Asymmetric conformal prediction with penalized kernel sum-of-squares

Curves and Surfaces 2026
Saint-Malo

Louis Allain[1,2] Sébastien Da Veiga[2] Brian Staber[1]

[1]Safran Tech, Digital Sciences & Technologies, 78114 Magny-Les-Hameaux, France
[2]Univ Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France
{louis.allain,brian.staber}@safrangroup.com
sebastien.da-veiga@ensai.fr

Table of Contents

1. Introduction
   1. Conformal prediction
   2. Kernel sum-of-squares

2. Our learning problem

3. Numerical experiments

4. Conclusion

5. References and appendix

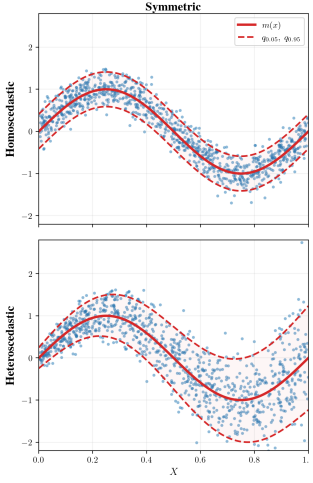


Motivation: bands around machine learning predictions

- Reliable prediction bands are essential in high-stakes ML: aeronautics, medicine, finance...

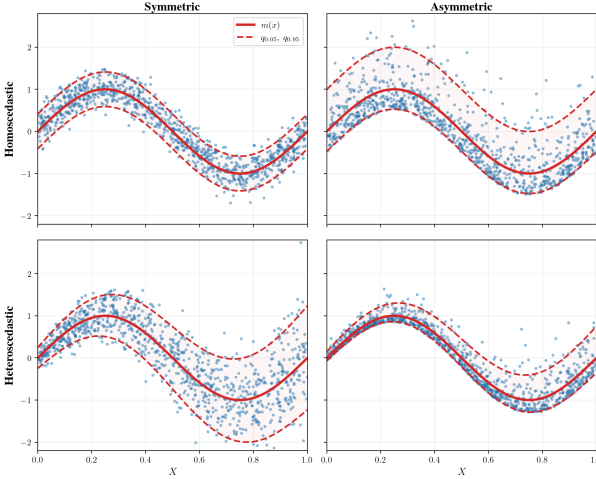
Motivation: bands around machine learning predictions

- Reliable prediction bands are essential in high-stakes ML: aeronautics, medicine, finance...
- Key feature: **adaptivity**
  - Homo- vs. heteroscedastic noise
  - Wide bands when uncertain and narrow when confident



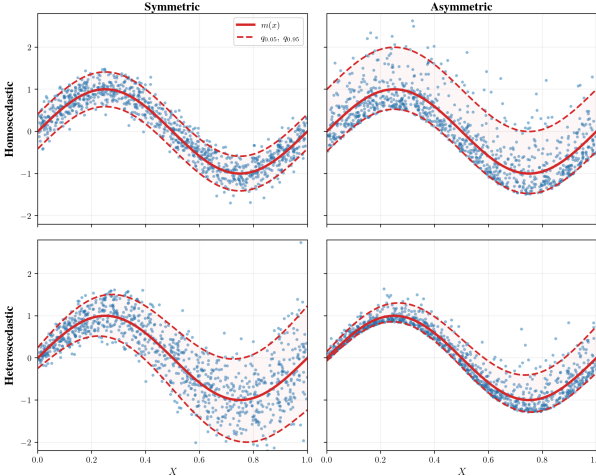
Motivation: bands around machine learning predictions

- Reliable prediction bands are essential in high-stakes ML: aeronautics, medicine, finance...
- Key feature: **adaptivity**
  - Homo- vs. heteroscedastic noise
  - Wide bands when uncertain and narrow when confident
- Additional features needed in practice:
  - Residuals are often **asymmetric** (Pouplin et al. 2024)



Motivation: bands around machine learning predictions

- Reliable prediction bands are essential in high-stakes ML: aeronautics, medicine, finance...
- Key feature: **adaptivity**
  - Homo- vs. heteroscedastic noise
  - Wide bands when uncertain and narrow when confident
- Additional features needed in practice:
  - Residuals are often **asymmetric** (Pouplin et al. 2024)
  - Intervals **around** a point prediction is a must have for engineers





Motivations: bands around machine learning predictions

Existing methods for prediction bands:

- **Symmetric**: incorrect coverage or prediction bands too large

  Even if the noise is symmetric, small sample settings or model bias can lead to artificial asymmetry

- **Asymmetric**: do not account for all information and may be wider

We need a method that **automatically adapts to the data**

Motivations: bands around machine learning predictions

Existing methods for prediction bands:

- **Symmetric**: incorrect coverage or prediction bands too large

  Even if the noise is symmetric, small sample settings or model bias can lead to artificial asymmetry

- **Asymmetric**: do not account for all information and may be wider

We need a method that **automatically adapts to the data**

✓ Our proposal:

A method that transitions smoothly between symmetric and asymmetric bands, built on **conformal prediction** for its distribution-free coverage guarantees

Conformal Prediction

- Conformal Prediction (CP): a rigorous method to construct prediction intervals satisfying the **marginal coverage** guarantee at level α:

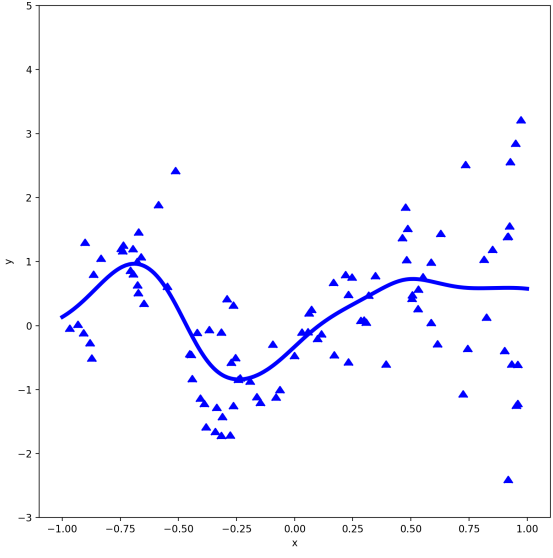
$$\mathbb{P}(Y \in \widehat{C}(X)) \geq 1 - \alpha$$

  with the following properties ✓ Finite sample ✓ Distribution free ✓ Model agnostic

- We focus on the split CP variant based on two independent datasets
  - $\mathcal{D}_n$, the *pre-training* set
  - $\mathcal{D}_m$, the *calibration* set

Conformal Prediction

1. Train predictive model $\widehat{m}$ on $\mathcal{D}_n$

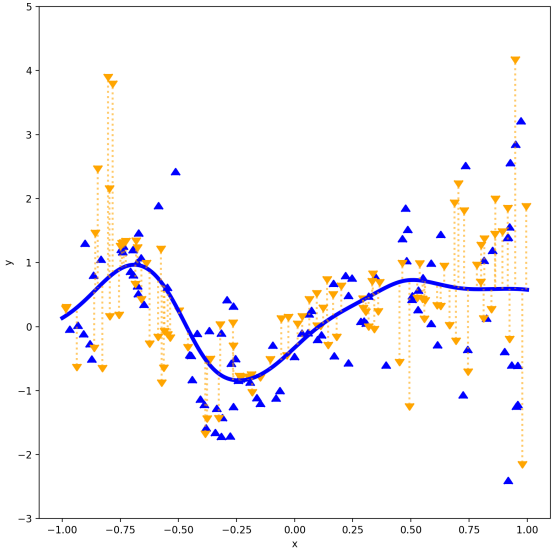


Conformal Prediction

1. Train predictive model $\widehat{m}$ on $\mathcal{D}_n$

2.1. Evaluate the quality of $\widehat{m}$ on $\mathcal{D}_m$

2.2. Compute the quantile $\widehat{q}_\alpha$ of these quality measures



Conformal Prediction

1. Train predictive model $\widehat{m}$ on $\mathcal{D}_n$

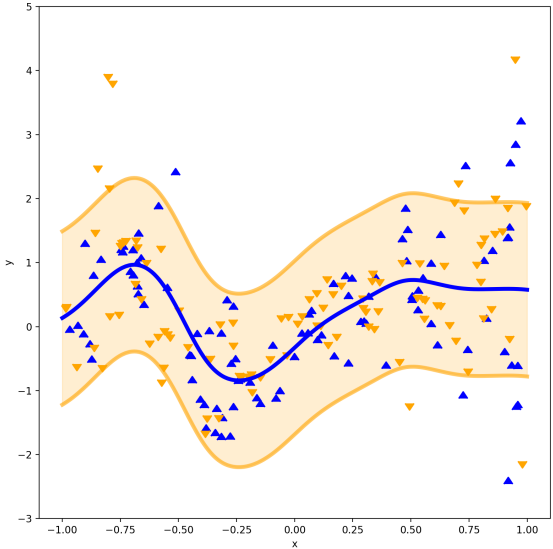
2.1. Evaluate the quality of $\widehat{m}$ on $\mathcal{D}_m$

2.2. Compute the adjusted quantile $\widehat{q}_\alpha$ of these quality measures

3. Under the assumption of data exchangeability, the prediction interval

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

satisfies all the desired properties



Scores

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

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

- Prediction intervals constructed with such scores:

$$\widehat{C}(X) = \left[\widehat{m}(X) - \widehat{q}_\alpha,\ \widehat{m}(X) + \widehat{q}_\alpha\right]$$

