Asymmetric conformal prediction with penalized kernel sum-of-squares

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

Abstract

Conformal prediction (CP) is a distribution-free method to construct reliable prediction intervals that has gained significant attention in recent years. Despite its success and various proposed extensions, a significant practical feature which has been overlooked in previous research is the potential skewed nature of the noise, or of the residuals when the predictive model exhibits bias. In this work, we leverage recent developments in CP to propose a new asymmetric procedure that bridges the gap between skewed and non-skewed noise distributions, while still maintaining adaptivity of the prediction intervals. We introduce a new statistical learning problem to construct adaptive and asymmetric prediction bands, with a unique feature based on a penalty which promotes symmetry: when its intensity varies, the intervals smoothly change from symmetric to asymmetric ones. This learning problem is based on reproducing kernel Hilbert spaces and the recently introduced kernel sum-of-squares framework. First, we establish representer theorems to make our problem tractable in practice, and derive dual formulations which are essential for scalability to larger datasets. Second, the intensity of the penalty is chosen using a novel data-driven method which automatically identifies the symmetric nature of the noise. We show that consenting to some asymmetry can let the learned prediction bands better adapt to small sample regimes or biased predictive models.

1. Introduction

Quantifying the prediction uncertainty of machine learning models has become a major concern for machine learning adoption in high stakes industries such as healthcare, aeronautics, financial forecasting and autonomous driving, where predictions help make important decisions. In practice, such prediction intervals should provide at least marginal coverage guarantees that hold in finite sample and without making any distributional assumptions on the data. Conformal prediction (see, e.g., Gammerman et al. (1998); Papadopoulos et al. (2002); Shafer & Vovk (2008); Angelopoulos & Bates (2023)) has emerged recently as a simple yet very powerful tool to provide such uncertainty quantification in the form of prediction intervals, with guaranteed marginal coverage in finite sample, while being distribution-free. CP also distinguished itself for its simplicity as it is no more complex than computing a quantile.

However in real-word applications, additional features are now commonly sought for. First adaptivity, which ensures that prediction bands are wider when the model lacks confidence or if the variability in the data is high, and narrower when both the model is confident and the variability is low. Quickly identified as an important bottleneck, adaptivity has been the subject of intensive research and advances in recent years (Lei & Wasserman, 2014; Romano et al., 2019; Hore & Barber, 2024; Gibbs et al., 2025; Allain et al., 2026). Second, the noise distribution may be skewed: this implies that symmetric intervals (as produced by standard split CP) would exhibit under- or over-coverage below or above the predictive model, even if they are calibrated to achieve marginal coverage. The same situation also arises if the noise is symmetric but the predictive model is biased. Contrary to adaptivity, accounting for asymmetric noise is a problem that has gone slightly unnoticed in recent research, and is a common obstacle in real-word applications (Pouplin et al., 2024), see Appendix B.3.1 for an in-depth discussion. If the noise is by essence asymmetric, symmetric prediction bands were shown to perform poorly in terms of coverage (Linusson et al., 2014).

Mainly two alternatives have been introduced to account for asymmetry in CP. Updating the calibration step was first proposed by Linusson et al. (2014), with coverage guarantees outside the interval. This calibration procedure has the appealing property of being applicable directly to most CP methods (see e.g., Romano et al. (2019); Barber et al. (2021b); Pion & Vazquez (2025)). Concurrently, new score functions adapted to asymmetry were proposed: Confor-

[1]Univ Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France [2]Safran Tech, Digital Sciences & Technologies, 78114 Magny-Les-Hameaux, France. Correspondence to: Louis Allain <louis.allain@safrangroup.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

malized Quantile Regression (CQR) (Romano et al., 2019), which builds upon quantile regression, and Distributional Conformal Prediction (DCP) (Chernozhukov et al., 2021) based on the estimation of the conditional density function. Unfortunately, if the noise is symmetric, relying on asymmetric intervals may be detrimental since they are built without using all available information and may be ultimately wider. Adapting to the noise empirical distribution is thus essential and more robust than assuming a specific symmetry structure of the residuals. This flexibility is crucial, because the empirical distribution often diverges from the true noise distribution due to sampling artifacts: small sample settings can induce artificial asymmetry, while bias in the predictive model estimation can skew the residuals (Cheung et al., 2026). Fully asymmetric bands may overfit this bias, while symmetric bands remain overly conservative. Penalizing asymmetry would allow the method to recover symmetry when appropriate, without enforcing it a priori. To adapt to these empirical imperfection, there is thus the need of a new paradigm that can transition between symmetric and asymmetric bands in a purely data-driven way.

Very recently, Allain et al. (2026) suggested to learn a score function to tackle adaptivity with symmetric prediction bands. Focusing on a normalized score function, they rely on reproducing kernel Hilbert spaces (RKHS) and kernel-sum-of-squares (kSoS) (Marteau-Ferey et al., 2020) to handle the positivity constraint of the normalization function. Interestingly, they also discuss an extension of their work to learn an asymmetric score function, but only as a numerical illustration. Following their initial findings, we build upon their convenient learning framework to design a method that can handle symmetric and asymmetric bands seamlessly.

Contributions. We start by presenting CP in Section 2 and recent methods to account for asymmetric noise. In Section 3, we first formalize the underlying theory to learn an asymmetric score function in the kSoS framework, by providing a representer theorem and detailing its dual formulation to construct scalable asymmetric prediction bands. Second, we introduce two penalized versions to bridge the gap between asymmetric and symmetric prediction bands, for which we also prove a representer theorem and a dual formulation to enable faster computation on large datasets. These new problems can provide intermediate settings where information is shared between upper and lower prediction bands. Finally, we propose a new data-driven strategy to tune the penalty intensity and other critical hyperparameters, such as the kernel lengthscales. This strategy is based on an adaptivity criterion, for which we also provide warm-start strategies to speed up the hyperparameter search. This allows to automatically detect the amount of asymmetry needed for any dataset. In Section 4, we conduct extensive experiments to compare our methods to usual conformal prediction methods and asymmetric focused ones. In particular, we illustrate that the proposed method dynamically adapts to asymmetry induced by data defects, such as limited sample sizes. Crucially, it counterbalances estimation errors in the predictive model, that would otherwise degrade the performance of strictly symmetric models.

2. Conformal prediction and asymmetry

Split conformal prediction. The full CP setting was introduced by Gammerman et al. (1998), but we focus here on the split variant (Papadopoulos et al., 2002). We suppose we have a training dataset $\mathcal{D}_N = \{(X_i, Y_i)\}_{i=1}^N$ from a pair $(X, Y) \sim P_{XY}$ where $X \in \mathcal{X} \subset \mathbb{R}^d$ and $Y \in \mathcal{Y} \subset \mathbb{R}$. This dataset is split in two parts: a *pre-training* dataset $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ and a *calibration* one $\mathcal{D}_m = \{(X_i, Y_i)\}_{i=1}^m$ with $N = n + m$.

The pre-training dataset $\mathcal{D}_n$ is used to train a predictive model $\widehat{m}_n(\cdot)$, which can be any machine learning algorithm. Then, the performance of the model is evaluated through so-called *scores* on the hold-out calibration dataset $\mathcal{D}_m$: the most common score in the literature is defined as the absolute errors $S(X_i, Y_i) := S_i = |Y_i - \widehat{m}_n(X_i)|$ for $i \in \mathcal{D}_m$. These scores are used to compute the quantile $\widehat{q}_\alpha$ of the set $\{S_i\}_{i\in\mathcal{D}_m}$ with an adjusted level $\lceil (1-\alpha)(m+1) \rceil / m$, where α is the desired error rate. Finally, for a new observation X_{N+1}, the split CP prediction bands are $\widehat{C}_{\mathcal{D}_N}(X_{N+1}) = [\widehat{m}_n(X_{N+1}) \pm \widehat{q}_\alpha]$, which satisfy the marginal coverage guarantee

$$\mathbb{P}\left(Y_{N+1} \in \widehat{C}_{\mathcal{D}_N}(X_{N+1})\right) \geq 1 - \alpha \tag{1}$$

for any N if $(X_1, Y_1), \ldots, (X_N, Y_N), (X_{N+1}, Y_{N+1})$ are exchangeable. Importantly, these prediction bands are symmetric around the point prediction $\widehat{m}_n(X_{N+1})$ and are not adaptive, i.e. they do not depend on X_{N+1}.

Accounting for asymmetry. Asymmetry in the prediction bands can be accommodated at different stages within CP. Modifying the calibration step was first proposed by Linusson et al. (2014): they consider the signed scores $S_i = Y_i - \widehat{m}_n(X_i)$ and compute the lower and upper quantiles $\widehat{q}_{\alpha_{\text{low}}}$ and $\widehat{q}_{\alpha_{\text{up}}}$ of the set $\{S_i\}_{i\in\mathcal{D}_m}$ at adjusted levels $\lceil \alpha_{\text{low}}(m+1) \rceil / m$ and $\lceil (1-\alpha_{\text{up}})(m+1) \rceil / m$. For $\alpha_{\text{low}} + \alpha_{\text{up}} = \alpha$, the new prediction bands $\widehat{C}_{\mathcal{D}_N}(X_{N+1}) = \left[\widehat{m}_n(X_{N+1}) - \widehat{q}_{\alpha_{\text{low}}}, \widehat{m}_n(X_{N+1}) + \widehat{q}_{\alpha_{\text{up}}}\right]$ satisfy the marginal coverage as in Equation (1). But this calibration scheme further guarantees a lower and upper coverage with respective probability $1 - \alpha_{\text{low}}$ and $1 - \alpha_{\text{up}}$, and it can easily be adopted in many conformal prediction method (see e.g., Romano et al. (2019); Barber et al. (2021b); Han et al. (2023); Pion & Vazquez (2025)). Unfortunately, those supplementary guarantees may come at the cost of inflating the width of the prediction bands (Romano et al., 2019).

The second way to deal with asymmetric noise in CP is to modify the score function itself. The popular CQR (Romano et al., 2019) relies on quantile regression: instead of using an interval built around an estimate $\widehat{m}_n(\cdot)$ of the regression function, they rely on estimates $\widehat{q}_n^{\alpha_{\text{low}}}(\cdot)$ and $\widehat{q}_n^{\alpha_{\text{up}}}(\cdot)$ of the conditional quantiles, and build the interval $\widehat{C}_{\mathcal{D}_N}(X_{N+1}) = \left[\widehat{q}_n^{\alpha_{\text{low}}}(X_{N+1}) - \widehat{q}_\alpha, \widehat{q}_n^{\alpha_{\text{up}}}(X_{N+1}) + \widehat{q}_\alpha\right]$ with the score function $S(X,Y) = \max\left(\widehat{q}_n^{\alpha_{\text{low}}}(X) - Y, Y - \widehat{q}_n^{\alpha_{\text{up}}}(X)\right)$ and where $\widehat{q}_\alpha$ is the adjusted quantile of the set $\{\max\left(\widehat{q}_n^{\alpha_{\text{low}}}(X_i) - Y_i, Y_i - \widehat{q}_n^{\alpha_{\text{up}}}(X_i)\right)\}_{i\in\mathcal{D}_m}$. By design, CQR builds asymmetric prediction bands with a symmetric calibration procedure. Although appealing, CQR suffers from two well known practical limitations: (a) in high stakes problems, decision makers usually prefer a point estimate with an interval around that point and (b) quantile regression in small data regime and/or in high dimensional problems can be quite challenging. In a parallel line of work, DCP (Chernozhukov et al., 2021) considers $\widehat{F}_{Y|X}$ an estimate of the conditional CDF with scores $S_i = |\widehat{F}_{Y_i|X_i} - 1/2|$, but with the same limitations as CQR, see also Sesia & Romano (2021).

Learning a score function for adaptivity. Recently, several authors proposed to *learn* the score function in order to target adaptivity. The core idea is to parameterize the score with unknown functions, which are estimated on the pre-training set. For example, Xie et al. (2024) define a task-specific loss (e.g. conditional coverage or minimum width), and consider a score function given by $S(X,Y) = \max\left(\mu_1(X) - Y, Y - \mu_2(X)\right)/\sigma(X)$. This score is parameterized by three unknown functions (μ_1, μ_2, σ), which are iteratively optimized with a boosting algorithm. In a similar vein, Allain et al. (2026) focus on a score function $S(X,Y) = (Y - m(X))^2/f(X)$, where f is a positive function parameterized using kernel sum-of-squares. Their learning problem is defined through several ingredients: first, minimization of intervals mean width and second, 100% coverage constraints on the pre-training set to uncover the band shape, which is later adjusted with the CP calibration step. For medium-size datasets, their approach showed better adaptivity than traditional competitors. However, their method is only limited to symmetric intervals, even if they incidentally suggest a possible generalization to asymmetric bands but do not provide theory, optimization, tuning, or empirical validation. This is the starting point of our proposal, which we tackle in the next section.

3. Regularized kernel SoS for asymmetric prediction bands

Let us consider two RKHSs $\mathcal{H}_{\text{low}}$ and $\mathcal{H}_{\text{up}}$ with respective kernels k_{low}, k_{up}, kernel lengthscales θ_{low}, θ_{up} and feature maps ϕ_{low}, ϕ_{up}. For $\mathcal{A}_{\text{low}} \in \mathcal{S}_+(\mathcal{H}_{\text{low}})$ and $\mathcal{A}_{\text{up}} \in \mathcal{S}_+(\mathcal{H}_{\text{up}})$ two positive semi-definite (PSD) operators from $\mathcal{H}_{\text{low}}$ (resp. $\mathcal{H}_{\text{up}}$) to $\mathcal{H}_{\text{low}}$ (resp. $\mathcal{H}_{\text{up}}$), we define two non-negative functions $f_{\text{low}}(X) = \langle\phi_{\text{low}}(X), \mathcal{A}_{\text{low}}\phi_{\text{low}}(X)\rangle_{\mathcal{H}_{\text{low}}}$ and $f_{\text{up}}(X) = \langle\phi_{\text{up}}(X), \mathcal{A}_{\text{up}}\phi_{\text{up}}(X)\rangle_{\mathcal{H}_{\text{up}}}$, called *kernel sum-of-squares*. For brevity, we will subsequently use the notation $(\cdot)$ when objects can be assessed for both $_{\text{low}}$ and $_{\text{up}}$. These two functions, thanks to their non-negativity property, are key components of our proposed new asymmetric score:

$$S(X,Y) = \max\big(\widehat{m}_n(X) - f_{\text{low}}(X) - Y, \quad Y - \widehat{m}_n(X) - f_{\text{up}}(X)\big). \tag{2}$$

This is a variant of the CQR score function, anchored at a predictive model $\widehat{m}_n(X)$, where $\widehat{q}_n^{\alpha_{\text{low}}}(X)$ and $\widehat{q}_n^{\alpha_{\text{up}}}(X)$ are replaced by $\widehat{m}_n(X) - f_{\text{low}}(X)$ and $\widehat{m}_n(X) + f_{\text{up}}(X)$, respectively. Our framework is agnostic to the choice of point predictor $\widehat{m}_n(X)$, allowing the use of diverse off-the-shelf models best suited to the underlying data distribution. From there, we propose to estimate the functions $f_{\text{low}}(X)$ and $f_{\text{up}}(X)$ defining the prediction bands by solving the following learning problem:

$$\begin{aligned} \inf_{\substack{\mathcal{A}_{\text{low}}\in\mathcal{S}_+(\mathcal{H}_{\text{low}}) \\ \mathcal{A}_{\text{up}}\in\mathcal{S}_+(\mathcal{H}_{\text{up}})}} \quad & \frac{b}{n}\sum_{i=1}^{n}\left(f_{\text{low}}(X_i) + f_{\text{up}}(X_i)\right) \\ & + \Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}}) \\ \text{s.t.} \quad & \widehat{m}_n(X_i) - Y_i - f_{\text{low}}(X_i) \le 0,\ i \in [n] \\ & Y_i - \widehat{m}_n(X_i) - f_{\text{up}}(X_i) \le 0,\ i \in [n] \end{aligned} \tag{3}$$

where $\Omega_{(\cdot)}(\mathcal{A}) = \lambda_{(\cdot)1}\|\mathcal{A}\|_\star + \lambda_{(\cdot)2}\|\mathcal{A}\|_F^2$ is a regularization function with nuclear and Frobenius norms. Such penalty controls the complexity of functions $f_{(\cdot)}$ and the bands adaptivity, while the first term in the objective function promotes tighter intervals. Importantly, the constraints impose 100% coverage on the pre-training set, which helps learn an adaptive shape and makes the problem convex.

Once the non-negative functions are estimated, we apply the usual split CP procedure using the asymmetric score defined in Equation (2), with final calibrated prediction intervals

$$\begin{aligned} \widehat{C}_{\mathcal{D}_N}(X) = [&\widehat{m}_n(X) - \widehat{f}_{\text{low}}(X) - \widehat{q}_\alpha, \\ &\widehat{m}_n(X) + \widehat{f}_{\text{up}}(X) + \widehat{q}_\alpha], \end{aligned} \tag{4}$$

where $\widehat{q}_\alpha$ is the adjusted quantile of the set $\{S_i\}_{i\in\mathcal{D}_m}$. Here the predictive model $\widehat{m}_n$ and the functions are estimated sequentially on the pre-training dataset $\mathcal{D}_n$.

Problem 3 is infinite dimensional, but by noticing it is separable we can derive a useful representer theorem. Before stating it, let us introduce additional notations. $\mathbf{K}_{(\cdot)}$ and $\mathbf{k}_{(\cdot)}(X) = \left(k_{(\cdot)}(X_1, X), \ldots, k_{(\cdot)}(X_n, X)\right)^\top$ denote the kernel matrix and vector associated to kernel $k_{(\cdot)}$.

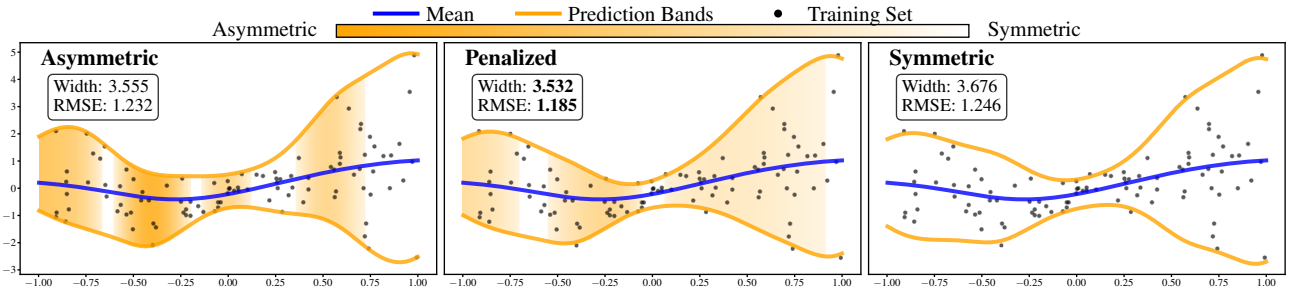


Figure 1. Penalized kSoS with varying penalty (dataset 1, symmetric noise). Left: asymmetric prediction bands produce tighter bands. Right: symmetric prediction bands tend to be overly conservative. An intermediate penalty value in the middle achieves tighter bands without being overly conservative and is closer to the oracle. Asymmetry of prediction bands is magnified with orange color.

