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

Louis Allain^{1,2} Sébastien Da Veiga² Brian Staber¹

¹Safran Tech, Digital Sciences & Technologies, 78114 Magny-Les-Hameaux, France

²Univ Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France

{louis.allain, brian.staber}@safrangroup.com

sebastien.da-veiga@ensai.fr

A Proofs

A.1 Proof of Theorem 2 - representer theorem

To begin, we first introduce some notation. When considering objects related to functions m and f , we will use the superscripts m and f to differentiate them. Associated to each function we consider a RKHS $\mathcal{H}^{(\cdot)}$, a kernel $k^{(\cdot)}$, a kernel matrix $\mathbf{K}^{(\cdot)}$, a feature map $\phi^{(\cdot)}$ and a column vector such that for $x \in \mathcal{X}$

$$\mathbf{k}^{(\cdot)}(X) = \left(k^{(\cdot)}(X_1, X), \dots, k^{(\cdot)}(X_n, X) \right)^T.$$

For the kernel matrix $[\mathbf{K}^f]_{ij} = k^f(X_i, X_j)$ only, we consider its Cholesky decomposition and empirical feature map

$$\mathbf{K}^f = \mathbf{V}^T \mathbf{V} \quad \text{and} \quad \Phi(X) = \mathbf{V}^{-T} \mathbf{k}^f(X).$$

Next, for a Hilbert space \mathcal{H} we write $\mathcal{S}(\mathcal{H})$ the set of bounded Hermitian linear operators from \mathcal{H} to \mathcal{H} and $\mathcal{S}_+(\mathcal{H})$ those that are positive-definite. We also write $\mathbb{S}_+^n = \mathbb{S}_+(\mathbb{R}^{n \times n})$ the set of real, symmetric and positive-definite square matrices of size n .

To prove our representer theorem, we first rewrite (4) as

$$\begin{aligned} \inf_{m \in \mathcal{H}^m, \mathcal{A} \in \mathcal{S}_+(\mathcal{H}^f)} \quad & a \cdot \frac{1}{n} \sum_{i=1}^n (Y_i - m(X_i))^2 + b \cdot \frac{1}{n} \sum_{i=1}^n f_{\mathcal{A}}(X_i) + \lambda_1 \|\mathcal{A}\|_* + \lambda_2 \|\mathcal{A}\|_F^2 \quad (13) \\ \text{s.t.} \quad & (Y_i - m(X_i))^2 - f_{\mathcal{A}}(X_i) \leq 0, \quad i \in [n], \\ & \|m\|_{\mathcal{H}^m}^2 - s \leq 0. \end{aligned}$$

In order to show that there exists a finite-dimensional representation for both m^* and \mathcal{A}^* , we will first show that for any fixed $m \in \mathcal{H}^m$, the optimal \mathcal{A} has a finite-dimensional representation. Indeed if $m \in \mathcal{H}^m$ is fixed, the problem writes

$$\begin{aligned} \inf_{\mathcal{A} \in \mathcal{S}_+(\mathcal{H}^f)} \quad & \frac{b}{n} \sum_{i=1}^n f_{\mathcal{A}}(X_i) + \lambda_1 \|\mathcal{A}\|_* + \lambda_2 \|\mathcal{A}\|_F^2 \\ \text{s.t.} \quad & (Y_i - m(X_i))^2 - f_{\mathcal{A}}(X_i) \leq 0, \quad i \in [n] \end{aligned}$$

or equivalently

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

where

$$L_m(f_{\mathcal{A}}(X_1), \dots, f_{\mathcal{A}}(X_n)) = \begin{cases} \frac{b}{n} \sum_{i=1}^n f_{\mathcal{A}}(X_i) & \text{if } (Y_i - m(X_i))^2 - f_{\mathcal{A}}(X_i) \leq 0, i \in [n], \\ +\infty & \text{else,} \end{cases}$$

and

$$\Omega(\mathcal{A}) = \lambda_1 \|\mathcal{A}\|_{\star} + \lambda_2 \|\mathcal{A}\|_F^2.$$

Since $L_m : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous and bounded below (notice that it is linear and bounded below by 0), we can apply Theorem 1 and Proposition 3 from Marteau-Ferey et al. [2020] to deduce that the solution is entirely characterized by a PSD matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ given by

$$\begin{aligned} \inf_{\mathbf{A} \in \mathbb{S}_+^n} \quad & \frac{b}{n} \sum_{i=1}^n \tilde{f}_{\mathbf{A}}(X_i) + \lambda_1 \|\mathbf{A}\|_{\star} + \lambda_2 \|\mathbf{A}\|_F^2 \\ \text{s.t.} \quad & (Y_i - m(X_i))^2 - \tilde{f}_{\mathbf{A}}(X_i) \leq 0, i \in [n], \end{aligned}$$

with

$$\tilde{f}_{\mathbf{A}}(X_i) = \langle \Phi(X_i), \mathbf{A} \Phi(X_i) \rangle \quad \text{and} \quad \Phi(X_i) = \mathbf{V}^{-T} \mathbf{k}^f(X_i).$$

Since this representation is valid for any $m \in \mathcal{H}^m$, Problem (13) is thus equivalent to

$$\begin{aligned} \inf_{m \in \mathcal{H}^m, \mathbf{A} \in \mathbb{S}_+^n} \quad & \frac{a}{n} \sum_{i=1}^n (Y_i - m(X_i))^2 + \frac{b}{n} \sum_{i=1}^n \tilde{f}_{\mathbf{A}}(X_i) + \lambda_1 \|\mathbf{A}\|_{\star} + \lambda_2 \|\mathbf{A}\|_F^2 \\ \text{s.t.} \quad & (Y_i - m(X_i))^2 - \tilde{f}_{\mathbf{A}}(X_i) \leq 0, i \in [n], \\ & \|m\|_{\mathcal{H}^m}^2 - s \leq 0. \end{aligned} \tag{14}$$

Now, to show that $m^* \in \mathcal{H}^m$ has a finite-dimensional representation, we will rely upon the dual problem. To do so, we first need to show that strong duality holds. Since our problem is convex, it is enough to check Slater's constraint qualification, *i.e.* we simply need to exhibit a strictly feasible point (m_0, \mathbf{A}_0) . We first take $m_0 = 0 \in \mathcal{H}^m$ which satisfies $\|m_0\|_{\mathcal{H}^m}^2 - s = -s < 0$. If we assume that \mathbf{K}^f is of full rank (this assumption is always satisfied if $k^{(f)}$ is universal and all training points X_i are distinct), it is invertible and we can define $\alpha = (\mathbf{K}^f)^{-1} \mathbf{Y}$ such that $Y_i = \sum_{j=1}^n \alpha_j k^{(f)}(X_i, X_j)$. Denote $\mathbf{B}_0 \succ 0$ the **positive-definite** matrix with elements $[\mathbf{B}_0]_{ii} = \alpha_i^2 + \epsilon \mathbb{1}_{\alpha_i=0}$ and $[\mathbf{B}_0]_{ij} = 0$ if $i \neq j$ where $\epsilon > 0$ (note that in general all α_i will be non-zeros, but if it is not the case we introduce a small ϵ). Then by Cauchy-Schwartz we have for all $i = 1, \dots, n$

$$\begin{aligned} Y_i^2 &= \left(\sum_{j=1}^n \alpha_j k^{(f)}(X_i, X_j) \right)^2 \\ &\leq n \sum_{j=1}^n \alpha_j^2 (k^{(f)}(X_i, X_j))^2 \\ &\leq n \sum_{j=1}^n (\alpha_j^2 + \epsilon \mathbb{1}_{\alpha_j=0}) (k^{(f)}(X_i, X_j))^2 \\ &= n \sum_{j=1}^n [\mathbf{B}_0]_{jj} (k^{(f)}(X_i, X_j))^2 = n \tilde{f}_{\mathbf{A}_0}(X_i) \end{aligned}$$

with $\mathbf{A}_0 = \mathbf{V} \mathbf{B}_0 \mathbf{V}^T \succ 0$, which implies $(Y_i - m_0(X_i))^2 = Y_i^2 < \tilde{f}_{\mathbf{A}_0}(X_i)$ as soon as $n \geq 1$. Consequently (m_0, \mathbf{A}_0) is a strictly feasible point and strong duality holds. We thus introduce $\Gamma \in \mathbb{R}_+^n$ the Lagrangian multipliers associated to the first n constraints and $\theta \in \mathbb{R}_+$ the Lagrangian

multiplier associated to the constraint on the norm of m . The Lagrangian function writes:

$$\begin{aligned}
\mathcal{L}(m, \mathbf{A}, \mathbf{\Gamma}, \theta) &= \frac{a}{n} \sum_{i=1}^n (Y_i - m(X_i))^2 + \frac{b}{n} \sum_{i=1}^n \tilde{f}_{\mathbf{A}}(X_i) + \lambda_1 \|\mathbf{A}\|_{\star} + \lambda_2 \|\mathbf{A}\|_F^2 \\
&\quad + \sum_{i=1}^n \Gamma_i \left[(Y_i - m(X_i))^2 - \tilde{f}_{\mathbf{A}}(X_i) \right] + \theta (\|m\|_{\mathcal{H}^m}^2 - s) \\
&= \sum_{i=1}^n \left(\frac{a}{n} + \Gamma_i \right) (Y_i - m(X_i))^2 + \theta (\|m\|_{\mathcal{H}^m}^2 - s) + \lambda_1 \|\mathbf{A}\|_{\star} + \lambda_2 \|\mathbf{A}\|_F^2 \\
&\quad - \sum_{i=1}^n \left(\Gamma_i - \frac{b}{n} \right) \tilde{f}_{\mathbf{A}}(X_i).
\end{aligned}$$

Following Muzellec et al. [2022] Appendix C.5, the optimality conditions of the Lagrangian function w.r.t. m are

$$\begin{aligned}
\nabla_m \mathcal{L}(m, \mathbf{A}, \mathbf{\Gamma}, \theta) &= \sum_{i=1}^n 2 \left(\Gamma_i + \frac{a}{n} \right) k^m(X_i, \cdot) (\langle m(\cdot), k^m(X_i, \cdot) \rangle_{\mathcal{H}^m} - Y_i) + 2\theta m(\cdot) \\
&= 2 \sum_{i=1}^n \left(\Gamma_i + \frac{a}{n} \right) [k^m(X_i, \cdot) \otimes k^m(X_i, \cdot)](m) + 2\theta m(\cdot) \\
&\quad - 2 \sum_{i=1}^n \left(\Gamma_i + \frac{a}{n} \right) Y_i k^m(X_i, \cdot)
\end{aligned}$$

and setting this gradient to 0 leads to

$$\left[\sum_{i=1}^n \left(\Gamma_i + \frac{a}{n} \right) (k^m(X_i, \cdot) \otimes k^m(X_i, \cdot)) + \theta \mathbf{I}_n \right] (m) = \sum_{i=1}^n \left(\Gamma_i + \frac{a}{n} \right) Y_i k^m(X_i, \cdot).$$

To conclude, we denote

$$\begin{aligned}
\mathbf{C}(\mathbf{\Gamma}, \theta) &:= \text{Diag}(\mathbf{\Gamma}_{\mathbf{a}}) \mathbf{K}^m + \theta \mathbf{I}_n, \\
\boldsymbol{\gamma}(\mathbf{\Gamma}, \theta) &:= \mathbf{C}(\mathbf{\Gamma}, \theta)^{-1} \text{Diag}(\mathbf{\Gamma}_{\mathbf{a}}) \mathbf{Y}, \\
\text{Diag}(\mathbf{\Gamma}_{\mathbf{a}}) &= \text{Diag}(\mathbf{\Gamma}) + \frac{a}{n} \mathbf{I}_n
\end{aligned}$$

such that m^* has the finite-dimensional representation

$$m^*(X) = \sum_{i=1}^n \gamma_i k^m(X_i, X) = \boldsymbol{\gamma}^T \mathbf{k}^m(X).$$