Scores

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

	Absolute errors	Normalization
$s(X_i, Y_i)$	$\lvert Y_i - \widehat{m}(X_i)\rvert$	$\frac{(Y_i - \widehat{m}(X_i))^2}{\widehat{f}(X_i)}$
Width	Constant	Adaptive
Type	Symmetric	Symmetric

- Prediction intervals constructed with such scores:

$$\widehat{C}(X) = \left[\widehat{m}(X) - \widehat{q}_\alpha \sqrt{\widehat{f}(X)},\ \widehat{m}(X) + \widehat{q}_\alpha \sqrt{\widehat{f}(X)}\right]$$

Scores

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

	Absolute errors	Normalization	Quantile regression
$s(X_i, Y_i)$	$\lvert Y_i - \widehat{m}(X_i)\rvert$	$\frac{(Y_i - \widehat{m}(X_i))^2}{\widehat{f}(X_i)}$	$\max(\widehat{q}_{\mathrm{l}}(X_i) - Y_i, Y_i - \widehat{q}_{\mathrm{u}}(X_i))$
Width	Constant	Adaptive	Adaptive
Type	Symmetric	Symmetric	**Asymmetric**

- Prediction intervals constructed with such scores:

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

Scores

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

	Absolute errors	Normalization	Quantile regression
$s(X_i, Y_i)$	$\lvert Y_i - \widehat{m}(X_i)\rvert$	$\frac{(Y_i - \widehat{m}(X_i))^2}{\widehat{f}(X_i)}$	$\max(\widehat{q}_{\mathrm{l}}(X_i) - Y_i, Y_i - \widehat{q}_{\mathrm{u}}(X_i))$
Width	Constant	Adaptive	Adaptive
Type	Symmetric	Symmetric	**Asymmetric**

- Prediction intervals constructed with such scores:

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

✗ Purely asymmetric

✗ Not built around a mean prediction

A new score function

- Taking inspiration from CQR scores we can define the following new score function

$$s(X_i, Y_i) = \max(\widehat{q}_{\text{low}}(X_i) - Y_i,\ Y_i - \widehat{q}_{\text{up}}(X_i))$$

A new score function

- Taking inspiration from CQR scores we can define the following new score function

$$s(X_i, Y_i) = \max(\widehat{m}(X_i) - \widehat{f}_{\text{low}}(X_i) - Y_i,\ Y_i - \widehat{q}_{\text{up}}(X_i))$$

A new score function

- Taking inspiration from CQR scores we can define the following new score function

$$s(X_i, Y_i) = \max(\widehat{m}(X_i) - \widehat{f}_{\text{low}}(X_i) - Y_i,\ Y_i - (\widehat{m}(X_i) + \widehat{f}_{\text{up}}(X_i)))$$

A new score function

- Taking inspiration from CQR scores we can define the following new score function

$$s(X_i, Y_i) = \max(\widehat{m}(X_i) - \widehat{f}_{\text{low}}(X_i) - Y_i,\ Y_i - (\widehat{m}(X_i) + \widehat{f}_{\text{up}}(X_i)))$$

- f_{low} and f_{up} measure the noise as **positive** displacement around the mean

A new score function

- Taking inspiration from CQR scores we can define the following new score function

$$s(X_i, Y_i) = \max(\widehat{m}(X_i) - \widehat{f}_{\text{low}}(X_i) - Y_i,\ Y_i - (\widehat{m}(X_i) + \widehat{f}_{\text{up}}(X_i)))$$

- f_{low} and f_{up} measure the noise as **positive** displacement around the mean

- **Contribution**: learn this score function in a goal-oriented way

A new score function

- Taking inspiration from CQR scores we can define the following new score function

$$s(X_i, Y_i) = \max(\widehat{m}(X_i) - \widehat{f}_{\text{low}}(X_i) - Y_i,\ Y_i - (\widehat{m}(X_i) + \widehat{f}_{\text{up}}(X_i)))$$

- f_{low} and f_{up} measure the noise as **positive** displacement around the mean
- **Contribution**: learn this score function in a goal-oriented way

Advantages of supervised learning framework

- Penalize wide intervals
- Control band regularity
- Target adaptivity
- Natural way to include symmetry transition

Kernel sum-of-squares (kSoS)

Positivity of f_{low} and f_{up} is imposed **by construction** using the kSoS framework

Definition 1 (Marteau-Ferey et al. 2020)

Consider a RKHS $\mathcal{H}$ with a feature map $\phi\colon \mathcal{X} \to \mathcal{H}$, a kSoS function is defined as

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

NB: f can be written as $f_{\mathcal{A}}(X) = \sum_{l \geq 0} \mu_l u_l(X) u_l(X)^\top$ for eigenfunctions $u_l \in \mathcal{H}$ and eigenvalues $\mu_l > 0$ of the SDP operator $\mathcal{A}$, hence the **sum-of-squares** name

Kernel sum-of-squares

Assumption 1 (Marteau-Ferey et al. 2020)

The regularization $\Omega(\mathcal{A})$ must write as a increasing function q of the operator singular values σ_k

$$\Omega(\mathcal{A}) = \sum_k q(\sigma_k)$$

The elastic-net regularization $\Omega(\mathcal{A}) = \lambda_1 \|\mathcal{A}\|_\star + \lambda_2 \|\mathcal{A}\|_{\mathrm{F}}^2$ satisfies Assumption 1.

- $\lambda_1 > 0$: lasso-type penalty, promotes **low-rank** operators
- $\lambda_2 > 0$: ridge-type penalty, makes Ω **strongly convex**

Kernel sum-of-squares

Learning problem with kSoS functions: **Representer theorem**

Theorem 1 (Marteau-Ferey et al. 2020)

Let L be a lower semi-continuous and bounded below function and Ω satisfy Assumption 1. The problem

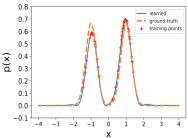
$$\inf_{\mathcal{A} \in \mathcal{S}_+(\mathcal{H})} L(f_{\mathcal{A}}(X_1), \dots, f_{\mathcal{A}}(X_n)) + \Omega(\mathcal{A})$$

admits a solution $\mathcal{A}^\star$: $\mathcal{A}^\star = \sum_{i,j=1}^n \boldsymbol{B}_{ij} \phi(X_i)\phi(X_j)^\top$ for some matrix $\boldsymbol{B} \in \mathbb{S}_+^n$, with $f_{\mathcal{A}^\star}(X) = \sum_{i,j=1}^n \boldsymbol{B}_{ij} k(X, X_i) k(X, X_j)$.

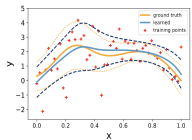
NB: extension to several kSoS functions with operators on **same** RKHS in their paper

Kernel sum-of-squares

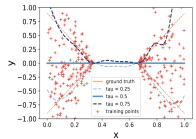
- The kSoS framework has seen great success in multiple applications



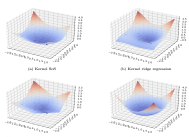
Density estimation
[Marteau-Ferey et al. 2020]



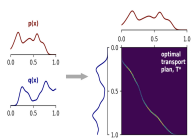
Heteroscedastic regression
[Marteau-Ferey et al. 2020]



Multi-quantile regression
[Marteau-Ferey et al. 2020]



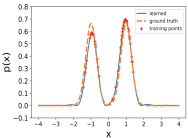
Convex regression
[Muzellec et al. 2021]



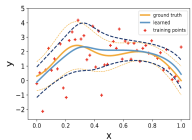
Optimal transport
[Vacher et al. 2024]

Kernel sum-of-squares

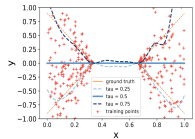
- The kSoS framework has seen great success in multiple applications



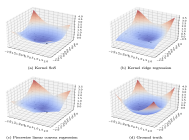
Density estimation
[Marteau-Ferey et al. 2020]



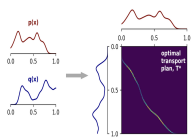
Heteroscedastic regression
[Marteau-Ferey et al. 2020]



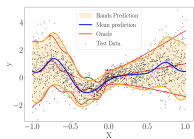
Multi-quantile regression
[Marteau-Ferey et al. 2020]



Convex regression
[Muzellec et al. 2021]



Optimal transport
[Vacher et al. 2024]



Sym. prediction bands
[Allain et al. 2025]

Table of Contents

1. Introduction
   1. Conformal prediction
   2. Kernel sum-of-squares

2. Our learning problem

3. Numerical experiments

4. Conclusion

5. References and appendix

From scores to prediction bands

- We are given the pre-training dataset $\mathcal{D}_n$, on which we want to learn $\widehat{f}_{\text{low}}$ and $\widehat{f}_{\text{up}}$
- Once the score function is learned, the split CP procedure yields the prediction band

$$\widehat{C}(X) = \left[\widehat{m}(X) - \widehat{f}_{\text{low}}(X) - \widehat{q}_\alpha,\ \widehat{m}(X) + \widehat{f}_{\text{up}}(X) + \widehat{q}_\alpha\right]$$

- It remains to design a learning procedure for $\widehat{f}_{\text{low}}$ and $\widehat{f}_{\text{up}}$

