Y_{N+1} \in \widehat{C}(X_{N+1})) \geq 1 - \alpha$$

for the true unknown value of the output Y_{N+1} at an unobserved point X_{N+1}

In practice

- Limitations of traditional approaches:
  - Prediction bands are model-specific, with significant variation between models
  - Guarantees only valid as $n \to +\infty$ or under strong assumptions that cannot be verified
  - **No coverage guarantee** for practical applications

- A recent promising candidate: conformal prediction

Conformal Prediction

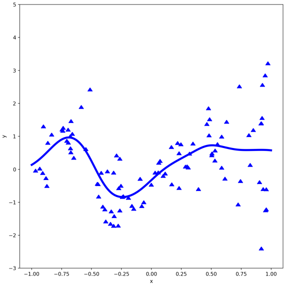
- Conformal Prediction (CP): a rigorous method to construct prediction intervals with the following properties:
  - ✓ Coverage guarantees
  - ✓ Finite sample
  - ✓ Distribution free
  - ✓ Model agnostic

- Several variants:
  - — Full CP
  - — Split CP
  - — Resampling strategies, e.g. jackknife+, CV+

Let us illustrate split CP, which is based on two independent datasets $\mathcal{D}_n$ (pre-training set) and $\mathcal{D}_m$ (calibration set)

Conformal Prediction

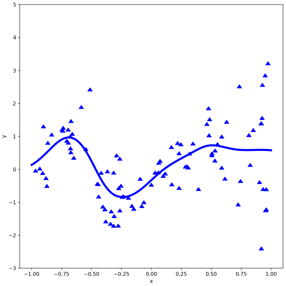
- The prediction model $\widehat{m}$ is trained on $\mathcal{D}_n$



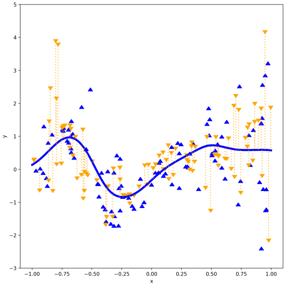
(a) Learn $\widehat{m}$ on $\mathcal{D}_n$

Conformal Prediction

- The prediction model $\widehat{m}$ is trained on $\mathcal{D}_n$
- $\mathcal{D}_m$ is used to evaluate some prediction quality of $\widehat{m}$, here for example the absolute residuals
- The quantile $\widehat{q}_\alpha$ of these quality measures is computed



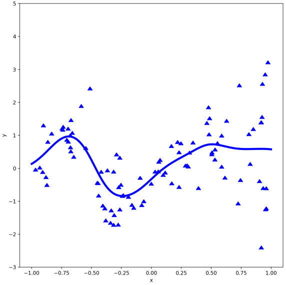
(a) Learn $\widehat{m}$ on $\mathcal{D}_n$



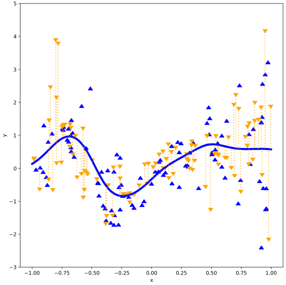
(b) Compute $|Y_i - \widehat{m}(X_i)|$ on $\mathcal{D}_m$

Conformal Prediction

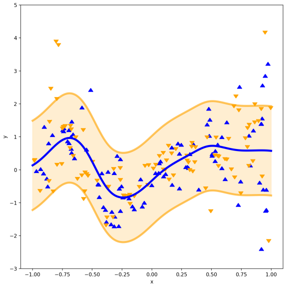
- The prediction model $\widehat{m}$ is trained on $\mathcal{D}_n$
- $\mathcal{D}_m$ is used to evaluate some prediction quality of $\widehat{m}$, here for example the absolute residuals
- The quantile $\widehat{q}_\alpha$ of these quality measures is computed
- The prediction interval $\widehat{C}(X) = [\widehat{m}(X) \pm \widehat{q}_\alpha]$ satisfies all the desired properties under the assumption of data exchangeability



(a) Learn $\widehat{m}$ on $\mathcal{D}_n$



(b) Compute $|Y_i - \widehat{m}(X_i)|$ on $\mathcal{D}_m$

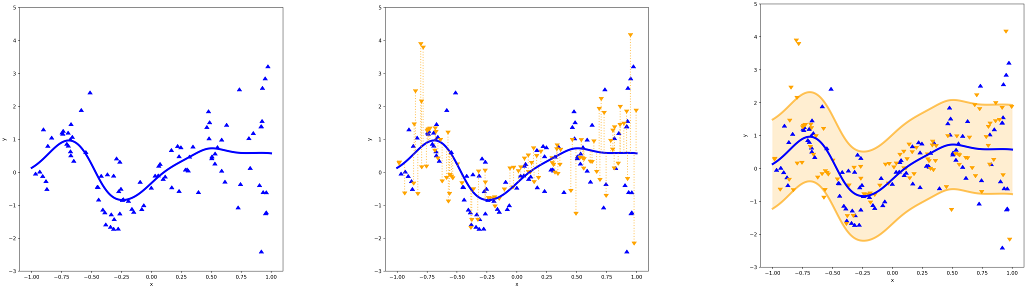


(c) $\widehat{C}(X) = [\widehat{m}(X) \pm \widehat{q}_\alpha]$

Scores

- Such evaluation of the prediction quality is performed by a **score function** s