In the end, plugging this expression in Problem (14) yields the semi-definite problem from Equation (7):

$$\begin{aligned}
&\inf_{\boldsymbol{\gamma} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{S}_+^n} \quad \frac{a}{n} \sum_{i=1}^n (Y_i - m(X_i))^2 + \frac{b}{n} \sum_{i=1}^n \tilde{f}_{\mathbf{A}}(X_i) + \lambda_1 \|\mathbf{A}\|_{\star} + \lambda_2 \|\mathbf{A}\|_F^2 \\
&\text{s.t.} \quad (Y_i - \boldsymbol{\gamma}^T \mathbf{k}^m(X_i))^2 - \tilde{f}_{\mathbf{A}}(X_i) \leq 0, \quad i \in [n], \\
&\quad \boldsymbol{\gamma}^T \mathbf{K}^m \boldsymbol{\gamma} - s \leq 0.
\end{aligned}$$

A.2 Proof of Proposition 2 - dual formulation

Proof. The dual problem of Equation (7) is defined as

$$d = \sup_{\substack{\mathbf{\Gamma} \in \mathbb{R}_+^n \\ \theta \in \mathbb{R}_+}} \inf_{\substack{m \in \mathcal{H}^m \\ \mathbf{A} \in \mathbb{S}_+^n}} \mathcal{L}(m, \mathbf{A}, \mathbf{\Gamma}, \theta) = \sup_{\substack{\mathbf{\Gamma} \in \mathbb{R}_+^n \\ \theta \in \mathbb{R}_+}} D(\mathbf{\Gamma}, \theta)$$

where the dual function is $D(\Gamma, \theta) := \inf_{m \in \mathcal{H}^m, \mathbf{A} \in \mathbb{S}_+^n} \mathcal{L}(m, \mathbf{A}, \Gamma, \theta)$. Remark first that in the previous section we already introduced the Lagrangian function

$$\begin{aligned} \mathcal{L}(m, \mathbf{A}, \Gamma, \theta) &= \sum_{i=1}^n \left(\frac{a}{n} + \Gamma_i \right) (Y_i - m(X_i))^2 + \theta (\|m\|_{\mathcal{H}^m}^2 - s) + \lambda_1 \|\mathbf{A}\|_* + \lambda_2 \|\mathbf{A}\|_F^2 \\ &\quad - \sum_{i=1}^n \left(\Gamma_i - \frac{b}{n} \right) \tilde{f}_{\mathbf{A}}(X_i), \end{aligned}$$

with optimality condition for m given by $m^*(X) = \gamma^T \mathbf{k}^m(X)$. Now, we need to derive the optimality conditions for \mathbf{A} :

$$\begin{aligned} &\inf_{\mathbf{A} \in \mathbb{S}_+^n} \lambda_1 \|\mathbf{A}\|_* + \lambda_2 \|\mathbf{A}\|_F^2 - \sum_{i=1}^n \left(\Gamma_i - \frac{b}{n} \right) \tilde{f}_{\mathbf{A}}(X_i) \\ &= \inf_{\mathbf{A} \in \mathbb{S}_+^n} \lambda_1 \|\mathbf{A}\|_* + \lambda_2 \|\mathbf{A}\|_F^2 - \langle \mathbf{A}, \mathbf{V} \text{Diag}(\Gamma_{-\mathbf{b}}) \mathbf{V}^T \rangle \end{aligned} \quad (15)$$

$$\begin{aligned} &= - \sup_{\mathbf{A} \in \mathbb{S}_+^n} \langle \mathbf{A}, \mathbf{V} \text{Diag}(\Gamma_{-\mathbf{b}}) \mathbf{V}^T \rangle - \lambda_1 \|\mathbf{A}\|_* - \lambda_2 \|\mathbf{A}\|_F^2 \\ &= - \Omega^* (\mathbf{V} \text{Diag}(\Gamma_{-\mathbf{b}}) \mathbf{V}^T) \end{aligned} \quad (16)$$

where Ω^* is the Fenchel conjugate of Ω . Equality (15) comes from the fact that $\sum_{i=1}^n \Gamma_i \tilde{f}_{\mathbf{A}}(X_i) = \langle \mathbf{A}, \mathbf{V} \text{Diag}(\Gamma) \mathbf{V}^T \rangle$. Indeed, recall that $\tilde{f}_{\mathbf{A}}(X_i) = \langle \Phi(X_i), \mathbf{A} \Phi(X_i) \rangle$ and $\Phi(X_i) = \mathbf{V}^{-T} \mathbf{k}^f(X_i)$, which yields

$$\begin{aligned} \sum_{i=1}^n \Gamma_i \tilde{f}_{\mathbf{A}}(X_i) &= \sum_{i=1}^n \Gamma_i \langle \Phi(X_i), \mathbf{A} \Phi(X_i) \rangle = \sum_{i=1}^n \Gamma_i \text{Tr}(\Phi(X_i)^T \mathbf{A} \Phi(X_i)) \\ &= \sum_{i=1}^n \Gamma_i \text{Tr}(\mathbf{A} \Phi(X_i) \Phi(X_i)^T) = \sum_{i=1}^n \Gamma_i \langle \mathbf{A}, \Phi(X_i) \Phi(X_i)^T \rangle \\ &= \langle \mathbf{A}, \sum_{i=1}^n \Gamma_i \Phi(X_i) \Phi(X_i)^T \rangle \end{aligned}$$

where the second term inside the brackets can be expressed as

$$\begin{aligned} \sum_{i=1}^n \Gamma_i \Phi(X_i) \Phi(X_i)^T &= \sum_{i=1}^n \Gamma_i \mathbf{V}^{-T} \mathbf{k}^f(X_i) \mathbf{k}^f(X_i)^T \mathbf{V}^{-1} \\ &= \mathbf{V}^{-T} \left[\sum_{i=1}^n \Gamma_i \mathbf{k}^f(X_i) \mathbf{k}^f(X_i)^T \right] \mathbf{V}^{-1} \\ &= \mathbf{V}^{-T} \mathbf{K}^f \text{Diag}(\Gamma) (\mathbf{K}^f)^T \mathbf{V}^{-1} \\ &= \mathbf{V} \text{Diag}(\Gamma) \mathbf{V}^T. \end{aligned}$$

Equation (16) is simply the definition of the Fenchel conjugate of $\Omega(\mathbf{A}) = \lambda_1 \|\mathbf{A}\|_* + \lambda_2 \|\mathbf{A}\|_F^2$. Finally, replacing m by its optimal value and the regularizations involving \mathbf{A} by the above expression, the dual function is

$$\begin{aligned} D(\Gamma, \theta) &= \inf_{\substack{m \in \mathcal{H}^m \\ \mathbf{A} \in \mathbb{S}_+^n}} \mathcal{L}(m, \mathbf{A}, \Gamma, \theta) \\ &= \underbrace{\mathbf{r}(\Gamma, \theta)^T \text{Diag}(\Gamma_{\mathbf{a}}) \mathbf{r}(\Gamma, \theta)}_{\text{first term}} + \underbrace{\theta(\gamma(\Gamma, \theta)^T \mathbf{K}^m \gamma(\Gamma, \theta) - s)}_{\text{second term}} - \underbrace{\Omega^*(\mathbf{V} \text{Diag}(\Gamma_{-\mathbf{b}}) \mathbf{V}^T)}_{\text{third term}} \end{aligned}$$

where $\mathbf{r}(\Gamma, \theta) := \mathbf{Y} - \mathbf{K}^m \gamma(\Gamma, \theta)$, which corresponds to Proposition 2.

Gradient computation. To solve the dual problem

$$d = \sup_{\substack{\Gamma \in \mathbb{R}_+^n \\ \theta \in \mathbb{R}_+}} D(\Gamma, \theta)$$

we propose to use an accelerated gradient algorithm, which requires the gradient of the dual function w.r.t. the Lagrange multipliers. We now provide the explicit computations for this gradient. In what follows, \mathbf{J}_{jj} denotes the matrix filled with zeros except for a 1 at row j and column j .

Gradient of first term $\mathbf{r}^\top \text{Diag}(\Gamma_{\mathbf{a}})\mathbf{r}$.

It is straightforward to show that

$$\begin{aligned} \frac{\partial \mathbf{C}}{\partial \Gamma_j} &= \frac{\partial \text{Diag}(\Gamma_{\mathbf{a}})}{\partial \Gamma_j} \mathbf{K}^m = \mathbf{J}_{jj} \mathbf{K}^m, \\ \frac{\partial \mathbf{C}^{-1}}{\partial \Gamma_j} &= -\mathbf{C}^{-1} \frac{\partial \mathbf{C}}{\partial \Gamma_j} \mathbf{C}^{-1} = -\mathbf{C}^{-1} \mathbf{J}_{jj} \mathbf{K}^m \mathbf{C}^{-1}, \\ \frac{\partial \gamma}{\partial \Gamma_j} &= \frac{\partial \mathbf{C}^{-1}}{\partial \Gamma_j} \text{Diag}(\Gamma_{\mathbf{a}}) \mathbf{Y} + \mathbf{C}^{-1} \frac{\partial \text{Diag}(\Gamma)}{\partial \Gamma_j} \mathbf{Y} \\ &= \mathbf{C}^{-1} \mathbf{J}_{jj} [-\mathbf{K}^m \mathbf{C}^{-1} \text{Diag}(\Gamma_{\mathbf{a}}) + \mathbf{I}_n] \mathbf{Y} \\ &= \mathbf{C}^{-1} \mathbf{J}_{jj} \mathbf{u}, \\ \frac{\partial \mathbf{r}}{\partial \Gamma_j} &= -\mathbf{K}^m \frac{\partial \gamma}{\partial \Gamma_j} = -\mathbf{K}^m \mathbf{C}^{-1} \mathbf{J}_{jj} \mathbf{u} \end{aligned}$$

where $\mathbf{u} := [-\mathbf{K}^m \mathbf{C}^{-1} \text{Diag}(\Gamma_{\mathbf{a}}) + \mathbf{I}_n] \mathbf{Y}$. We then get

$$\begin{aligned} \frac{\partial \mathbf{r}^\top \text{Diag}(\Gamma_{\mathbf{a}}) \mathbf{r}}{\partial \Gamma_j} &= r_j^2 + 2\mathbf{r}^\top \text{Diag}(\Gamma_{\mathbf{a}}) \frac{\partial \mathbf{r}}{\partial \Gamma_j} \\ &= r_j^2 - 2\mathbf{r}^\top \text{Diag}(\Gamma_{\mathbf{a}}) \mathbf{K}^m \mathbf{C}^{-1} \mathbf{J}_{jj} \mathbf{u} \\ &= r_j^2 - 2\mathbf{s}^\top \mathbf{J}_{jj} \mathbf{u}, \\ \frac{\partial \mathbf{r}^\top \text{Diag}(\Gamma_{\mathbf{a}}) \mathbf{r}}{\partial \Gamma} &= \mathbf{r}^{\odot 2} - 2\mathbf{s} \odot \mathbf{u} \end{aligned}$$

where $\mathbf{s} := \mathbf{r}^\top \text{Diag}(\Gamma_{\mathbf{a}}) \mathbf{K}^m \mathbf{C}^{-1}$. Similarly,

$$\begin{aligned} \frac{\partial \mathbf{C}}{\partial \theta} &= \mathbf{I}_n, \\ \frac{\partial \mathbf{C}^{-1}}{\partial \theta} &= -\mathbf{C}^{-1} \frac{\partial \mathbf{C}}{\partial \theta} \mathbf{C}^{-1} = -\mathbf{C}^{-2}, \\ \frac{\partial \gamma}{\partial \theta} &= \frac{\partial \mathbf{C}^{-1}}{\partial \theta} \text{Diag}(\Gamma_{\mathbf{a}}) \mathbf{Y} = -\mathbf{C}^{-2} \text{Diag}(\Gamma_{\mathbf{a}}) \mathbf{Y}, \\ \frac{\partial \mathbf{r}}{\partial \theta} &= -\mathbf{K}^m \frac{\partial \gamma}{\partial \theta} = \mathbf{K}^m \mathbf{C}^{-2} \text{Diag}(\Gamma_{\mathbf{a}}) \mathbf{Y} := \mathbf{t}, \end{aligned}$$

such that

$$\begin{aligned} \frac{\partial \mathbf{r}^\top \text{Diag}(\Gamma_{\mathbf{a}}) \mathbf{r}}{\partial \theta} &= 2\mathbf{r}^\top \text{Diag}(\Gamma_{\mathbf{a}}) \frac{\partial \mathbf{r}}{\partial \theta} \\ &= 2\mathbf{r}^\top \text{Diag}(\Gamma_{\mathbf{a}}) \mathbf{t}. \end{aligned}$$