Learning problem for asymmetric kSoS

$$\inf_{\substack{\mathcal{A}_{\text{low}} \in \mathcal{S}_+(\mathcal{H}_{\text{low}}) \\ \mathcal{A}_{\text{up}} \in \mathcal{S}_+(\mathcal{H}_{\text{up}})}} \quad \frac{b}{n} \sum_{i=1}^{n} (f_{\mathcal{A}_{\text{low}}}(X_i) + f_{\mathcal{A}_{\text{up}}}(X_i)) + \Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}})$$

$$\text{s.t.} \quad \widehat{m}_n(X_i) - Y_i - f_{\text{low}}(X_i) \leq 0, \ i \in [n]$$

$$Y_i - \widehat{m}_n(X_i) - f_{\text{up}}(X_i) \leq 0, \ i \in [n]$$

Learning problem for asymmetric kSoS

$$\inf_{\substack{\mathcal{A}_{\text{low}} \in \mathcal{S}_+(\mathcal{H}_{\text{low}}) \\ \mathcal{A}_{\text{up}} \in \mathcal{S}_+(\mathcal{H}_{\text{up}})}} \quad \frac{b}{n} \sum_{i=1}^{n} \left(f_{\mathcal{A}_{\text{low}}}(X_i) + f_{\mathcal{A}_{\text{up}}}(X_i) \right) + \Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}})$$

$$\text{s.t.} \quad \widehat{m}_n(X_i) - Y_i - f_{\text{low}}(X_i) \leq 0, \ i \in [n]$$

$$Y_i - \widehat{m}_n(X_i) - f_{\text{up}}(X_i) \leq 0, \ i \in [n]$$

i) Minimization of mean interval width

Learning problem for asymmetric kSoS

$$\inf_{\substack{\mathcal{A}_{\text{low}} \in \mathcal{S}_+(\mathcal{H}_{\text{low}}) \\ \mathcal{A}_{\text{up}} \in \mathcal{S}_+(\mathcal{H}_{\text{up}})}} \quad \frac{b}{n} \sum_{i=1}^{n} \left(f_{\mathcal{A}_{\text{low}}}(X_i) + f_{\mathcal{A}_{\text{up}}}(X_i) \right) + \Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}})$$

$$\text{s.t.} \quad \widehat{m}_n(X_i) - Y_i - f_{\text{low}}(X_i) \leq 0, \ i \in [n]$$

$$Y_i - \widehat{m}_n(X_i) - f_{\text{up}}(X_i) \leq 0, \ i \in [n]$$

i) Minimization of mean interval width

ii) Control the regularity of f_{low} and f_{up}

Learning problem for asymmetric kSoS

$$\inf_{\substack{\mathcal{A}_{\text{low}} \in \mathcal{S}_+(\mathcal{H}_{\text{low}}) \\ \mathcal{A}_{\text{up}} \in \mathcal{S}_+(\mathcal{H}_{\text{up}})}} \quad \frac{b}{n} \sum_{i=1}^{n} \left(f_{\mathcal{A}_{\text{low}}}(X_i) + f_{\mathcal{A}_{\text{up}}}(X_i) \right) + \Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}})$$

$$\text{s.t.} \quad \widehat{m}_n(X_i) - Y_i - f_{\text{low}}(X_i) \leq 0, \ i \in [n]$$

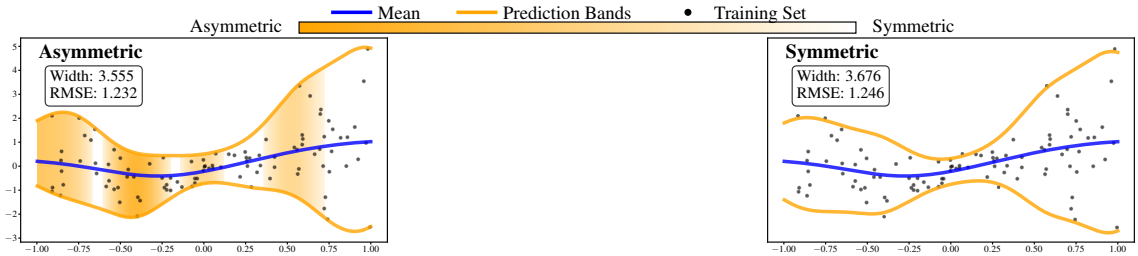
$$Y_i - \widehat{m}_n(X_i) - f_{\text{up}}(X_i) \leq 0, \ i \in [n]$$

i) Minimization of mean interval width

ii) Control the regularity of f_{low} and f_{up}

iii) 100% coverage on the training sample — **convex constraint**
(later adjusted with split CP)

Transition between symmetry and asymmetry?



Width	**3.555**	3.676
RMSE	**1.232**	1.246

Could an intermediate model be even better than the asymmetric bands?

Learning problem for penalized kSoS

$$\inf_{\substack{\mathcal{A}_{\text{low}} \in \mathcal{S}_+(\mathcal{H}_{\text{low}}) \\ \mathcal{A}_{\text{up}} \in \mathcal{S}_+(\mathcal{H}_{\text{up}})}} \quad \frac{b}{n} \sum_{i=1}^{n} (f_{\mathcal{A}_{\text{low}}}(X_i) + f_{\mathcal{A}_{\text{up}}}(X_i)) + \Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}}) \boxed{+ \boldsymbol{\lambda_{\text{pen}}\, \mathcal{P}(f_{\text{low}}, f_{\text{up}})}}$$

$$\text{s.t.} \quad \widehat{m}_n(X_i) - Y_i - f_{\text{low}}(X_i) \leq 0, \; i \in [n]$$
$$Y_i - \widehat{m}_n(X_i) - f_{\text{up}}(X_i) \leq 0, \; i \in [n]$$

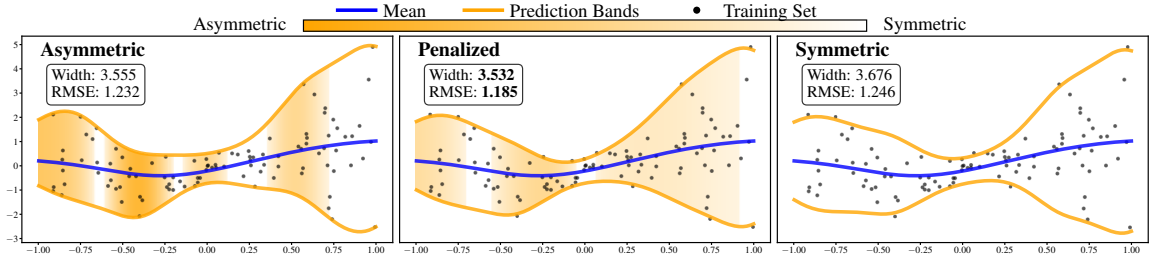
i) Minimization of mean interval width

ii) Control the regularity of f_{low} and f_{up}

iii) 100% coverage on the training sample

iv) **Symmetry penalty $\mathcal{P}$**: pushes $f_{\text{low}} \approx f_{\text{up}}$ when λ_{pen} is large (symmetric limit)

Impact of the penalty



Width	3.555	**3.532**	3.676
RMSE	1.232	**1.185**	1.246

Even with true symmetric noise, allowing some asymmetry can lead to better prediction bands

Asymmetric penalties

- We propose two possible penalties $\mathcal{P}(f_{\text{low}}, f_{\text{up}})$:

Operator penalty

$$\mathcal{P} = \|\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}}\|_\star + \|\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}}\|_F^2$$

Training set penalty

$$\mathcal{P} = \frac{1}{n}\sum_{i=1}^{n} \left(f_{\mathcal{A}_{\text{low}}}(X_i) - f_{\mathcal{A}_{\text{up}}}(X_i)\right)^2$$

Asymmetric penalties

- We propose two possible penalties $\mathcal{P}(f_{\text{low}}, f_{\text{up}})$:

Operator penalty

$$\mathcal{P} = \|\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}}\|_\star + \|\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}}\|_F^2$$

Training set penalty

$$\mathcal{P} = \frac{1}{n}\sum_{i=1}^{n} \left(f_{\mathcal{A}_{\text{low}}}(X_i) - f_{\mathcal{A}_{\text{up}}}(X_i)\right)^2$$

- Theorem 1 does not apply if $\mathcal{A}_{\text{low}}$ and $\mathcal{A}_{\text{up}}$ act on different RKHSs

- We consider a collection of p operators

$$\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_p) \in \mathcal{S}_+(\mathcal{H}_1) \otimes \ldots \otimes \mathcal{S}_+(\mathcal{H}_p)$$

Generalized representer theorem

Theorem 2 (Allain et al. 2026)

Let L be a lower semi-continuous and bounded below function and let each Ω_s, $s = 1, \dots, p$ satisfies 1. The problem

$$\inf_{\mathcal{A} \in \mathcal{S}_+(\mathcal{H}_1) \otimes \dots \otimes \mathcal{S}_+(\mathcal{H}_p)} L(f_{\mathcal{A}}(X_1), \dots, f_{\mathcal{A}}(X_n)) + \sum_{s=1}^{p} \Omega_s(\mathcal{A}_s)$$

admits a solution $\mathcal{A}^\star$ which can be written as

$$\mathcal{A}^\star = \left(\sum_{i,j=1}^{n} \mathbf{B}_{sij}\, \phi_s(X_i) \phi_s(X_j)^\top \right)_{1 \le s \le p} \quad \text{for } p \text{ matrices } \mathbf{B}_s \in \mathbb{S}_+^n,$$

with $f_{\mathcal{A}^\star}(X) = \left(\sum_{i,j=1}^{n} \mathbf{B}_{sij}\, k_s(X, X_i) k_s(X, X_j) \right)_{1 \le s \le p}$.