	Absolute errors
$s(X_i, Y_i)$	$\lvert Y_i - \widehat{m}(X_i) \rvert$



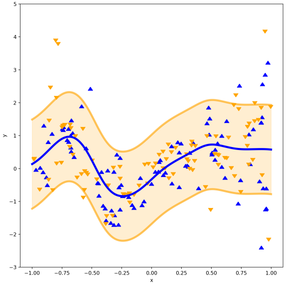
(a) Learn $\widehat{m}$ on $\mathcal{D}_n$ (b) Compute $\lvert Y_i - \widehat{m}(X_i)\rvert$ on $\mathcal{D}_m$ (c) $\widehat{C}(X) = [\widehat{m}(X) \pm \widehat{q}_\alpha]$

Scores

- Such evaluation of the prediction quality is performed by a **score function** s

	Absolute errors
$s(X_i, Y_i)$	$\lvert Y_i - \widehat{m}(X_i) \rvert$

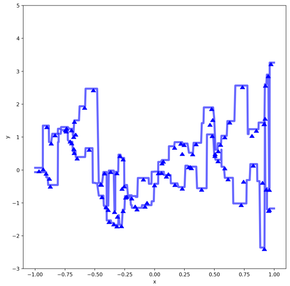
But you may have noticed that choosing the absolute errors leads to prediction intervals with **constant width**



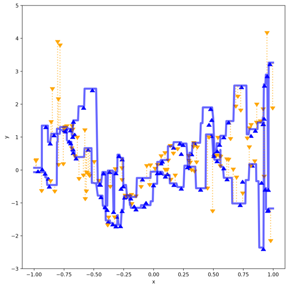
(c) $\widehat{C}(X) = [\widehat{m}(X) \pm \widehat{q}_\alpha]$

Towards adaptivity

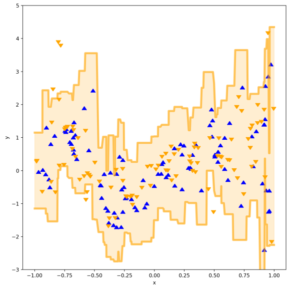
	Absolute errors	Quantile regression
$s(X_i, Y_i)$	$\lvert Y_i - \widehat{m}(X_i) \rvert$	$\max(\widehat{q}_{\mathrm{l}}(X_i) - Y_i, Y_i - \widehat{q}_{\mathrm{u}}(X_i))$



(a) Learn $\widehat{q}_{\mathrm{l}}, \widehat{q}_{\mathrm{u}}$ on $\mathcal{D}_n$



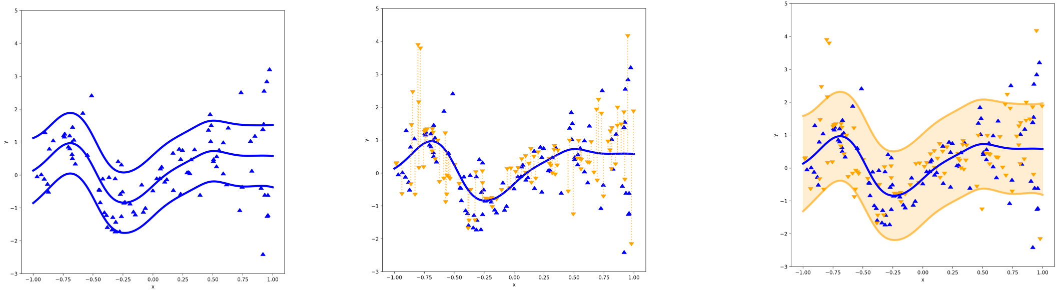
(b) Compute $\max(\widehat{q}_{\mathrm{l}}(X_i) - Y_i, Y_i - \widehat{q}_{\mathrm{u}}(X_i))$ on $\mathcal{D}_m$



(c) $\widehat{C} = [\widehat{q}_{\mathrm{l}}(X) - \widehat{q}_{\alpha}, \widehat{q}_{\mathrm{u}}(X) + \widehat{q}_{\alpha}]$

Towards adaptivity

	Absolute errors	Quantile regression	Normalization
$s(X_i, Y_i)$	$\lvert Y_i - \widehat{m}(X_i)\rvert$	$\max(\widehat{q}_{\mathrm{l}}(X_i) - Y_i, Y_i - \widehat{q}_{\mathrm{u}}(X_i))$	$\dfrac{(Y_i - \widehat{m}(X_i))^2}{\widehat{f}(X_i)}$