We further consider the Cholesky decomposition of the kernel matrix $\mathbf{K}_{(\cdot)} = \mathbf{V}_{(\cdot)}^\top \mathbf{V}_{(\cdot)}$ and $r_{\text{low}}(X_i, Y_i) = \widehat{m}_n(X_i) - Y_i$, $r_{\text{up}}(X_i, Y_i) = Y_i - \widehat{m}_n(X_i)$ the residuals. Finally, for a symmetric matrix $\mathbf{A}$ with eigendecomposition $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$, its positive part is defined as $[\mathbf{A}]_+ = \mathbf{U}\max(0, \mathbf{D})\mathbf{U}^\top$ and we write $\forall x \in \mathbb{R}$, $\text{Diag}((\cdot)_x) := \text{Diag}((\cdot)) + \frac{x}{n}\mathbf{I}_n$.

Theorem 3.1 (Representer theorem)**.** *Let* $(b, \lambda_{(\cdot)1}) \in \mathbb{R}_+^2$ *and* $\lambda_{(\cdot)2} > 0$. *Then Equation* (3) *admits a unique solution* $(\tilde{f}_{\mathbf{A}^\star_{\text{low}}}, \tilde{f}_{\mathbf{A}^\star_{\text{up}}})$ *of the form* $\tilde{f}_{\mathbf{A}^\star_{(\cdot)}}(X) = \mathbf{\Phi}_{(\cdot)}(X)^\top \mathbf{A}^\star_{(\cdot)} \mathbf{\Phi}_{(\cdot)}(X)$ *for some matrix* $\mathbf{A}^\star_{(\cdot)} \in \mathbb{S}_+^n$ *solution of the semi-definite programming (SDP) problem*

$$\begin{aligned} \inf_{\mathbf{A}_{(\cdot)} \in \mathbb{S}_+^n} \quad & \frac{b}{n} \sum_{i=1}^n \tilde{f}_{\mathbf{A}_{(\cdot)}}(X_i) + \lambda_{(\cdot)1} \|\mathbf{A}_{(\cdot)}\|_\star + \lambda_{(\cdot)2} \|\mathbf{A}_{(\cdot)}\|_F^2 \\ \text{s.t.} \quad & r_{(\cdot)}(X_i, Y_i) - \tilde{f}_{\mathbf{A}_{(\cdot)}}(X_i) \leq 0, \ i \in [n]\,. \end{aligned} \tag{5}$$

The proof can be found in Appendix A.1. In practice, the associated SDP problem can be solved efficiently up to 200 samples using off-the-shelves solvers (O'Donoghue et al., 2016). Crucially, to scale up to larger training sizes, we obtain a dual formulation for both functions.

Proposition 3.2 (Dual formulation)**.** *Let* $(b, \lambda_{(\cdot)1}) \in \mathbb{R}_+^2$ *and* $\lambda_{(\cdot)2} > 0$. *Equation* (5) *admits the dual formulation*

$$\sup_{\mathbf{\Gamma}_{(\cdot)} \in \mathbb{R}_+^n} \mathbf{\Gamma}_{(\cdot)} \mathbf{r}_{(\cdot)}^\top(\mathbf{X}, Y) - \Omega^\star_{+,(\cdot)}(\mathbf{V}_{(\cdot)} \text{Diag}(\mathbf{\Gamma}_{(\cdot)-b}) \mathbf{V}_{(\cdot)}^\top)$$

where $\Omega^\star_{+,(\cdot)}(\mathbf{B}) = \frac{1}{4\lambda_{(\cdot)2}} \|\left[\mathbf{B} - \lambda_{(\cdot)1}\mathbf{I}_n\right]_+\|_F^2$. *Moreover, if* $\widehat{\mathbf{\Gamma}}_{(\cdot)}$ *is a solution of the dual formulation, a solution of Equation* (5) *can be retrieved as*

$$\widehat{\mathbf{A}}_{(\cdot)} = \frac{1}{2\lambda_{(\cdot)2}} \left[\mathbf{V}_{(\cdot)} \text{Diag}(\widehat{\mathbf{\Gamma}}_{(\cdot)-b}) \mathbf{V}_{(\cdot)}^\top - \lambda_{(\cdot)1}\mathbf{I}_n\right]_+ .$$

As with CQR, the prediction bands obtained with Equation (4) are asymmetric and flexible: lower and upper bands can have different kernel functions.

3.1. Symmetric penalization

To go further, the nature of the noise is often unknown and assuming asymmetry can be detrimental since upper and lower bands are estimated separately. Taking inspiration from the literature on penalization for supervised learning, our central idea is to add a penalty $\mathcal{P}(f_{\mathcal{A}_{\text{low}}}, f_{\mathcal{A}_{\text{up}}})$ in the objective function of Equation (3), so that the larger the penalty, the more symmetric the prediction bands:

$$\begin{aligned} \inf_{\substack{\mathcal{A}_{\text{low}} \in \mathcal{S}_+(\mathcal{H}_{\text{low}}) \\ \mathcal{A}_{\text{up}} \in \mathcal{S}_+(\mathcal{H}_{\text{up}})}} \quad & \frac{b}{n} \sum_{i=1}^n \left(f_{\mathcal{A}_{\text{low}}}(X_i) + f_{\mathcal{A}_{\text{up}}}(X_i)\right) \\ & + \Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}}) \\ & + \mathcal{P}(f_{\mathcal{A}_{\text{low}}}, f_{\mathcal{A}_{\text{up}}}) \\ \text{s.t.} \quad & \widehat{m}_n(X_i) - Y_i - f_{\mathcal{A}_{\text{low}}}(X_i) \leq 0, \ i \in [n] \\ & Y_i - \widehat{m}_n(X_i) - f_{\mathcal{A}_{\text{up}}}(X_i) \leq 0, \ i \in [n]\,. \end{aligned} \tag{6}$$

Introducing such a continuum between asymmetric and symmetric prediction bands can be advantageous in practice: it allows to mitigate the impact of small samples and compensate biased predictive models. As an illustration, in Figure 1 we show that even for a test case with true symmetric noise, penalized asymmetric prediction bands can lead to better local coverage. Importantly, these new problems bridge the gap between our asymmetric Problem (5) and the symmetric one proposed by Allain et al. (2026). In the following, we discuss two different penalties that can achieve this behavior.

Operator penalty. First, let us consider that lower and upper bounds are associated to the same RKHS $\mathcal{H}_{\text{low}} = \mathcal{H}_{\text{up}} = \mathcal{H}$; the functions are thus only parameterized by their operators, defined on the same space. In this case, a natural way to incorporate a symmetric penalty in our problem is to enforce equality of the operators by considering a regularization on their difference:

$$\begin{aligned} \mathcal{P}(f_{\mathcal{A}_{\text{low}}}, f_{\mathcal{A}_{\text{up}}}) = & \lambda_{\text{pen1}} \|\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}}\|_\star \\ & + \lambda_{\text{pen2}} \|\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}}\|_F^2 . \end{aligned} \tag{7}$$

Intuitively, as $\lambda_{\text{pen1}}, \lambda_{\text{pen2}} \to \infty$, the operators will tend to be equal and thus the lower and upper bands will coincide on

their whole domain of definition (see Proposition 3.8). This will yield symmetric prediction bands, although different from the ones initially proposed by Allain et al. (2026) (see Appendix B.3.4 for a detailed comparison).

Penalization on the training set. Another way to impose symmetry is to control the difference of the lower and upper bands on the training points only:

$$\mathcal{P}(f_{\mathcal{A}_{\text{low}}}, f_{\mathcal{A}_{\text{up}}}) = \lambda_{\text{pen}} \sum_{i=1}^{n} \big(f_{\mathcal{A}_{\text{low}}}(X_i) - f_{\mathcal{A}_{\text{up}}}(X_i)\big)^2. \tag{8}$$

As $\lambda_{\text{pen}} \to \infty$, the lower and upper functions will coincide on all training points. Contrary to the previous operator penalty, this no longer imposes equality on their domain of definition. But we show in Proposition 3.8 that, under mild assumptions, their difference can be upper bounded.

Representer theorems and dual formulations. To make the penalized problems tractable, we prove the following representer theorems, see Sections A.2 and A.3 for the proofs and the resulting SDP problems.

Theorem 3.3 (Representer theorems with penalty)**.** *Let $(b, \lambda_{\text{low}1}, \lambda_{\text{up}1}, \lambda_{\text{pen}1}) \in \mathbb{R}^4_+$ and $\lambda_{\text{low}2}, \lambda_{\text{up}2}, \lambda_{\text{pen}2} > 0$. Then Problem (6) with Penalty (7) or Penalty (8) admits a unique solution $(\tilde{f}_{\mathbf{A}^\star_{\text{low}}}(X), \tilde{f}_{\mathbf{A}^\star_{\text{up}}}(X)) = \big(\mathbf{\Phi}(X)^\top \mathbf{A}^\star_{\text{low}} \mathbf{\Phi}(X),\ \mathbf{\Phi}(X)^\top \mathbf{A}^\star_{\text{up}} \mathbf{\Phi}(X)\big)$ for some matrices $\big(\mathbf{A}^\star_{\text{low}}, \mathbf{A}^\star_{\text{up}}\big) \in \mathbb{S}^n_+ \times \mathbb{S}^n_+$.*

For scalability, we also derive dual formulations which allow to consider datasets up to thousand samples.

Proposition 3.4 (Dual formulations with penalty)**.** *Let $(b, \lambda_{\text{low}1}, \lambda_{\text{up}1}, \lambda_{\text{pen}1}) \in \mathbb{R}^4_+$ and $\lambda_{\text{low}2}, \lambda_{\text{up}2}, \lambda_{\text{pen}2} > 0$. Problem (6) with Penalty (7) admits the dual formulation*

$$\begin{aligned}\sup_{\substack{(\mathbf{\Gamma}_{\text{low}}, \mathbf{\Gamma}_{\text{up}}) \in \mathbb{R}^{2n}_+ \\ \mathbf{W} \in \mathbb{S}^n}} & (\mathbf{\Gamma}_{\text{up}} - \mathbf{\Gamma}_{\text{low}})\mathbf{r}^\top - \Omega^\star_{\text{pen}}(\mathbf{W}) \\ & - \Omega^\star_{+,\text{low}}(\mathbf{V}_{\text{low}} \text{Diag}(\mathbf{\Gamma}_{\text{low}-\mathbf{b}}) \mathbf{V}_{\text{low}}^\top - \mathbf{W}) \\ & - \Omega^\star_{+,\text{up}}(\mathbf{V}_{\text{up}} \text{Diag}(\mathbf{\Gamma}_{\text{up}-\mathbf{b}}) \mathbf{V}_{\text{up}}^\top + \mathbf{W})\end{aligned}$$

where $\mathbf{r}$ is the vector of residuals $r_i = Y_i - \widehat{m}_n(X_i)$, $\Omega^\star_{+,(\cdot)}(\mathbf{B}) = \frac{1}{4\lambda_{(\cdot)2}} \|\big[\mathbf{B} - \lambda_{(\cdot)1}\mathbf{I}_n\big]_+\|_F^2$, $\Omega^\star_{\text{pen}}(\mathbf{B}) = (1/4\lambda_{\text{pen}2}) \sum_{i=1}^n \max(0, |\lambda_i(\mathbf{B})| - \lambda_{\text{pen}1})^2$, while Problem (6) with Penalty (8) admits the dual formulation

$$\begin{aligned}\sup_{\substack{(\mathbf{\Gamma}_{\text{low}}, \mathbf{\Gamma}_{\text{up}}) \in \mathbb{R}^{2n}_+ \\ \boldsymbol{\alpha}_0 \in \mathbb{R}^n}} & (\mathbf{\Gamma}_{\text{up}} - \mathbf{\Gamma}_{\text{low}})\mathbf{r}^\top - \frac{1}{4\lambda_{\text{pen}}} \boldsymbol{\alpha}_0 \boldsymbol{\alpha}_0^\top \\ & - \Omega^\star_{+,\text{low}}(\mathbf{V}_{\text{low}} \text{Diag}((\mathbf{\Gamma}_{\text{low}} + \boldsymbol{\alpha}_0)_{-\mathbf{b}}) \mathbf{V}_{\text{low}}^\top) \\ & - \Omega^\star_{+,\text{up}}(\mathbf{V}_{\text{up}} \text{Diag}((\mathbf{\Gamma}_{\text{up}} - \boldsymbol{\alpha}_0)_{-\mathbf{b}}) \mathbf{V}_{\text{up}}^\top).\end{aligned}$$

Contrary to the asymmetric case, observe that the dual formulation for the operator penalty involves an optimization problem with $\mathcal{O}(n^2)$ unknowns. Interestingly, the training set penalty however scales linearly, and is thus more suited to larger datasets. The only downside is that controlling equality of lower and upper bands everywhere requires additional assumptions, as elaborated in the next paragraph.

Moreover, both dual formulations naturally accommodate low-rank kernel approximations, enabling them to handle datasets with up to $n = 10000$ samples, see Section 4.3.

Errors bounds. We now give theoretical insights on these two novel penalties. We place ourselves in the ideal setting $\mathcal{H}_{\text{low}} = \mathcal{H}_{\text{up}} = \mathcal{H}$ where we can reach strict equality of lower and upper bands, and introduce assumptions on $\mathcal{H}$ and $\mathcal{X}$, the domain of definition of the inputs X.

Assumption 3.5. For a bounded open set $\mathcal{X} \in \mathbb{R}^d$, the RKHS $\mathcal{H}$ of functions on $\mathcal{X}$ with norm $\|\cdot\|_{\mathcal{H}}$ satisfies $f|\mathcal{X} \in \mathcal{H}, \forall f \in C^\infty(\mathbb{R}^d)$. Moreover $\forall u, v \in \mathcal{H}$, $u \cdot v \in \mathcal{H}$ and $\exists M \geq 1$ such that $\|u \cdot v\|_{\mathcal{H}} \leq \text{M}\|u\|_{\mathcal{H}}\|v\|_{\mathcal{H}}$.

Assumption 3.6. For a bounded open set $\mathcal{X} \in \mathbb{R}^d$, the kernel k of $\mathcal{H}$ satisfies $\max_{|\alpha|=1} \sup_{x,y \in \mathcal{X}} |\partial_x^\alpha \partial_y^\alpha k(x,y)| \leq \text{D}^2 < \infty$ for some $\text{D} \geq 1$.

Assumption 3.7. $\mathcal{X} = \cup_{x \in S} \text{B}_r(x)$, where S is a bounded subset of $\mathbb{R}^d$ and $\text{B}_r(x)$ is the ball of center x and radius r.

Assumptions 3.5 and 3.6 are mild assumptions that hold for classic kernels such as the Matérn $5/2$ one. As noted by Rudi et al. (2025), Assumption 3.7 can be relaxed to $\mathcal{X}$ having Lipschitz continuous boundaries, which will typically hold for most datasets in practice. We can now state our result which provides an upper bound on $\Delta_f := \sup_{x \in \mathcal{X}} |\tilde{f}_{\mathbf{A}_{\text{low}}}(x) - \tilde{f}_{\mathbf{A}_{\text{up}}}(x)|$, the difference between lower and upper bands, see Appendix A.4 for the proof.

Proposition 3.8 (Error bounds)**.** *Let $\mathcal{H}$ be a RKHS with associated kernel k which satisfies Assumption 3.5. If k is bounded such that $\|k\|_\infty := \sup_{x \in \mathcal{X}} \sqrt{k(X,X)} < \infty$, then for any two PSD matrices $\mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}} \in \mathbb{S}^n_+$, we have*

$$\Delta_f \leq \text{M}\|k\|_\infty \, \|\mathbf{A}_{\text{low}} - \mathbf{A}_{\text{up}}\|_\star.$$

If furthermore $\mathcal{X}$ satisfies Assumption 3.7 and $\mathcal{H}$ satisfies Assumption 3.6, then for any finite subset $\widehat{X} = \{X_i\}_{i=1}^n$ of $\mathcal{X}$ and any two PSD matrices $\mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}} \in \mathbb{S}^n_+$, we have

$$\begin{aligned}\Delta_f \leq\ & 2\, C_{\tilde{f}_{\mathbf{A}_{\text{low}}}, \tilde{f}_{\mathbf{A}_{\text{up}}}} \rho_{\widehat{X}, \mathcal{X}} \\ & + \sqrt{\sum_{i=1}^n (\tilde{f}_{\mathbf{A}_{\text{low}}}(X_i) - \tilde{f}_{\mathbf{A}_{\text{up}}}(X_i))^2}\end{aligned}$$

where $C_{\tilde{f}_{\mathbf{A}_{\text{low}}}, \tilde{f}_{\mathbf{A}_{\text{up}}}} = 2d\text{DM}\|\mathbf{A}_{\text{low}} - \mathbf{A}_{\text{up}}\|_\star$ with D and M constants depending on kernel k. $\rho_{\widehat{X}, \mathcal{X}}$ is the so-called fill-in distance defined by $\rho_{\widehat{X}, \mathcal{X}} = \sup_{x \in \mathcal{X}} \min_{X_i \in \widehat{X}} \|x - X_i\|$.

The first part of Proposition 3.8 gives an upper bound which involves the penalty on the operators, while the second part relates to the penalty on the training set.

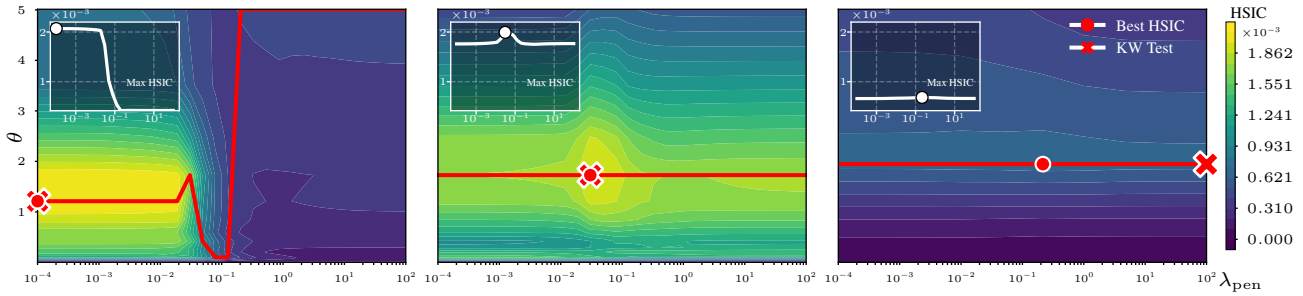


Figure 2. HSIC contour plots for asymmetric noise distribution (left) and symmetric ones (middle, right). Left: lower λ_{pen} values are clearly favored by HSIC. Middle: allowing for some asymmetry produces more adaptive bands than asymmetric or symmetric ones. Right: all λ_{pen} values achieve similar adaptivity. According to the KW test, the symmetric model is preferred over the highest HSIC model.

3.2. Hyperparameter tuning

In our experiments, we observe that $\lambda_{(\cdot)1}$ and $\lambda_{(\cdot)2}$ have small impact on the estimated bands, which confirms the findings of Allain et al. (2026): we thus propose to fix their value at $\lambda_{(\cdot)1} = \lambda_{(\cdot)2} = 1$. On the other hand, b, λ_{pen} and kernel lengthscales $(\theta_{\text{low}}, \theta_{\text{up}})$ play an important role on the shape of the prediction bands. Actually, the width penalty b is compensated by θ_{low} and θ_{up} (see Appendix B.2), such that it is sufficient to fix b and optimize the lengthscales. We thus focus now exclusively on $(\theta_{\text{low}}, \theta_{\text{up}}, \lambda_{\text{pen}})$.