Gradient of second term $\theta(\gamma^\top \mathbf{K}^m \gamma - s)$.

$$\begin{aligned} \frac{\partial \theta(\gamma^\top \mathbf{K}^m \gamma - s)}{\partial \Gamma_j} &= 2\theta \gamma^\top \mathbf{K}^m \frac{\partial \gamma}{\partial \Gamma_j} \\ &= 2\theta \gamma^\top \mathbf{K}^m \mathbf{C}^{-1} \mathbf{J}_{jj} \mathbf{u}, \\ &= 2\theta \mathbf{p}^\top \mathbf{J}_{jj} \mathbf{u} \\ \frac{\partial \theta(\gamma^\top \mathbf{K}^m \gamma - s)}{\partial \Gamma} &= 2\theta \mathbf{p} \odot \mathbf{u} \end{aligned}$$

where $\mathbf{p} := \gamma^\top \mathbf{K}^m \mathbf{C}^{-1}$. We also have

$$\begin{aligned} \frac{\partial \theta (\gamma^\top \mathbf{K}^m \gamma - s)}{\partial \theta} &= (\gamma^\top \mathbf{K}^m \gamma - s) + 2\theta \gamma^\top \mathbf{K}^m \frac{\partial \gamma}{\partial \theta} \\ &= (\gamma^\top \mathbf{K}^m \gamma - s) - 2\theta \gamma^\top \mathbf{K}^m \mathbf{C}^{-2} \text{Diag}(\Gamma_{\mathbf{a}}) \mathbf{Y} \\ &= (\gamma^\top \mathbf{K}^m \gamma - s) - 2\theta \gamma^\top \mathbf{t}. \end{aligned}$$

Gradient of third term $\Omega^*(\mathbf{V} \text{Diag}(\Gamma_{-\mathbf{b}}) \mathbf{V}^T)$.

From Marteau-Ferey et al. [2020] Lemma 5, we have

$$\nabla \Omega^*(\mathbf{A}^*) = \frac{1}{2\lambda_2} [\mathbf{A}^* - \lambda_1 \mathbf{I}_n]_+.$$

We then use the chain rule following Equation (137) from Petersen and Pedersen [2012]:

$$\begin{aligned} \frac{\partial \Omega^*(\mathbf{V} \text{Diag}(\Gamma_{-\mathbf{b}}) \mathbf{V}^T)}{\partial \Gamma_j} &= \text{Tr} \left[[\nabla \Omega^*(\mathbf{V} \text{Diag}(\Gamma_{-\mathbf{b}}) \mathbf{V}^T)]^T \mathbf{V} \mathbf{J}_{jj} \mathbf{V}^T \right] \\ &= \left(\mathbf{V}^T [\nabla \Omega^*(\mathbf{V} \text{Diag}(\Gamma_{-\mathbf{b}}) \mathbf{V}^T)]^T \mathbf{V} \right)_{jj}, \\ \frac{\partial \Omega^*(\mathbf{V} \text{Diag}(\Gamma_{-\mathbf{b}}) \mathbf{V}^T)}{\partial \Gamma} &= \text{Diag} \left(\mathbf{V}^T [\nabla \Omega^*(\mathbf{V} \text{Diag}(\Gamma_{-\mathbf{b}}) \mathbf{V}^T)]^T \mathbf{V} \right). \end{aligned}$$

Recovering the solution from optimal Lagrange multipliers. Once the dual problem is solved (in practice the convergence of our accelerated gradient algorithm is checked with some small relative tolerances on the constraints and the duality gap, *e.g.* 10^{-2} , see Appendix B.3), we need to recover the optimal solutions of the primal problem. Denoting $\hat{\Gamma} \in \mathbb{R}_+^n$ and $\hat{\theta} \in \mathbb{R}_+$ the optimal Lagrange multipliers, the approximated mean function is recovered by

$$\hat{\gamma} = \left(\text{Diag}(\hat{\Gamma}_{\mathbf{a}}) \mathbf{K}^m + \hat{\theta} \mathbf{I}_n \right)^{-1} \text{Diag}(\hat{\Gamma}_{\mathbf{a}}) \mathbf{Y}.$$

On the other hand, to reconstruct the matrix \mathbf{A} , we follow Theorem 8 from Marteau-Ferey et al. [2020]:

$$\begin{aligned} \hat{\mathbf{A}} &= \nabla \Omega^* \left(\mathbf{V} \text{Diag}(\hat{\Gamma}_{-\mathbf{b}}) \mathbf{V}^T \right) \\ &= \frac{1}{2\lambda_2} \left[\mathbf{V} \text{Diag}(\hat{\Gamma}_{-\mathbf{b}}) \mathbf{V}^T - \lambda_1 \mathbf{I}_n \right]_+. \end{aligned}$$

A.3 Proof of Proposition 3 - marginal coverage

Proposition 3 is a trivial extension of standard split CP: the kernel SoS procedure is conducted on pre-training data only. The scores are then computed on an independent calibration data, and thus training conditional coverage is straightforward, yielding marginal coverage (see for example Lei et al. [2018]). Algorithm 1 gives a detailed overview of the full procedure.

Improved calibration from Liang [2022] and Fan et al. [2024]. The split CP calibration procedure and Fan et al. [2024] are really close. Indeed, the calibration procedure in Fan et al. [2024] is also based on a splitting strategy as for split CP, but without adjusting the quantile level. Their calibration method is thus almost equivalent to split CP, and from a theoretical perspective they recover the target coverage level but with different tools (which allows them to derive explicit coverage bounds for their bands before calibration). The main advantage of using split CP, is that we can benefit from all previous and future extensions of CP to handle specific situations (*e.g.* covariate shift, jackknife+, ...) at almost no cost, unlike Fan et al. [2024]’s strategy. On the other hand, Liang [2022] also provides pre-calibration coverage bounds, while relying on sample splitting. The main variation with respect to split CP or Fan et al. [2024] is that the hyperparameter which must be optimized on the calibration set no longer appears as a multiplicative constant in front of the band width, but as an additive constant inside a square root, thus implying a more complex dyadic search (in contrast to

split CP framework or Fan et al. [2024]’s one, where it is simply given as a sample quantile). This simple difference partly explains why they require other proof strategies to get the target coverage. Once again, we believe that the traditional split CP is more straightforward and easier to apprehend, with no obvious disadvantage (to the best of our knowledge) with respect to Fan et al. [2024] or Liang [2022] procedures.

A.4 Proof of Proposition 4 - local coverage

Before giving a detailed proof of our bounds, we first recall the definitions of the maximum mean discrepancy and the Hilbert-Schmidt independence criterion.

MMD and HSIC.

Definition 1 (Maximum Mean Discrepancy [Smola et al., 2007]) *Let X and Y be random vectors defined on a topological space \mathcal{Z} , with respective Borel probability measures P_X and P_Y . Let $k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ be a kernel function and let $\mathcal{H}(k)$ be the associated reproducing kernel Hilbert space. The maximum mean discrepancy between P_X and P_Y is defined as*

$$\text{MMD}_k(P_X, P_Y) = \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} |\mathbb{E}_{X \sim P_X}[f(x)] - \mathbb{E}_{Y \sim P_Y}[f(y)]|.$$

The squared MMD admits the following closed-form expression:

$$\begin{aligned} \text{MMD}_k(P_X, P_Y)^2 &= \mathbb{E}_{X \sim P_X, X' \sim P_X}[k(X, X')] + \mathbb{E}_{Y \sim P_Y, Y' \sim P_Y}[k(Y, Y')] \\ &\quad - 2\mathbb{E}_{X \sim P_X, Y \sim P_Y}[k(X, Y)], \end{aligned}$$

which can be estimated thanks to U- or V-statistics.

Now given a pair of random vectors $(U, V) \in \mathcal{X} \times \mathcal{Y}$ with probability distribution P_{UV} , we define the product RKHS $\mathcal{H} = \mathcal{F} \times \mathcal{G}$ with kernel $k_{\mathcal{H}}((u, v), (u', v')) = k_{\mathcal{X}}(u, u')k_{\mathcal{Y}}(v, v')$. A measure of the dependence between U and V can then be defined as the distance between the mean embedding of P_{UV} and $P_U \otimes P_V$, the joint distribution with independent marginals P_U and P_V :

$$\text{MMD}^2(P_{UV}, P_U \otimes P_V) = \|\mu_{P_{UV}} - \mu_{P_U} \otimes \mu_{P_V}\|_{\mathcal{H}}^2.$$

This measure is the so-called *Hilbert-Schmidt independence criterion* (HSIC, see Gretton et al. [2005]) and can be expanded as

$$\begin{aligned} \text{HSIC}(U, V) &= \text{MMD}^2(P_{UV}, P_U \otimes P_V) \\ &= \mathbb{E}_{U, U', V, V'} k_{\mathcal{X}}(U, U') k_{\mathcal{Y}}(V, V') \\ &\quad + \mathbb{E}_{U, U'} k_{\mathcal{X}}(U, U') \mathbb{E}_{V, V'} k_{\mathcal{Y}}(V, V') \\ &\quad - 2\mathbb{E}_{U, V} [\mathbb{E}_{U'} k_{\mathcal{X}}(U, U') \mathbb{E}_{V'} k_{\mathcal{Y}}(V, V')] \end{aligned}$$

where (U', V') is an independent copy of (U, V) . Once again, the reproducing property implies that HSIC can be expressed as expectations of kernels, which facilitates its estimation when compared to other dependence measures such as the mutual information.

Bounds on local coverage. To lighten notations, we denote $X = X_{N+1}$, $Y = Y_{N+1}$, $R = |Y - \hat{m}_{\mathcal{D}_N}(X)|$, $V = \hat{f}_{\mathcal{D}_N}(X)$ and $S = R/\sqrt{V}$. In the frame of Proposition 4, we work conditionally on \mathcal{D}_N , which means that in what follows $\hat{m}_{\mathcal{D}_N}(\cdot)$ and $\hat{f}_{\mathcal{D}_N}(\cdot)$ are deterministic functions. The chain rule for mutual information gives

$$\begin{aligned} \text{MI}((X, V), R) &= \text{MI}(X, R) + \text{MI}(V, R|X) \\ &= \text{MI}(V, R) + \text{MI}(X, R|V). \end{aligned}$$

Conditionally on X , V is constant and then R and V are independent. This implies $\text{MI}(V, R|X) = 0$ and

$$\text{MI}(X, R) - \text{MI}(V, R) = \text{MI}(X, R|V).$$

We now write

$$\begin{aligned}
\text{MI}(X, S) &= \text{MI}\left(X, \frac{R}{\sqrt{V}}\right) \\
&\leq \text{MI}(X, (R, V)) \quad (\forall g, \text{MI}(g(X), Y) \leq \text{MI}(X, Y)) \\
&\leq \text{MI}((X, V), (R, V)) \quad (\text{MI}((X_1, X_2), Y) \geq \text{MI}(X_1, Y)) \\
&= \text{MI}(X, R|V) + H(V) \quad (\text{MI}(X, Y|Z) = \text{MI}(X, Z), (Y, Z) - H(Z)) \\
&\leq \text{MI}(X, R|V) + H(X) \quad (\forall g, H(g(X)) \leq H(X)) \\
&= \text{MI}(X, R) - \text{MI}(V, R) + H(X)
\end{aligned}$$

and we can observe that only $\text{MI}(V, R)$ depends on V . We thus deduce that

$$1 - \exp(-\text{MI}(X, S)) \leq 1 - \alpha_1 \exp(\text{MI}(V, R))$$

where $\alpha_1 = \exp(-\text{MI}(X, R) - H(X))$ is independent from V , which proves Equation (11).