(a) Learn $\widehat{m}$, $\widehat{f}$ on $\mathcal{D}_n$

(b) Compute $\frac{(Y_i - \widehat{m}(X_i))^2}{\widehat{f}(X_i)}$ on $\mathcal{D}_m$

(c) $\widehat{C} = [\widehat{m}(X) \pm \sqrt{\widehat{q}_\alpha \widehat{f}(X)}]$

Towards adaptivity

	Absolute errors	Quantile regression[1]	Normalization[2]
$s(X_i, Y_i)$	$\lvert Y_i - \widehat{m}(X_i)\rvert$	$\max(\widehat{q}_{\mathrm{l}}(X_i) - Y_i, Y_i - \widehat{q}_{\mathrm{u}}(X_i))$	$\frac{(Y_i - \widehat{m}(X_i))^2}{\widehat{f}(X_i)}$

- $\widehat{f}$ is **any** estimate of the errors of $\widehat{m}$ (e.g. other ML models trained on the absolute residuals, resampling procedure, Bayesian approach such as GPs, ...)
- Key fact: this estimation is made without any consideration for coverage nor adaptivity

We then propose to *learn* the score function in a way that targets both adaptivity and coverage

[1] [Romano et al. 2019]
[2] [Lei et al. 2014; Johansson et al. 2014; Papadopoulos 2024; Jaber et al. 2024]

Table of Contents

Learning problem for a score function

- We consider a normalized score: $\frac{(Y-m(X))^2}{f(X)}$, with $\boldsymbol{f \geq 0}$
- As for all learning problems, we must:
  - Specify the criterion to minimize, to be discussed later
  - Choose a search space for our functions, here we rely on **kernel methods**
    - m lives in the Reproducible Kernel Hilbert Space (RKHS) $\mathcal{H}^m$ with kernel k^m and lengthscales θ^m
    - f is a ***kernel sum-of-squares*** function, in order to impose its **positivity**

Kernel sum-of-squares (kSoS)

- Consider a RKHS $\mathcal{H}^f$ with a feature map $\phi\colon \mathcal{X} \to \mathcal{H}^f$, a kernel SoS function is defined as

$$f(X) = \phi(X)^\top \mathcal{A} \phi(X), \quad \text{with} \quad \mathcal{A} \in \mathcal{S}_+(\mathcal{H}^f)$$

- f can be written as

$$f_{\mathcal{A}}(X) = \sum_{l \geq 0} \lambda_l u_l(X) u_l(X)^\top$$

for functions $u_l \in \mathcal{H}^f$ with λ_l the eigenvalues of the operator $\mathcal{A}$, hence the **sum-of-squares** name

Learning the score function amounts to simultaneously learning

$$m \in \mathcal{H}^m,\ f \in \mathcal{S}o\mathcal{S}(\mathcal{H}^f) \quad \Leftrightarrow \quad m \in \mathcal{H}^m,\ \mathcal{A} \in \mathcal{S}_+(\mathcal{H}^f)$$

Learning problem for a score function

$$
\begin{aligned}
\inf_{m\in\mathcal{H}^m,\, \mathcal{A}\in\mathcal{S}_+(\mathcal{H}^f)} \quad & \frac{a}{n}\sum_{i=1}^{n}\left(Y_i - m(X_i)\right)^2 + \frac{b}{n}\sum_{i=1}^{n} f_{\mathcal{A}}(X_i) + \lambda_1\|\mathcal{A}\|_\star + \lambda_2\|\mathcal{A}\|_F^2 && (1)\\
\text{s.t.} \quad & f_{\mathcal{A}}(X_i) \geq \left(Y_i - m(X_i)\right)^2,\ i\in[n], && (2)\\
& \|m\|_{\mathcal{H}^m}^2 \leq s && (3)
\end{aligned}
$$

Learning problem for a score function

$$\inf_{m \in \mathcal{H}^m,\, \mathcal{A} \in \mathcal{S}_+(\mathcal{H}^f)} \quad \frac{a}{n} \sum_{i=1}^{n} (Y_i - m(X_i))^2 + \frac{b}{n} \sum_{i=1}^{n} f_{\mathcal{A}}(X_i) + \lambda_1 \|\mathcal{A}\|_\star + \lambda_2 \|\mathcal{A}\|_F^2 \tag{1}$$

$$\text{s.t.} \quad f_{\mathcal{A}}(X_i) \geq (Y_i - m(X_i))^2 ,\ i \in [n] , \tag{2}$$

$$\|m\|_{\mathcal{H}^m}^2 \leq s \tag{3}$$

i) Faithful estimation of the mean function

Learning problem for a score function

$$\inf_{m\in\mathcal{H}^m,\,\mathcal{A}\in\mathcal{S}_+(\mathcal{H}^f)} \quad \frac{a}{n}\sum_{i=1}^{n}(Y_i - m(X_i))^2 + \frac{b}{n}\sum_{i=1}^{n} f_{\mathcal{A}}(X_i) + \lambda_1\|\mathcal{A}\|_\star + \lambda_2\|\mathcal{A}\|_F^2 \tag{1}$$

$$\text{s.t.} \quad f_{\mathcal{A}}(X_i) \geq (Y_i - m(X_i))^2\,,\ i\in[n]\,, \tag{2}$$

$$\|m\|^2_{\mathcal{H}^m} \leq s \tag{3}$$

i) Faithful estimation of the mean function
ii) 100% coverage on the training sample - **convex** constraint (later adjusted with split CP)