Given that adaptivity is a critical property, it can be used as a principled objective for hyperparameter selection. Perfectly adaptive intervals would satisfy the conditional coverage $\mathbb{P}(Y_{N+1} \in \widehat{C}_{\mathcal{D}_N}(X_{N+1})|X_{N+1} = x) \geq 1 - \alpha$, but such coverage is unfortunately impossible to achieve in a distribution-free setting (Vovk, 2012; Barber et al., 2021a). Alternatively, we consider a relaxed version of conditional coverage $p_{\mathcal{D}_N} = \mathbb{P}(Y_{N+1} \in \widehat{C}_{\mathcal{D}_N}(X_{N+1})|\mathcal{D}_N, X_{N+1} \in \omega_X)$, where we condition on X being in a small neighborhood $\omega_X \in \mathcal{F}_X$ from the event space $\mathcal{F}_X$ such that $\mathbb{P}(X \in \omega_X) \geq \delta$. Deutschmann et al. (2024) recently proved that such coverage with split CP can be controlled with the mutual information between the inputs and the score function. Their bound was generalized in Allain et al. (2026) for normalized scores, with the Hilbert Schmidt Independence Criterion (HSIC, Gretton et al. (2005)) between the residuals and the width of the intervals. The dependence measure is now between one-dimensional variables and quantified with HSIC to improve numerical stability. Interestingly, we show below that it is possible to generalize further their result for asymmetric scores as in Equation (2).

Proposition 3.9. *Let $\widehat{C}_{\mathcal{D}_N}$ be prediction intervals built from a score function $S(X,Y) = \max\big(l(X) - Y, Y - u(X)\big)$ through split CP with $\mathcal{D}_N = \mathcal{D}_n \cup \mathcal{D}_m$. Then:*

$$p_{\mathcal{D}_N} \geq 1 - \alpha - \frac{1}{\delta}\sqrt{1 - \frac{\alpha_1}{1 - \alpha_2 \text{HSIC}(\tilde{r}_{\mathcal{D}_n}, W_{\mathcal{D}_n})}}$$

where $\tilde{r}_{\mathcal{D}_n} = |Y - (\widehat{u}_{\mathcal{D}_n}(X) + \widehat{l}_{\mathcal{D}_n}(X))/2|$ are the absolute centered residuals and $W_{\mathcal{D}_n} = \widehat{u}_{\mathcal{D}_n}(X) - \widehat{l}_{\mathcal{D}_n}(X)$ is the width of the prediction bands. α_1 is a constant and α_2 only depends on the kernel used for HSIC.

The proof can be found in Appendix A.5. Proposition 3.9 provides an intuitive framework: a stronger dependence between the width of the intervals and the absolute centered residuals promotes neighborhood coverage, a relaxed version of conditional coverage. To target local coverage specifically, we thus propose to maximize $\text{HSIC}(\tilde{r}_{\mathcal{D}_n}, W_{\mathcal{D}_n})$. This allows for the precise tuning of $(\theta_{\text{low}}, \theta_{\text{up}})$ according to this criterion, where HSIC is estimated with a cross-validation procedure. To address potentially very small values of HSIC, Allain et al. (2026) advocated the use of a test of independence to determine if its value is significantly different from 0. If not, the simpler homoscedastic model with arbitrary large $(\theta_{\text{low}}, \theta_{\text{up}})$ is chosen.

The final hyperparameter λ_{pen} determines the choice between symmetric and asymmetric prediction bands. A natural approach is to select the level of symmetry that maximizes adaptivity by identifying $(\theta_{\text{low}}, \theta_{\text{up}})$ with highest HSIC for each λ_{pen}. However, maximum HSIC values may remain close across different λ_{pen} values. This raises a question similar to the homoscedastic case, where we ask if the variation in HSIC is statistically significant across multiple values of λ_{pen}. To answer this, we perform a Kruskal-Wallis rank test over bootstrapped HSIC values. If significant differences exist, we select $(\theta_{\text{low}}, \theta_{\text{up}}, \lambda_{\text{pen}})$ that maximizes HSIC. Conversely, if the differences are not significant, we conclude that the asymmetric bands do not offer a clear advantage over symmetric ones. In that case, we default to the simpler, symmetric model: we show this phenomenon in Figure 2. Finally, identifying the best penalty can be computationally expensive, as it requires evaluating multiple λ_{pen}. But we can take full advantage of our dual formulations, by using a warm-start approach which reduces computational cost by $\sim 65\%$ when optimizing over multiple λ_{pen} values, making hyperparameter search practical even for $n \sim 1000$, see Appendix B.2. Algorithm 1 in Appendix B.2 details our full training pipeline.

4. Experiments

We compare our method against established baselines representative of different approaches to CP: CQR as the standard asymmetric method using quantile regression [1], homoscedastic GP as a symmetric adaptive baseline, and heteroscedastic GP (Binois et al., 2018) as an adaptive baseline that can capture varying noise scales. We also consider the symmetric kSoS of Allain et al. (2026). While additional recent methods exist (e.g., DCP, locally weighted CP), these baselines cover the key trade-offs: CQR provides asymmetric intervals but relies on quantile estimation which can be challenging in small samples, GPs provide probabilistic predictions with theoretical guarantees but typically assume Gaussian noise. Our method aims to combine the strengths of both while automatically detecting when asymmetry is beneficial. For all experiments we set $\alpha = 0.1$ and homGP as the predictive model except for asymmetric datasets $3-5$ where we use cubic splines.

4.1. Synthetic datasets

Table 1 details the 4 datasets considered here: datasets 1 and 2 (symmetric) and datasets 3 and 4 (asymmetric).

Table 1. Synthetic datasets, $n = 100$, 20 repetitions. Target is $Y = \mu(X) + \sigma(X)\epsilon$, see Appendix B.3.6 for details.

Dataset	Input X	Mean $\mu(X)$	Scale $\sigma(X)$	Noise ϵ
1	$\mathcal{U}(-1,1)$	$\omega(X)\cdot\mathbb{I}_{X\leq c} + X\cdot\mathbb{I}_{X>c}$	$\sqrt{0.1+2X^2}$	$\mathcal{N}(0,1)$
2	$\mathcal{N}(0,1)$	$0.5X$	$\lvert\sin(X)\rvert$	$\mathcal{N}(0,1)$
3	$\mathcal{U}(-1,1)$	$\sin(5X)$	X	LogNormal(1)
4	$\mathcal{U}(-1,1)$	see B.3.6	see B.3.6	see B.3.6

To assess the quality of prediction intervals, we rely on mean width and introduce two global measures of local coverage. The natural measure of adaptivity is local coverage, which can be estimated for synthetic datasets with conditional samples. We could then compare methods with the *absolute coverage gap*, the distance to the target level α: $\mathrm{ACG} = \frac{1}{n_X}\sum_{i=1}^{n_X} |\hat{p}(X_i) - (1-\alpha)|$ where $\hat{p}$ is an estimate of the local coverage obtained with conditional samples $\{(X_i, \{Y_{ij}\}_{j=1}^{n_Y})\}_{i=1}^{n_X}$. However, in an asymmetric case, local coverage can be misleading as it does not account for the distribution tails: instead, we seek bands that satisfy lower and upper local coverage at $1-\alpha/2$. This readily implies that local coverage will be at least $1-\alpha$, but the converse does not hold.

We thus consider low and up alternate versions of the absolute coverage gap: $\mathrm{ACG}^{(\cdot)} = \frac{1}{n_X}\sum_{i=1}^{n_X} |\hat{p}^{(\cdot)}(X_i) - (1 - \alpha/2)|$ where $\hat{p}^{(\cdot)}$ are low/up local coverage estimations. These metrics allow to capture defects in both tails of the noise distribution, see Appendix B.3.2 for an in-depth discussion. We denote by ACG^c the combination of the low/up variants.

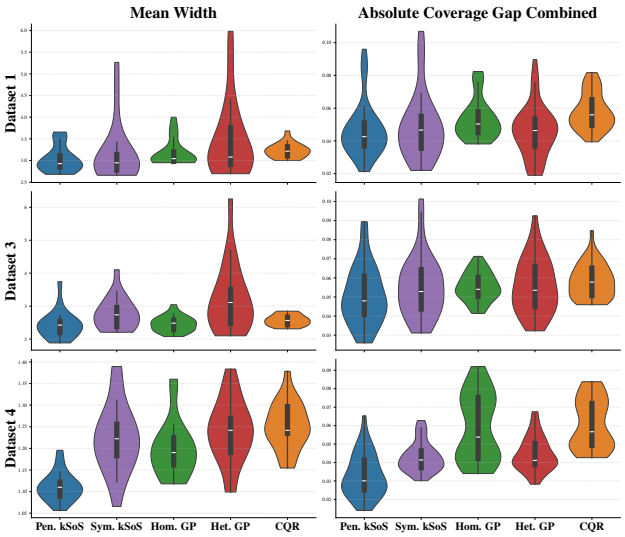


Figure 3. Mean width (left) and absolute coverage gap combined (right) for datasets 1, 3 and 4 with $b = 10$, $n = 100$, 20 repetitions.

For dataset 1 with symmetric noise, Figure 3 top row, shows that both CQR and homGP produce intervals with poor ACG^c. On the other hand, penalized kSoS achieves slightly lower ACG^c than hetGP and symmetric kSoS, but with smaller mean width than hetGP and equivalent to symmetric kSoS. The superior performance of penalized kSoS over symmetric kSoS possibly comes from a small sample counterbalancing effect and the biased predictive model, see Figure 4 for a discussion.

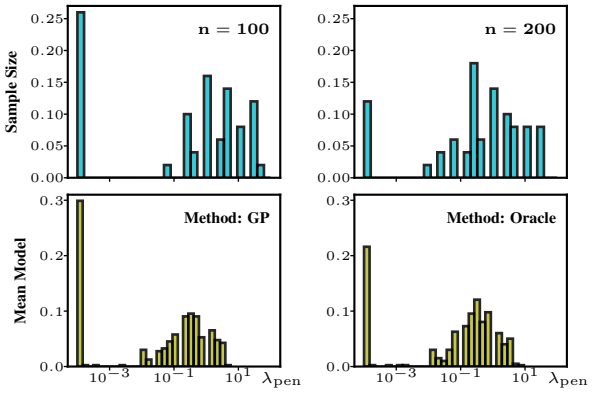


Figure 4. Histogram of selected λ_{pen} among 500 repetitions. When the sample size increases from $n = 100$ to $n = 200$ (top row, dataset 2), and when the predictive model changes from a Gaussian Process to the oracle (bottom row, dataset 1), the purely asymmetric model is selected less often. Our hyperparameter tuning method favors a symmetric model when the sample size increases and when the learned predictive model is more accurate.

[1] We implement CQR with random forests following the original paper's recommendations for tabular data. While other base learners (gradient boosting, neural networks) could be used, random forests provide a strong baseline for our dataset sizes.

For the asymmetric noise distribution in dataset 3, we observe in Figure 3 middle row, that homGP and penalized kSoS are the only methods to produce tight intervals. However, homGP has much higher ACG^c while penalized kSoS achieves the lowest ACG^c among all methods.

Finally, for dataset 4, penalized kSoS has both mean width and ACG^c much lower than all competitors. Penalized kSoS is the only method to achieve good adaptivity while maintaining small mean width in both types of noise distribution. Appendix B.3.6 contains results and discussion for additional test cases.

4.2. Real-word datasets

Finally, we consider 13 real-word datasets commonly used for regression. To measure the performance of prediction bands we again consider mean width, but ACG is now out of reach. To measure adaptivity in practice, we rely on the worst-set coverage introduced by Thurin et al. (2025): $\min_{l=1,\dots,L} \mathbb{P}(Y_{N+1} \in \widehat{C}_{\mathcal{D}_N}(X_{N+1}) | X_{N+1} \in \mathcal{R}_l)$ where $\{\mathcal{R}_l\}_{l=1,\dots,L}$ is a partition of $\mathcal{X}$. As before, we consider low/up variants and denote WSC^c their combination.

Figure 5 reports the obtained mean widths and WSC^c. On Concrete, homGP has the smallest mean width, however only penalized kSoS achieves WSC^c close to $1-\alpha/2$. On Diabetes, homGP and penalized kSoS attain similar WSC^c, but penalized kSoS has much smaller mean width. On this dataset, CQR exhibits similar mean width as penalized kSoS, but underperforms for WSC^c. On Housing, penalized kSoS has a slightly larger mean width but is, by far, the closest one to $1-\alpha/2$ in terms of WSC^c, meaning that it has much better adaptivity. On the remaining datasets, penalized kSoS usually performs better in terms of WSC^c, and when ties occur, it reaches at least similar mean width, if not smaller. Detailed results on all datasets and comparisons with symmetric kSoS are available in Appendix B.3.7.

These findings demonstrate that penalized kSoS is always first or close second when measured by WSC^c and mean width, a robustness never achieved by any other method.

4.3. Low-rank approximation

In our dual formulations from Proposition 3.4, the Fenchel conjugate function $\Omega_+^\star$ requires to compute the positive part and hence the eigendecomposition of a $n \times n$ matrix of the form $\mathbf{V}_{(\cdot)}\mathrm{Diag}(\mathbf{\Gamma})\mathbf{V}_{(\cdot)}^\top$ for some dual variable $\mathbf{\Gamma}$ depending on the penalty type and where $\mathbf{V}_{(\cdot)}$ is the Cholesky decomposition of the kernel matrix $\mathbf{K}_{(\cdot)}$. This means that each evaluation of the dual objective function and its gradient has complexity $\mathcal{O}(n^3)$. To scale our method further, we now show how standard kernel low-rank approximations fit naturally into our dual framework, requiring no modification to the algorithm or theory.

The full rank Cholesky decomposition of $\mathbf{K}_{(\cdot)}$ can be replaced by a low-rank approximation $\mathbf{K}_{(\cdot)} \approx \widetilde{\mathbf{V}}_{(\cdot)}^\top \widetilde{\mathbf{V}}_{(\cdot)}$ where $\widetilde{\mathbf{V}}_{(\cdot)} \in \mathbb{R}^{r \times n}$ with rank $r \ll n$. This reduces the computational complexity of $\Omega_+^\star$ from $\mathcal{O}(n^3)$ to $\mathcal{O}(r^3)$. Several methods can be used to compute a low-rank kernel approximation, such as Nyström (Williams & Seeger, 2000), Random Fourier Features (Rahimi & Recht, 2007) or randomized SVD (Halko et al., 2011). In Figure 6, we illustrate how our dual formulation scales when using such approximation, with a rank deduced to target a reconstruction error of the kernel matrix less than 1% (measured by the Frobenius norm). The dual formulation can be solved with $n = 10000$ samples in 10s, whereas the full rank version requires more than 100s for $n = 1000$.

Importantly, the low-rank dual approach scales without degrading the performance of our method. Table 2 compares penalized kSoS with a low-rank approximation against CQR on some analytical datasets and two real-world ones where the available dataset is large enough to consider a training

Dataset	CQR	Het. GP	Hom. GP	Pen. kSoS
Concrete	22.66 ± 1.06	21.53 ± 1.58	$\mathbf{21.01} \pm 1.58$	21.74 ± 1.19
Bike	217.98 ± 6.54	199.23 ± 14.27	169.57 ± 7.83	$\mathbf{162.47} \pm 5.86$
Diabetes	$\mathbf{188.65} \pm 12.59$	196.25 ± 16.01	193.21 ± 15.80	$\mathbf{190.46} \pm 14.63$
Housing log	0.967 ± 0.039	0.855 ± 0.041	$\mathbf{0.768} \pm 0.040$	0.836 ± 0.037
Housing	1.77 ± 0.055	1.70 ± 0.14	$\mathbf{1.60} \pm 0.082$	1.80 ± 0.14
MPG	9.82 ± 1.06	9.76 ± 1.31	$\mathbf{9.33} \pm 1.02$	$\mathbf{9.32} \pm 1.14$
Boston	12.16 ± 1.18	10.85 ± 0.92	$\mathbf{9.76} \pm 0.80$	11.05 ± 1.23
Energy	1.50 ± 0.16	1.46 ± 0.098	1.72 ± 0.079	$\mathbf{1.39} \pm 0.12$
Miami	$31.4e^4 \pm 1.17e^4$	$36.4e^4 \pm 8.23e^4$	$29.5e^4 \pm 2.29e^4$	$\mathbf{26.9e^4} \pm 1.45e^4$
Sulfur	$\mathbf{0.051} \pm 0.003$	0.055 ± 0.009	$\mathbf{0.050} \pm 0.003$	$\mathbf{0.050} \pm 0.002$
Power	13.29 ± 0.30	$\mathbf{12.88} \pm 0.40$	13.05 ± 0.36	$\mathbf{12.99} \pm 0.29$
Yacht	0.645 ± 0.085	$\mathbf{0.595} \pm 0.086$	0.609 ± 0.082	$\mathbf{0.599} \pm 0.105$
Kin8nm	0.607 ± 0.020	0.425 ± 0.023	$\mathbf{0.386} \pm 0.009$	0.393 ± 0.011

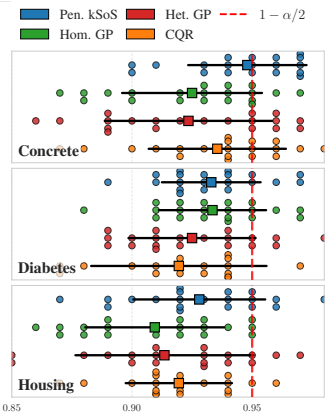


Figure 5. Left: mean width of prediction intervals on the test set for thirteen real-world datasets (mean±sd on 10 repetitions, values within 1% of the minimum in bold). Right: worst-set coverage low/up combined for three datasets.

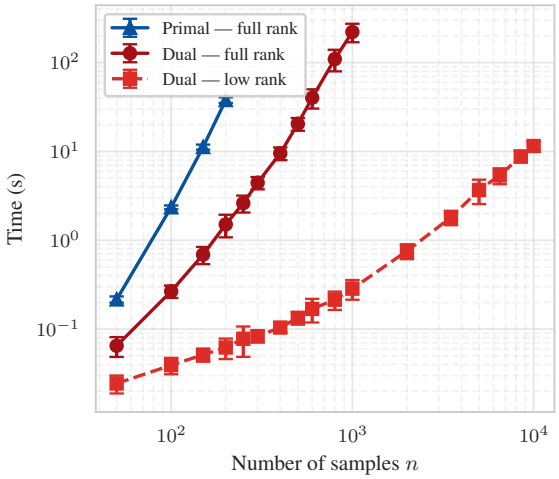


Figure 6. Dataset 1: computing time for solving primal, dual and low-rank dual (reconstruction error $\leq 1\%$) formulations as a function of n (mean±sd over 20 repetitions).

size n up to 10000. Our results show that penalized kSoS still achieves better or equivalent WSC^c in all instances except one, and systematic better mean width.

5. Conclusion

In this work, we introduce a flexible framework for asymmetric conformal prediction bands using kernel sum-of-squares. By incorporating two novel symmetric penalties, our approach seamlessly transitions between symmetric and asymmetric bands. We establish representer theorems that reduce these infinite-dimensional problems to SDP formulations, and derive dual versions to ensure scalability to larger datasets.

Beyond the core optimization, we contribute two significant tuning strategies: an HSIC-based approach for optimizing kernel lengthscales to enhance adaptivity, and a data-driven method for calibrating symmetric penalization levels. Our results demonstrate that this flexibility enhances resilience against small sample sizes, and compensates for bias in predictive models. Crucially, our experiments illustrate that while our framework can automatically detect symmetric noise distributions, allowing for controlled asymmetry can often yield superior adaptivity even when the noise itself is symmetric.