Elements of proof

Key ideas:

1. **Projection onto the data subspace.**
   For each RKHS $\mathcal{H}_s$, let $\mathcal{H}_s^n = \text{span}\{\phi_s(X_i)\}_{i=1}^n$ and Π_s^n the orthogonal projection. Since $f_{\mathcal{A}_s} = f_{\Pi_s^n \mathcal{A}_s \Pi_s^n}$, and the regularizer satisfies $\Omega_s(\Pi_s^n \mathcal{A}_s \Pi_s^n) \leq \Omega_s(\mathcal{A}_s)$, **projecting never increases the objective** $\Rightarrow$ search over $\bigotimes_{s=1}^p \mathcal{S}_+(\mathcal{H}_s^n)$.

2. **Boundedness of the search space.**
   Assumption 1 for Ω_s implies that minimizing sequences have **bounded Frobenius norm**: restriction to a compact ball $\mathcal{K}_{R_0}$.

3. **Existence by compactness.**
   $\mathcal{K}_{R_0}$ is compact (finite-dimensional, closed and bounded). Minimizing a lower semi-continuous function on a non-empty compact set **always admits a minimum**.

Application to the penalized problem

- **Loss function:** L is **lower semi-continuous and bounded below**.
- **Regularization:** $\Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}})$, where each Ω is an **elastic-net** regularizer satisfying Assumption 1.

$\Rightarrow$ The generalized representer theorem applies: the solution takes the form

$$(f_{\mathcal{A}^\star_{\text{low}}}(X),\ f_{\mathcal{A}^\star_{\text{up}}}(X)) = (\mathbf{k}_{\text{low}}(X)^\top \mathbf{B}^\star_{\text{low}}\, \mathbf{k}_{\text{low}}(X),\ \mathbf{k}_{\text{up}}(X)^\top \mathbf{B}^\star_{\text{up}}\, \mathbf{k}_{\text{up}}(X))$$



Semi Definite Program

This theorem leads to a semi-definite program (SDP)

- ✓ Solvable using off-the-shelves solvers such as SCS
- ✗ **Problem:** SDP do not scale past ∼200 samples

Semi Definite Program

This theorem leads to a semi-definite program (SDP)

- ✓ Solvable using off-the-shelves solvers such as SCS
- ✗ **Problem:** SDP do not scale past ∼200 samples
- ✓ **Solution:** derive a dual form solvable with up to ∼1000 samples

Scale up with dual formulation

Theorem 3 (Allain et al. 2026)

Let $(b, \lambda_{\text{low}1}, \lambda_{\text{up}1}, \lambda_{\text{pen}}) \in \mathbb{R}_+$ and $\lambda_{\text{low}2}, \lambda_{\text{up}2} > 0$. The training set penalty problem admits a dual formulation:

$$\sup_{\substack{\boldsymbol{\Gamma}_{\text{low}}, \boldsymbol{\Gamma}_{\text{up}} \in \mathbb{R}^n_+ \\ \boldsymbol{\alpha}_0 \in \mathbb{R}^n}} (\boldsymbol{\Gamma}_{\text{up}} - \boldsymbol{\Gamma}_{\text{low}})\boldsymbol{r}^\top - \frac{1}{4\lambda_{\text{pen}}}\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^\top - \Omega^\star_{+,\text{low}}\big(\boldsymbol{V}_{\text{low}}\,\text{Diag}((\boldsymbol{\Gamma}_{\text{low}} + \boldsymbol{\alpha}_0)_{-b})\boldsymbol{V}_{\text{low}}^\top\big) - \Omega^\star_{+,\text{up}}\big(\boldsymbol{V}_{\text{up}}\,\text{Diag}((\boldsymbol{\Gamma}_{\text{up}} - \boldsymbol{\alpha}_0)_{-b})\boldsymbol{V}_{\text{up}}^\top\big)$$

where $r_i = Y_i - \widehat{m}(X_i)$ and $\Omega^\star_{+,(\cdot)}(\boldsymbol{B}) = \frac{1}{4\lambda_{(\cdot)2}}\|\big[\boldsymbol{B} - \lambda_{(\cdot)1}\boldsymbol{I}_n\big]_+\|^2_{\text{F}}$.

- Optimisation over $3n$ variables
- Solvable using gradient descent

$\Omega^\star_{+,(\cdot)}$: positive part $\rightarrow$ eigendecomposition

Elements of proof

Main steps:

1. **Lagrangian fromulation.**
   - The $2n$ coverage constraints introduce positive multipliers $\boldsymbol{\Gamma}_{\text{low}}, \boldsymbol{\Gamma}_{\text{up}} \in \mathbb{R}^n_+$.
   - The dual function $D(\boldsymbol{\Gamma}_{\text{low}}, \boldsymbol{\Gamma}_{\text{up}})$ involves a joint infimum over $(\mathcal{A}_{\text{low}}, \mathcal{A}_{\text{up}})$ that is **not separable** due to the penalty coupling term.

2. **Fenchel duality.**
   - Applying Theorem 3.3.5 from Borwein et al. 2006 to D introduces a new dual variable $\boldsymbol{\alpha}_0 \in \mathbb{R}^n$.
   - The conjugate of the new loss function yields $\frac{1}{4\lambda_{\text{pen}}} \boldsymbol{\alpha}_0 \boldsymbol{\alpha}_0^\top$
   - Fenchel conjugates of the regularizers take the same closed form as in Marteau-Ferey et al. 2020.

$$\Omega^\star_{+,(\cdot)}(\boldsymbol{B}) = \frac{1}{4\lambda_{(\cdot)2}} \| \left[\boldsymbol{B} - \lambda_{(\cdot)1} \boldsymbol{I}_n \right]_+ \|^2_{\text{F}}$$

Scale up with low-rank approximation

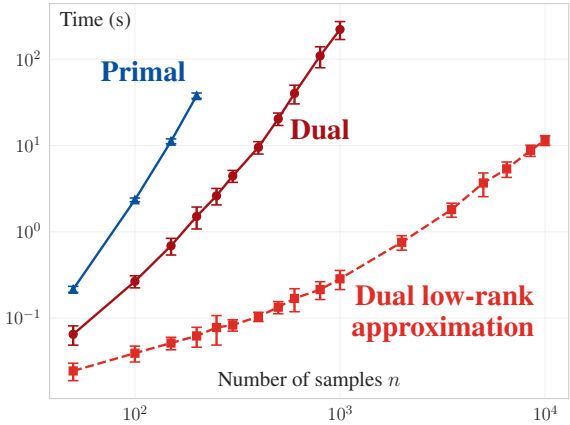
Bottleneck: evaluating $\Omega_+^\star$ requires an eigendecomposition which costs $\mathcal{O}(n^3)$.

- The argument of $\Omega_+^\star$ always has the form $\boldsymbol{V}\,\mathrm{Diag}(\boldsymbol{\alpha})\,\boldsymbol{V}^\top$ with $\boldsymbol{\alpha} \in \mathbb{R}^n$, where $\boldsymbol{V}$ is the Cholesky factor of the kernel matrix $\boldsymbol{K} = \boldsymbol{V}^\top \boldsymbol{V}$

Low-rank approximation of $\boldsymbol{V}$:

- Decompose $\boldsymbol{K}$ such that $\boldsymbol{K} = \widetilde{\boldsymbol{V}}^\top \widetilde{\boldsymbol{V}}$ with $\widetilde{\boldsymbol{V}} \in \mathbb{R}^{r \times n}$
- Substituting $\widetilde{\boldsymbol{V}}$ for $\boldsymbol{V}$, the argument of $\Omega_+^\star$ becomes $\widetilde{\boldsymbol{V}}\,\mathrm{Diag}(\boldsymbol{\alpha})\,\widetilde{\boldsymbol{V}}^\top \in \mathbb{R}^{r \times r}$, whose eigendecomposition costs only $\mathcal{O}(r^3)$

Scale up with low rank approximation



✓ Cost per iteration **drops from $\mathcal{O}(n^3)$ to $\mathcal{O}(r^3)$**, scaling up to ∼10000 samples.

Hyperparameter tuning

Main hyperparameters of our learning problem:

- $\lambda_{\text{pen}} \rightarrow$ controls the quantity of asymmetry in the bands
- $\theta_{\text{low}}, \theta_{\text{up}} \rightarrow$ kernel lengthscales associated to f_{low} and f_{up}

Hyperparameter tuning

Main hyperparameters of our learning problem:

- $\lambda_{\text{pen}} \rightarrow$ controls the quantity of asymmetry in the bands
- $\theta_{\text{low}}, \theta_{\text{up}} \rightarrow$ kernel lengthscales associated to f_{low} and f_{up}

Not a traditional supervised learning problem: which **criterion**?

Hyperparameter tuning

Main hyperparameters of our learning problem:

- $\lambda_{\text{pen}} \rightarrow$ controls the quantity of asymmetry in the bands
- $\theta_{\text{low}}, \theta_{\text{up}} \rightarrow$ kernel lengthscales associated to f_{low} and f_{up}

Not a traditional supervised learning problem: which **criterion**?

- Ideally, remember that we target adaptivity, i.e. local coverage $\geq 1 - \alpha$ for all $\boldsymbol{x}$

Hyperparameter tuning

Main hyperparameters of our learning problem:

- $\lambda_{\text{pen}} \rightarrow$ controls the quantity of asymmetry in the bands
- $\theta_{\text{low}}, \theta_{\text{up}} \rightarrow$ kernel lengthscales associated to f_{low} and f_{up}

Not a traditional supervised learning problem: which **criterion**?