Learning problem for a score function

$$\inf_{m \in \mathcal{H}^m,\, \mathcal{A} \in \mathcal{S}_+(\mathcal{H}^f)} \quad \frac{a}{n}\sum_{i=1}^{n}(Y_i - m(X_i))^2 + \frac{b}{n}\sum_{i=1}^{n} f_{\mathcal{A}}(X_i) + \lambda_1 \|\mathcal{A}\|_\star + \lambda_2 \|\mathcal{A}\|_F^2 \tag{1}$$

$$\text{s.t.} \quad f_{\mathcal{A}}(X_i) \geq (Y_i - m(X_i))^2\,,\ i \in [n]\,, \tag{2}$$

$$\|m\|_{\mathcal{H}^m}^2 \leq s \tag{3}$$

i) Faithful estimation of the mean function
ii) 100% coverage on the training sample - **convex** constraint (later adjusted with split CP)
iii) Minimization of the interval mean width

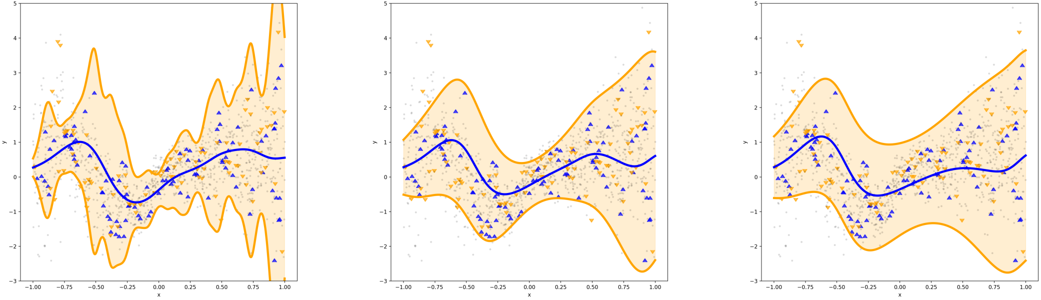
Learning problem for a score function

$$
\begin{aligned}
\inf_{m \in \mathcal{H}^m,\, \mathcal{A} \in \mathcal{S}_+(\mathcal{H}^f)} \quad & \frac{a}{n} \sum_{i=1}^{n} (Y_i - m(X_i))^2 + \frac{b}{n} \sum_{i=1}^{n} f_{\mathcal{A}}(X_i) + \lambda_1 \|\mathcal{A}\|_\star + \lambda_2 \|\mathcal{A}\|_F^2 && (1) \\
\text{s.t.} \quad & f_{\mathcal{A}}(X_i) \geq (Y_i - m(X_i))^2 \,,\; i \in [n]\,, && (2) \\
& \|m\|_{\mathcal{H}^m}^2 \leq s && (3)
\end{aligned}
$$

i) Faithful estimation of the mean function
ii) 100% coverage on the training sample - **convex** constraint (later adjusted with split CP)
iii) Minimization of the interval mean width
iv) Control of the regularity of the bands
   - lasso-type norm $\|\mathcal{A}\|_\star$
   - ridge-type norm $\|\mathcal{A}\|_F$

Representer theorem

- We proved a representer theorem for this infinite dimensional problem
- It becomes a Semi-Definite Program (SDP) problem, solvable using off-the-shelves solvers



(a) $\theta^f = 0.1$ (b) $\theta^f = 0.5$ (c) $\theta^f = 0.9$

Note: θ^f is the vector of lengthscales for k^f, the kernel corresponding to $\mathcal{H}^f$

Scalability

- The SDP problem is not scalable past 200 samples
- We proved that it admits a dual representation if $\lambda_2 > 0$, which is solvable using **accelerated gradient descent**