While effective, the current approach faces two primary constraints. First, consistent with standard kernel methods, performance is best suited for dimensions up to approximately 15. However, the framework remains compatible with specific kernels for structured objects such as time series or graphs, which would allow for higher-dimensional applications. Furthermore, the current approach scales efficiently to $n \sim 1000$ via dual formulations and warm-start optimization. For larger datasets, the framework remains compatible with kernel approximation techniques (Nyström, random features), but these extensions require additional theoretical analysis of their impact on bands, which is a promising direction for future work.

All our experiments can be reproduced using the code available at gitlab.com/drti/ksos-bands-asym.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

Table 2. WSC^c and mean width of prediction intervals on the test set for five datasets using low-rank approximation with reconstruction error $\leq 1\%$ (mean±sd on 10 repetitions, values within 1% of best in bold).

Dataset	**Competitor**	$n = 1000$		$n = 2000$		$n = 5000$		$n = 10000$	
		WSC^c	MW	WSC^c	MW	WSC^c	MW	WSC^c	MW
Kin8nm	Pen. kSoS	**0.92** ±0.02	**0.39** ±0.01	**0.92** ±0.02	**0.38** ±0.01	**0.93** ±0.02	**0.36** ±0.01	—	—
	CQR	0.91 ±0.02	0.61 ±0.02	0.89 ±0.03	0.57 ±0.02	0.90 ±0.02	0.52 ±0.01	—	—
Power	Pen. kSoS	**0.89** ±0.03	**13.05** ±0.27	**0.90** ±0.03	**12.66** ±0.41	0.90 ±0.02	**11.13** ±0.58	—	—
	CQR	0.88 ±0.03	13.30 ±0.31	0.89 ±0.04	**12.63** ±0.28	**0.91** ±0.02	11.60 ±0.45	—	—
Dataset 3	Pen. kSoS	**0.86** ±0.03	**2.04** ±0.28	**0.87** ±0.03	**2.00** ±0.10	**0.87** ±0.03	**1.97** ±0.10	**0.88** ±0.02	**1.96** ±0.07
	CQR	0.84 ±0.04	2.41 ±0.13	0.84 ±0.04	2.44 ±0.11	0.84 ±0.03	2.42 ±0.11	0.83 ±0.03	2.39 ±0.13
Dataset 4	Pen. kSoS	**0.91** ±0.03	**1.12** ±0.04	**0.91** ±0.03	**1.12** ±0.03	**0.91** ±0.03	**1.13** ±0.03	**0.91** ±0.03	**1.12** ±0.03
	CQR	0.88 ±0.08	1.25 ±0.06	0.89 ±0.07	1.22 ±0.06	0.90 ±0.08	1.24 ±0.04	0.89 ±0.07	1.24 ±0.06
Dataset 5	Pen. kSoS	**0.89** ±0.05	**3.12** ±0.25	**0.90** ±0.04	**3.38** ±0.30	**0.88** ±0.04	**3.45** ±0.22	**0.89** ±0.04	**3.39** ±0.13
	CQR	0.86 ±0.04	3.52 ±0.21	0.87 ±0.04	3.49 ±0.21	0.86 ±0.04	**3.46** ±0.26	0.86 ±0.04	**3.42** ±0.21

Allain, L., Da Veiga, S., and Staber, B. Scalable and adaptive prediction bands with kernel sum-of-squares. *Advances in Neural Information Processing Systems*, 38:132590–132609, 2026.

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4): 494–591, March 2023. ISSN 1935-8237. doi: 10.1561/2200000101.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(4):455–482, aug 2021a. doi: 10.1093/imaiai/iaaa017.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021b. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS1965.

Becker, S., Fadili, J., and Ochs, P. On quasi-newton forward-backward splitting: proximal calculus and convergence. *SIAM Journal on Optimization*, 29(4):2445–2481, 2019.

Binois, M., Gramacy, R. B., and Ludkovski, M. Practical heteroscedastic gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821, 2018.

Borwein, J. and Lewis, A. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, New York, 2006. doi: https://doi.org/10.1007/978-0-387-31256-9.

Braun, S., Aolaritei, L., Jordan, M. I., and Bach, F. Minimum volume conformal sets for multivariate regression. *arXiv preprint arXiv:2503.19068*, 2025. URL https://arxiv.org/abs/2503.19068.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

Cheung, M. Y., Netherton, T. J., Court, L. E., Veeraraghavan, A., and Balakrishnan, G. Bias-aware conformal prediction for metric-based imaging pipelines. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pp. 01–05. IEEE, 2026.

Chung, Y., Neiswanger, W., Char, I., and Schneider, J. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. *Advances in Neural Information Processing Systems*, 34:10971–10984, 2021.

Deutschmann, N., Rigotti, M., and Martinez, M. R. Adaptive conformal regression with split-jackknife+ scores. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of statistics*, 32(2):407–451, 2004.

Fanaee-T, H. Bike Sharing. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C5W894.

Fortuna, L., Graziani, S., Rizzo, A., and Xibilia, M. G. *Soft sensors for monitoring and control of industrial processes*. Springer, 2007.

Gammerman, A., Vovk, V., and Vapnik, V. Learning by transduction. In *Conference on Uncertainty in Artificial Intelligence*, 1998.

Gerritsma, J., Onnink, R., and Versluis, A. Yacht Hydrodynamics. UCI Machine Learning Repository, 1981. DOI: https://doi.org/10.24432/C5XG7R.

Ghahramani, Z. The Kin family of datasets. DELVE Repository, University of Toronto, 1996.

Gibbs, I., Cherian, J. J., and Candès, E. J. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1100–1126, 03 2025. ISSN 1369-7412. doi: 10.1093/jrsssb/qkaf008.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Gramacy, R. B. and Lee, H. K. H. Adaptive design and analysis of supercomputer experiments. *Technometrics*, 51(2):130–145, May 2009. ISSN 1537-2723. doi: 10.1198/tech.2009.0015.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Halko, N., Martinsson, P.-G., and Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

Han, X., Tang, Z., Ghosh, J., and Liu, Q. Split Localized Conformal Prediction. *arXiv preprint arXiv:2206.13092*, 2023. URL http://arxiv.org/abs/2206.13092.

Harrison Jr, D. and Rubinfeld, D. L. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

Hore, R. and Barber, R. F. 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.

Kivaranovic, D., Johnson, K. D., and Leeb, H. Adaptive, distribution-free prediction intervals for deep networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 4346–4356. PMLR, 2020.

Lei, J. and Wasserman, L. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1): 71–96, jan 2014. doi: 10.1111/rssb.12021.

Lewis, A. S. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1):173–183, 1995.

Lewis, A. S. Derivatives of spectral functions. *Mathematics of Operations Research*, 21(3):576–588, 1996.

Linusson, H., Johansson, U., and Löfström, T. Signed-Error Conformal Regression. In *Advances in Knowledge Discovery and Data Mining*, volume 8443, pp. 224–236. Springer International Publishing, 2014. doi: 10.1007/978-3-319-06608-0_19.

Liu, D. C. and Nocedal, J. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Marteau-Ferey, U., Bach, F., and Rudi, A. Non-parametric models for non-negative functions. *Advances in neural information processing systems*, 33:12816–12826, 2020.

Mayer, M., Bourassa, S. C., Hoesli, M., and Scognamiglio, D. Machine learning applications to land and structure valuation. *Journal of Risk and Financial Management*, 15(5):193, 2022.

O'Donoghue, B. Operator splitting for a homogeneous embedding of the linear complementarity problem. *SIAM Journal on Optimization*, 31:1999–2023, August 2021.

O'Donoghue, B., Chu, E., Parikh, N., and Boyd, S. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, June 2016.

O'Donoghue, B., Chu, E., Parikh, N., and Boyd, S. SCS: Splitting conic solver, version 3.2.7, November 2023.

Pace, R. K. and Barry, R. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *European Conference on Machine Learning*, 2002.

Pedersen, G. K. *Analysis now*, volume 118. Springer Science & Business Media, 2012.

Pion, A. and Vazquez, E. Gaussian process interpolation with conformal prediction: Methods and comparative analysis. In Nicosia, G., Ojha, V., Giesselbach, S., Pardalos, M. P., and Umeton, R. (eds.), *Machine Learning, Optimization, and Data Science*, pp. 218–228, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-82484-5.

Pouplin, T., Jeffares, A., Seedat, N., and Van Der Schaar, M. Relaxed quantile regression: Prediction intervals for asymmetric noise. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40951–40981. PMLR, 21–27 Jul 2024.

Quinlan, R. Auto MPG. UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5859H.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Rockafellar, R. T. *Convex Analysis*. Princeton University Press, Princeton, 1970. doi: doi:10.1515/9781400873173.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Rudi, A., Marteau-Ferey, U., and Bach, F. Finding global minima via kernel approximations. *Mathematical Programming*, 209(1):703–784, 2025.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pp. 2263–2291, 2013.

Sesia, M. and Romano, Y. Conformal prediction using conditional histograms. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6304–6315. Curran Associates, Inc., 2021.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A hilbert space embedding for distributions. In *International conference on algorithmic learning theory*, pp. 13–31. Springer, 2007.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Tfekci, P. and Kaya, H. Combined Cycle Power Plant. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5002N.

Thurin, G., Nadjahi, K., and Boyer, C. Optimal transport-based conformal prediction. *PMLR*, 267:59509–59527, 13–19 Jul 2025.

Tsanas, A. and Xifara, A. Energy Efficiency. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C51307.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Vovk, V. Conditional validity of inductive conformal predictors. In Hoi, S. C. H. and Buntine, W. (eds.), *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR.

Wang, C. X. and Tay, W. P. Semi-nonparametric estimation of distribution divergence in non-euclidean spaces. *arXiv preprint arXiv:2204.02031*, 2023. URL `https://arxiv.org/abs/2204.02031`.

Wendland, H. and Rieger, C. Approximate interpolation with applications to selecting smoothing parameters. *Numerische Mathematik*, 101(4):729–748, 2005.

Williams, C. and Seeger, M. Using the nyström method to speed up kernel machines. In Leen, T., Dietterich, T., and Tresp, V. (eds.), *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

Xie, R., Barber, R. F., and Candès, E. J. Boosted conformal prediction intervals. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 71868–71899. Curran Associates, Inc., 2024.

Yeh, I.-C. Concrete Compressive Strength. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C5PK67.

This appendix is organized as follows:

- **Section A: Theoretical Proofs**
  - A.1: Asymmetric problem without penalty (Theorem 3.1, Proposition 3.2)
  - A.2: Asymmetric problem with operator penalty (Theorem 3.3, Proposition 3.4)
  - A.3: Asymmetric problem with training set penalty (Theorem 3.3, Proposition 3.4)
  - A.4: Error bounds for both penalties (Proposition 3.8)
  - A.5: HSIC-based coverage bounds (Proposition 3.9)
- **Section B: Additional Experiments**
  - B.1: Cross-validation and Kruskal-Wallis test details
  - B.2: Implementation details, computational complexity, warm-start
  - B.3: Extended experimental results
    - Illustration of asymmetry and adaptivity of noise distribution
    - Evaluation metrics discussion and justification
    - Differences between operator and training set penalties
    - Comparison with symmetric prediction bands
    - Symmetric versus asymmetric calibration
    - Additional synthetic test cases (datasets 1–5)
    - Complete real-world dataset descriptions and results

Key results we reference from the main text:

- For the discussion on adaptivity and asymmetry: Section B.3.1
- For scalability: Section B.2 (warm-start), Figure 9
- For algorithm description: Section B.2
- For adaptivity metric justification: Section B.3.2 (Figure 11)
- For penalty selection patterns: Section B.2 (Figure 8), B.3 (Figure 26)

A. Proofs

To begin, we first introduce some notation common to several proofs. Since we are working with kernel SoS functions depending on operators defined on different RKHSs, we use subscripts $(\cdot)$ to differentiate them. Associated to each kernel SoS function, we thus 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

$$\mathbf{k}_{(\cdot)}(X) = \big(k_{(\cdot)}(X_1, X), \ldots, k_{(\cdot)}(X_n, X)\big)^\top.$$

defined for all $X \in \mathcal{X}$.

For kernel matrices $[\mathbf{K}_{(\cdot)}]_{ij} = k_{(\cdot)}(X_i, X_j)$, we further consider their Cholesky decomposition and empirical feature map

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

Depending on the theorems, $(\cdot)$ will refer to either $_{\mathrm{low}}$ and $_{\mathrm{up}}$, or will be used to enumerate a collection of kernel SoS functions $s = 1, \ldots, p$.

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 square matrices of size n and $\mathbb{S}^n_+ = \mathbb{S}_+(\mathbb{R}^{n\times n})$ the set of real, symmetric and positive-definite square matrices of size n. We also consider $m(\cdot)$ to be a fixed predictive model which has been trained separately.

In some of the following proofs, we will rely on a generalization of Theorem 1 from Marteau-Ferey et al. (2020) to different operators on p different spaces, that we will particularize to our setting. We then start by proving this extension.

Let us consider a collection of p operators

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

and denote the following multivariate function

$$f_{\mathcal{A}}(X) = f_{\mathcal{A}_1,\ldots,\mathcal{A}_p}(X) = \big(f_{\mathcal{A}_1}(X), \ldots, f_{\mathcal{A}_p}(X)\big) \in \mathbb{R}^p_+, \ \forall X \in \mathcal{X}$$

where $f_{\mathcal{A}_s}(X) = \langle \phi_s(X), \mathcal{A}_s \phi_s(X) \rangle_{\mathcal{H}_s}$, $s = 1, \ldots, p$. We also introduce below a specific class of regularizers.

Assumption A.1. Let $\mathcal{H}$ be a Hilbert space, for any $\mathcal{A} \in \mathcal{S}(\mathcal{H})$, $\Omega(\mathcal{A})$ is of the form

$$\Omega(\mathcal{A}) = \begin{cases} \mathrm{Tr}(q(\mathcal{A})) = \sum_k q(\sigma_k) & \text{if } \mathcal{A} = U\mathrm{Diag}(\sigma)U^\top \in \mathcal{S}_\infty(\mathcal{H}),\ \sum_k q(\sigma_k) < +\infty \\ +\infty & \text{otherwise,} \end{cases} \tag{9}$$

where $q\colon \mathbb{R} \to \mathbb{R}_+$ is:

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

We can define for each operator $\mathcal{A}_s$ a penalty function Ω_s, and consider the aggregated penalty function

$$\Omega_{\mathrm{agg}}(\mathcal{A}) = \sum_{s=1}^p \Omega_s(\mathcal{A}_s). \tag{10}$$

The following theorem is the main result and shows that Equation (11) admits a solution that has a finite dimensional representation.

Theorem A.2. *Let L be a lower semi-continuous and bounded below function and let Ω_{agg} be defined as in Equation* (10) *where each Ω_s, $s = 1, \ldots, p$ satisfies Assumption A.1. The problem*

$$\inf_{\mathcal{A} \in \mathcal{K}^C(\mathcal{H}_1,\ldots,\mathcal{H}_p)} L\left(f_{\mathcal{A}}(X_1), \ldots, f_{\mathcal{A}}(X_n)\right) + \Omega_{\mathrm{agg}}(\mathcal{A}) \tag{11}$$

admits a solution $\mathcal{A}^\star = \left(\sum_{i,j=1}^n \mathbf{B}_{sij} \phi_s(X_i)\phi_s(X_j)^\top\right)_{1 \le s \le p}$, such that $f_{\mathcal{A}^\star}$ can be written as

$$f_{\mathcal{A}^\star}(X) = \Big(\sum_{i,j=1}^n \mathbf{B}_{sij} \phi_s(X_i)\phi_s(X_j)^\top\Big)_{1 \le s \le p} \tag{12}$$

for p matrices $\mathbf{B}_s \in \mathbb{R}^{n \times n}$, $\mathbf{B}_s \succeq 0$.

Proof. We follow the proof of Marteau-Ferey et al. (2020), Section B.3. The first step is to prove the following lemma, which generalizes Lemma 2 of Marteau-Ferey et al. (2020).

Lemma A.3. *The function Ω_{agg} in Equation* (10) *satisfies the following properties:*

1. *For any collection of Hilbert spaces $\left(\mathcal{H}_1^1, \ldots, \mathcal{H}_p^1\right)$ and $\left(\mathcal{H}_1^2, \ldots, \mathcal{H}_p^2\right)$ and any linear isometries $O_s\colon \mathcal{H}_s^1 \to \mathcal{H}_s^2$, $s = 1, \ldots, p$, it holds that for all $\mathcal{A}_1, \ldots, \mathcal{A}_p \in \mathcal{K}^C\left(\mathcal{H}_1^1, \ldots, \mathcal{H}_p^1\right)$:*

$$\sum_{s=1}^p \Omega_s(O_s \mathcal{A}_s O_s^*) = \sum_{s=1}^p \Omega_s(\mathcal{A}_s).$$

2. *For any collection of Hilbert spaces $(\mathcal{H}_1, \dots, \mathcal{H}_p)$ and any orthogonal projections $\Pi_s \in \mathcal{S}(\mathcal{H}_s)$, $s = 1, \dots, p$, it holds that for all $\mathcal{A}_1, \dots, \mathcal{A}_p \in \mathcal{K}^C(\mathcal{H}_1, \dots, \mathcal{H}_p)$:*

$$\sum_{s=1}^{p} \Omega_s(\Pi_s \mathcal{A}_s \Pi_s) \leq \sum_{s=1}^{p} \Omega_s(\mathcal{A}_s)$$

3. *For any collection of finite dimensional Hilbert spaces $(\mathcal{H}_{n1}, \dots, \mathcal{H}_{np})$, it holds that:*

$$\Omega_{\text{agg}} \text{ is lower semi} - \text{continuous} \quad \text{and} \quad \Omega_{\text{agg}}(\mathcal{A}) = \sum_{s=1}^{p} \Omega_s(\mathcal{A}_s) \underset{\sup_s \|\mathcal{A}_s\|_{\text{op}} \to +\infty}{\longrightarrow} +\infty \tag{13}$$

Proof. Ω_{agg} is defined as a separable sum of $\Omega_s, s = 1, \dots, p$ which all satisfy Lemma 2 of Marteau-Ferey et al. (2020) under Assumption A.1. Properties 1, 2 and continuity in 3 are thus straightforward by applying equalities and inequalities element-wise on each $\Omega_s, s = 1, \dots, p$. The last part of Property 3 comes from Lemma 2, (iii) from Marteau-Ferey et al. (2020) applied to $\Omega_{s^*}(\mathcal{A}_{s^*})$ for s^* corresponding to $\sup_s \|\mathcal{A}_s\|_{\text{op}}$. □

Now, the first part of Theorem A.2 comes from the following proposition, which shows that the solution of the infinite-dimensional problem of Equation (11) with p operators and penalty function in Equation (10) lies in a space of finite dimension. For every Hilbert space $\mathcal{H}_s$, we write $\mathcal{H}_{ns}$ for the finite dimensional subspace of $\mathcal{H}_s$ generated by $\{(\phi_s(X_i))_{1 \leq i \leq n}\}$, and define Π_{ns} the orthogonal projection on $\mathcal{H}_{ns}$ such that:

$$\Pi_{ns} \in \mathcal{S}(\mathcal{H}_s), \quad \Pi_{ns}^2 = \Pi_{ns}, \quad \text{range}(\Pi_{ns}) = \mathcal{H}_{ns}.$$

Finally, we denote

$$\mathcal{K}_n^C(\mathcal{H}_1, \dots, \mathcal{H}_p) \coloneqq \left\{(\Pi_{n1}\mathcal{A}_1\Pi_{n1}, \dots, \Pi_{np}\mathcal{A}_p\Pi_{np}) : \ (\mathcal{A}_1, \dots, \mathcal{A}_p) \in \mathcal{K}^C(\mathcal{H}_1, \dots, \mathcal{H}_p)\right\}.$$

Since $\forall 1 \leq s \leq p, \ \Pi_{ns}\mathcal{A}_s\Pi_{ns} \in \mathcal{S}(\mathcal{H}_s)$ and $\Pi_{ns}\mathcal{A}_s\Pi_{ns} \geq 0$, we have the inclusion:

$$\mathcal{K}_n^C(\mathcal{H}_1, \dots, \mathcal{H}_p) \subset \mathcal{K}^C(\mathcal{H}_1, \dots, \mathcal{H}_p). \tag{14}$$

Proposition A.4. *Let L be a lower semi-continuous function, bounded below and let Ω_{agg} be defined as in Equation* (10) *where each $\Omega_s, s = 1, \dots, p$ satisfies Assumption A.1. The problem*

$$\inf_{\mathcal{A} \in \mathcal{K}^C(\mathcal{H}_1, \dots, \mathcal{H}_p)} L\left(f_{\mathcal{A}}(X_1), \dots, f_{\mathcal{A}}(X_n)\right) + \Omega_{\text{agg}}(\mathcal{A})$$

admits a solution $\mathcal{A}^\star \in \mathcal{K}_n^C(\mathcal{H}_1, \dots, \mathcal{H}_p)$.