For the second part of the proposition, from Equation (15) in Wang and Tay [2023], we have the bound

$$\text{TV}(\mathbb{P}, \mathbb{Q}) \geq \frac{1}{2\sqrt{M_k}} \text{MMD}_k(\mathbb{P}, \mathbb{Q})$$

where $\text{TV}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|$ for \mathbb{P}, \mathbb{Q} defined on a measurable space (Ω, \mathcal{F}) and $\text{MMD}_k(\mathbb{P}, \mathbb{Q})$ are the total variation and the maximum mean discrepancy between probability distributions \mathbb{P} and \mathbb{Q} , respectively. Here, the MMD depends on the choice of a kernel k , which is bounded by $M_k = \sup_{x \in \mathcal{X}} k(x, x)$, and must be characteristic for the inequality to hold. We then apply this inequality to $\mathbb{P} = P_{VR}$ the joint distribution of (V, R) and $\mathbb{Q} = P_V \otimes P_R$ the joint distribution with independent marginals P_V and P_R , to get

$$1 - \exp(-\text{MI}(V, R)) \geq \text{TV}^2(P_{VR}, P_V \otimes P_R) \geq \alpha_2 \text{HSIC}(V, R),$$

where the inequality on the left is the Bretagnolle-Huber inequality, the inequality on the right comes from the HSIC definition $\text{HSIC}(X, Y) = \text{MMD}^2(P_{XY}, P_X \otimes P_Y)$ and we denote $\alpha_2 = 1/(4M_k)$ with k the kernel used in HSIC. We finally have

$$1 - \alpha_1 \exp(\text{MI}(V, R)) \leq 1 - \frac{\alpha_1}{1 - \alpha_2 \text{HSIC}(V, R)}$$

and Equation (12) follows.

Remark 1 *In our initial mutual information bound, we replace $H(V)$ by $H(X)$ in order to obtain in the end a bound which can be expressed only with HSIC. This may seem crude, and of course we could easily incorporate $H(V)$ in our criterion to get a sharper bound: however, our numerical experiments show that even without this term the criterion yields satisfying adaptivity.*

In order to illustrate the reason why we advocate using HSIC over MI, that is numerical stability, we reproduce here our experiment on test case 2. For several values of b , we compute both criteria for a grid of θ^f candidate values, with $n = 100$ and 10-fold cross-validation. Figure 4 shows that HSIC (left) allows to clearly discriminate the lengthscales, while MI (right) suffers from estimation instability and is thus unusable in practice to identify a satisfying lengthscale.

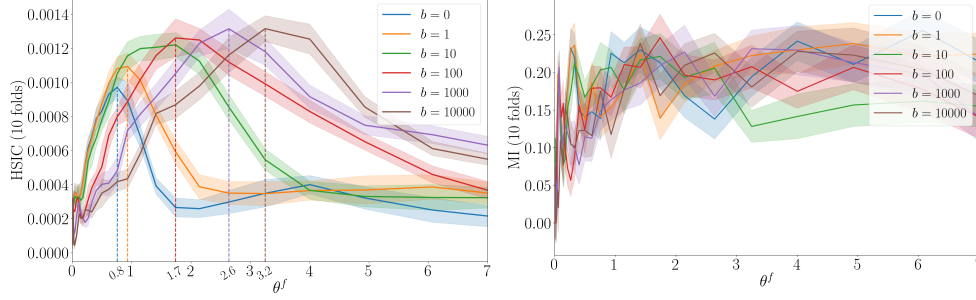


Figure 4: Test case 2 with $n = 100$. HSIC (left) and MI (right) criteria between $r(X, Y)$ and $f(X)$ as a function of b and θ^f (confidence intervals obtained by bootstrap and optimal values of θ^f in dashed lines).

B Additional experiments and details

B.1 Hyperparameter influence

In order to quantitatively evaluate the influence of each hyperparameter in our kernel SoS formulation, we perform the following extensive study:

1. We select test cases 1, 2 and 3 (see Appendix B.4 for details)
2. For each test case:
 - (a) We generate a training dataset of size $n = 50, 100$ for 3 different random seeds and 10 different values for θ^f
 - (b) For each hyperparameter of interest a, b, λ_1 and λ_2
 - i. We fix all hyperparameters to 1 except the one of interest which varies between 10^{-3} and 10^7
 - ii. We solve the kernel SoS problem
 - iii. We compute the root mean-squared error, the mean width, the nuclear norm and the Frobenius norm of the optimal solution
 - (c) We normalize these indicators with their median over all combinations to be able to compare the different test cases on a common ground

Boxplots of the combined indicators are given in Figure 5. We first observe that on average a does not have a large impact on the root mean-squared errors (top left). On the other hand, λ_2 does not influence at all the optimal Frobenius norm or the interval mean width (top right). However, the bottom row provides interesting insights. Indeed at first glance, since b and λ_1 both appear explicitly in the formula for the PSD matrix A , we may have intuitively thought that they should have a similar effect. Our results actually show that λ_1 has a very small influence, contrary to b which, when it increases, greatly reduces the interval width at a cost of inflating the nuclear norm.

Note however that our conclusions related to a only apply to test cases with symmetric noise (which is the case for test cases 1, 2 and 3). In presence of asymmetry, it may be required to select $a > 0$ to ensure a satisfying estimation of the mean function. We give an illustration in Appendix B.4.

B.2 Cross-validation for kernel hyperparameter estimation and HSIC test of independence

Cross-validation. Let K be the number of folds. For $k \in [K]$, we write \mathcal{D}_k the fold dataset k and $\mathcal{D}_{-k} = \mathcal{D}_n \setminus \mathcal{D}_k$. We denote by $\hat{m}_{-k}, \hat{f}_{-k}$ the mean and scaling functions trained on \mathcal{D}_{-k} according to Equation (8). Define two sets,

$$R_K = \bigcup_{k=1}^K \{(Y_i - \hat{m}_{-k}(X_i))^2\}_{i \in \mathcal{D}_k} \quad \text{and} \quad F_K = \bigcup_{k=1}^K \{\hat{f}_{-k}(X_i)\}_{i \in \mathcal{D}_k}.$$

We seek

$$\max_{\theta^f \in \mathbb{R}^d} \widehat{\text{HSIC}}(R, F), \tag{17}$$

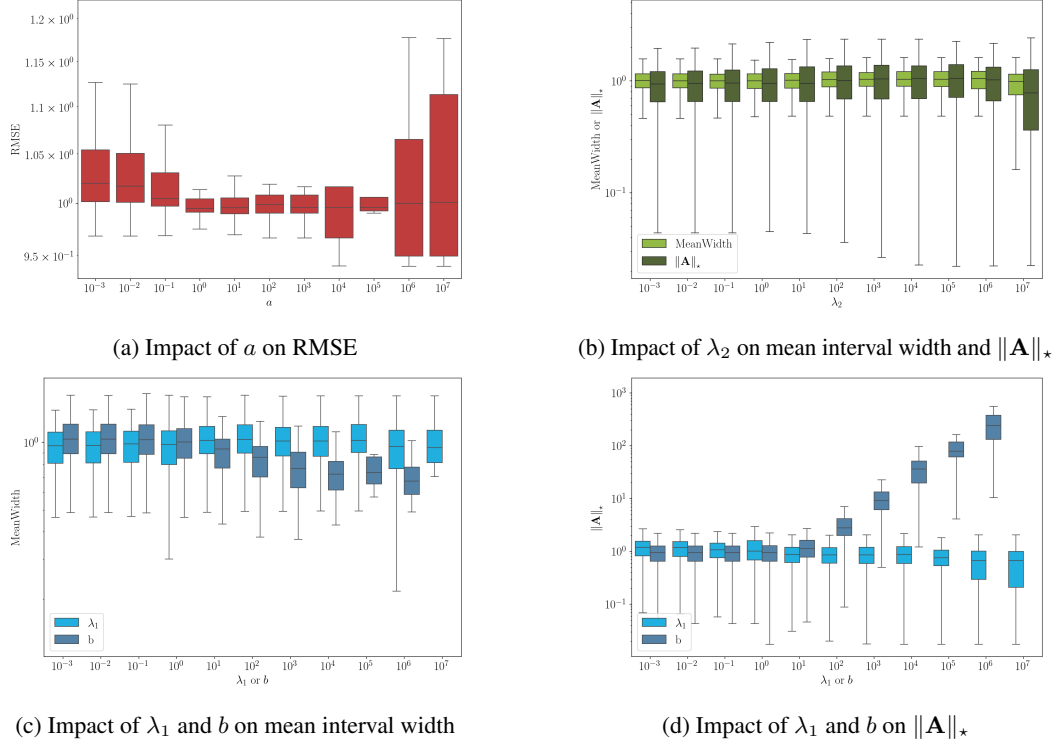


Figure 5: Marginal impact of hyperparameters a , b , λ_1 and λ_2 over several values of θ^f , test cases and random seeds on rmse, mean interval width and regularization norms.

where $\widehat{\text{HSIC}}(R, F)$ is estimated with samples R_K and F_K . In all our experiments, we use the energy distance kernel $k(x, x') = |x| + |x'| - |x - x'|$, which has been shown to be characteristic by Sejdinovic et al. [2013]. In practice, since HSIC is estimated with samples, the objective function of this optimization problem is noisy: consequently we resort to the BOBYQA optimization algorithm [Powell et al., 2009].

Once the lengthscales θ^f are selected, we train the scores on the whole pre-training dataset to learn $\widehat{m}_{\mathcal{D}_n}$ and $\widehat{f}_{\mathcal{D}_n}$. From there, we use the calibration set following the split conformal procedure. Our complete algorithm is summarized in Algorithm 1.

Algorithm 1: Split CP with adaptive kernel SoS

Data: $\mathcal{D}_N = \mathcal{D}_n \cup \mathcal{D}_m$, the pre-training and calibration datasets.

Input: $(a, b, \lambda_1) \in \mathbb{R}_+^3$, $\lambda_2 > 0$, the error-rate $\alpha \in]0, 1[$ and choose kernels k^m and k^f .

- 1 Fit a GP on \mathcal{D}_n with estimated nugget effect. Retrieve estimated lengthscales θ^m and set $s = \|m_{\text{GP}}\|^2$
- 2 Solve Equation (17) to estimate θ^f
- 3 Solve Equation (8) to get $\widehat{m}_{\mathcal{D}_n}$ and $\widehat{f}_{\mathcal{D}_n}$
- 4 Compute the scores $\{s_i\}_{i \in \mathcal{D}_m} = \{(Y_i - \widehat{m}_{\mathcal{D}_n}(X_i))^2 / \widehat{f}_{\mathcal{D}_n}(X_i)\}_{i \in \mathcal{D}_m}$
- 5 Compute the adjusted level quantile $\widehat{q}_\alpha(\{s_i\}_{i \in \mathcal{D}_m})$ of these scores

Output: $\widehat{C}_{\mathcal{D}_N}(\cdot) = \left[\widehat{m}_{\mathcal{D}_n}(\cdot) \pm \sqrt{\widehat{q}_\alpha \widehat{f}_{\mathcal{D}_n}(\cdot)} \right]$

HSIC test of independence. Solving Equation (17) will always yield an optimized value for θ^f . However in practice, the maximum attainable value of $\widehat{\text{HSIC}}(R, F)$ may be small, thus suggesting that the residuals and the width of the most adaptive prediction bands are independent. In such

situations, assuming that the noise is homoscedastic is natural, which leads us to design constant prediction bands by choosing an arbitrary large value for θ^f (we use $\theta^f = 20$ in our experiments). The key point is to detect that $\widehat{\text{HSIC}}(R, F)$ is sufficiently close to 0 in a meaningful statistical way: to do this we use the HSIC test of independence introduced by Gretton et al. [2007]. The null distribution of this test needs to be approximated, we use a permutation procedure. From a practical viewpoint and for numerical stability, we repeat the test 10 times and compute the median of the p-values.

Experiments with naive lengthscales values. We also investigated the performance of naive choices for θ^f . The first one is to choose $\theta^f = \theta^m$, where θ^m is the lengthscales for the mean function. The second naive choice is to take θ^f equal to the median of the feature distances (a usual rule-of-thumb - ROT below - for kernel methods). We report below, in Table 1, the mean width for the two naive choices and for the lengthscales that maximize the HSIC (Opt. HSIC). We can see that HSIC-optimized lengthscales always perform the best in terms of mean width, although with ties with naive choices in some instances. We also report the local coverage for all cases in Table 1. HSIC-optimized lengthscales are the only one to consistently perform well across all cases.