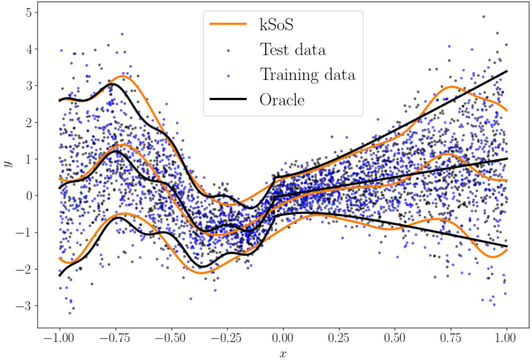


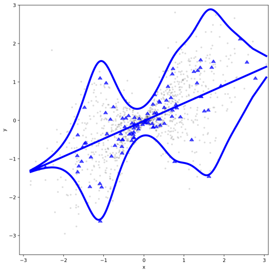
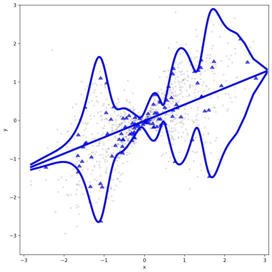
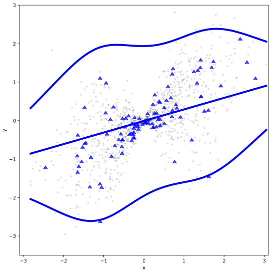
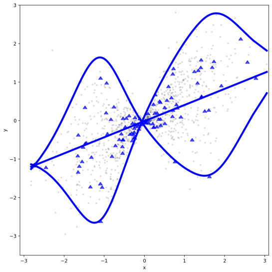
Figure 5: Dual solver with 2000 samples

Hyperparameter tuning

- Hyperparameters a, λ_1, λ_2 do not have a huge impact on the prediction bands
- We fix θ^m and s using a preliminary Gaussian Process model

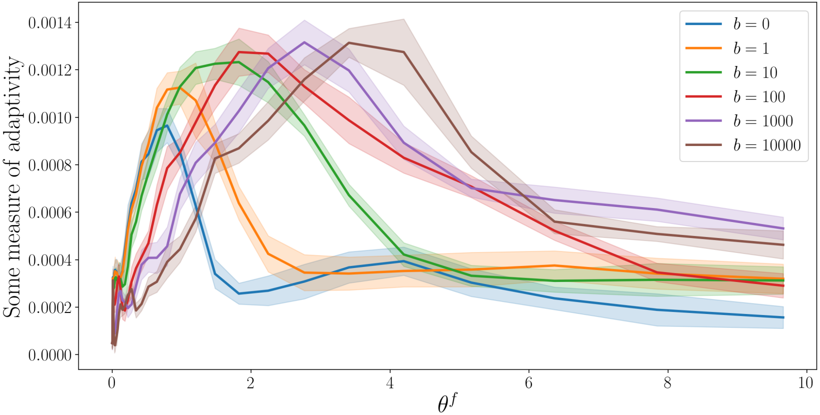
- We focus on the two most important hyperparameters
  - $\boldsymbol{b}$: mean width
  - $\boldsymbol{\theta^f}$: lengthscales associated to f, control complexity of f

Hyperparameter tuning

	$b = 1$	$b = 100$
$\theta^f = 0.5$		
$\theta^f = 1.74$		

Before calibration

Hyperparameter tuning



For each b, we can find an optimal value for θ^f that maximizes adaptivity

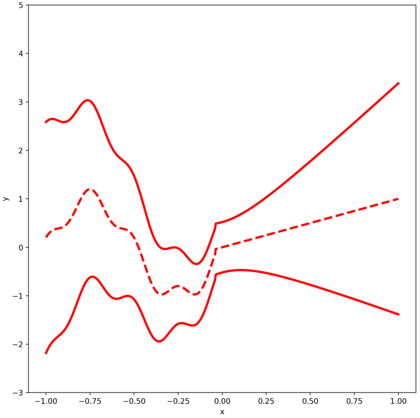
Hyperparameter tuning

- Which adaptivity measure can we use to choose θ^f?
- We propose the Hilbert-Schmidt Independence Criterion (HSIC), an independence measure between random variables
- What is the link between independence of random variables and adaptivity?

Hyperparameter tuning

- Perfectly adaptive bands guarantee local coverage

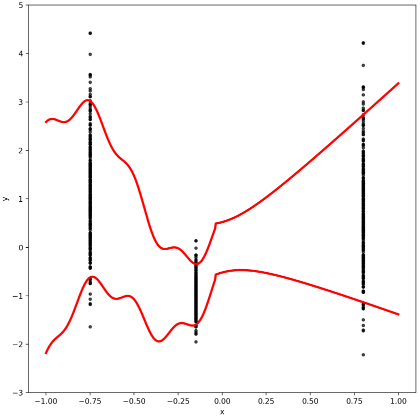
$$\mathbb{P}(Y_{N+1} \in \widehat{C}(X_{N+1}) \mid X_{N+1} = x) \geq 1 - \alpha$$



Hyperparameter tuning

- Perfectly adaptive bands guarantee local coverage

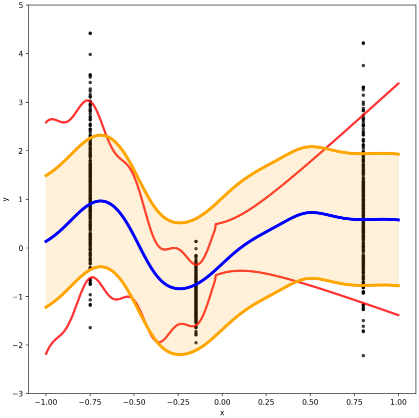
$$\mathbb{P}(Y_{N+1} \in \widehat{C}(X_{N+1}) \mid X_{N+1} = x) \geq 1 - \alpha$$