Proof. Let us first write for any $\mathcal{A} \in \bigotimes_{s=1}^{p} \mathcal{S}(\mathcal{H}_s), \ J(\mathcal{A}) \coloneqq L\left((f_{\mathcal{A}}(X_i))_{1 \leq i \leq n}\right) + \Omega_{\text{agg}}(\mathcal{A})$. The goal here is to show that $\inf_{\mathcal{A} \in \mathcal{K}^C(\mathcal{H}_1, \dots, \mathcal{H}_p)} J(\mathcal{A})$ admits a solution in $\mathcal{K}_n^C(\mathcal{H}_1, \dots, \mathcal{H}_p)$. The proof consists of three main steps:

- we first show that
$$\inf_{\mathcal{A} \in \mathcal{K}_n^C(\mathcal{H}_1, \dots, \mathcal{H}_p)} J(\mathcal{A}) = \inf_{\mathcal{A} \in \mathcal{K}^C(\mathcal{H}_1, \dots, \mathcal{H}_p)} J(\mathcal{A}) \tag{15}$$
- then, we show that if $\inf_{\mathcal{A} \in \mathcal{K}_n^C(\mathcal{H}_1, \dots, \mathcal{H}_p)} J(\mathcal{A})$ exists, it is attained for operators with bounded Frobenius norm
- finally, we show that such minimum exists.

First step. This is a direct generalization of the proof of Proposition 7 in Marteau-Ferey et al. (2020). Using the projections $\Pi_{ns}, s = 1, \dots, p$, it is easy to show that for all $1 \leq i \leq n$, $f_{\mathcal{A}}(X_i) = f_{(\Pi_{ns}\mathcal{A}_s\Pi_{ns})_{1 \leq s \leq p}}(X_i)$ and that $\Omega_{\text{agg}}(\mathcal{A}) = \Omega_{\text{agg}}\left((\Pi_{ns}\mathcal{A}_s\Pi_{ns})_{1 \leq s \leq p}\right)$ (using Property 2 from Lemma A.3). Putting those results together and using the inclusion in Equation (14), the first step follows.

Second step. We rigorously mirror the proof of Proposition 7 in Marteau-Ferey et al. (2020). The key idea is to show that we can replace $\mathcal{K}_n^C(\mathcal{H}_1, \ldots, \mathcal{H}_p)$ with

$$\mathcal{K}_{R_0} \coloneqq \left\{ \tilde{\mathcal{A}} \in \mathcal{K}^C(\mathcal{H}_{n1}, \ldots, \mathcal{H}_{np}) : \forall 1 \leq s \leq p, \|\tilde{\mathcal{A}}_s\|_F \leq R_0 \right\} \tag{16}$$

where R_0 is a constant. First, by defining for all $1 \leq s \leq p$ the injections $V_{ns} \colon \mathcal{H}_{ns} \to \mathcal{H}_s$ (s.t. $V_{ns}V_{ns}^* = \Pi_{ns}$ and $V_{ns}^* V_{ns} = I_{H_{ns}}$), we have:

$$\inf_{\mathcal{A} \in \mathcal{K}_n^C(\mathcal{H}_1, \ldots, \mathcal{H}_p)} J(\mathcal{A}) = \inf_{\tilde{\mathcal{A}}_1, \ldots, \tilde{\mathcal{A}}_p \in \mathcal{K}^C(\mathcal{H}_{n1}, \ldots, \mathcal{H}_{np})} J\left((V_{ns}\tilde{\mathcal{A}}_s V_{ns}^*)_{1 \leq s \leq p} \right).$$

Next, using Property 1 from Lemma A.3 it is immediate to show that for any $\tilde{\mathcal{A}}_1, \ldots, \tilde{\mathcal{A}}_p \in \mathcal{K}^C(\mathcal{H}_{n1}, \ldots, \mathcal{H}_{np})$,

$$J\left((V_{ns}\tilde{\mathcal{A}}_s V_{ns}^*)_{1 \leq s \leq p} \right) = L\left(f_{(V_{ns}\tilde{\mathcal{A}}_s V_{ns}^*)_{1 \leq s \leq p}}(X_i) \right) + \sum_{s=1}^{p} \Omega_s(\tilde{\mathcal{A}}_s). \tag{17}$$

Now, let $(\tilde{\mathcal{A}}_{0s})_{1 \leq s \leq p} \in \mathcal{K}^{\mathcal{C}}(\mathcal{H}_{n1}, \ldots, \mathcal{H}_{np})$ be a point such that $J_0 \coloneqq J\left((\tilde{\mathcal{A}}_{0s})_{1 \leq s \leq p} \right) < +\infty$, and let c_0 be a lower bound for L. By Property 3 of Lemma A.3, there exists a radius R_0 such that $\forall (\tilde{\mathcal{A}}_s)_{1 \leq s \leq p} \in \mathcal{K}^{\mathcal{C}}(\mathcal{H}_{n1}, \ldots, \mathcal{H}_{np})$, $\exists s_0 \in [p]$ satisfying:

$$\|\tilde{\mathcal{A}}_{s_0}\|_F > R_0 \Longrightarrow \sum_{s=1}^{p} \Omega_s(\tilde{\mathcal{A}}_s) > J_0 - c_0.$$

This means that the infimum of Equation (17), if it exists, lies in the space of operators with bounded Frobenius norm. Finally, c_0 being a lower bound for L, this implies that

$$\inf_{\tilde{\mathcal{A}}_1, \ldots, \tilde{\mathcal{A}}_p \in \mathcal{K}^C(\mathcal{H}_{n1}, \ldots, \mathcal{H}_{np})} J\left((V_{ns}\tilde{\mathcal{A}}_s V_{ns}^*)_{1 \leq s \leq p} \right) = \inf_{\substack{\tilde{\mathcal{A}}_1, \ldots, \tilde{\mathcal{A}}_p \in \mathcal{K}^C(\mathcal{H}_{n1}, \ldots, \mathcal{H}_{np}) \\ \forall s,\ \|\tilde{\mathcal{A}}_s\| \leq R_0}} J\left((V_{ns}\tilde{\mathcal{A}}_s V_{ns}^*)_{1 \leq s \leq p} \right).$$

Third step. We finally show that our problem admits a solution, because we minimize a lower semi-continuous function on a non empty compact space. Observe first that $\mathcal{K}_{R_0}$ is a non empty compact space. This follows directly from the fact that $\mathcal{H}_{n1}, \ldots, \mathcal{H}_{np}$ are finite dimensional: the space $\mathcal{K}_{R_0}$ is compact (closed and bounded) and non empty (it contains $(\tilde{\mathcal{A}}_{0s})_{1 \leq s \leq p}$). Next, the function

$$(\tilde{\mathcal{A}}_s)_{1 \leq s \leq p} \longrightarrow J\left((V_{ns}\tilde{\mathcal{A}}_s V_{ns}^*)_{1 \leq s \leq p} \right) \tag{18}$$

is lower semi-continuous, as a composition of the linear semi-continuous function L with the linear (thus continuous) function $(\tilde{\mathcal{A}}_s)_{1 \leq s \leq p} \longrightarrow (f_{(V_{ns}\tilde{\mathcal{A}}_s V_{ns}^*)_{1 \leq s \leq p}})_{1 \leq i \leq n}$, plus the linear semi-continuous penalty function Ω (Assumption A.1).

Thus, the function in Equation (18) admits a minimum on any non empty compact set, and in particular $\mathcal{K}_{R_0}$:

$$\exists \tilde{\mathcal{A}}^\star \in \mathcal{K}_{R_0} \colon\ J\left((V_{ns}\tilde{\mathcal{A}}_s^\star V_{ns}^*)_{1 \leq s \leq p} \right) = \inf_{\substack{\tilde{\mathcal{A}}_1, \ldots, \tilde{\mathcal{A}}_p \in \mathcal{K}^C(\mathcal{H}_{n1}, \ldots, \mathcal{H}_{np}) \\ \forall s,\ \|\tilde{\mathcal{A}}_s\| \leq R_0}} J\left((V_{ns}\tilde{\mathcal{A}}_s V_{ns}^*)_{1 \leq s \leq p} \right).$$

By going back up to the previous equalities, we have that

$$J(\mathcal{A}^\star) = \inf_{\mathcal{A} \in \mathcal{K}^C(\mathcal{H}_1, \ldots, \mathcal{H}_p)} J(\mathcal{A})$$

with $\mathcal{A}^\star = (V_{ns}\tilde{\mathcal{A}}_s V_{ns}^*)_{1 \leq s \leq p}$, which completes the proof of Proposition A.4. □

The last part is to show that $\mathcal{A}^\star \in \mathcal{K}_n^C(\mathcal{H}_1, \ldots, \mathcal{H}_p)$ the solution from Proposition A.4, leads to a function f as in Equation (12). This is achieved with the following lemma.

Lemma A.5. *The set $\mathcal{K}_n^C(\mathcal{H}_1, \dots, \mathcal{H}_p)$ can be represented as*

$$\mathcal{K}_n^C(\mathcal{H}_1, \dots, \mathcal{H}_p) = \left\{ \Big(\sum_{i,j=1}^n \mathbf{B}_{sij} \phi_s(X_i)\phi_s(X_j)^\top \Big)_{1 \le s \le p} \; : \; \forall 1 \le s \le p, \; \mathbf{B}_s \in \mathbb{R}^{n \times n}, \mathbf{B}_s \succeq 0 \right\}.$$

In particular, for any $\mathcal{A} \in \mathcal{K}_n^C(\mathcal{H}_1, \dots, \mathcal{H}_p)$, there exists p matrices $\mathbf{B}_s \in \mathbb{R}^{n \times n}$, $\mathbf{B}_s \succeq 0$ such that

$$\mathcal{A} = \Big(\sum_{i,j=1}^n \mathbf{B}_{sij} \phi_s(X_i)\phi_s(X_j)^\top \Big)_{1 \le s \le p} \implies \forall X \in \mathcal{X}, f_{\mathcal{A}}(X) = \Big(\sum_{i,j=1}^n \mathbf{B}_{sij} k_s(X_i, X) k_s(X_j, X) \Big)_{1 \le s \le p}.$$

Proof. We apply Lemma 3 of Marteau-Ferey et al. (2020) to each individual operators $\mathcal{A}_s \in \mathcal{S}_+(\mathcal{H}_s)$ to retrieve the result on the collection of operators $(A_s)_{1 \le s \le p}$. □

Applying Lemma A.5 to the solution $\mathcal{A}^\star \in \mathcal{K}_n^C(\mathcal{H}_1, \dots, \mathcal{H}_p)$ concludes the proof of Theorem A.2. □

Theorem A.2 provides a finite-dimensional equivalent problem which involves unknown PSD matrices $\mathbf{B}_1, \dots, \mathbf{B}_p$. Marteau-Ferey et al. (2020) and Allain et al. (2026) also propose an equivalent formulation with a different parameterization in terms of PSD matrices $\mathbf{A}_1, \dots, \mathbf{A}_p$ based on the Cholesky decomposition of the kernels. More precisely, define

$$\tilde{f}_{\mathbf{A}_s}(X) = \mathbf{\Phi}_s(x)^\top \mathbf{A}_s \, \mathbf{\Phi}_s(x) \tag{19}$$

for $s = 1, \dots, p$. With these notations, the following proposition shows that we obtain the same solution if we optimize the PSD matrices $\mathbf{A}_1, \dots, \mathbf{A}_p$ instead of $\mathbf{B}_1, \dots, \mathbf{B}_p$.

Proposition A.6. *Under the assumptions of Theorem A.2, the following problem has at least one solution, which is unique if for all $1 \le s \le p$, $\lambda_{s2} > 0$ and L is convex:*

$$\inf_{\mathbf{A}_1, \dots, \mathbf{A}_p \in \mathbb{S}_+^n} L\left(\tilde{f}_{(\mathbf{A}_s)_{1 \le s \le p}}(X_i) \right) + \sum_{s=1}^p \Omega_s(\mathbf{A}_s). \tag{20}$$

Moreover, for any given solution $\mathbf{A}_1^\star, \dots, \mathbf{A}_p^\star \in \mathbb{S}_+^n$ of Equation (20)*, the function $\tilde{f}_{(\mathbf{A}_s^\star)_{1 \le s \le p}}$ is a minimizer of Equation* (11)*.*

Proof. For each $s \in \{1, \dots, p\}$, define the operator

$$S_{ns} : \mathcal{H}_s \to \mathbb{R}^n, \quad S_{ns}(h) = \big(\langle h, \phi_s(x_i) \rangle_{\mathcal{H}_s} \big)_{1 \le i \le n}$$

and its adjoint $S_{ns}^* : \mathbb{R}^n \to \mathcal{H}_s$ is then given by

$$S_{ns}^* \alpha = \sum_{i=1}^n \alpha_i \, \phi_s(x_i), \quad \forall \alpha \in \mathbb{R}^n.$$

It follows that $\mathbf{K}_s = S_{ns} S_{ns}^*$. Then, for each $s \in \{1, \dots, p\}$, we define

$$O_{ns} : \mathbb{R}^n \to \mathcal{H}_s, \quad O_{ns} = S_{ns}^* \mathbf{V}_s^\top \big(\mathbf{V}_s \mathbf{V}_s^\top \big)^{-1},$$

where we assume that $\mathbf{K}_s$ is full rank (this is the case when using a universal kernel as the Matérn one, and if all training points X_i are distinct). Further note that all O_{ns} satisfy Lemma 4 from Marteau-Ferey et al. (2020).

Now, for each $s \in \{1, \dots, p\}$, define:

$$\tilde{f}_{\mathbf{A}_s}(x) = \mathbf{\Phi}_s(x)^\top \mathbf{A}_s \, \mathbf{\Phi}_s(x), \quad \mathbf{A}_s \in \mathbb{R}^{n \times n}, \quad \mathbf{A}_s \succeq 0,$$

where the feature map $\mathbf{\Phi}_s : \mathcal{X} \to \mathbb{R}^n$ is defined as $\mathbf{\Phi}_s(x) = O_{ns}^* \, \phi_s(x)$, such that

$$f_{O_{ns} \mathbf{A}_s O_{ns}^*} = \tilde{f}_{\mathbf{A}_s}. \tag{21}$$

Moreover, as in Marteau-Ferey et al. (2020) proof of Proposition 3, Equation (b),

$$\{\, O_{ns}\mathbf{A}_s O_{ns}^* : \mathbf{A}_s \in \mathbb{R}^{n\times n},\ \mathbf{A}_s \succeq 0 \,\} = \mathcal{S}_n(\mathcal{H}_s)_+ \tag{22}$$

where $\mathcal{S}_n(\mathcal{H}_s)_+$ corresponds to $\mathcal{K}_n^C(\mathcal{H}_1,\dots,\mathcal{H}_p) = \bigotimes_{s=1}^p \mathcal{S}_n(\mathcal{H}_s)_+$.

Finally, since O_{ns} is an isometry, we have

$$\Omega_s(O_{ns}\mathbf{A}_s O_{ns}^*) = \Omega_s(\mathbf{A}_s), \tag{23}$$

and using Equations 21 and 23 it holds that

$$J\left((O_{ns}\mathbf{A}_s O_{ns}^*)_{1\le s\le p}\right) = J\left((\mathbf{A}_s)_{1\le s\le p}\right).$$

Finally, it follows from Equation (22) that

$$\inf_{\mathbf{A}_1,\dots,\mathbf{A}_p\in\mathbb{S}_+^n} J\left((\mathbf{A}_s)_{1\le s\le p}\right) = \inf_{\mathcal{B}_1,\dots,\mathcal{B}_p\in\mathcal{K}_n^C(\mathcal{H}_1,\dots,\mathcal{H}_p)} J\left((\mathcal{B}_s)_{1\le s\le p}\right) = \inf_{\mathcal{B}\in\mathcal{K}^C(\mathcal{H}_1,\dots,\mathcal{H}_p)} J(\mathcal{B})$$

where the last equality comes from Equation (15). This shows that the solution of the left hand side problem (i.e. Equation (20)) is a solution of the right hand side problem (i.e. Equation (11)), which concludes the proof. □

A.1. Asymmetric problem with different kernels

For any linear operator $\mathcal{A} \in \mathcal{S}(\mathcal{H})$, let us first define the penalty function

$$\Omega_{(\cdot)}(\mathcal{A}) = \lambda_{(\cdot)1}\|\mathcal{A}\|_\star + \lambda_{(\cdot)2}\|\mathcal{A}\|_F^2 \tag{24}$$

with Fenchel conjugate $\Omega_{(\cdot)}^\star$, and the residuals

$$r_{(\cdot)}(X_i, Y_i) = \begin{cases} m(X_i) - Y_i & \text{if } (\cdot) \text{ is low} \\ Y_i - m(X_i) & \text{if } (\cdot) \text{ is up.} \end{cases}$$

Note that $\Omega_{(\cdot)}(\cdot)$ satisfies Assumption A.1 as long as $\lambda_{(\cdot)1} > 0$ or $\lambda_{(\cdot)2} > 0$.