- Ideally, remember that we target adaptivity, i.e. local coverage $\geq 1 - \alpha$ for all $\boldsymbol{x}$
- To do this, we show that a *relaxed* version of local coverage is controlled with

$$\textbf{HSIC}(\textbf{res}, \textbf{width}),$$

a dependence measure between residuals and interval width (this is intuitive!)

Hyperparameter tuning

Main hyperparameters of our learning problem:

- $\lambda_{\text{pen}} \rightarrow$ controls the quantity of asymmetry in the bands
- $\theta_{\text{low}}, \theta_{\text{up}} \rightarrow$ kernel lengthscales associated to f_{low} and f_{up}

Not a traditional supervised learning problem: which **criterion**?

- Ideally, remember that we target adaptivity, i.e. local coverage $\geq 1 - \alpha$ for all $\boldsymbol{x}$
- To do this, we show that a *relaxed* version of local coverage is controlled with

  $$\textbf{HSIC}(\textbf{res}, \textbf{width}),$$

  a dependence measure between residuals and interval width (this is intuitive!)
- We then select $(\lambda_{\text{pen}}, \theta_{\text{low}}, \theta_{\text{up}})$ to maximize this dependence

Table of Contents

1. Introduction
   1. Conformal prediction
   2. Kernel sum-of-squares

2. Our learning problem

3. Numerical experiments

4. Conclusion

5. References and appendix

Adaptivity Metrics

✗ Usual metrics hide systematic under/over-coverage in presence of asymmetric noise

✓ Instead, we seek bands satisfying **both lower and upper** local coverage at $1 - \alpha/2$

New adaptivity metrics: **analytical datasets**

Adaptivity Metrics

✗ Usual metrics hide systematic under/over-coverage in presence of asymmetric noise

✓ Instead, we seek bands satisfying **both lower and upper** local coverage at $1 - \alpha/2$

New adaptivity metrics: **analytical datasets**

Distance to perfect conditional coverage: Absolute Coverage Gap

$$\mathrm{ACG}^{\mathrm{low}} = \frac{1}{n_X} \sum_{i=1}^{n_X} |\hat{p}^{\mathrm{low}}(X_i) - (1 - \alpha/2)| \qquad \mathrm{ACG}^{\mathrm{upp}} = \frac{1}{n_X} \sum_{i=1}^{n_X} |\hat{p}^{\mathrm{upp}}(X_i) - (1 - \alpha/2)|$$

$\hat{p}^{\mathrm{low}}(X_i)$, $\hat{p}^{\mathrm{upp}}(X_i)$ estimated with conditional samples at X_i

Adaptivity Metrics

✗ Usual metrics hide systematic under/over-coverage in presence of asymmetric noise

✓ Instead, we seek bands satisfying **both lower and upper** local coverage at $1 - \alpha/2$

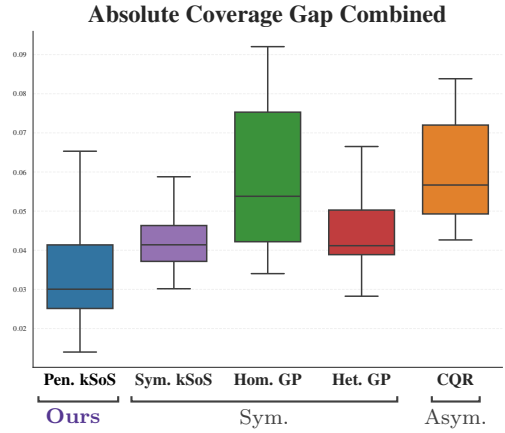
New adaptivity metrics: **analytical datasets**

Distance to perfect conditional coverage: Absolute Coverage Gap

$$\mathrm{ACG}^{\mathrm{low}} = \frac{1}{n_X} \sum_{i=1}^{n_X} |\hat{p}^{\mathrm{low}}(X_i) - (1 - \alpha/2)| \qquad \mathrm{ACG}^{\mathrm{upp}} = \frac{1}{n_X} \sum_{i=1}^{n_X} |\hat{p}^{\mathrm{upp}}(X_i) - (1 - \alpha/2)|$$

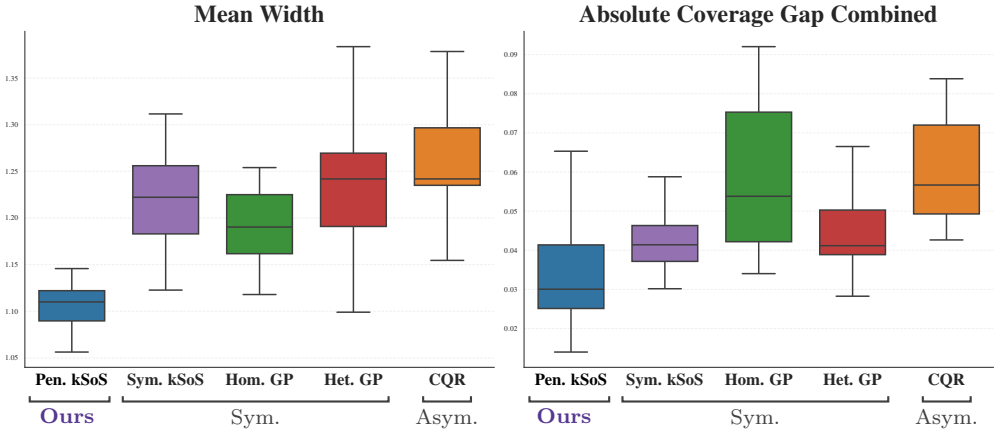
- $\mathbf{ACG}^{c} \rightarrow$ we combine low and upp variants
- Smaller is better

Asymmetric and heteroscedastic analytical dataset



✓ Penalized kSoS performs **better than or comparably to any competitor** on analytical datasets, regardless of noise structure

Asymmetric and heteroscedastic analytical dataset



✓ Penalized kSoS performs **better than or comparably to any competitor** on analytical datasets, regardless of noise structure

Adaptivity Metrics

✗ Usual metrics hide systematic under/over-coverage in presence of asymmetric noise

✓ Instead, we seek bands satisfying **both lower and upper** local coverage at $1 - \alpha/2$

Adaptivity Metrics

✗ Usual metrics hide systematic under/over-coverage in presence of asymmetric noise

✓ Instead, we seek bands satisfying **both lower and upper** local coverage at $1 - \alpha/2$

New adaptivity metrics: **real-world datasets**

Adaptivity Metrics

✗ Usual metrics hide systematic under/over-coverage in presence of asymmetric noise

✓ Instead, we seek bands satisfying **both lower and upper** local coverage at $1 - \alpha/2$

New adaptivity metrics: **real-world datasets**

Alternative to conditional coverage: Worst Set Coverage (Thurin et al. 2025)

$$\mathrm{WSC}^{\mathrm{low}} = \min_{l=1,\dots,L} \hat{p}^{\mathrm{low}}(\mathcal{R}_l) \qquad \mathrm{WSC}^{\mathrm{upp}} = \min_{l=1,\dots,L} \hat{p}^{\mathrm{upp}}(\mathcal{R}_l)$$

$\hat{p}^{\mathrm{low}}(\mathcal{R}_l)$, $\hat{p}^{\mathrm{upp}}(\mathcal{R}_l)$ estimated with samples X_i in each region $\mathcal{R}_l$ of a partition

Adaptivity Metrics

✗ Usual metrics hide systematic under/over-coverage in presence of asymmetric noise

✓ Instead, we seek bands satisfying **both lower and upper** local coverage at $1 - \alpha/2$

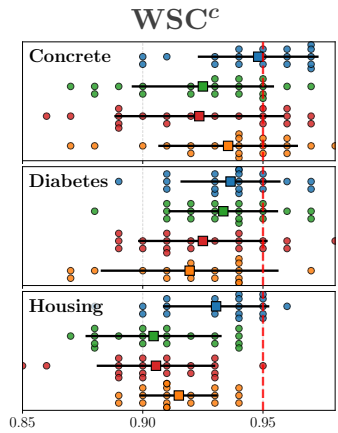
New adaptivity metrics: **real-world datasets**

Alternative to conditional coverage: Worst Set Coverage (Thurin et al. 2025)

$$\mathrm{WSC}^{\mathrm{low}} = \min_{l=1,\dots,L} \hat{p}^{\mathrm{low}}(\mathcal{R}_l) \qquad \mathrm{WSC}^{\mathrm{upp}} = \min_{l=1,\dots,L} \hat{p}^{\mathrm{upp}}(\mathcal{R}_l)$$

- $\mathbf{WSC}^c \rightarrow$ we combine low and upp variants
- Closer to the target $1 - \alpha/2$ is better

Real World datasets

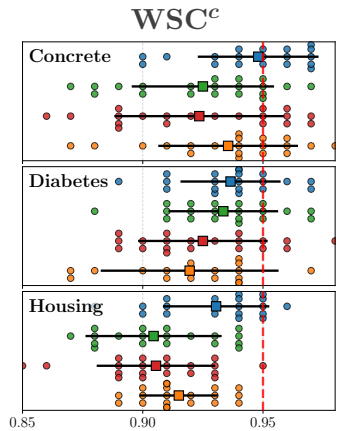


✓ On real world datasets, Pen. kSoS is always **first or close second** when measured by WSC^c and mean width, a robustness never achieved by any other method

Real World datasets

Mean width

Dataset	CQR	Het GP	Hom GP	Pen. kSoS
Concrete	22.66	21.53	**21.01**	21.74
Bike	217.98	199.23	169.57	**162.47**
Diabetes	**188.65**	196.25	193.21	**190.46**
Housing log	0.967	0.855	**0.768**	0.836
Housing	1.77	1.70	**1.60**	1.80
MPG	9.82	9.76	**9.33**	**9.32**
	Asym.	Sym.		**Ours**