Hyperparameter tuning

- Perfectly adaptive bands guarantee local coverage

$$\mathbb{P}(Y_{N+1} \in \widehat{C}(X_{N+1}) \mid X_{N+1} = x) \geq 1 - \alpha$$



Hyperparameter tuning

- Without hypothesis on the data, satisfying this local coverage leads to infinitely wide prediction bands [Vovk 2012; Barber et al. 2021]
- We can relax the local coverage by considering X in a small neighbourhood ω_X, such that $\forall x \in \mathcal{X},\ \mathbb{P}(x \in \omega_X) \geq \delta$:

$$p_{\mathcal{D}_N} := \mathbb{P}(Y_{N+1} \in \widehat{C}(X_{N+1}) | X_{N+1} \in \omega_X) \geq 1 - \alpha$$

- Deutschmann et al. 2024 proved a lower bound for $p_{\mathcal{D}_N}$, which involves $\mathrm{MI}(X, S_{\theta^f}(X, Y))$, but MI is not robust numerically

Hyperparameter tuning

- Using information theory results and recent inequalities result between the TV distance and the MMD, we proved a new bound

$$p_{\mathcal{D}_N} \geq 1 - \alpha - \frac{1}{\delta}\sqrt{1 - \frac{\alpha_1}{1 - \alpha_2 \mathrm{HSIC}(r_{\mathcal{D}_n}(X_{N+1}, Y_{N+1}), \widehat{f}_{\theta^f}(X_{N+1}))}}$$

- HSIC is much more robust than MI

Hyperparameter tuning

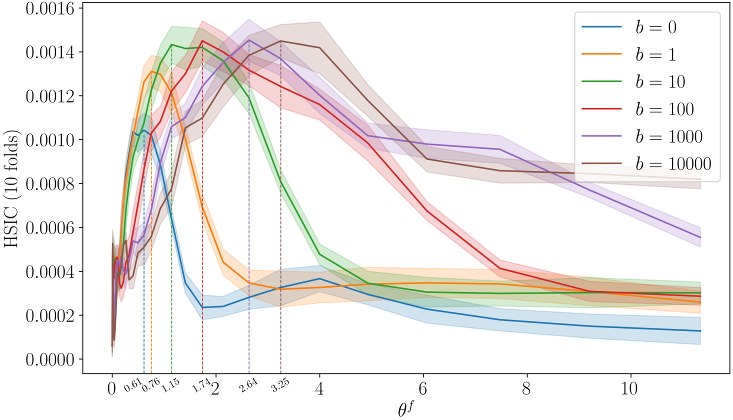
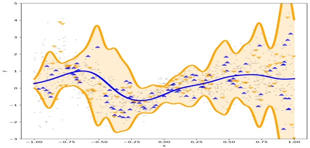


Figure 6: $\mathrm{HSIC}(r_{\mathcal{D}_n}(X,Y), \widehat{f}_{\theta^f}(X))$

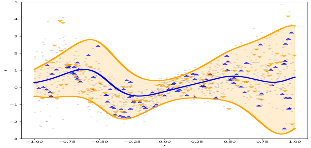
Maximizing this HSIC, i.e. the dependence between the residuals and the interval widths, allows to target better local coverage

Table of Contents

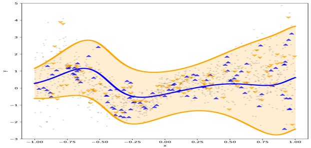
Analytical cases



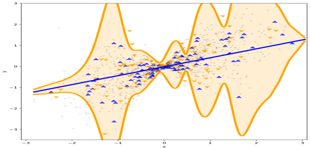
(a) $\theta^f = 0.1$



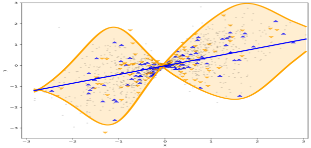
(b) $\boldsymbol{\theta^f_{\mathrm{HSIC}}} = 0.5$



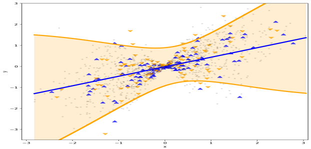
(c) $\theta^f = 0.9$



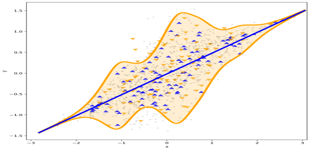
(d) $\theta^f = 0.5$



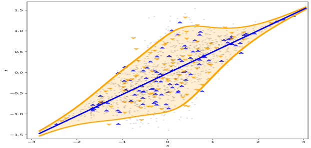
(e) $\boldsymbol{\theta^f_{\mathrm{HSIC}}} = 1.74$