For the asymmetric setting without penalty, our infinite dimensional problem writes:

$$\begin{aligned} \inf_{\substack{\mathcal{A}_{\text{low}}\in\mathcal{S}_+(\mathcal{H}_{\text{low}}) \\ \mathcal{A}_{\text{up}}\in\mathcal{S}_+(\mathcal{H}_{\text{up}})}} \quad & \frac{b}{n}\sum_{i=1}^n \left(f_{\mathcal{A}_{\text{low}}}(X_i) + f_{\mathcal{A}_{\text{up}}}(X_i)\right) + \Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}}) \\ \text{s.t.} \quad & r_{\text{low}}(X_i, Y_i) - f_{\mathcal{A}_{\text{low}}}(X_i) \le 0,\ i \in [n] \\ & r_{\text{up}}(X_i, Y_i) - f_{\mathcal{A}_{\text{up}}}(X_i) \le 0,\ i \in [n]. \end{aligned} \tag{25}$$

Since this problem is separable in both objectives and constraints, it is sufficient to prove a representer theorem for each problem independently:

$$\begin{aligned} \inf_{\mathcal{A}_{(\cdot)}\in\mathcal{S}_+(\mathcal{H}_{(\cdot)})} \quad & \frac{b}{n}\sum_{i=1}^n f_{\mathcal{A}_{(\cdot)}}(X_i) + \Omega_{(\cdot)}(\mathcal{A}_{(\cdot)}) \\ \text{s.t.} \quad & r_{(\cdot)}(X_i, Y_i) - f_{\mathcal{A}_{(\cdot)}}(X_i) \le 0,\ i \in [n]. \end{aligned} \tag{26}$$

Theorem A.7 (Representer theorem)**.** *Let $(b, \lambda_{(\cdot)1}) \in \mathbb{R}_+^2$ and $\lambda_{(\cdot)2} > 0$. Then Equation* (26) *admits a unique solution $f_{\mathbf{A}_{(\cdot)}^\star}$ of the form $f_{\mathbf{A}_{(\cdot)}^\star}(X) = \mathbf{\Phi}_{(\cdot)}(X)^\top \mathbf{A}_{(\cdot)}^\star \mathbf{\Phi}_{(\cdot)}(X)$ for some matrix $\mathbf{A}_{(\cdot)}^\star \in \mathbb{S}_+^n$.*

Proof. We follow the proof of Theorem 2 from Allain et al. (2026) with a fixed mean function, by replacing their residuals r^2 with $r_{(\cdot)}$. In the end, Problem (26) admits a finite representation entirely characterized by a PSD matrix $\mathbf{A}_{(\cdot)} \in \mathbb{S}_+^n$ given by

$$\begin{aligned} \inf_{\mathbf{A}_{(\cdot)}\in\mathbb{S}_+^n} \quad & \frac{b}{n}\sum_{i=1}^n \tilde{f}_{\mathbf{A}_{(\cdot)}}(X_i) + \lambda_1\|\mathbf{A}_{(\cdot)}\|_\star + \lambda_2\|\mathbf{A}_{(\cdot)}\|_F^2 \\ \text{s.t.} \quad & r_{(\cdot)}(X_i, Y_i) - \tilde{f}_{\mathbf{A}_{(\cdot)}}(X_i) \le 0,\ i \in [n], \end{aligned} \tag{27}$$

with $\tilde{f}_{\mathbf{A}_{(\cdot)}}(X_i) = \mathbf{\Phi}_{(\cdot)}(X_i)^\top \mathbf{A}_{(\cdot)} \mathbf{\Phi}_{(\cdot)}(X_i)$. □

The optimal solution of Problem (25) is thus recovered as $f_{\mathbf{A}^\star_{\text{low}}}(X) = \mathbf{\Phi}_{\text{low}}(X)^\top \mathbf{A}^\star_{\text{low}} \mathbf{\Phi}_{\text{low}}(X)$ and $f_{\mathbf{A}^\star_{\text{up}}}(X) = \mathbf{\Phi}_{\text{up}}(X)^\top \mathbf{A}^\star_{\text{up}} \mathbf{\Phi}_{\text{up}}(X)$ for matrices $\mathbf{A}^\star_{\text{low}}, \mathbf{A}^\star_{\text{up}} \in \mathbb{S}^n_+$ solutions of Equation (27) with $(\cdot)$ equal to $_{\text{low}}$ and $_{\text{up}}$, respectively.

Now, for each independent problem $_{\text{low}}$ and $_{\text{up}}$, we exhibit a dual formulation. Let us first introduce, for any symmetric matrix $\mathbf{A} \in \mathbb{S}^n$, the penalty function $\Omega_{+,(\cdot)}$ defined as

$$\Omega_{+,(\cdot)}(\mathbf{A}) = \begin{cases} \Omega_{(\cdot)}(\mathbf{A}) & \text{if} \quad \mathbf{A} \succeq 0 \\ +\infty & \text{otherwise} \end{cases} \tag{28}$$

where $\Omega_{(\cdot)}(\mathbf{A})$ is the matrix equivalent of Equation (24).

Proposition A.8 (Dual formulation)**.** *Let $(b, \lambda_{(\cdot)1}) \in \mathbb{R}^2_+$ and $\lambda_{(\cdot)2} > 0$. Equation* (27) *admits a dual formulation of the form*

$$\sup_{\mathbf{\Gamma}_{(\cdot)} \in \mathbb{R}^n_+} \mathbf{\Gamma}_{(\cdot)} \mathbf{r}_{(\cdot)}(\mathbf{X}, Y)^\top - \Omega^\star_{+,(\cdot)}(\mathbf{V}_{(\cdot)} \text{Diag}(\mathbf{\Gamma}_{(\cdot)_{-\mathbf{b}}}) \mathbf{V}_{(\cdot)}^\top) \tag{29}$$

where $\Omega^\star_{+,(\cdot)}(\mathbf{B}) = \frac{1}{4\lambda_{(\cdot)2}} \| \left[\mathbf{B} - \lambda_{(\cdot)1} \mathbf{I}_n\right]_+ \|_F^2$ and $\forall x \in \mathbb{R},\ \text{Diag}\left(\mathbf{\Gamma}_{(\cdot)x}\right) := \text{Diag}\left(\mathbf{\Gamma}_{(\cdot)}\right) + \frac{x}{n}\mathbf{I}_n$. Moreover, if $\widehat{\mathbf{\Gamma}}_{(\cdot)}$ is solution of Equation (29)*, a solution of Equation* (27) *can be retrieved as*

$$\widehat{\mathbf{A}}_{(\cdot)} = \frac{1}{2\lambda_{(\cdot)2}} \left[\mathbf{V}_{(\cdot)} \text{Diag}(\widehat{\mathbf{\Gamma}}_{(\cdot)_{-\mathbf{b}}}) \mathbf{V}_{(\cdot)}^\top - \lambda_{(\cdot)1} \mathbf{I}_n \right]_+$$

where $[\mathbf{A}]_+$ denotes the positive part of $\mathbf{A}$. For a symmetric matrix $\mathbf{A}$ with eigendecomposition $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$, its positive part is defined as $[\mathbf{A}]_+ = \mathbf{U} \max(0, \mathbf{D}) \mathbf{U}^\top$.

Proof. Again, we apply the proof of Proposition 2 from Allain et al. (2026) with a fixed mean function and by replacing r^2 with $r_{(\cdot)}$.

Finally, to solve such dual formulation, we rely on explicit gradients given below.

Gradient computation.

$$\frac{\partial \left(\mathbf{\Gamma}_{(\cdot)} \mathbf{r}_{(\cdot)}(\mathbf{X}, Y)^\top \right)}{\partial \mathbf{\Gamma}_{(\cdot)}} = \mathbf{r}_{(\cdot)}(\mathbf{X}, Y).$$

$$\frac{\partial \Omega^\star_{+,(\cdot)}(\mathbf{V}_{(\cdot)} \text{Diag}(\mathbf{\Gamma}_{(\cdot)_{-\mathbf{b}}}) \mathbf{V}_{(\cdot)}^\top)}{\partial \mathbf{\Gamma}_{(\cdot)}} = \text{Diag}\left(\mathbf{V}_{(\cdot)}^\top \left[\nabla \Omega^\star_{+,(\cdot)} \left(\mathbf{V}_{(\cdot)} \text{Diag}(\mathbf{\Gamma}_{(\cdot)_{-\mathbf{b}}}) \mathbf{V}_{(\cdot)}^\top \right) \right]^\top \mathbf{V}_{(\cdot)} \right).$$

More details can be found in Allain et al. (2026).

Recovering the solution from optimal Lagrange multipliers. Once the dual problem is solved (see Appendix Appendix B.2), we need to recover the optimal solutions of the primal problem. Denoting $\widehat{\mathbf{\Gamma}}_{(\cdot)} \in \mathbb{R}^n_+$ the optimal Lagrange multipliers for both problems, to reconstruct the matrices $\mathbf{A}_{(\cdot)}$ we follow Allain et al. (2026):

$$\begin{aligned} \widehat{\mathbf{A}}_{(\cdot)} &= \nabla \Omega^\star_{+,(\cdot)} \left(\mathbf{V}_{(\cdot)} \text{Diag}(\widehat{\mathbf{\Gamma}}_{(\cdot)_{-\mathbf{b}}}) \mathbf{V}_{(\cdot)}^\top \right) \\ &= \frac{1}{2\lambda_{(\cdot)2}} \left[\mathbf{V}_{(\cdot)} \text{Diag}(\widehat{\mathbf{\Gamma}}_{(\cdot)_{-\mathbf{b}}}) \mathbf{V}_{(\cdot)}^\top - \lambda_{(\cdot)1} \mathbf{I}_n \right]_+ . \end{aligned}$$

□

A.2. Asymmetric problem with operator penalty

For the asymmetric setting with the penalty on operators, our infinite dimensional problem writes:

$$\begin{aligned} \inf_{\substack{\mathcal{A}_{\text{low}} \in \mathcal{S}_+(\mathcal{H}) \\ \mathcal{A}_{\text{up}} \in \mathcal{S}_+(\mathcal{H})}} \quad & \frac{b}{n} \sum_{i=1}^{n} \left(f_{\mathcal{A}_{\text{low}}}(X_i) + f_{\mathcal{A}_{\text{up}}}(X_i) \right) + \Psi(\mathcal{A}_{\text{low}}, \mathcal{A}_{\text{up}}) \\ \text{s.t.} \quad & r_{\text{low}}\left(X_i, Y_i\right) - f_{\mathcal{A}_{\text{low}}}(X_i) \leq 0, \ i \in [n] \\ & r_{\text{up}}\left(X_i, Y_i\right) - f_{\mathcal{A}_{\text{up}}}(X_i) \leq 0, \ i \in [n] \end{aligned} \tag{30}$$

where $\Psi(\mathcal{A}_{\text{low}}, \mathcal{A}_{\text{up}}) = \Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}}) + \Omega_{\text{pen}}(\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}})$. This time, the problem is no longer separable because of the penalty term.

Theorem A.9 (Representer theorem with operator penalty)**.** *Let* $(b, \lambda_{\text{low}1}, \lambda_{\text{up}1}, \lambda_{\text{pen}1}) \in \mathbb{R}_+^4$ *and* $\lambda_{\text{low}2}, \lambda_{\text{up}2}, \lambda_{\text{pen}2} > 0$*. Then Problem 30 admits a unique solution* $(f_{\mathbf{B}^\star_{\text{low}}}, f_{\mathbf{B}^\star_{\text{up}}})$ *of the form* $(f_{\mathbf{B}^\star_{\text{low}}}(X), f_{\mathbf{B}^\star_{\text{up}}}(X)) = \left(\mathbf{k}(X)^\top \mathbf{B}^\star_{\text{low}} \mathbf{k}(X), \ \mathbf{k}(X)^\top \mathbf{B}^\star_{\text{up}} \mathbf{k}(X)\right)$ *for some matrices* $\left(\mathbf{B}^\star_{\text{low}}, \mathbf{B}^\star_{\text{up}}\right) \in (\mathbb{S}_+^n)^2$.

Proof. We apply Theorem A.2 to the case $p = 2$ with $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{H}$ with kernel k and loss function

$$L((f_{\mathcal{A}_{\text{low}}}(X_i), f_{\mathcal{A}_{\text{up}}}(X_i))_{1 \leq i \leq n}) = \begin{cases} \frac{b}{n} \sum_{i=1}^{n} \left(f_{\mathcal{A}_{\text{low}}}(X_i) + f_{\mathcal{A}_{\text{up}}}(X_i) \right) & \text{if } (\mathcal{A}_{\text{low}}, \mathcal{A}_{\text{up}}) \in C \\ +\infty & \text{otherwise} \end{cases}$$

where C is the set of all positive definite operators $(\mathcal{A}_{\text{low}}, \mathcal{A}_{\text{up}})$ such that

$$\begin{cases} r_{\text{low}}\left(X_i, Y_i\right) - f_{\mathcal{A}_{\text{low}}}(X_i) \leq 0, \ i \in [n] \\ r_{\text{up}}\left(X_i, Y_i\right) - f_{\mathcal{A}_{\text{up}}}(X_i) \leq 0, \ i \in [n] \end{cases}$$

and the new penalty function $\Psi((\mathcal{A}_{\text{low}}, \mathcal{A}_{\text{up}})) = \Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}}) + \Omega_{\text{pen}}(\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}})$, where Ω_{low}, Ω_{up} and Ω_{pen} are defined as in Equation (24).

First, notice that $\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}}$ is not a positive semi-definite operator anymore but Remark 3 in Marteau-Ferey et al. (2020) still applies since $\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}} \in \mathcal{S}(\mathcal{H})$. Furthermore, Ω_{pen} verifies Assumption A.1. However, Ψ no longer writes as a separable sum as in Equation (10), and we need to show that it satisfies the three properties of Lemma A.3 to apply Theorem A.2.

For the first and second property, we can directly apply Lemma 2 from Marteau-Ferey et al. (2020) individually to $\Omega_{\text{low}}, \Omega_{\text{up}}$ and Ω_{pen}, since it requires the operators to be symmetric only (not necessarily positive semi-definite), and thus also applies to Ω_{pen}. For the third property, Ψ is continuous as a composition of the linear (hence continuous) function $(\mathcal{A}_{\text{low}}, \mathcal{A}_{\text{up}}) \to (\mathcal{A}_{\text{low}}, \mathcal{A}_{\text{up}}, \mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}})$ and lower-semi continuous functions ($\Omega_{\text{low}}, \Omega_{\text{up}}$ and Ω_{pen}). The last part of property 3 holds because if at least one of $\|\mathcal{A}_{\text{low}}\|_{\text{op}}, \ \|\mathcal{A}_{\text{up}}\|_{\text{op}}$ goes to infinity, we have:

$$\Psi((\mathcal{A}_{\text{low}}, \mathcal{A}_{\text{up}})) \underset{\max(\|\mathcal{A}_{\text{low}}\|_{\text{op}}, \|\mathcal{A}_{\text{up}}\|_{\text{op}}) \to +\infty}{\longrightarrow} +\infty.$$

Finally, since $L \colon \mathbb{R}^{2n} \to \mathbb{R}$ is lower semi-continuous (notice that it is linear and bounded below by 0) and Ψ satisfies all properties from Lemma A.3, we can apply Theorem A.2 with $p = 2$ to deduce that the solution is entirely characterized by two PSD matrices $\left(\mathbf{B}^\star_{\text{low}}, \mathbf{B}^\star_{\text{up}}\right) \in (\mathbb{S}_+^n)^2$ and

$$(f_{\mathbf{B}^\star_{\text{low}}}(X), f_{\mathbf{B}^\star_{\text{up}}}(X)) = \left(\mathbf{k}(X)^\top \mathbf{B}^\star_{\text{low}} \mathbf{k}(X), \ \mathbf{k}(X)^\top \mathbf{B}^\star_{\text{up}} \mathbf{k}(X)\right).$$

□

Theorem A.9 also has an equivalent with matrices $\mathbf{A}_{\text{low}}$, $\mathbf{A}_{\text{up}}$ instead of $\mathbf{B}_{\text{low}}$, $\mathbf{B}_{\text{up}}$ involving $\tilde{f}_{\mathbf{A}_{(\cdot)}}$ as in Equation (19).

Proposition A.10 (Representer theorem with operator penalty, formulation **A**)**.** *Let* $(b, \lambda_{\text{low}1}, \lambda_{\text{up}1}, \lambda_{\text{pen}1}) \in \mathbb{R}^4_+$ *and* $\lambda_{\text{low}2}, \lambda_{\text{up}2}, \lambda_{\text{pen}2} > 0$*. The following problem admits a unique solution if* L *is convex:*

$$\begin{aligned} \inf_{\substack{\mathbf{A}_{\text{low}} \in \mathbb{S}^n_+ \\ \mathbf{A}_{\text{up}} \in \mathbb{S}^n_+}} \quad & \frac{b}{n} \sum_{i=1}^n \left(\tilde{f}_{\mathbf{A}_{\text{low}}}(X_i) + \tilde{f}_{\mathbf{A}_{\text{up}}}(X_i) \right) + \Psi(\mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}}) \\ \text{s.t.} \quad & r_{\text{low}}\left(X_i, Y_i\right) - \tilde{f}_{\mathbf{A}_{\text{low}}}(X_i) \leq 0, \ i \in [n] \\ & r_{\text{up}}\left(X_i, Y_i\right) - \tilde{f}_{\mathbf{A}_{\text{up}}}(X_i) \leq 0, \ i \in [n]\,. \end{aligned} \tag{31}$$

Moreover, for any given solution $\mathbf{A}^\star_{\text{low}}, \mathbf{A}^\star_{\text{up}} \in \mathbb{S}^n_+$ *of Equation* (31)*, the function* $(\tilde{f}_{\mathbf{A}^\star_{\text{low}}}, \tilde{f}_{\mathbf{A}^\star_{\text{up}}})$ *is a minimizer of Equation* (30)*.*

Proof. Similarly to the proof of Theorem A.9, we apply Proposition A.6 to the case $p = 2$ with $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{H}$, but Ψ does not write as a separable sum: we thus must check if it satisfies the isometry invariance property to conclude. This is trivial, since

$$\begin{aligned} \Psi((O_n \mathbf{A}_{\text{low}} O_n^*, O_n \mathbf{A}_{\text{up}} O_n^*)) &= \Omega_{\text{low}}(O_n \mathbf{A}_{\text{low}} O_n^*) + \Omega_{\text{up}}(O_n \mathbf{A}_{\text{up}} O_n^*) + \Omega_{\text{pen}}(O_n \mathbf{A}_{\text{low}} O_n^* - O_n \mathbf{A}_{\text{up}} O_n^*) \\ &= \Omega_{\text{low}}(\mathbf{A}_{\text{low}}) + \Omega_{\text{up}}(\mathbf{A}_{\text{up}}) + \Omega_{\text{pen}}(O_n(\mathbf{A}_{\text{low}} - \mathbf{A}_{\text{up}})O_n^*) \\ &= \Omega_{\text{low}}(\mathbf{A}_{\text{low}}) + \Omega_{\text{up}}(\mathbf{A}_{\text{up}}) + \Omega_{\text{pen}}(\mathbf{A}_{\text{low}} - \mathbf{A}_{\text{up}}) \\ &= \Psi((\mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}})). \end{aligned}$$