	Mean-Width		
	θ^m	ROT	Opt. HSIC
Case 1	2.862 \pm 0.123	2.898 \pm 0.155	2.877 \pm 0.209
Case 2	2.260 \pm 0.076	2.032 \pm 0.110	2.052 \pm 0.103
Case 3	1.482 \pm 0.061	1.399 \pm 0.095	1.385 \pm 0.080
Case 4	4.185 \pm 0.239	4.227 \pm 0.254	4.063 \pm 0.159

Table 1: Comparison of mean width for different choices of θ^f .

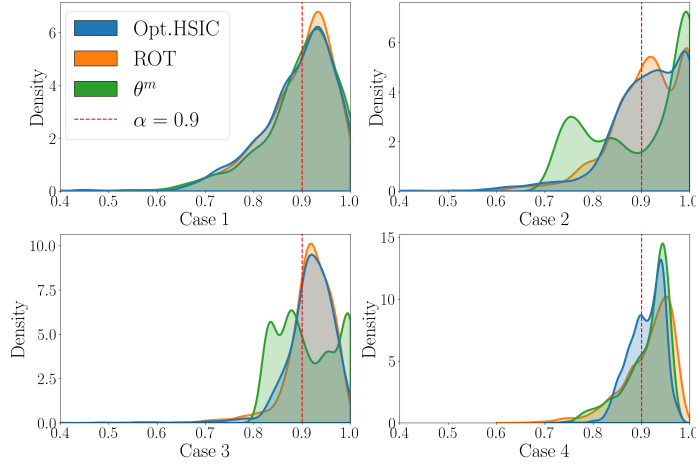


Figure 6: Local coverage for different choices of θ^f .

B.3 Implementation details

This section provides additional details about the SDP and dual solvers used throughout the experiments.

SDP solver. The SDP problem is solved with the SCS algorithm O’Donoghue [2021], O’Donoghue et al. [2023] available in the convex optimization software CVXPY Diamond and Boyd [2016], Agrawal et al. [2018]. An example of script that implements the SDP problem defined by Equation (7) is shown below. Here, `vector_variable` and `matrix_variable` denote the unknowns $\gamma \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{S}_+^n$. The remaining variables, such as the Gram matrix `Km_train` (\mathbf{K}^m) and the feature vector `phi_variance_gram_matrix` ($\Phi(\mathbf{X})$) have to be computed by the user using the training set and the chosen kernel functions.

```

import numpy as np
import cvxpy as cp
import scs

# Variables
matrix_variable = cp.Variable((n, n), symmetric=True)
vector_variable = cp.Variable((n, d))

# Expressions
f_A = cp.reshape(
    cp.diag(phi_variance_gram_matrix.T @ matrix_variable @
        phi_variance_gram_matrix),
    (-1, 1),
    order="C",
)
mean_estimator = cp.matmul(Km_train, vector_variable)

# Objective
objective = cp.Minimize(
    (a/n) * cp.sum_squares(y_train - mean_estimator)
    + (b/n) * cp.sum(f_A)
    + lambda1 * cp.trace(matrix_variable)
    + lambda2 * cp.square(cp.norm(matrix_variable, "fro"))
)

# Constraints
constraints = [
    matrix_variable >> 0,
    f_A >= (y_train - mean_estimator)**2,
    cp.quad_form(vector_variable, Km_train, assume_PSD=True) <= s,
]

# Problem definition
problem = cp.Problem(objective, constraints)

# Solve
problem.solve(solver=cp.SCS, verbose=False)

```

Listing 1: Solving the SDP problem with CVXPY and SCS.

Dual formulation algorithm. We optimize the dual objective (8) using a projected gradient method with Nesterov acceleration and leveraging forward-backward backtracking line-search on the smoothness parameter, see Barré et al. [2022], Truong and Nguyen [2021], enforcing feasibility by clipping the Lagrange multipliers and applying convergence checks on constraint satisfaction and duality gap. The gradient is computed analytically and exploits problem structure, including sparsity and Cholesky factorization for efficiency. In Figure 3 left, the time reported are computed with a tolerance of 0.01 for both the convergence and duality gap checks. The simulations were performed on an AMD Ryzen 7 9700X 8-Core Processor (3.80 GHz).

B.4 Additional numerical experiments

Formulation B versus formulation A. As mentioned in Section 2.2, Marteau-Ferey et al. [2020] provide two equivalent formulations of the kernel SoS problems in terms of a PSD matrix \mathbf{B} or \mathbf{A} . Though theoretically equivalent, we propose to investigate their respective numerical efficiency. For different test cases, different lengthscales and different random seeds, we record the computational time of the SCS solver to obtain the optimal solution with the \mathbf{B} formulation, the \mathbf{A} formulation and with or without Frobenius regularization for both, see Figure 7. The latter serves as a numerical illustration of its interest from a computational perspective, beyond the theoretical one related to strong convexity.

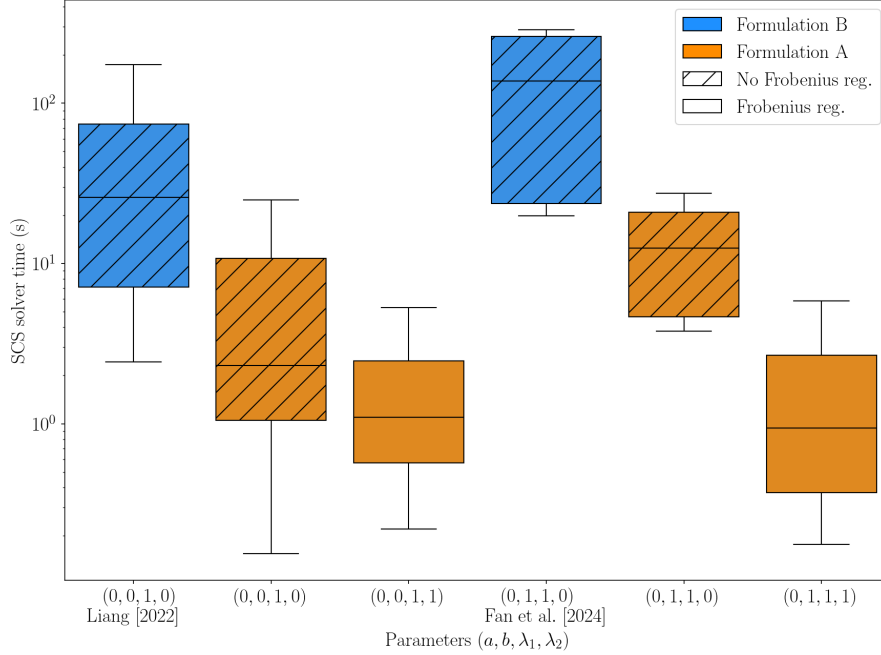


Figure 7: Time comparison between formulation A, B and with or without Frobenius regularization. Each model has been run for test cases 1, 2 and 3 with three random seeds, two sample sizes $n = 50$ and $n = 100$ and ten lengthscales between 0.1 and 1.0. The y-axis is plotted on a logarithmic scale.

In average for different hyperparameter combinations, we observe a significant reduction between both formulations, with formulation A yielding 90% computational savings. Setting $\lambda_2 > 0$ also improves resolution times, in particular when $b > 0$.

Evaluation metrics. For our experiments on analytical test cases, we compute the following metrics:

- The mean width of the prediction intervals on the test set
- The mutual information between X and $S(X, Y)$ on the test set
- The R_{SQI}^2 criterion proposed in Deutschmann et al. [2024]. We first discretize the interval widths of test samples with quantiles, leading to a partition of the test set. For each partition, we compute the $(1 - \alpha)$ -quantile of the absolute residuals and the median width: the R_{SQI}^2 is the R^2 determination coefficient of the linear regression without intercept between these two
- The local coverage, obtained by approximating $\mathbb{P}(Y_{N+1} \in \hat{C}(X_{N+1}) \mid X_{N+1} = x)$ by its empirical counterpart with samples from Y_{N+1} (of size n_Y) at different random locations X_{N+1} (of size n_X)

For real-world datasets, we do not compute the mutual information for numerical robustness due to the dimension of X , and cannot estimate the local coverage since we do not have access to multiple samples from Y_{N+1} for a given X_{N+1} . We then consider:

- The mean width of the prediction intervals on the test set as previously
- The worst-set coverage introduced by Thurin et al. [2025]. Starting from a partition $\{\mathcal{R}_l\}_{l=1,\dots,L}$ of the input space, we compute the marginal coverage in each region $\mathbb{P}(Y_{N+1} \in \hat{C}_{\mathcal{D}_N}(X_{N+1}) | X_{N+1} \in \mathcal{R}_l)$. The worst-set coverage is defined as the minimum of all these coverages: the closer it is to the target α , the more adaptive the intervals. In practice, we follow the ideas of Thurin et al. [2025] to define the regions, with a procedure that may not yield a partition: we randomly select $L = 10$ samples from the test set, and for each of them we identify the 100-th closest neighbors in the feature space to estimate the marginal coverage.

We discuss in further detail the R_{SQI}^2 criterion proposed in Deutschmann et al. [2024], using Figure 8 as a visual aid. This illustration excludes the CQR model due to the lack of a predictor for the mean function.

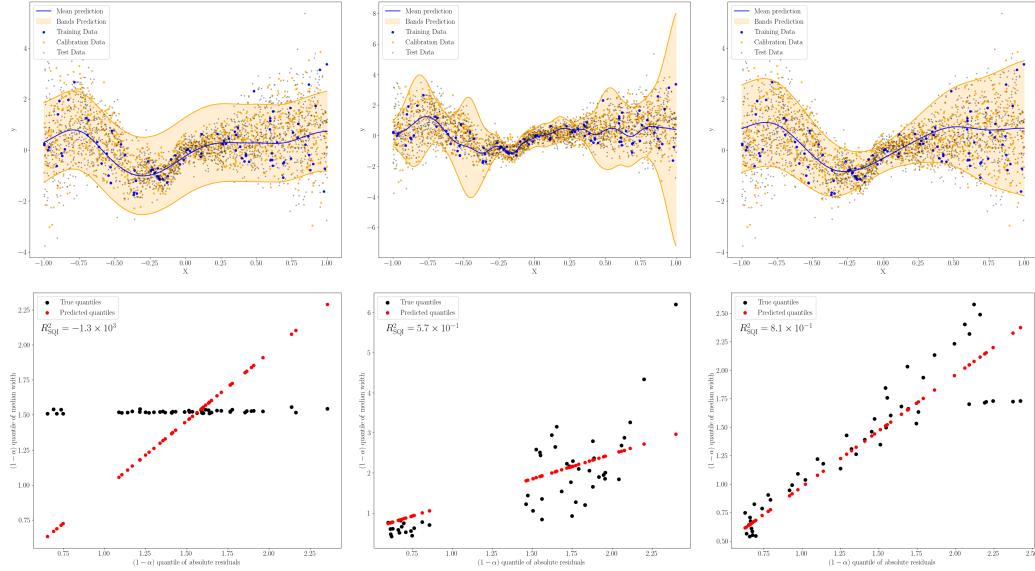


Figure 8: Test case 1 with $n = 100$, $n_X = 100$ and $n_Y = 1000$. We showcase prediction bands (top row) and the absolute residuals quantiles versus the median width quantiles (bottom row) for three models, homoscedastic GP (left column), heteroscedastic GP (middle column) and kernel SoS (right column). We use 50 quantiles to compute R_{SQI}^2 .

For the homoscedastic GP model (left) we observe constant width of the prediction bands. Consequently, the median width quantiles form a constant line with respect to the absolute error quantiles. The linear model without intercept thus fits these data very poorly and results in a negative determination coefficient. This behavior is observed throughout cases 1, 2 and 3, where the homoscedastic GP always outputs constant prediction bands.

For the heteroscedastic GP model, we see that the prediction bands are more adaptive than the homoscedastic one, resulting in a positive determination coefficient. However, the linear regression is not perfect: we can spot some bands that overcover (top row, middle column), for $X \in (0.75, 1.0]$ and around $X = 0.5$ and $X = -0.5$. On the corresponding quantile plot, these points correspond to the mean width quantiles that are above the regression line in red. We see that the model mostly overcovers for points with high residuals. In addition, the model also undercovers in some regions, for instance at $X = -0.2$, which corresponds to the black points in the lower left of the quantile regression plot, where the absolute errors are very low, or at $X = -0.7$ and $X = -1$ where the absolute errors are larger. The latter correspond to the black points in the middle of the figure under the red regression line.