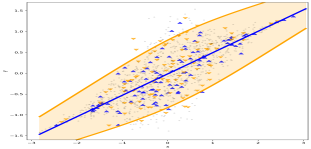
(f) $\theta^f = 10$



(g) $\theta^f = 0.4$

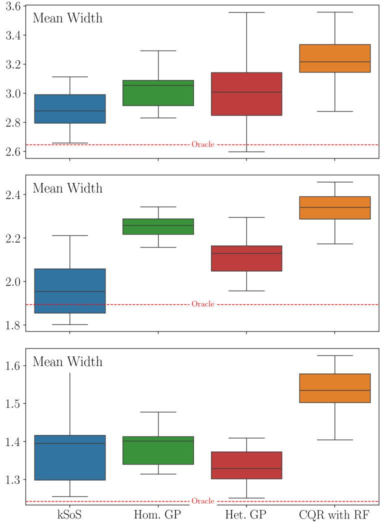


(h) $\boldsymbol{\theta^f_{\mathrm{HSIC}}} = 0.9$



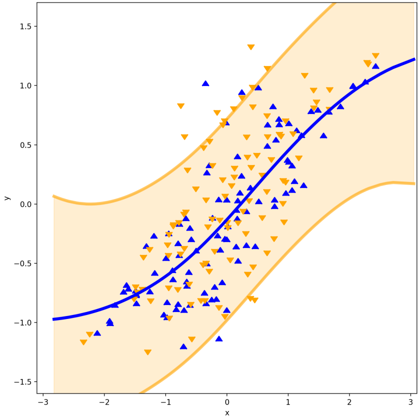
(i) $\theta^f = 3$

Mean width metric

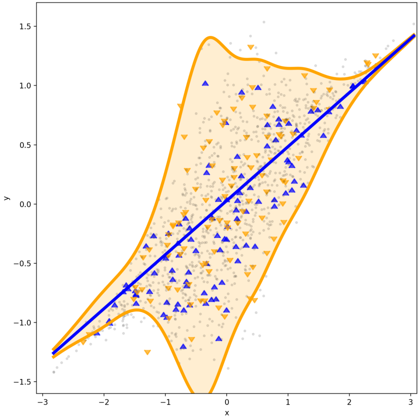


- A common measure for adaptive prediction bands in the literature is **mean width**, which should be minimized
- kSoS leads to better or as good mean width as competitors
- **However, mean width does not always tell the full story**

Mean width metric

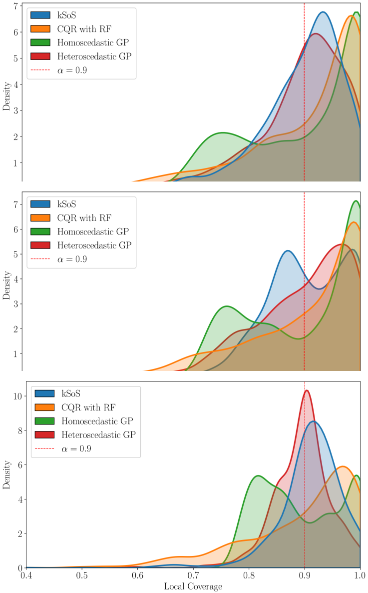


(a) Homoscedastic GP
MW = 1.712



(b) kSoS with Opt. HSIC
MW = 1.759

Local coverage metric



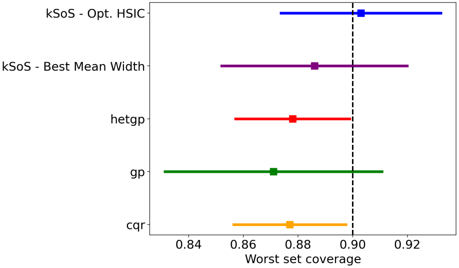
- The best measure of adaptivity is **local coverage**
- The target for local coverage is a Dirac at $1 - \alpha = 0.9$
- kSoS leads to better concentrated local coverage in general