□

Proposition A.11 (Dual formulation with operator penalty)**.** *Let* $(b, \lambda_{\text{low}1}, \lambda_{\text{up}1}, \lambda_{\text{pen}1}) \in \mathbb{R}^4_+$ *and* $\lambda_{\text{low}2}, \lambda_{\text{up}2}, \lambda_{\text{pen}2} > 0$*. Problem 31 admits a dual formulation of the form*

$$\begin{aligned} \sup_{\substack{\mathbf{\Gamma}_{\text{low}} \in \mathbb{R}^n_+, \mathbf{\Gamma}_{\text{up}} \in \mathbb{R}^n_+ \\ \mathbf{W} \in \mathbb{S}^n}} & (\mathbf{\Gamma}_{\text{up}} - \mathbf{\Gamma}_{\text{low}})\mathbf{r}^\top - \Omega^\star_{\text{pen}}(\mathbf{W}) \\ & - \Omega^\star_{+,\text{low}}(\mathbf{V}_{\text{low}} \text{Diag}(\mathbf{\Gamma}_{\text{low}-\mathbf{b}}) \mathbf{V}_{\text{low}}^\top - \mathbf{W}) - \Omega^\star_{+,\text{up}}(\mathbf{V}_{\text{up}} \text{Diag}(\mathbf{\Gamma}_{\text{up}-\mathbf{b}}) \mathbf{V}_{\text{up}}^\top + \mathbf{W}) \end{aligned} \tag{32}$$

where $\mathbf{r}$ *is the vector of residuals* $r_i = Y_i - m(X_i)$*,* $\Omega^\star_{+,(\cdot)}(\mathbf{B}) = \frac{1}{4\lambda_{(\cdot)2}} \| \left[\mathbf{B} - \lambda_{(\cdot)1} \mathbf{I}_n\right]_+ \|_F^2$*,* $\Omega^\star_{\text{pen}}(\mathbf{B}) = (1/4\lambda_{\text{pen}2}) \sum_{i=1}^n \max(0, |\lambda_i(\mathbf{B})| - \lambda_{\text{pen}1})^2$ *and* $\forall x \in \mathbb{R}$*,* $\text{Diag}\left((\cdot)_x\right) := \text{Diag}\left((\cdot)\right) + \frac{x}{n}\mathbf{I}_n$*. Moreover, if* $(\widehat{\mathbf{\Gamma}}_{\text{low}}, \widehat{\mathbf{\Gamma}}_{\text{up}}, \widehat{\mathbf{W}})$ *is a solution of Equation* (32)*, a solution of Problem 31 can be retrieved as*

$$\begin{aligned} \widehat{\mathbf{A}}_{\text{low}} &= \frac{1}{2\lambda_{\text{low}2}} \left[\mathbf{V}_{\text{low}} \text{Diag}(\widehat{\mathbf{\Gamma}}_{\text{low}})_{-\mathbf{b}} \mathbf{V}_{\text{low}}^\top - \widehat{\mathbf{W}} - \lambda_{\text{low}1} \mathbf{I}_n \right]_+ \\ \widehat{\mathbf{A}}_{\text{up}} &= \frac{1}{2\lambda_{\text{up}2}} \left[\mathbf{V}_{\text{up}} \text{Diag}(\widehat{\mathbf{\Gamma}}_{\text{up}})_{-\mathbf{b}} \mathbf{V}_{\text{up}}^\top + \widehat{\mathbf{W}} - \lambda_{\text{up}1} \mathbf{I}_n \right]_+ . \end{aligned}$$

Proof. The dual problem is defined as

$$d = \sup_{\substack{\mathbf{\Gamma}_{\text{low}} \in \mathbb{R}^n_+ \\ \mathbf{\Gamma}_{\text{up}} \in \mathbb{R}^n_+}} \inf_{\substack{\mathbf{A}_{\text{low}} \in \mathbb{S}^n \\ \mathbf{A}_{\text{up}} \in \mathbb{S}^n}} \mathcal{L}(\mathbf{\Gamma}_{\text{low}}, \mathbf{\Gamma}_{\text{up}}, \mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}}) = \sup_{\substack{\mathbf{\Gamma}_{\text{low}} \in \mathbb{R}^n_+ \\ \mathbf{\Gamma}_{\text{up}} \in \mathbb{R}^n_+}} (\mathbf{\Gamma}_{\text{up}} - \mathbf{\Gamma}_{\text{low}})\mathbf{r}^\top + D(\mathbf{\Gamma}_{\text{low}}, \mathbf{\Gamma}_{\text{up}}) \tag{33}$$

where we denote

$$D(\mathbf{\Gamma}_{\text{low}}, \mathbf{\Gamma}_{\text{up}}) \coloneqq \inf_{\mathbf{A}_{\text{low}} \in \mathbb{S}^n,\ \mathbf{A}_{\text{up}} \in \mathbb{S}^n} \overline{\mathcal{L}}(\mathbf{\Gamma}_{\text{low}}, \mathbf{\Gamma}_{\text{up}}, \mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}}) \tag{34}$$

with $\overline{\mathcal{L}}$ defined as

$$\begin{aligned} \overline{\mathcal{L}}(\mathbf{\Gamma}_{\text{low}}, \mathbf{\Gamma}_{\text{up}}, \mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}}) = & \sum_{i=1}^n (\frac{b}{n} - \Gamma_{\text{low},i}) \tilde{f}_{\mathbf{A}_{\text{low}}}(X_i) + \sum_{i=1}^n (\frac{b}{n} - \Gamma_{\text{up},i}) \tilde{f}_{\mathbf{A}_{\text{up}}}(X_i) \\ & + \Omega_{+,\text{low}}(\mathbf{A}_{\text{low}}) + \Omega_{+,\text{up}}(\mathbf{A}_{\text{up}}) + \Omega_{\text{pen}}(\mathbf{A}_{\text{low}} - \mathbf{A}_{\text{up}}). \end{aligned}$$

This time, to derive the optimality conditions for $(\mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}})$, we follow Allain et al. (2026) Appendix A.2, Equation (15). First, observe that

$$\overline{\mathcal{L}} = \Psi_+(\mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}}) - \langle (\mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}}), (\mathbf{V}_{\text{low}}\text{Diag}(\widehat{\mathbf{\Gamma}}_{\text{low}})_{-\mathbf{b}}\mathbf{V}_{\text{low}}^\top, \mathbf{V}_{\text{up}}\text{Diag}(\widehat{\mathbf{\Gamma}}_{\text{up}})_{-\mathbf{b}}\mathbf{V}_{\text{up}}^\top)\rangle_{\mathbb{S}^n \times \mathbb{S}^n}$$

where $\Psi_+(\mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}}) = \Omega_{+,\text{low}}(\mathbf{A}_{\text{low}}) + \Omega_{+,\text{up}}(\mathbf{A}_{\text{up}}) + \Omega_{\text{pen}}(\mathbf{A}_{\text{low}} - \mathbf{A}_{\text{up}})$. Then, by definition, $D(\mathbf{\Gamma}_{\text{low}}, \mathbf{\Gamma}_{\text{up}})$ writes as the Fenchel conjugate function of Ψ_+ evaluated at specific matrices

$$D(\mathbf{\Gamma}_{\text{low}}, \mathbf{\Gamma}_{\text{up}}) = -\Psi_+^\star\Big(\mathbf{V}_{\text{low}}\text{Diag}(\mathbf{\Gamma}_{\text{low}})_{-\mathbf{b}}\mathbf{V}_{\text{low}}^\top, \mathbf{V}_{\text{up}}\text{Diag}(\mathbf{\Gamma}_{\text{up}})_{-\mathbf{b}}\mathbf{V}_{\text{up}}^\top\Big). \tag{35}$$

The explicit formulation of $\Psi_+^\star$ is given in the following lemma.

Lemma A.12. *The Fenchel conjugate of* $\Psi_+(\mathbf{A}_1, \mathbf{A}_2) = \Omega_{+,\text{low}}(\mathbf{A}_1) + \Omega_{+,\text{up}}(\mathbf{A}_2) + \Omega_{\text{pen}}(\mathbf{A}_1 - \mathbf{A}_2)$ *writes*

$$\Psi_+^\star(\mathbf{B}_1, \mathbf{B}_2) = -\sup_{\mathbf{W} \in \mathbb{S}^n} -\Omega_{+,\text{low}}^\star(\mathbf{B}_1 - \mathbf{W}) - \Omega_{+,\text{up}}^\star(\mathbf{B}_2 + \mathbf{W}) - \Omega_{\text{pen}}^\star(\mathbf{W}).$$

Proof. The proof relies on the fact that Ψ_+ can be written as a composition $\Psi(\mathbf{A}_1, \mathbf{A}_2) = f \circ L(\mathbf{A}_1, \mathbf{A}_2)$ of the linear operator $L\colon (\mathbf{A}_1, \mathbf{A}_2) \in (\mathbb{S}^n)^2 \mapsto (\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_1 - \mathbf{A}_2) \in (\mathbb{S}^n)^3$ and the function $f\colon (\mathbf{U}, \mathbf{V}, \mathbf{W}) \in (\mathbb{S}^n)^3 \mapsto \Omega_{+,\text{low}}(\mathbf{U}) + \Omega_{+,\text{up}}(\mathbf{V}) + \Omega_{\text{pen}}(\mathbf{W}) \in \mathbb{R}$. The Fenchel conjugate of Ψ_+ then writes (see, e.g., Theorem 16.3 in Rockafellar (1970)):

$$\Psi_+^\star(\mathbf{B}_1, \mathbf{B}_2) = \inf_{\substack{\mathbf{U},\mathbf{V},\mathbf{W} \\ \text{s.t. } L^\star(\mathbf{U},\mathbf{V},\mathbf{W}) = (\mathbf{B}_1,\mathbf{B}_2)}} \Omega_{+,\text{low}}^\star(\mathbf{U}) + \Omega_{+,\text{up}}^\star(\mathbf{V}) + \Omega_{\text{pen}}^\star(\mathbf{W}) \tag{36}$$

where $L^\star$, the adjoint of L, can be deduced from

$$\begin{aligned}\langle L(\mathbf{A}_1, \mathbf{A}_2), (\mathbf{U}, \mathbf{V}, \mathbf{W})\rangle_{(\mathbb{S}^n)^3} &= \langle \mathbf{A}_1, \mathbf{U}\rangle + \langle \mathbf{A}_2, \mathbf{V}\rangle + \langle \mathbf{A}_1 - \mathbf{A}_2, \mathbf{W}\rangle \\ &= \langle \mathbf{A}_1, \mathbf{U} + \mathbf{W}\rangle + \langle \mathbf{A}_2, \mathbf{V} - \mathbf{W}\rangle \\ &= \langle (\mathbf{A}_1, \mathbf{A}_2), L^\star(\mathbf{U}, \mathbf{V}, \mathbf{W})\rangle_{(\mathbb{S}^n)^2}\end{aligned}$$

such that $L^\star(\mathbf{U}, \mathbf{V}, \mathbf{W}) = (\mathbf{U} + \mathbf{W}, \mathbf{V} - \mathbf{W}) \in (\mathbb{S}^n)^2$. Now Equation (36) writes

$$\Psi_+^\star(\mathbf{B}_1, \mathbf{B}_2) = \inf_{\substack{\mathbf{U}+\mathbf{W}=\mathbf{B}_1 \\ \mathbf{V}-\mathbf{W}=\mathbf{B}_2 \\ \mathbf{W}}} \Omega_{+,\text{low}}^\star(\mathbf{U}) + \Omega_{+,\text{up}}^\star(\mathbf{V}) + \Omega_{\text{pen}}^\star(\mathbf{W})$$

and the result follows by replacing $\mathbf{U}$ and $\mathbf{V}$ by $\mathbf{B}_1 - \mathbf{W}$ and $\mathbf{B}_2 + \mathbf{W}$, respectively. □

Before concluding the proof, we need to derive the dual function $\Omega_{\text{pen}}^\star$ (this is not the same as $\Omega_{(\cdot)}^\star$, because it is defined on symmetric matrices and not positive semi-definite ones), which is addressed in the following lemma.

Lemma A.13. *For any symmetric matrix* $\mathbf{X} \in \mathbb{S}^n$ *such that* $\mathbf{X} = \mathbf{U}^\top\text{Diag}(\lambda(\mathbf{X}))\mathbf{U}$ *where* $\lambda(\mathbf{X}) = (\lambda_1(\mathbf{X}), \ldots, \lambda_n(\mathbf{X}))$ *denotes the eigenvalues of* X*, the function* $f(\mathbf{X}) = \lambda_1\|\mathbf{X}\|_\star + \lambda_2\|\mathbf{X}\|_F^2$ *admits a Fenchel conjugate given by*

$$f^\star(\mathbf{Y}) = \frac{1}{4\lambda_2}\sum_{i=1}^n \max(0, |\lambda_i(\mathbf{Y})| - \lambda_1)^2$$

with gradient

$$\nabla f^\star(\mathbf{Y}) = \frac{1}{2\lambda_2}\mathbf{U}^\top\text{Diag}(\text{sign}(\lambda(\mathbf{Y})) \times \max(0, |\lambda(\mathbf{Y})| - \lambda_1))\mathbf{U}.$$

Proof. We first write f as a function φ of the singular values of $\mathbf{X}$. The nuclear and Frobenius norms are the 1- and 2-Schatten norms, respectively, which implies that

$$f(\mathbf{X}) = \lambda_1\sum_{i=1}^n \sigma_i(\mathbf{X}) + \lambda_2\sum_{i=1}^n \sigma_i(\mathbf{X})^2 = \varphi(\sigma(\mathbf{X}))$$

where $\varphi((s_1, \dots, s_n)) = \lambda_1 \sum_{i=1}^n s_i + \lambda_2 \sum_{i=1}^n s_i^2$ and $\sigma(\mathbf{X})$ are the singular values of $\mathbf{X}$. Because f is unitary invariant, from Theorem 2.4 in Lewis (1995), we know that

$$f^\star(\mathbf{Y}) = \varphi^\star(\sigma(\mathbf{Y})).$$

Now, let us compute the dual function $\varphi^\star$. By definition, it is given by

$$\begin{aligned}\varphi^\star(t) &= \sup_{s \geq 0} \langle t, s \rangle - \lambda_1 \sum_{i=1}^n s_i - \lambda_2 \sum_{i=1}^n s_i^2 \\ &= \sup_{(s_1,\dots,s_n) \geq 0} \sum_{i=1}^n (t_i s_i - \lambda_1 s_i - \lambda_2 s_i^2).\end{aligned}$$

Since this optimization problem is separable, we can maximize each term individually. Let $g(s) = ts - \lambda_1 s - \lambda_2 s^2$, which attains its maximum for $s = (t - \lambda_1)/2\lambda_2$. We now have two cases: either $t - \lambda_1 \leq 0$ and the supremum is reached for $s = 0$ and its value is 0, or $t - \lambda_1 > 0$, in which case the supremum is reached for $s = (t - \lambda_1)/2\lambda_2$ and its value is $(t - \lambda_1)^2/4\lambda_2$. All in one, $\sup_{s \geq 0} g(s) = \max(0, (t - \lambda_1)/4\lambda_2)^2$ and $\varphi^\star(t) = \frac{1}{4\lambda_2} \sum_{i=1}^n \max(0, (t_i - \lambda_1))^2$.

This means that the Fenchel conjugate of f is $f^\star(\mathbf{Y}) = \frac{1}{4\lambda_2} \sum_{i=1}^n \max(0, \sigma_i(\mathbf{Y}) - \lambda_1)^2$, and since $\mathbf{X}$ and $\mathbf{Y}$ are symmetric matrices, we have $\sigma_i(\mathbf{Y}) = |\lambda_i(\mathbf{Y})|$ from which we deduce

$$f^\star(\mathbf{Y}) = \frac{1}{4\lambda_2} \sum_{i=1}^n \max(0, |\lambda_i(\mathbf{Y})| - \lambda_1)^2.$$

Now, to compute its gradient, observe that $f^\star(\mathbf{Y})$ writes as some function h applied to the eigenvalues of $\mathbf{Y}$ with $h(u) = \frac{1}{4\lambda_2} \sum_{i=1}^n \max(0, |u_i| - \lambda_1)^2$. Because h is permutation invariant, we can apply Theorem 1.1 from Lewis (1996) to $f^\star$ to get

$$\nabla f^\star(\mathbf{Y}) = \mathbf{U}^\top \mathrm{Diag}(h'(\lambda(\mathbf{Y})))\mathbf{U}. \tag{37}$$

The last step is to compute the gradient of h. For all $1 \leq i \leq n$, we have

$$\frac{\partial h(u)}{\partial u_i} = \frac{1}{4\lambda_2} \frac{d}{du_i} \max(0, |u_i| - \lambda_1)^2$$

and observe that:

- If $|u_i| \leq \lambda_1$, then $\max(0, |u_i| - \lambda_1)^2 = 0$ and the derivative is 0.
- If $u_i > \lambda_1$, then $\max(0, |u_i| - \lambda_1)^2 = (u_i - \lambda_1)^2$ and the derivative is $2(u_i - \lambda_1) \underset{u_i \to \lambda_1^+}{\to} 0$.
- If $u_i < -\lambda_1$, then $\max(0, |u_i| - \lambda_1)^2 = (u_i + \lambda_1)^2$ and the derivative is $2(u_i + \lambda_1) \underset{u_i \to -\lambda_1^-}{\to} 0$.

Consequently, $\max(0, |u_i| - \lambda_1)^2$ is differentiable with derivative $2\,\mathrm{sign}(u_i) \times \max(0, |u_i| - \lambda_1)$. Plugging this into Equation (37) gives

$$\nabla f^\star(\mathbf{Y}) = \frac{1}{2\lambda_2} \mathbf{U}^\top \mathrm{Diag}(\mathrm{sign}(\lambda(\mathbf{Y})) \times \max(0, |\lambda(\mathbf{Y})| - \lambda_1))\mathbf{U}$$

which concludes the proof of Lemma A.13. □

Finally, applying Lemma A.12 together with the explicit formulation of $\Omega^\star_{\mathrm{pen}}$ from Lemma A.13 in Equation (33), our main dual problem writes

$$\begin{aligned} d = &\sup_{\substack{\mathbf{\Gamma}_{\mathrm{low}} \in \mathbb{R}^n_+, \mathbf{\Gamma}_{\mathrm{up}} \in \mathbb{R}^n_+ \\ \mathbf{W} \in \mathbb{S}^n}} (\mathbf{\Gamma}_{\mathrm{up}} - \mathbf{\Gamma}_{\mathrm{low}})\mathbf{r}^\top - \Omega^\star_{\mathrm{pen}}(\mathbf{W}) \\ &\qquad - \Omega^\star_{+,\mathrm{low}}(\mathbf{V}_{\mathrm{low}} \mathrm{Diag}(\mathbf{\Gamma}_{\mathrm{low}-\mathbf{b}})\mathbf{V}_{\mathrm{low}}^\top - \mathbf{W}) - \Omega^\star_{+,\mathrm{up}}(\mathbf{V}_{\mathrm{up}} \mathrm{Diag}(\mathbf{\Gamma}_{\mathrm{up}-\mathbf{b}})\mathbf{V}_{\mathrm{up}}^\top + \mathbf{W}) \\ = &\sup_{\substack{\mathbf{\Gamma}_{\mathrm{low}} \in \mathbb{R}^n_+, \mathbf{\Gamma}_{\mathrm{up}} \in \mathbb{R}^n_+ \\ \mathbf{W} \in \mathbb{S}^n}} g(\mathbf{\Gamma}_{\mathrm{low}}, \mathbf{\Gamma}_{\mathrm{up}}, \mathbf{W}) \end{aligned}$$

Gradient computation.

$$\frac{\partial g}{\partial \mathbf{\Gamma}_{\mathrm{low}}} = -\mathbf{r} - \mathrm{Diag}\left(\mathbf{V}_{\mathrm{low}}^\top \left[\nabla\Omega_{+,\mathrm{low}}^\star\left(\mathbf{V}_{\mathrm{low}}\mathrm{Diag}(\mathbf{\Gamma}_{\mathrm{low}_{-\mathbf{b}}})\mathbf{V}_{\mathrm{low}}^\top - \mathbf{W}\right)\right]^\top \mathbf{V}_{\mathrm{low}}\right)$$

$$\frac{\partial g}{\partial \mathbf{\Gamma}_{\mathrm{up}}} = \mathbf{r} - \mathrm{Diag}\left(\mathbf{V}_{\mathrm{up}}^\top \left[\nabla\Omega_{+,\mathrm{up}}^\star\left(\mathbf{V}_{\mathrm{up}}\mathrm{Diag}(\mathbf{\Gamma}_{\mathrm{up}_{-\mathbf{b}}})\mathbf{V}_{\mathrm{up}}^\top + \mathbf{W}\right)\right]^\top \mathbf{V}_{\mathrm{up}}\right)$$

$$\begin{aligned}\frac{\partial g}{\partial \mathbf{W}} = &-\nabla\Omega_{\mathrm{pen}}^\star(\mathbf{W}) + \nabla\Omega_{+,\mathrm{low}}^\star\left(\mathbf{V}_{\mathrm{low}}\mathrm{Diag}(\mathbf{\Gamma}_{\mathrm{low}_{-\mathbf{b}}})\mathbf{V}_{\mathrm{low}}^\top + \mathbf{W}\right)\\ &- \nabla\Omega_{+,\mathrm{up}}^\star\left(\mathbf{V}_{\mathrm{up}}\mathrm{Diag}(\mathbf{\Gamma}_{\mathrm{up}_{-\mathbf{b}}})\mathbf{V}_{\mathrm{up}}^\top + \mathbf{W}\right)\end{aligned}$$

These gradients come from elementary computations as in previous sections, and $\nabla\Omega_{\mathrm{pen}}^\star$ is given in Lemma A.13.