Finally, for the kernel SoS model, we observe a better linear regression model. In particular, prediction bands (top row, right column) are indeed narrower in the middle where the mean predictor makes smaller errors and the data are less noisy, and wider at both ends, where the data have more noise which results in a mean predictor with larger errors.

Additional analytical test cases and results. The complete list of test cases we consider is given below. Note that we only provide R_{SQI}^2 for kernel SoS and the heteroscedastic GP: homoscedastic GP typically yields constant bands with largely negative determination coefficient, and CQR does not provide a mean function predictor. For all experiments related to adaptivity metrics, we perform 20 replications with different random seeds, and local coverage is estimated with $n_X = 100$ independent random locations X_{N+1} for which we generate $n_Y = 1000$ independent samples from Y_{N+1} . R_{SQI}^2 , MI and mean width are estimated with a test set of size $n_{\text{test}} = 1000$. For kSoS, we train first an initial model with $\theta^f = 5$ and extract the values of the mean-width and the norms, which serve as a normalization before setting the hyperparameter values $\lambda_1 = \lambda_2 = 1$ and b (depending on the test case).

Case 1. Inspired from Gramacy and Lee [2009].

$$d = 1, \quad X \sim \mathcal{U}[-1, 1], \quad Y = m(X) + \sigma(X)\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1)$$

$$m(X) = \begin{cases} \sin(\pi(2X + 1/5)) + 0.2 \cos(4\pi(2X + 1/5)) & \text{if } 10X + 1 \leq 9.6 \\ X - 9/10 & \text{otherwise} \end{cases}$$

$$\sigma(X) = \sqrt{0.1 + 2X^2}$$

We report below in Figure 9 and 10 the results from the main paper, with the additional R_{SQI}^2 .

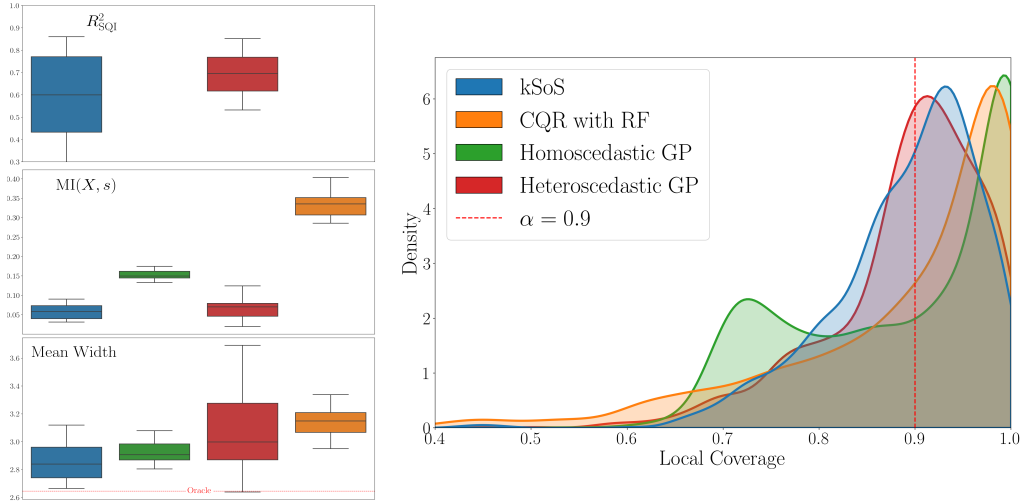


Figure 9: Test case 1 with $d = 1$ and $n = 100$. Adaptivity metrics and density of local coverage for $a = 0$ and $b = 10$.

Case 2. Corresponds to setting 1 in Hore and Barber [2024].

$$X \sim \mathcal{N}_d(0, I_d), \quad Y = m(X) + \sigma(X)\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1)$$

$$m(X) = 0.5 \sum_{i=1}^d X^{(i)}$$

$$\sigma(X) = \sum_{i=1}^d |\sin(X^{(i)})|$$

We start by setting $d = 1$ and display the adaptivity metrics in Figure 11. For this test case, we observe that all methods tend to produce prediction bands that are too large (hence a local coverage

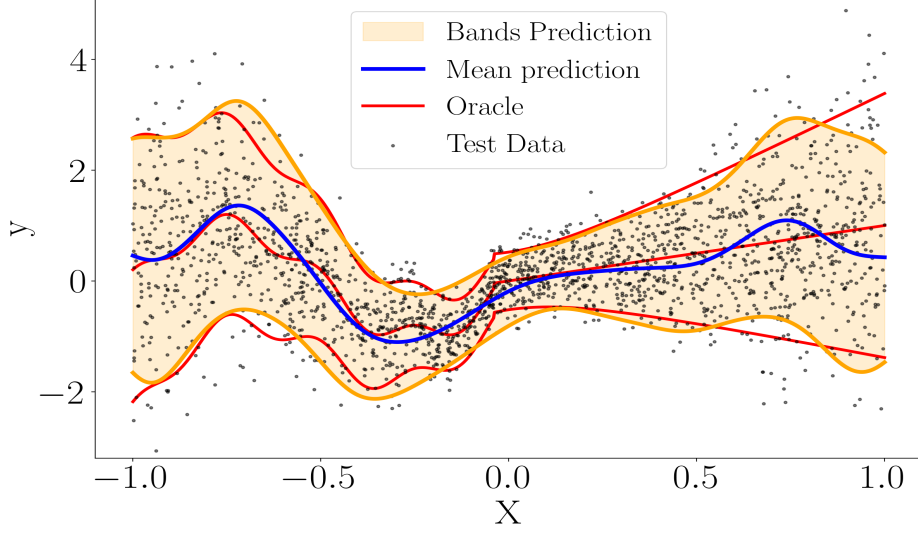


Figure 10: Test case 1 with $d = 1$ and $n = 2000$. Optimal solution of dual formulation for $a = b = 0$, $\lambda_1 = \lambda_2 = 1$ and $\theta^f = 0.3$.

density leaning towards 1), with a notable exception for kernel SoS. Although MI and R_{SoI}^2 are similar for the heteroscedastic GP and kernel SoS, the latter has intervals with much smaller mean width.

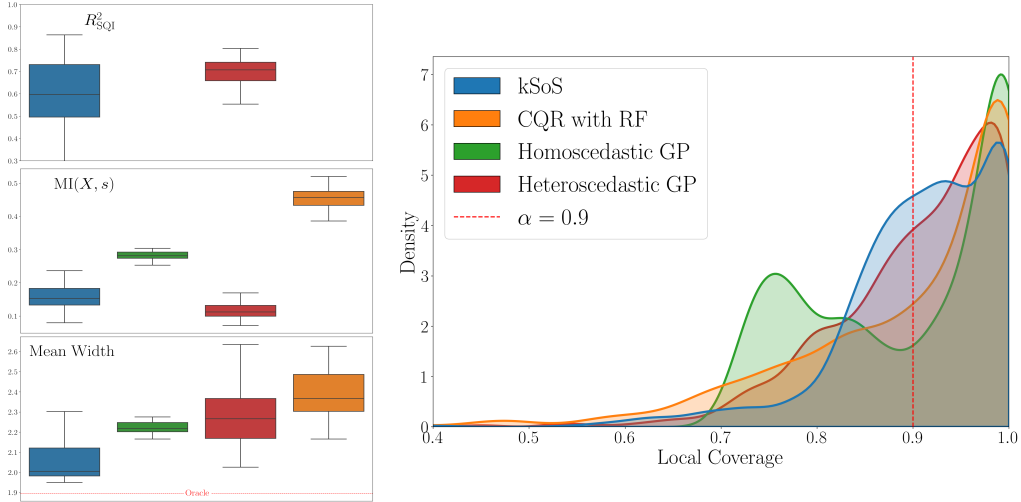


Figure 11: Test case 2 with $d = 1$ and $n = 100$. Adaptivity metrics and density of local coverage for $a = 0$ and $b = 10$.

Figure 12 shows the optimal solution of our dual formulation for $n = 2000$.

Remark 2 We do not investigate this test case in larger dimension, since by concentration due to the additive structure, the intervals tend to be constant when d increases. The same comment applies to test case 3. For brevity, we thus postpone the investigation of such a setting to test case 4.

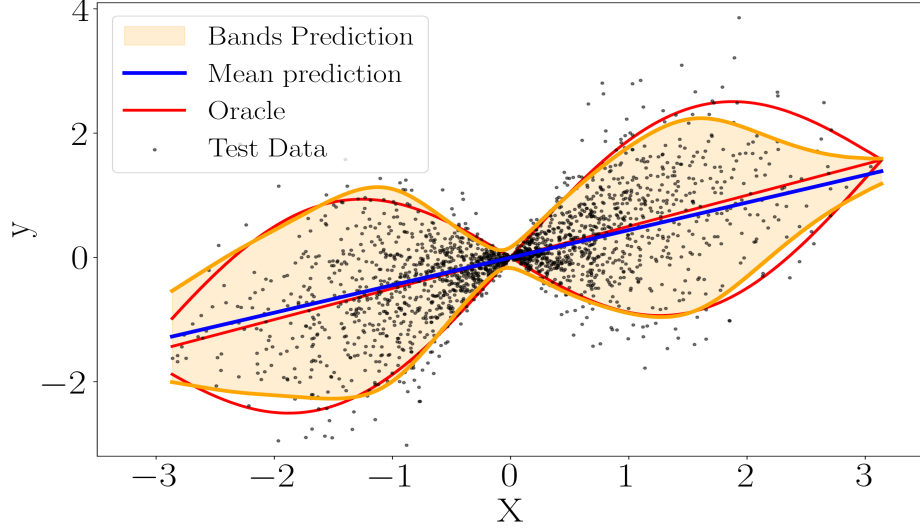


Figure 12: Test case 2 with $d = 1$ and $n = 2000$. Optimal solution of dual formulation for $a = 0$, $b = 1000$, $\lambda_1 = \lambda_2 = 1$ and $\theta^f = 1.2$.

Case 3. Corresponds to setting 2 in Hore and Barber [2024].

$$X \sim \mathcal{N}_d(0, I_d), \quad Y = m(X) + \sigma(X)\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1)$$

$$m(X) = 0.5 \sum_{i=1}^d X^{(i)}$$

$$\sigma(X) = \sum_{i=1}^d \frac{4}{3} \phi\left(\frac{2X}{3}\right), \quad \phi : \text{pdf of standard Gaussian}$$

For $d = 1$, the homoscedastic GP has poor local coverage, but with small prediction bands, as can be seen in Figure 13. Kernel SoS and the heteroscedastic GP have highly similar characteristics, with excellent local coverage as well as low MI and intervals with small mean width.

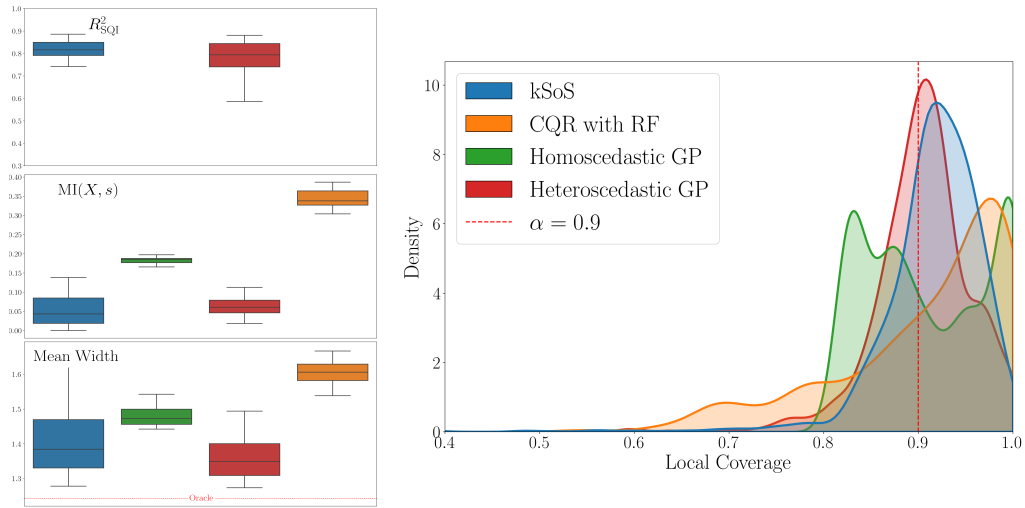


Figure 13: Test case 3 with $d = 1$ and $n = 100$. Adaptivity metrics and density of local coverage for $a = 0$ and $b = 10$.