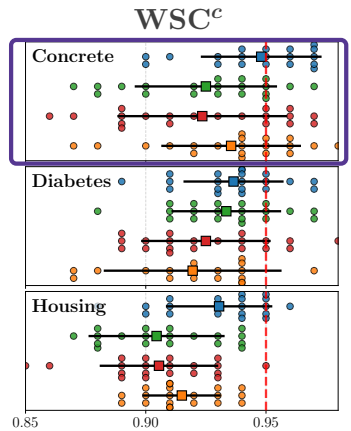


✓ On real world datasets, Pen. kSoS is always **first or close second** when measured by WSC^c and mean width, a robustness never achieved by any other method

Real World datasets

Mean width

Dataset	CQR	Het GP	Hom GP	Pen. kSoS
Concrete	22.66	21.53	**21.01**	21.74
Bike	217.98	199.23	169.57	**162.47**
Diabetes	**188.65**	196.25	193.21	**190.46**
Housing log	0.967	0.855	**0.768**	0.836
Housing	1.77	1.70	**1.60**	1.80
MPG	9.82	9.76	**9.33**	**9.32**
	Asym.	Sym.		**Ours**

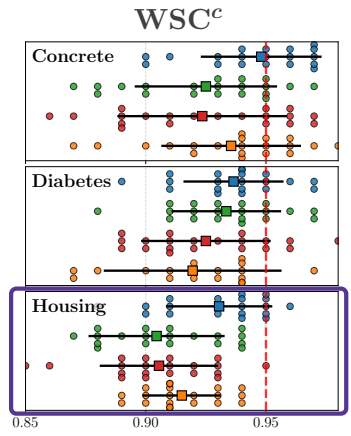


✓ On real world datasets, Pen. kSoS is always **first or close second** when measured by WSC^c and mean width, a robustness never achieved by any other method

Real World datasets

Mean width

Dataset	CQR	Het GP	Hom GP	Pen. kSoS
Concrete	22.66	21.53	**21.01**	21.74
Bike	217.98	199.23	169.57	**162.47**
Diabetes	**188.65**	196.25	193.21	**190.46**
Housing log	0.967	0.855	**0.768**	0.836
Housing	1.77	1.70	**1.60**	1.80
MPG	9.82	9.76	**9.33**	**9.32**
	Asym.	Sym.		**Ours**



✓ On real world datasets, Pen. kSoS is always **first or close second** when measured by WSC^c and mean width, a robustness never achieved by any other method

Table of Contents

1. Introduction
   1. Conformal prediction
   2. Kernel sum-of-squares

2. Our learning problem

3. Numerical experiments

4. Conclusion

5. References and appendix

Contributions

- New **continuum** between symmetric and asymmetric bands (with representer theorem, dual formulations and error bounds)
- **Data-driven** strategy, allowing the model to automatically adapt to the appropriate level of symmetry
- Based on new lower bound on conditional coverage with HSIC
- Penalization **compensates for small samples artifacts and model bias**, even when the true noise is symmetric

Paper accepted at ICML 2026
Preprint available on arXiv



Future work

- 100% coverage is a hard constraint which makes the problem convex, but is sensitive to outliers
  - We are investigating non-convex soft contraints and loss functions, with difference of convex methods
- Generalization to simultaneously learn the mean function in its own RKHS

ENSAI

Thank you for listening!

Table of Contents

1. Introduction
   1. Conformal prediction
   2. Kernel sum-of-squares

2. Our learning problem

3. Numerical experiments

4. Conclusion

5. References and appendix

References

- Allain, Louis et al. (2025). Scalable and adaptive prediction bands with kernel sum-of-squares. arXiv: 2505.21039 [stat.ML]. URL: https://arxiv.org/abs/2505.21039.
- Allain, Louis et al. (Jan. 2026). "Asymmetric Conformal Prediction with Penalized Kernel Sum-of-Squares". In.
- 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.
- Borwein, Jonathan et al. (2006). Convex Analysis and Nonlinear Optimization: Theoryand Examples. Springer.
- 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.

References

- Gretton, Arthur et al. (2005). “Measuring statistical dependence with Hilbert-Schmidt norms”. In: International conference on algorithmic learning theory. Springer, pp. 63–77.
- Marteau-Ferey, Ulysse et al. (July 8, 2020). Non-parametric Models for Non-negative Functions. DOI: 10.48550/arXiv.2007.03926.
- Muzellec, Boris et al. (2021). Learning PSD-valued functions using kernel sums-of-squares. arXiv: 2111.11306 [stat.ML]. URL: https://arxiv.org/abs/2111.11306.
- Pouplin, Thomas et al. (21–27 Jul 2024). “Relaxed Quantile Regression: Prediction Intervals for Asymmetric Noise”. In: Proceedings of the 41st International Conference on Machine Learning. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 40951–40981.

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.
- Vacher, Adrien et al. (June 2024). “Optimal Estimation of Smooth Transport Maps with Kernel SoS”. In: SIAM Journal on Mathematics of Data Science 6.2, pp. 311–342. ISSN: 2577-0187. DOI: 10.1137/22M1528847.
- 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.
- Wang, Chong Xiao et al. (2023). Semi-Nonparametric Estimation of Distribution Divergence in Non-Euclidean Spaces. URL: https://arxiv.org/abs/2204.02031.

Kernel sum-of-squares

We also consider the **elastic-net** regularization: $\Omega(\mathcal{A}) = \lambda_1 \|\mathcal{A}\|_\star + \lambda_2 \|\mathcal{A}\|_{\mathrm{F}}^2$

Assumption 2 (Marteau-Ferey et al. 2020)

Let $\mathcal{H}$ be a Hilbert space. For any $\mathcal{A} \in \mathcal{S}(\mathcal{H})$, $\Omega(\mathcal{A})$ is of the form

$$\Omega(\mathcal{A}) = \begin{cases} \mathrm{Tr}(q(\mathcal{A})) = \sum_k q(\sigma_k) & \text{if } \mathcal{A} = U \mathrm{Diag}(\sigma) U^\top \in \mathcal{S}_\infty(\mathcal{H}),\ \sum_k q(\sigma_k) < +\infty \\ +\infty & \text{otherwise,} \end{cases}$$

where $q \colon \mathbb{R} \to \mathbb{R}_+$ is non-decreasing on $\mathbb{R}_+$ with $q(0) = 0$, lower semi-continuous, and $q(\sigma) \underset{|\sigma| \to +\infty}{\longrightarrow} +\infty$.

The elastic-net regularization satisfies Assumption 1 with $q(\sigma) = \lambda_1 |\sigma| + \lambda_2 \sigma^2$.

- $\lambda_1 > 0$: lasso-type penalty, promotes **low-rank** operators
- $\lambda_2 > 0$: ridge-type penalty, makes Ω **strongly convex**

Application to the penalized problem

- **Loss function:**

$$L = \begin{cases} \frac{b}{n} \sum_{i=1}^{n} \big(f_{\mathcal{A}_{\text{low}}}(X_i) + f_{\mathcal{A}_{\text{up}}}(X_i) \big) + \lambda_{\text{pen}} \sum_{i=1}^{n} \big(f_{\mathcal{A}_{\text{low}}}(X_i) - f_{\mathcal{A}_{\text{up}}}(X_i) \big)^2 & \text{if } (\mathcal{A}_{\text{low}}, \mathcal{A}_{\text{up}}) \in C, \\ +\infty & \text{otherwise.} \end{cases}$$

where C encodes the 100% coverage constraints.
$\rightarrow L$ is **lower semi-continuous and bounded below**.

- **Regularization:** $\Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}})$, where each Ω is an **elastic-net** regularizer satisfying Assumption 1.

$\Rightarrow$ The generalized representer theorem applies: the solution takes the form

$$\big(f_{\mathcal{A}^\star_{\text{low}}}(X),\ f_{\mathcal{A}^\star_{\text{up}}}(X) \big) = \big(\mathbf{k}_{\text{low}}(X)^\top \mathbf{B}^\star_{\text{low}}\, \mathbf{k}_{\text{low}}(X),\ \mathbf{k}_{\text{up}}(X)^\top \mathbf{B}^\star_{\text{up}}\, \mathbf{k}_{\text{up}}(X) \big)$$

Error bound (training set penalty)

- The training set penalty controls $f_{\text{low}} \approx f_{\text{up}}$ only on training points $\mathcal{D}_n$. We want to control their difference **everywhere on** Ω.
- This requires additional assumptions on the RKHS $\mathcal{H}_{\text{low}} = \mathcal{H}_{\text{up}} = \mathcal{H}$ and the domain Ω.

Assumptions:

- **A.2**: For a bounded open set $\Omega \subset \mathbb{R}^d$, $\mathcal{H}$ satisfies $f|_\Omega \in \mathcal{H}$ for all $f \in C^\infty(\mathbb{R}^d)$, and $\exists\, M \geq 1$ s.t. $\|u \cdot v\|_{\mathcal{H}} \leq M \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}}$ for all $u, v \in \mathcal{H}$.
- **A.3**: For a bounded open set $\Omega \subset \mathbb{R}^d$, $\max_{|\alpha|=1} \sup_{x,y \in \Omega} |\partial_x^\alpha \partial_y^\alpha k(x,y)| \leq D^2 < \infty$ for some $D \geq 1$.
- **A.4**: $\Omega = \bigcup_{x \in S} B_r(x)$, where S is a bounded subset of $\mathbb{R}^d$ and $B_r(x)$ is the ball of center x and radius r.