Real world datasets - Comparison between mean widths

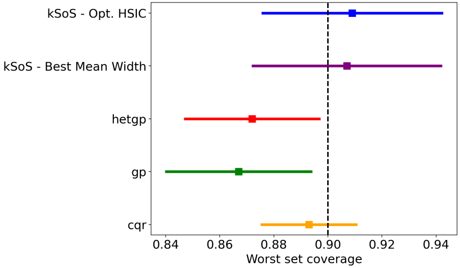
Dataset	CQR	Het GP	Hom GP	kSoS Best mean width	kSoS Opt. HSIC
Concrete	0.586 ± 0.032	$\mathbf{0.508} \pm 0.052$	0.543 ± 0.044	0.556 ± 0.044	0.568 ± 0.06
Bike	1.114 ± 0.062	1.000 ± 0.079	$\mathbf{0.809} \pm 0.024$	$\mathbf{0.804} \pm 0.032$	$\mathbf{0.803} \pm 0032$
Bio	$\mathbf{1.879} \pm 0.046$	2.21 ± 0.100	2.194 ± 0.119	2.03 ± 0.07	–
Diabetes	188.62 ± 9.33	191.24 ± 11.95	190.58 ± 11.19	$\mathbf{185.83} \pm 14.47$	187.6 ± 16.18
MPG	9.89 ± 0.82	9.70 ± 1.06	9.71 ± 0.73	$\mathbf{9.15} \pm 0.8$	9.36 ± 0.82
Housing	1.816 ± 0.045	1.585 ± 0.099	$\mathbf{1.453} \pm 0.099$	$\mathbf{1.468} \pm 0.094$	1.586 ± 0.104

- Mean width for six real-world datasets, kSoS with HSIC-optimized θ^f achieves best mean width on almost every datasets against competitors
- **Again, mean width does not tell the full story**

Real world datasets - Comparison with worst-set coverage



(a) Housing dataset



(b) Concrete dataset

- Worst set coverage is a substitute for local coverage for real datasets[3]
- **kSoS achieves better or equal worst-set coverage than competitors with better mean width**

[3] Thurin et al. 2025

Table of Contents

Contributions

- Learning setting for a score function in the context of split CP
- Representer theorem to make the problem tractable
- Solvable in practice with the primal (small n) using SDP or the dual (big n) using AGD
- Brand new adaptivity measure based on HSIC, that allows to automatically choose hyperparameters of the model

- **Paper accepted at NeurIPS 2025**, preprint available on arXiv, final version in the proceedings

arxiv.org/abs/2505.21039

Q&R

Thank you for listening!

Your feedback will be highly appreciated!

Table of Contents

References

- Barber, Rina Foygel et al. (2021). "The limits of distribution-free conditional predictive inference". In: Information and Inference: A Journal of the IMA. DOI: 10.1093/imaiai/iaaa017.
- Deutschmann, Nicolas et al. (2024). "Adaptive Conformal Regression with Split-Jackknife+ Scores". In: Transactions on Machine Learning Research. URL: https://openreview.net/forum?id=1fbTGC3BUD.
- Jaber, Edgar et al. (2024). Conformal Approach To Gaussian Process Surrogate Evaluation With Coverage Guar arXiv: 2401.07733 [stat.ML]. URL: https://arxiv.org/abs/2401.07733.
- Johansson, U. et al. (2014). "Regression conformal prediction with random forests". In: Machine Learning. URL: https://api.semanticscholar.org/CorpusID:14015369.

ENSAI

References

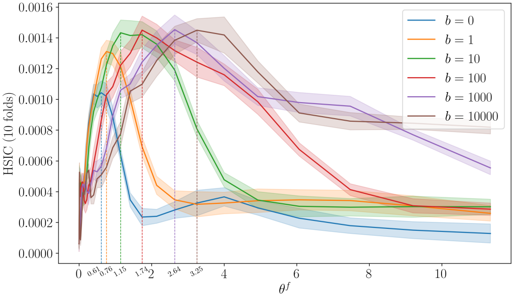
- Lei, Jing et al. (2014). "Distribution-free Prediction Bands for Non-parametric Regression". In: Journal of the Royal Statistical Society Series B: Statistical Methodology. DOI: 10.1111/rssb.12021.
- Papadopoulos, Harris (2024). "Guaranteed Coverage Prediction Intervals With Gaussian Process Regression". In: IEEE Transactions on Pattern Analysis and Machine Intelligence. DOI: 10.1109/tpami.2024.3418214.
- Romano, Yaniv et al. (2019). "Conformalized Quantile Regression". In: Advances in Neural Information Processing Systems. Ed. by H. Wallach et al. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf.

References

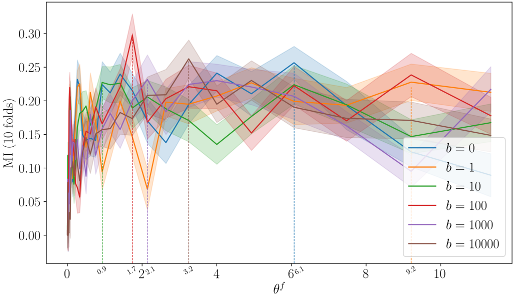
Thurin, Gauthier et al. (2025). "Optimal transport-based conformal prediction". In: Forty-second International Conference on Machine Learning. URL: https://openreview.net/forum?id=kEAyffH3tn.

Vovk, Vladimir (2012). "Conditional Validity of Inductive Conformal Predictors". In: Proceedings of the Asian Conference on Machine Learning. Proceedings of Machine Learning Research. PMLR. URL: https://proceedings.mlr.press/v25/vovk12.html.

Bound on conditional coverage



(a) HSIC



(b) MI