With $n = 2000$, we obtain in Figure 14 the following optimal solution of the dual formulation.

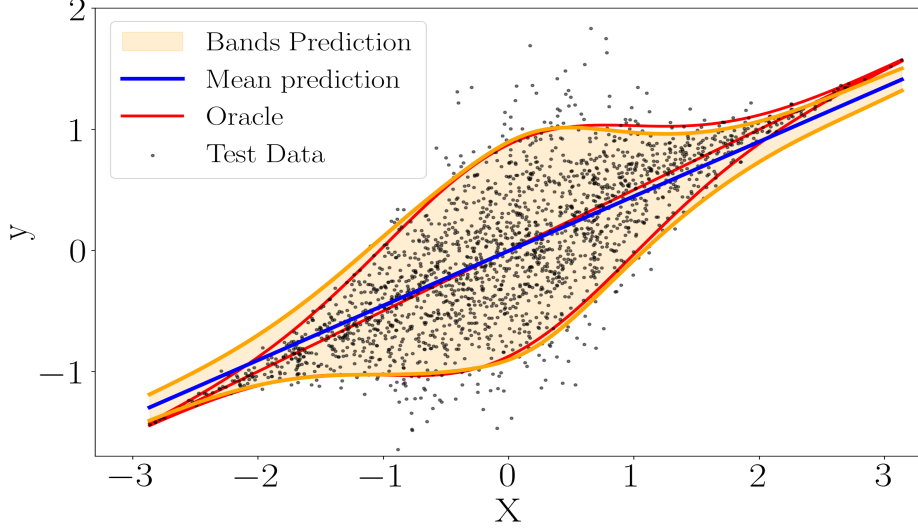


Figure 14: Test case 3 with $d = 1$ and $n = 2000$. Optimal solution of dual formulation for $a = b = 0$, $\lambda_1 = \lambda_2 = 1$ and $\theta^f = 0.9$.

Case 4. Inspired from Kivaranovic et al. [2020].

$$\begin{aligned}
 X &\sim \mathcal{U}[0, 1]^d, \quad Y = m(X) + \sigma(X)\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \\
 m(X) &= 2 \sin(\pi \beta^\top X) + \pi \beta^\top X \\
 \sigma(X) &= \sqrt{1 + (\beta^\top X)^2}
 \end{aligned}$$

For this test case, the oracle prediction bands are actually close to be constant: since we have seen previously that R_{SQI}^2 is not a relevant indicator in such situation, we discard it from our analysis. In dimension $d = 1$ we set $\beta = 1$ and obtain the remaining adaptivity metrics given in Figure 15.

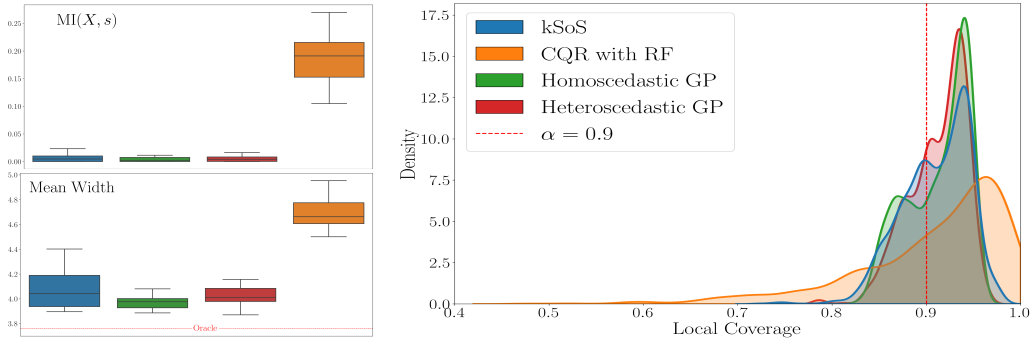


Figure 15: Test case 4 with $d = 1$ and $n = 100$. Adaptivity metrics and density of local coverage for $a = b = 0$.

As expected, homoscedastic GP, which produces almost constant intervals, performs the best in this setting. Except CQR, all methods yield similar MI, as well as similar satisfying local coverage. But this time, kSoS tend to select larger intervals than GPs: the good performance of heteroscedastic GP actually comes from the fact that it often prefers a homoscedastic model during fitting. We give in Figure 16 the optimal solution of the dual formulation obtained with $n = 2000$.

We now turn to the same test case in dimension $d = 5$ with $n = 150$ and $\beta = (1, 0.1, 0.1, 0.1, 0.1)$. Figure 17 shows a comparison with all competitors in terms of adaptivity metrics, but observe that we also do not compute MI since estimation is not robust due to the dimension.

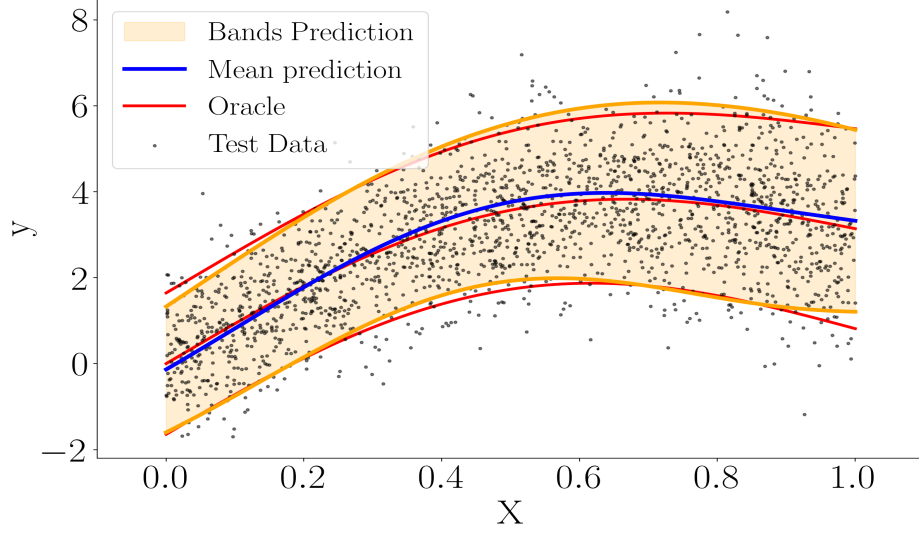


Figure 16: Test case 4 with $d = 1$ and $n = 2000$. Optimal solution of dual formulation for $a = b = 0$, $\lambda_1 = \lambda_2 = 1$ and $\theta^f = 1.6$.

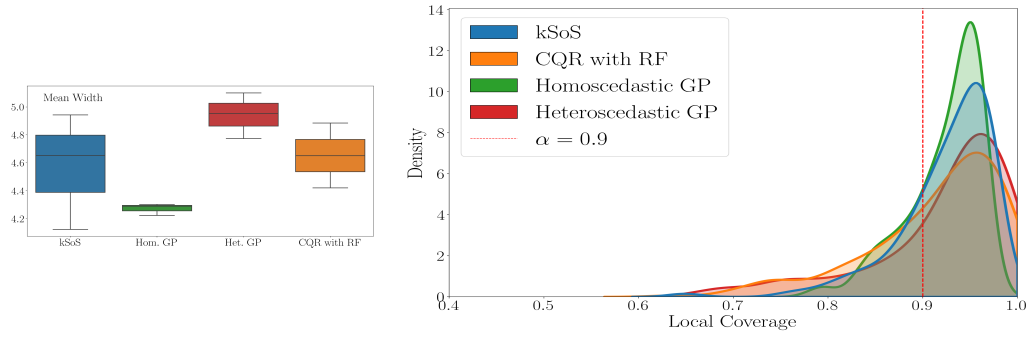


Figure 17: Test case 4 with $d = 5$ and $n = 150$. Adaptivity metrics and density of local coverage.

We can first remark that all methods tend to overcover more than in dimension $d = 1$. But similarly to $d = 1$, here the homoscedastic GP is also the best model in terms of local coverage and mean width. Interestingly, the performance of heteroscedastic GP and CQR degrades, while kSoS still achieves local coverage as good as homoscedastic GP, although with larger intervals on average.

Additional real-world datasets and results. We also investigate six real-world datasets:

1. Three of them (Concrete [Yeh, 1998], Bike [Fanaee-T, 2013] and Bio [Rana, 2013]) are taken from Romano et al. [2019].
2. The last three are standard regression datasets: Diabetes [Efron et al., 2004], Housing [Pace and Barry, 1997] and MPG [Quinlan, 1993].

They vary in terms of number of features and total sample size, see Table 2 for a detailed description. Note that for some of them, we perform a preliminary pre-processing step, by removing outliers (with a heteroscedastic GP model) and removing inactive features (identified with a heteroscedastic GP model and automatic relevance determination [Rasmussen and Williams, 2005]). For Housing specifically, we remove censored data ($\text{target} \geq 5$) and apply a logarithm transformation of the target.

Dataset	Nb features	Total sample size	n_{train}	n_{cal}	n_{test}	Outliers
Concrete	8	1030	412	412	206	30%
Bike	(18) 13	10886	1000	1000	1000	20%
Bio	(9) 4	45730	1000	1000	1000	NA
Diabetes	10	442	101	170	171	NA
MPG	7	398	100	146	146	NA
Housing	8	20640	1000	1000	1000	NA

Table 2: Description of six real-world datasets: number of features (before and after filtering, if applicable), total sample size, training set sample size, calibration set sample size, test set sample size, percentage of removed outliers in training set.

Each experiment is repeated 10 times, where we randomly sample the training, calibration and test datasets. As for analytical test cases, we compare CQR, homoscedastic GP, heteroscedastic GP, and an isotropic kSoS with two choices for θ^f : the optimal value identified with our HSIC criterion, and the value that yields the smallest mean width. See Table 3 for details on the selection of hyperparameters for both choices: since θ^f optimized with HSIC varies along the experiments, we do not report its value. Note also that for the Bio dataset, which exhibits asymmetric noise, we use the extended variant of kSoS presented in the next paragraph, and thus has not an HSIC-optimized version (since our criterion is only valid for symmetric intervals). We also use the kSoS formulation with known mean, given by a homoscedastic GP. To supplement the mean widths given in Table 1, we also provide in Table 4 the estimated marginal coverage on the test set. As expected from theory, all methods achieve the target coverage ($\alpha = 0.9$ here).

Dataset	kSoS		kSoS
	Best mean width θ^f	b	Opt. HSIC b
Concrete	3	10	10
Bike	7	10	10
Bio	2	10000	—
Diabetes	0.3	100	10
MPG	10	10	100
Housing	20	10	100

Table 3: Hyperparameters for six real-world datasets and the choices of lengthscales.

To complement the mean width analysis of each method in terms of adaptivity, we also compute in Figure 18 the worst-set coverage on these datasets except for MPG and Diabetes (very small test set size) and Bio (asymmetric noise). The mean widths are recalled below in Table 5.

Dataset	CQR	Het GP	Hom GP	kSoS Best mean width	kSoS Opt. HSIC
Concrete	0.902 ± 0.020	0.898 ± 0.020	0.898 ± 0.015	0.900 ± 0.034	0.899 ± 0.034
Bike	0.904 ± 0.012	0.903 ± 0.018	0.905 ± 0.012	0.901 ± 0.012	0.900 ± 0.012
Bio	0.902 ± 0.013	0.900 ± 0.013	0.900 ± 0.016	0.905 ± 0.012	—
Diabetes	0.899 ± 0.037	0.896 ± 0.032	0.904 ± 0.025	0.892 ± 0.027	0.886 ± 0.036
MPG	0.889 ± 0.037	0.913 ± 0.026	0.899 ± 0.028	0.895 ± 0.029	0.892 ± 0.03
Housing	0.897 ± 0.011	0.893 ± 0.012	0.900 ± 0.017	0.898 ± 0.016	0.899 ± 0.019

Table 4: Estimated marginal coverage on the test test for six real-world datasets (mean \pm sd on 10 repetitions).