The Matérn 5/2 kernel satisfies all three assumptions.

Error bound (training set penalty)

Theorem 4 (Allain et al. 2026)

Under Assumptions A.2, A.3, A.4, for any two PSD operators $\mathcal{A}_1, \mathcal{A}_2 \in \mathcal{S}_+(\mathcal{H})$*:*

$$\sup_{x\in\Omega} |f_{\mathcal{A}_1}(x) - f_{\mathcal{A}_2}(x)| \leq 2\, C_{f_1,f_2}\, \rho_{\hat{X},\Omega} + \sqrt{\sum_{i=1}^{n} (f_{\mathcal{A}_1}(X_i) - f_{\mathcal{A}_2}(X_i))^2}$$

where $C_{f_1,f_2} = 2dDM\|\mathcal{A}_1 - \mathcal{A}_2\|_\star$ *and* $\rho_{\widehat{X},\Omega} = \sup_{x\in\Omega} \min_{X_i\in\widehat{X}} \|x - X_i\|$ *is the fill-in distance.*

- **First term:** as $n \to \infty$, the **fill-in distance** $\rho_{\hat{X},\Omega} \to 0$
- **Second term:** directly controlled by the penalty $\mathcal{P}$

Hyperparameter tuning

Main hyperparameters of our learning problem:

- $\lambda_{\text{pen}} \rightarrow$ controls the quantity of asymmetry in the bands
- $\theta_{\text{low}}, \theta_{\text{up}} \rightarrow$ kernel lengthscales associated to f_{low} and f_{up}

Hyperparameter tuning

Main hyperparameters of our learning problem:

- $\lambda_{\text{pen}} \rightarrow$ controls the quantity of asymmetry in the bands
- $\theta_{\text{low}}, \theta_{\text{up}} \rightarrow$ kernel lengthscales associated to f_{low} and f_{up}

Recall our objective is to construct **adaptive** bands, which translates to **conditional coverage**:

$$\mathbb{P}(Y \in \widehat{C}(X) | X = x) \geq 1 - \alpha$$

Hyperparameter tuning

Main hyperparameters of our learning problem:

- $\lambda_{\text{pen}} \rightarrow$ controls the quantity of asymmetry in the bands
- $\theta_{\text{low}}, \theta_{\text{up}} \rightarrow$ kernel lengthscales associated to f_{low} and f_{up}

Recall our objective is to construct **adaptive** bands, which translates to **conditional coverage**:

$$\mathbb{P}(Y \in \widehat{C}(X) | X = x) \geq 1 - \alpha$$

✗ Without hypothesis on the data, satisfying this conditional coverage leads to infinitely wide prediction bands (Vovk 2012; Barber et al. 2021)

Hyperparameter tuning

We can relax the conditional coverage by considering X in a small neighbourhood ω_X, such that $\mathbb{P}(X \in \omega_X) \geq \delta$:

$$\mathbb{P}(Y \in \widehat{C}(X) | X \in \omega_X)$$

Hyperparameter tuning

We can relax the conditional coverage by considering X in a small neighbourhood ω_X, such that $\mathbb{P}(X \in \omega_X) \geq \delta$:

$$\mathbb{P}(Y \in \widehat{C}(X) | X \in \omega_X)$$

Theorem 5 (Deutschmann et al. 2024)

*If the mutual information (*MI*) between X and the score $s(X, Y)$ is finite, then for any neighbourhood ω_X such that $\mathbb{P}(X \in \omega_X) \geq \delta$:*

$$\mathbb{P}(Y \in \widehat{C}(X) | X \in \omega_X) \geq 1 - \alpha - \frac{1}{\delta}\sqrt{1 - e^{-\mathrm{MI}(X,s)}}$$

Hyperparameter tuning

We can relax the conditional coverage by considering X in a small neighbourhood ω_X, such that $\mathbb{P}(X \in \omega_X) \geq \delta$:

$$\mathbb{P}(Y \in \widehat{C}(X) | X \in \omega_X)$$

Theorem 5 (Deutschmann et al. 2024)

*If the mutual information (*MI*) between X and the score $s(X, Y)$ is finite, then for any neighbourhood ω_X such that $\mathbb{P}(X \in \omega_X) \geq \delta$:*

$$\mathbb{P}(Y \in \widehat{C}(X) | X \in \omega_X) \geq 1 - \alpha - \frac{1}{\delta}\sqrt{1 - e^{-\mathrm{MI}(X,s)}}$$

In theory, gives a principled way to optimize hyperparameters of the score by minimizing MI

Hyperparameter tuning

- Unfortunately
  - MI is **numerically instable**
  - X is potentially **high-dimensional**

- Our new bound:
  - Involves 1**-dimension variables** only
  - Replaces MI with HSIC (Hilbert-Shmidt Independence Criterion, Gretton et al. 2005), which is **numerically stable**

Hyperparameter tuning

Using asym. scores $s(X, Y) = \max(l(X) - Y, Y - u(X))$:

Theorem 6 (Allain et al. 2026)

If the interval $\widehat{C}$ is built with a split CP procedure using asymmetric scores, then

$$\mathbb{P}(Y \in \widehat{C}(X) | X \in \omega_X) \geq 1 - \alpha - \frac{1}{\delta}\sqrt{1 - \frac{\alpha_1}{1 - \alpha_2 \mathbf{HSIC}(\mathbf{res}, \mathbf{width})}}$$

with

- $\mathbf{res} = |Y - (u(X) + l(X))/2|$: ***centered residuals***
- $\mathbf{width} = (u(X) - l(X))/2$: ***interval length***

✓ HSIC is **numerically stable** ✓ only 1-**dimensional** variables

Elements of proof

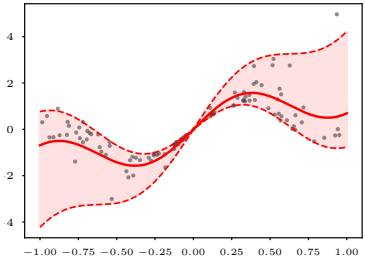
1. **Rewrite the score.** The score function can be written as a function of the width and the centered residuals
$$s(X, Y) = -\frac{W}{2} + R$$
2. **Information theory results.** Transform the d-dimensional bound into a 1-dimensional one
$$\mathrm{MI}(X, s) \leq \mathrm{H}(X, Y) + \mathrm{H}(X) - \mathrm{MI}(W, R)$$
3. **Relationships between distances of distributions.**
   - Bretagnolle-Huber inequality (Kullback-Leibler and Total Variation)
   - Wang et al. 2023 inequality (Total Variation and Maximum Mean Discrepancy)

$$\sqrt{1 - e^{-\mathrm{KL}(\mathbb{P}, \mathbb{Q})}} \geq \mathrm{TV}(\mathbb{P}, \mathbb{Q}) \geq \frac{1}{2\sqrt{M_k}} \mathrm{MMD}_k(\mathbb{P}, \mathbb{Q})$$

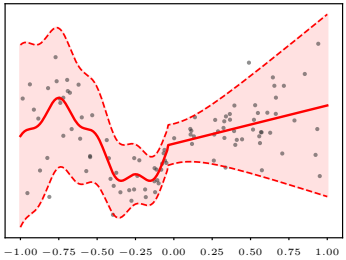
Hyperparameter tuning

From Theorem 6, we can target conditional coverage by **maximizing** HSIC(**res**, **width**).