Dataset	CQR	Het GP	Hom GP	kSoS Best mean width	kSoS Opt. HSIC
Concrete	0.586 ± 0.032	0.508 ± 0.052	0.543 ± 0.044	0.556 ± 0.044	0.568 ± 0.06
Bike	1.114 ± 0.062	1.000 ± 0.079	0.809 ± 0.024	0.803 ± 0.031	0.803 ± 0.032
Housing	1.816 ± 0.045	1.585 ± 0.099	1.453 ± 0.068	1.468 ± 0.094	1.586 ± 0.104

Table 5: Mean width of prediction intervals for three real-world datasets (mean \pm sd on 10 repetitions).

In average, CQR and HSIC-optimized kSoS have better worst-set coverage than both GPs, but CQR systematically yields intervals with much larger mean widths. As for the comparison between kSoS and GPs:

- Housing (Figure 18, left). Homoscedastic GP and kSoS focusing on mean width exhibit the lowest mean width, meaning that they clearly outperform heteroscedastic GP since they all have similar worst-set coverage. But they do not reach the target of $\alpha = 0.9$. On the contrary, kSoS optimized with HSIC stays closer to the target, with a mean width in line with heteroscedastic GP.
- Concrete (Figure 18, middle). In terms of mean width, heteroscedastic GP is the clear winner. However, its worst-set coverage is around 0.87. Both kSoS variants have a larger mean width, but their worst-set coverage reach the target. On this test case, homoscedastic GP has coverage similar to heteroscedastic GP but with a larger mean width.
- Bike (Figure 18, right). All methods have similar worst-set coverage in average, while the lowest mean width is obtained with homoscedastic GP and kSoS.

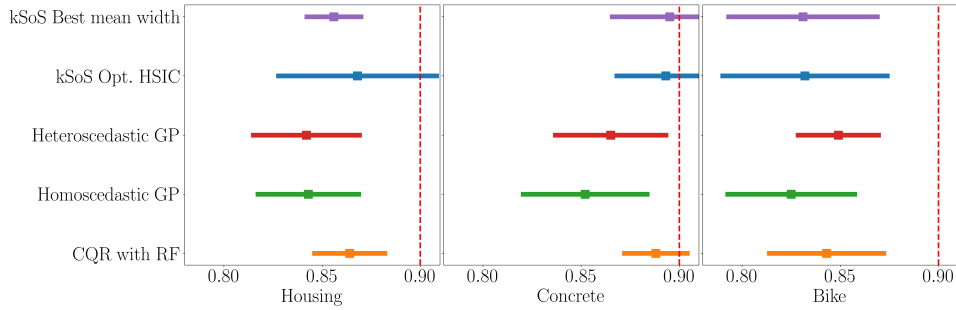


Figure 18: Housing ($b = 100$), Concrete ($b = 10$) and Bike ($b = 10$) dataset. Mean and standard deviation of worst-set coverage, 10 repetitions.

Non-symmetric intervals. We can readily extend our kernel SoS framework to non-symmetric prediction intervals of the form

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

with score function $S(X, Y) = \max(m(X) - f_{\mathbf{A}^{\text{low}}}(X) - Y, Y - m(X) - f_{\mathbf{A}^{\text{up}}}(X))$ inspired by CQR. In this setting, the kernel SoS problem writes

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

Contrary to the symmetric noise case, here the constraints no longer ensure a good fit of the regression function: this means that we need to set $a > 0$. Also remark that different RKHSs can be chosen for \mathcal{A}^{low} and \mathcal{A}^{up} in order to adapt to different regularities for the left and right tails of the conditional distribution.

As an illustration, we focus on a test case from Braun et al. [2025], which involves an exponentially distributed noise:

$$\begin{aligned} d = 1, \quad X &\sim \mathcal{U}[-1, 1], \quad Y = m(X) + \sigma(X)\epsilon, \quad \epsilon \sim \mathcal{E}(1) \\ m(X) &= \sin(2X) \\ \sigma(X) &= 0.5 + 2X \end{aligned}$$

The optimal solution of the kSOS problem is given in Figure 19 for symmetric and non-symmetric intervals with $a = 0$ and $a = 1000$, for $n = 100$. First observe that when $a = 0$ the mean function is biased, unlike for $a = 1000$. In addition, breaking the symmetry clearly improves adaptivity.

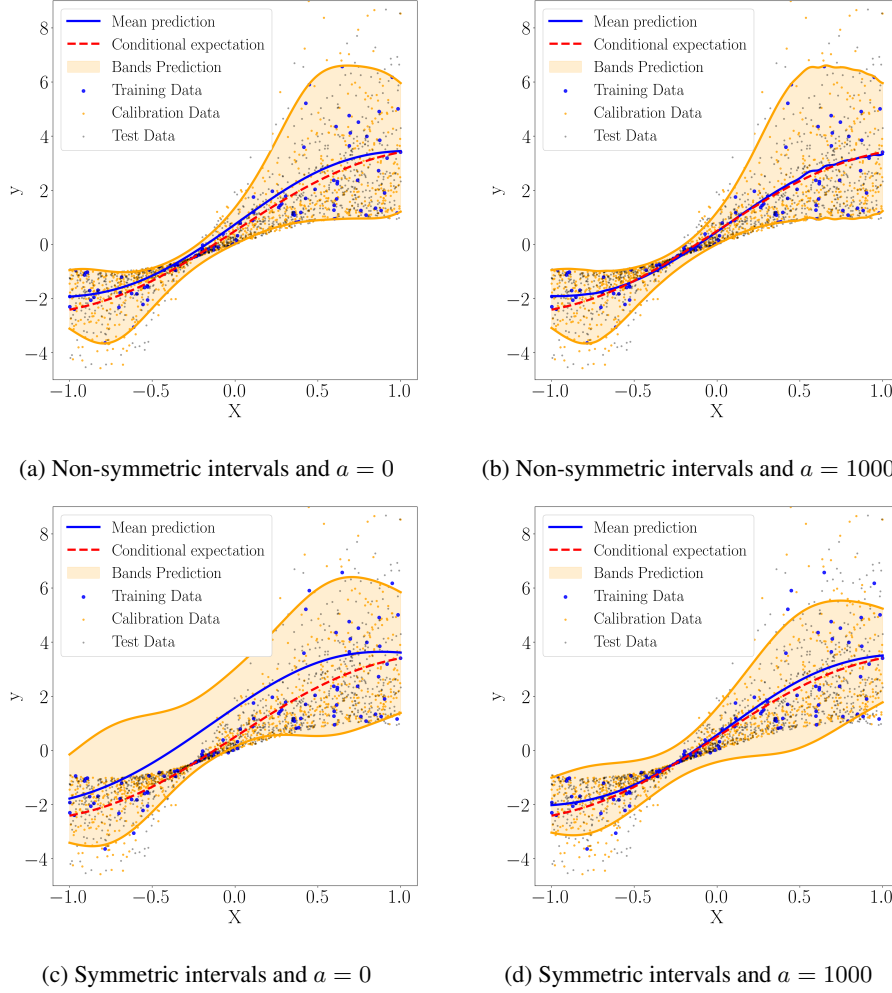


Figure 19: Test case 5. Optimal mean function and non-symmetric prediction intervals (up) and symmetric intervals (bottom) for $a = 0$ (left), $a = 1000$ (right), with parameters $b = 1$, $\lambda_1 = 1$, $\lambda_2 = 1$, $\theta^f = 0.7$.

References

- Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Non-parametric models for non-negative functions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12816–12826. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/968b15768f3d19770471e9436d97913c-Paper.pdf.
- Boris Muzellec, Francis Bach, and Alessandro Rudi. Learning psd-valued functions using kernel sums-of-squares, 2022. URL <https://arxiv.org/abs/2111.11306>.
- Kaare B. Petersen and Michael S. Pedersen. The Matrix Cookbook, 2012. URL <https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf>.
- Jing Lei, Max G’Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL <https://doi.org/10.1080/01621459.2017.1307116>.

- Tengyuan Liang. Universal prediction band via semi-definite programming. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1558–1580, August 2022. ISSN 1467-9868. doi: 10.1111/rssb.12542. URL <http://dx.doi.org/10.1111/rssb.12542>.
- Jianqing Fan, Jiawei Ge, and Debarghya Mukherjee. Utopia: Universally trainable optimal prediction intervals aggregation, 2024. URL <https://arxiv.org/abs/2306.16549>.
- Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International conference on algorithmic learning theory*, pages 13–31. Springer, 2007.
- Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.
- Chong Xiao Wang and Wee Peng Tay. Semi-nonparametric estimation of distribution divergence in non-euclidean spaces, 2023. URL <https://arxiv.org/abs/2204.02031>.
- D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.
- Michael JD Powell et al. The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06*, University of Cambridge, Cambridge, 26:26–46, 2009.
- Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.
- Brendan O’Donoghue. Operator splitting for a homogeneous embedding of the linear complementarity problem. *SIAM Journal on Optimization*, 31:1999–2023, August 2021.
- Brendan O’Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. SCS: Splitting conic solver, version 3.2.7. <https://github.com/cvxgrp/scs>, November 2023.
- Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.
- Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.
- Mathieu Barré, Adrien Taylor, and Francis Bach. A note on approximate accelerated forward-backward methods with absolute and relative errors, and possibly strongly convex objectives. *Open Journal of Mathematical Optimization*, 3:1, 2022. doi: 10.5802/ojmo.12. URL <https://ojmo.centre-mersenne.org/articles/10.5802/ojmo.12/>.
- Truong T. Truong and Huy T. Nguyen. Backtracking gradient descent method and some applications in large scale optimisation. part 2: Algorithms and experiments. *Applied Mathematics & Optimization*, 84:2557–2586, 2021. doi: 10.1007/s00245-020-09718-8. URL <https://doi.org/10.1007/s00245-020-09718-8>.
- Nicolas Deutschmann, Mattia Rigotti, and Maria Rodriguez Martinez. Adaptive conformal regression with split-jackknife+ scores. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=1fbTGC3BUD>.
- Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. *PMLR*, 267:59509–59527, 13–19 Jul 2025. URL <https://proceedings.mlr.press/v267/thurin25a.html>.
- Robert B. Gramacy and Herbert K. H. Lee. Adaptive design and analysis of supercomputer experiments. *Technometrics*, 51(2):130–145, May 2009. ISSN 1537-2723. doi: 10.1198/tech.2009.0015. URL <http://dx.doi.org/10.1198/TECH.2009.0015>.
- Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2024. doi: 10.1093/jrsssb/qkae103. URL <https://doi.org/10.1093/jrsssb/qkae103>.

- Danijel Kivaranovic, Kory D Johnson, and Hannes Leeb. Adaptive, distribution-free prediction intervals for deep networks. In *International Conference on Artificial Intelligence and Statistics*, pages 4346–4356. PMLR, 2020.
- I-Cheng Yeh. Concrete Compressive Strength. UCI Machine Learning Repository, 1998. DOI: <https://doi.org/10.24432/C5PK67>.
- Hadi Fanaee-T. Bike Sharing. UCI Machine Learning Repository, 2013. DOI: <https://doi.org/10.24432/C5W894>.
- Prashant Rana. Physicochemical Properties of Protein Tertiary Structure. UCI Machine Learning Repository, 2013. DOI: <https://doi.org/10.24432/C5QW3H>.
- Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5103c3584b063c431bd1268e9b5e76fb-Paper.pdf.
- Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–451, 2004.
- R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.
- R. Quinlan. Auto MPG. UCI Machine Learning Repository, 1993. DOI: <https://doi.org/10.24432/C5859H>.
- Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 11 2005. ISBN 9780262256834. doi: 10.7551/mitpress/3206.001.0001. URL <https://doi.org/10.7551/mitpress/3206.001.0001>.
- Sacha Braun, Liviu Aolaritei, Michael I Jordan, and Francis Bach. Minimum volume conformal sets for multivariate regression. *arXiv preprint arXiv:2503.19068*, 2025.