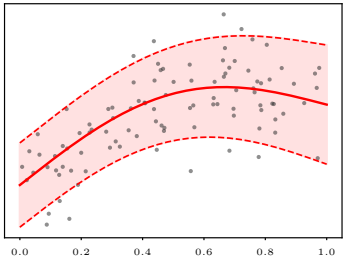
Hyperparameter tuning



(a) asymmetric noise

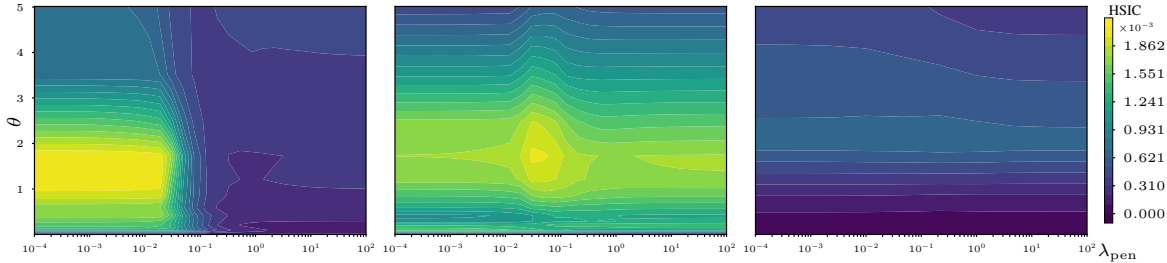


(b) symmetric noise



(c) symmetric noise

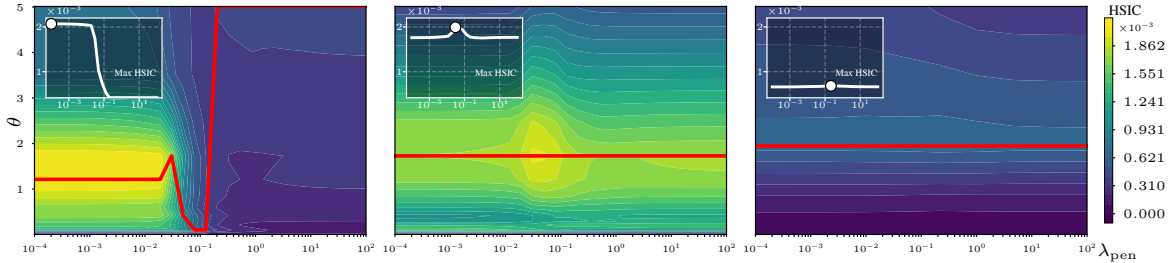
Hyperparameter tuning



(a) asymmetric noise **(b)** symmetric noise **(c)** symmetric noise

Hyperparameter tuning

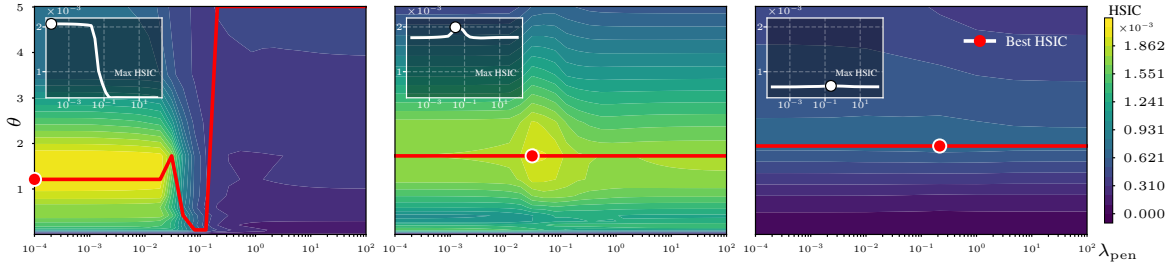
1. For each λ_{pen} we identify the best $(\theta_{\text{low}}, \theta_{\text{up}})$ with HSIC (—)



(a) asymmetric noise **(b)** symmetric noise **(c)** symmetric noise

Hyperparameter tuning

1. For each λ_{pen} we identify the best $(\theta_{\text{low}}, \theta_{\text{up}})$ with HSIC (—)
2. We select λ_{pen} with best opt. HSIC (•)



(a) asymmetric noise **(b)** symmetric noise **(c)** symmetric noise

Hyperparameter tuning

From Theorem 6, we can target conditional coverage by **maximizing** HSIC(**res**, **width**).

Two scenarios can arise when estimating the HSIC:

Hyperparameter tuning

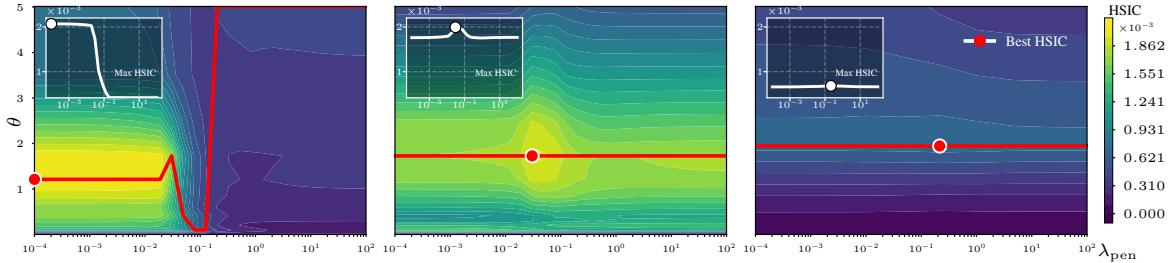
From Theorem 6, we can target conditional coverage by **maximizing** HSIC(**res**, **width**).

Two scenarios can arise when estimating the HSIC:

1. **Are the variations in HSIC significant between different λ_{pen}?**
   Kruskal-Wallis rank test:
   If not significantly different across λ_{pen}, we default to the symmetric model.

Hyperparameter tuning

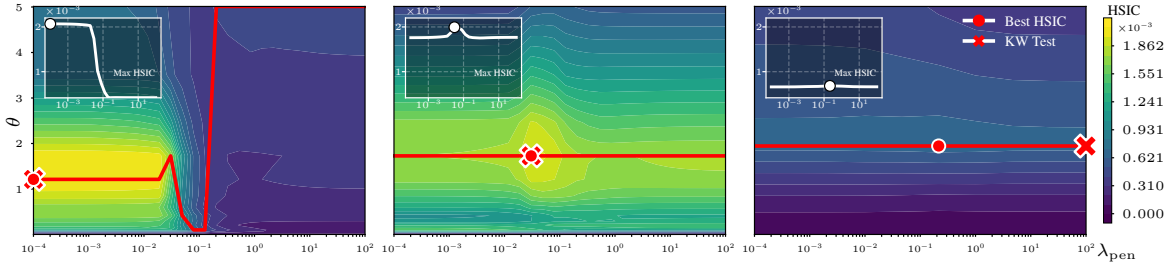
1. For each λ_{pen} we identify the best $(\theta_{\text{low}}, \theta_{\text{up}})$ with HSIC (—)
2. We select λ_{pen} with best opt. HSIC (•)



(a) asymmetric noise **(b)** symmetric noise **(c)** symmetric noise

Hyperparameter tuning

1. For each λ_{pen} we identify the best $(\theta_{\text{low}}, \theta_{\text{up}})$ with HSIC (—)
2. We select λ_{pen} with best opt. HSIC (•)
3. Except if we detect no significant difference in opt. HSIC along λ_{pen} (KW test), in which case we default to the simpler symmetric model (large λ_{pen}) (✕)



(a) asymmetric noise **(b)** symmetric noise **(c)** symmetric noise

Hyperparameter tuning

From Theorem 6, we can target conditional coverage by **maximizing** HSIC(**res**, **width**).

Two scenarios can arise when estimating the HSIC:

1. **Are the variations in HSIC significant between different λ_{pen}?**
   Kruskal-Wallis rank test:
   If not significantly different across λ_{pen}, we default to the symmetric model.

Hyperparameter tuning

From Theorem 6, we can target conditional coverage by **maximizing** HSIC(**res**, **width**).

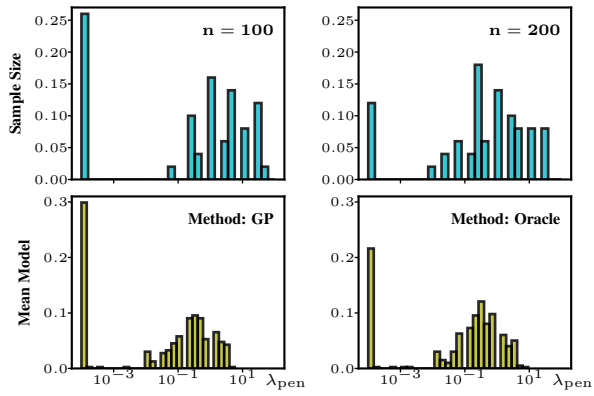
Two scenarios can arise when estimating the HSIC:

1. **Are the variations in HSIC significant between different λ_{pen}?**
   Kruskal-Wallis rank test:

   If not significantly different across λ_{pen}, we default to the symmetric model.
2. **Are the HSIC values significantly different from 0?**
   Permutation independence test Gretton et al. 2005:

   If the values are not significantly different from 0, we default to the simpler homoscedastic model.

Impact of the penalty



✓ The asymmetric continuum allows Pen. kSoS to **mitigate** small samples artifacts and biased predictive models

Adaptivity Metrics

Usual adaptivity metrics are:

Absolute Coverage Gap:

$$\mathrm{ACG} = \frac{1}{n_X} \sum_{i=1}^{n_X} \left| \hat{p}(X_i) - (1-\alpha) \right|$$

with $\hat{p}$ an estimate of the local coverage

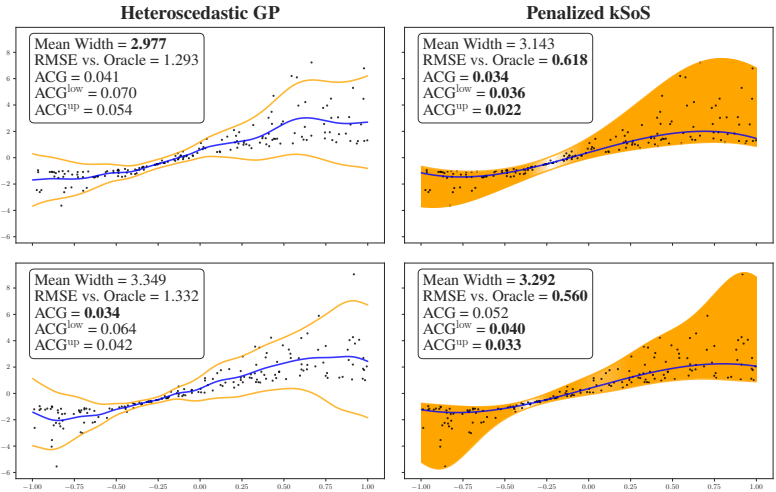
Worst Set Coverage:

$$\mathrm{WSC} = \min_{l=1,\dots,L} \mathbb{P}\left(Y \in \widehat{C}_{\mathcal{D}_N}(X) \mid X \in \mathcal{R}_l\right)$$

with $\{\mathcal{R}_l\}_{l=1,\dots,L}$ a partition of Ω

In the presence of asymmetric noise, those metrics perform poorly

Adaptivity Metrics



✗ Mean width alone is **misleading**

✗ Local coverage is **not meaningful**