Recovering the solution from optimal Lagrange multipliers. Similarly to Proposition A.8, by denoting $(\widehat{\mathbf{\Gamma}}_{\mathrm{low}}, \widehat{\mathbf{\Gamma}}_{\mathrm{up}}, \widehat{\mathbf{W}}) \in \mathbb{R}_+^n \times \mathbb{R}_+^n \times \mathbb{S}^n$ the optimal variables of the dual problem, to reconstruct the matrices $\widehat{\mathbf{A}}_{\mathrm{low}}, \widehat{\mathbf{A}}_{\mathrm{up}}$ we have:

$$\begin{aligned}\widehat{\mathbf{A}}_{\mathrm{low}} &= \nabla\Omega_{+,\mathrm{low}}^\star\left(\mathbf{V}_{\mathrm{low}}\mathrm{Diag}(\widehat{\mathbf{\Gamma}}_{\mathrm{low}_{-\mathbf{b}}})\mathbf{V}_{\mathrm{low}}^\top - \widehat{\mathbf{W}}\right)\\ &= \frac{1}{2\lambda_{\mathrm{low}2}}\left[\mathbf{V}_{\mathrm{low}}\mathrm{Diag}(\widehat{\mathbf{\Gamma}}_{\mathrm{low}_{-\mathbf{b}}})\mathbf{V}_{\mathrm{low}}^\top - \widehat{\mathbf{W}} - \lambda_{\mathrm{low}1}\mathbf{I}_n\right]_+\\ \widehat{\mathbf{A}}_{\mathrm{up}} &= \nabla\Omega_{+,\mathrm{up}}^\star\left(\mathbf{V}_{\mathrm{up}}\mathrm{Diag}(\widehat{\mathbf{\Gamma}}_{\mathrm{up}_{-\mathbf{b}}})\mathbf{V}_{\mathrm{up}}^\top + \widehat{\mathbf{W}}\right)\\ &= \frac{1}{2\lambda_{\mathrm{up}2}}\left[\mathbf{V}_{\mathrm{up}}\mathrm{Diag}(\widehat{\mathbf{\Gamma}}_{\mathrm{up}_{-\mathbf{b}}})\mathbf{V}_{\mathrm{up}}^\top + \widehat{\mathbf{W}} - \lambda_{\mathrm{up}1}\mathbf{I}_n\right]_+\end{aligned}$$

□

A.3. Asymmetric problem with training set penalty

For the asymmetric setting with training set penalty, our infinite dimensional problem writes:

$$\begin{aligned}\inf_{\substack{\mathcal{A}_{\text{low}}\in\mathcal{S}_+(\mathcal{H}_{\text{low}})\\ \mathcal{A}_{\text{up}}\in\mathcal{S}_+(\mathcal{H}_{\text{up}})}} \quad & \frac{b}{n}\sum_{i=1}^{n}\left(f_{\mathcal{A}_{\text{low}}}(X_i)+f_{\mathcal{A}_{\text{up}}}(X_i)\right)+\Omega_{\text{low}}(\mathcal{A}_{\text{low}})+\Omega_{\text{up}}(\mathcal{A}_{\text{up}})+\lambda_{\text{pen}}\sum_{i=1}^{n}\left(f_{\mathcal{A}_{\text{low}}}(X_i)-f_{\mathcal{A}_{\text{up}}}(X_i)\right)^2\\ \text{s.t.} \quad & r_{\text{low}}\left(X_i,Y_i\right)-f_{\mathcal{A}_{\text{low}}}(X_i)\leq 0,\ i\in[n]\\ & r_{\text{up}}\left(X_i,Y_i\right)-f_{\mathcal{A}_{\text{up}}}(X_i)\leq 0,\ i\in[n]\,.\end{aligned} \tag{38}$$

Once again, the problem is not separable because of the penalty term. To derive a representer theorem, we use Theorem A.2, from which the following theorem is a special case.

Theorem A.14 (Representer theorem with training set penalty)**.** *Let $(b,\lambda_{\text{low}1},\lambda_{\text{up}1})\in\mathbb{R}^3_+$ and $\lambda_{\text{low}2},\lambda_{\text{up}2},\lambda_{\text{pen}}>0$. Then Problem (38) admits a unique solution $(f_{\mathbf{B}^\star_{\text{low}}},f_{\mathbf{B}^\star_{\text{up}}})$ of the form $(f_{\mathbf{B}^\star_{\text{low}}}(X),f_{\mathbf{B}^\star_{\text{up}}}(X))=\left(\mathbf{k}_{\text{low}}(X)^\top\mathbf{B}^\star_{\text{low}}\mathbf{k}_{\text{low}}(X),\ \mathbf{k}_{\text{up}}(X)^\top\mathbf{B}^\star_{\text{up}}\mathbf{k}_{\text{up}}(X)\right)$ for some matrices $\left(\mathbf{B}^\star_{\text{low}},\mathbf{B}^\star_{\text{up}}\right)\in(\mathbb{S}^n_+)^2$.*

Proof. We apply Theorem A.2 to the case $p=2$, with the loss function

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

where C is the set of all positive definite operators $(\mathcal{A}_{\text{low}},\mathcal{A}_{\text{up}})$ such that

$$\begin{cases}r_{\text{low}}\left(X_i,Y_i\right)-f_{\mathcal{A}_{\text{low}}}(X_i)\leq 0,\ i\in[n]\\ r_{\text{up}}\left(X_i,Y_i\right)-f_{\mathcal{A}_{\text{up}}}(X_i)\leq 0,\ i\in[n]\end{cases}$$

and with the penalty function $\Omega((\mathcal{A}_{\text{low}},\mathcal{A}_{\text{up}}))=\Omega_{\text{low}}(\mathcal{A}_{\text{low}})+\Omega_{\text{up}}(\mathcal{A}_{\text{up}})$, where Ω_{low} and Ω_{up} are defined as in Equation (24).

First, from Remark 3 in Marteau-Ferey et al. (2020), Ω_{low} and Ω_{up} verify Assumption A.1. Then, since $L\colon\mathbb{R}^{2n}\to\mathbb{R}$ is lower semi-continuous (notice that it is linear and bounded below by 0) and Ω writes as a sum as in Equation (10), we can apply Theorem A.2 with $p=2$ to deduce that the solution is entirely characterized by two PSD matrices $\left(\mathbf{B}^\star_{\text{low}},\mathbf{B}^\star_{\text{up}}\right)\in(\mathbb{S}^n_+)^2$ and

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

□

Once again, a solution of the previous representer theorem can be recovered with matrices $\mathbf{A}_{\text{low}}$, $\mathbf{A}_{\text{up}}$ instead of $\mathbf{B}_{\text{low}}$, $\mathbf{B}_{\text{up}}$.

Proposition A.15 (Representer theorem with penalty 1, formulation **A**)**.** *Let $(b,\lambda_{\text{low}1},\lambda_{\text{up}1})\in\mathbb{R}^3_+$ and $\lambda_{\text{low}2},\lambda_{\text{up}2},\lambda_{\text{pen}}>0$. The following problem admits a unique solution if L is convex:*

$$\begin{aligned}\inf_{\substack{\mathbf{A}_{\text{low}}\in\mathbb{S}^n_+\\ \mathbf{A}_{\text{up}}\in\mathbb{S}^n_+}} \quad & \frac{b}{n}\sum_{i=1}^{n}\left(\tilde{f}_{\mathbf{A}_{\text{low}}}(X_i)+\tilde{f}_{\mathbf{A}_{\text{up}}}(X_i)\right)+\Omega_{\text{low}}(\mathbf{A}_{\text{low}})+\Omega_{\text{up}}(\mathbf{A}_{\text{up}})+\lambda_{\text{pen}}\sum_{i=1}^{n}\left(\tilde{f}_{\mathbf{A}_{\text{low}}}(X_i)-\tilde{f}_{\mathbf{A}_{\text{up}}}(X_i)\right)^2\\ \text{s.t.} \quad & r_{\text{low}}\left(X_i,Y_i\right)-\tilde{f}_{\mathbf{A}_{\text{low}}}(X_i)\leq 0,\ i\in[n]\\ & r_{\text{up}}\left(X_i,Y_i\right)-\tilde{f}_{\mathbf{A}_{\text{up}}}(X_i)\leq 0,\ i\in[n]\,.\end{aligned} \tag{39}$$

Moreover, for any given solution $\mathbf{A}^\star_{\text{low}},\mathbf{A}^\star_{\text{up}}\in\mathbb{S}^n_+$ of Equation (39)*, the function $(\tilde{f}_{\mathbf{A}^\star_{\text{low}}},\tilde{f}_{\mathbf{A}^\star_{\text{up}}})$ is a minimizer of Equation* (38)*.*

Proof. This is a direct application of Proposition A.6 with $p=2$, and the penalty and loss functions as in the proof of Theorem A.14. □

We now exhibit a dual formulation for Problem (39).

Proposition A.16 (Dual formulation with training set penalty)**.** *Let* $(b, \lambda_{\text{low}1}, \lambda_{\text{up}1}) \in \mathbb{R}^3_+$ *and* $\lambda_{\text{low}2}, \lambda_{\text{up}2}, \lambda_{\text{pen}} > 0$. *Problem (39) admits a dual formulation of the form*

$$\sup_{\substack{\boldsymbol{\Gamma}_{\text{low}} \in \mathbb{R}^n_+, \boldsymbol{\Gamma}_{\text{up}} \in \mathbb{R}^n_+ \\ \boldsymbol{\alpha}_0 \in \mathbb{R}^n}} (\boldsymbol{\Gamma}_{\text{up}} - \boldsymbol{\Gamma}_{\text{low}})\mathbf{r}^\top - \frac{1}{4\lambda_{\text{pen}}}\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^\top \tag{40}$$
$$- \Omega^\star_{+,\text{low}}(\mathbf{V}_{\text{low}}\text{Diag}((\boldsymbol{\Gamma}_{\text{low}} + \boldsymbol{\alpha}_0)_{-\mathbf{b}})\mathbf{V}_{\text{low}}^\top) - \Omega^\star_{+,\text{up}}(\mathbf{V}_{\text{up}}\text{Diag}((\boldsymbol{\Gamma}_{\text{up}} - \boldsymbol{\alpha}_0)_{-\mathbf{b}})\mathbf{V}_{\text{up}}^\top)$$

where $\mathbf{r}$ *is the vector of residuals* $r_i = Y_i - m(X_i)$, $\Omega^\star_{+,(\cdot)}(\mathbf{B}) = \frac{1}{4\lambda_{(\cdot)2}}\|\left[\mathbf{B} - \lambda_{(\cdot)1}\mathbf{I}_n\right]_+\|_F^2$ *and* $\forall x \in \mathbb{R},\ \text{Diag}\left((\cdot)_x\right) := \text{Diag}\left((\cdot)\right) + \frac{x}{n}\mathbf{I}_n$. *Moreover, if* $(\widehat{\boldsymbol{\Gamma}}_{\text{low}}, \widehat{\boldsymbol{\Gamma}}_{\text{up}}, \widehat{\boldsymbol{\alpha}}_0)$ *is a solution of Equation* (40)*, a solution of Problem A.15 can be retrieved as*

$$\widehat{\mathbf{A}}_{\text{low}} = \frac{1}{2\lambda_{\text{low}2}}\left[\mathbf{V}_{\text{low}}\text{Diag}((\widehat{\boldsymbol{\Gamma}}_{\text{low}} + \widehat{\boldsymbol{\alpha}}_0)_{-\mathbf{b}})\mathbf{V}_{\text{low}}^\top - \lambda_{\text{low}1}\mathbf{I}_n\right]_+$$
$$\widehat{\mathbf{A}}_{\text{up}} = \frac{1}{2\lambda_{\text{up}2}}\left[\mathbf{V}_{\text{up}}\text{Diag}((\widehat{\boldsymbol{\Gamma}}_{\text{up}} - \widehat{\boldsymbol{\alpha}}_0)_{-\mathbf{b}})\mathbf{V}_{\text{up}}^\top - \lambda_{\text{up}1}\mathbf{I}_n\right]_+.$$

Proof. As opposed to the separable case in Proposition A.8, the interaction between $\mathbf{A}_{\text{low}}$ and $\mathbf{A}_{\text{up}}$ in the penalty term requires more attention. The dual problem is defined as

$$d = \sup_{\substack{\boldsymbol{\Gamma}_{\text{low}} \in \mathbb{R}^n_+ \\ \boldsymbol{\Gamma}_{\text{up}} \in \mathbb{R}^n_+}} \inf_{\substack{\mathbf{A}_{\text{low}} \in \mathbb{S}^n \\ \mathbf{A}_{\text{up}} \in \mathbb{S}^n}} \mathcal{L}(\boldsymbol{\Gamma}_{\text{low}}, \boldsymbol{\Gamma}_{\text{up}}, \mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}}) = \sup_{\substack{\boldsymbol{\Gamma}_{\text{low}} \in \mathbb{R}^n_+ \\ \boldsymbol{\Gamma}_{\text{up}} \in \mathbb{R}^n_+}} (\boldsymbol{\Gamma}_{\text{up}} - \boldsymbol{\Gamma}_{\text{low}})\mathbf{r}^\top + D(\boldsymbol{\Gamma}_{\text{low}}, \boldsymbol{\Gamma}_{\text{up}}) \tag{41}$$

where we denote

$$D(\boldsymbol{\Gamma}_{\text{low}}, \boldsymbol{\Gamma}_{\text{up}}) \coloneqq \inf_{\mathbf{A}_{\text{low}} \in \mathbb{S}^n,\ \mathbf{A}_{\text{up}} \in \mathbb{S}^n} \overline{\mathcal{L}}(\boldsymbol{\Gamma}_{\text{low}}, \boldsymbol{\Gamma}_{\text{up}}, \mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}}) \tag{42}$$

with $\overline{\mathcal{L}}$ defined as

$$\begin{aligned}\overline{\mathcal{L}}(\boldsymbol{\Gamma}_{\text{low}}, \boldsymbol{\Gamma}_{\text{up}}, \mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}}) = &\sum_{i=1}^n (\frac{b}{n} - \Gamma_{\text{low},i})\tilde{f}_{\mathbf{A}_{\text{low}}}(X_i) + \sum_{i=1}^n (\frac{b}{n} - \Gamma_{\text{up},i})\tilde{f}_{\mathbf{A}_{\text{up}}}(X_i) \\ &+ \lambda_{\text{pen}}\sum_{i=1}^n (\tilde{f}_{\mathbf{A}_{\text{low}}}(X_i) - \tilde{f}_{\mathbf{A}_{\text{up}}}(X_i))^2 \\ &+ \Omega_{+,\text{low}}(\mathbf{A}_{\text{low}}) + \Omega_{+,\text{up}}(\mathbf{A}_{\text{up}}).\end{aligned}$$

Contrary to Allain et al. (2026) where $D()$ has an explicit expression, here we follow Marteau-Ferey et al. (2020); Allain et al. (2026) and use Theorem 3.3.5 from Borwein & Lewis (2006) to get the following lemma.

Lemma A.17. *The dual problem associated to Equation* (42) *writes*

$$D(\boldsymbol{\Gamma}_{\text{low}}, \boldsymbol{\Gamma}_{\text{up}}) = \sup_{\boldsymbol{\alpha}_0 \in \mathbb{R}^n} -\frac{1}{4\lambda_{\text{pen}}}\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^\top \tag{43}$$
$$- \Omega^\star_{+,\text{low}}(\mathbf{V}_{\text{low}}\text{Diag}((\boldsymbol{\Gamma}_{\text{low}} + \boldsymbol{\alpha}_0)_{-\mathbf{b}})\mathbf{V}_{\text{low}}^\top) \tag{44}$$
$$- \Omega^\star_{+,\text{up}}(\mathbf{V}_{\text{up}}\text{Diag}((\boldsymbol{\Gamma}_{\text{up}} - \boldsymbol{\alpha}_0)_{-\mathbf{b}}))\mathbf{V}_{\text{up}}^\top)$$

.

Proof. First notice that D can be decomposed as:

$$D(\boldsymbol{\Gamma}_{\text{low}}, \boldsymbol{\Gamma}_{\text{up}}) = \inf_{\mathbf{x} \in \mathbb{S}^n \times \mathbb{S}^n} L(R\mathbf{x}) + f(\mathbf{x})$$

where $\mathbf{x} = (\mathbf{A}_{\text{low}}, \mathbf{A}_{\text{up}})$, $R:\ \mathbf{x} \in \mathbb{S}^n \times \mathbb{S}^n \mapsto (\ldots, \tilde{f}_{\mathbf{A}_{\text{low}}}(X_i), \ldots, \tilde{f}_{\mathbf{A}_{\text{up}}}(X_i), \ldots)$, $f(\mathbf{x}) = \Omega_{+,\text{low}}(\mathbf{A}_{\text{low}}) + \Omega_{+,\text{up}}(\mathbf{A}_{\text{up}})$ and

$$L(R\mathbf{x}) = \sum_{i=1}^{n}(\frac{b}{n} - \Gamma_{\text{low},i})\tilde{f}_{\mathbf{A}_{\text{low}}}(X_i) + \sum_{i=1}^{n}(\frac{b}{n} - \Gamma_{\text{up},i})\tilde{f}_{\mathbf{A}_{\text{up}}}(X_i) + \lambda_{\text{pen}}\sum_{i=1}^{n}(\tilde{f}_{\mathbf{A}_{\text{low}}}(X_i) - \tilde{f}_{\mathbf{A}_{\text{up}}}(X_i))^2.$$

Applying Theorem 3.3.5 from Borwein & Lewis (2006) to Equation (42) gives

$$D(\mathbf{\Gamma}_{\text{low}}, \mathbf{\Gamma}_{\text{up}}) = \sup_{\boldsymbol{\alpha}^\star = (\boldsymbol{\alpha}^\star_{\text{low}}, \boldsymbol{\alpha}^\star_{\text{up}}) \in \mathbb{R}^{2n}} - L^\star(-\boldsymbol{\alpha}^\star) - f^\star(R^\star\boldsymbol{\alpha}^\star). \tag{45}$$

We first derive the expression for $L^\star$, by expressing L in quadratic form. Note that $L^\star(-\boldsymbol{\alpha}^\star) = [\tilde{L}(\mathbf{u})]^\star(\boldsymbol{\alpha}^\star)$, where $\tilde{L}(\mathbf{u}) = L(-\mathbf{u})$, such that:

$$\begin{aligned}
\tilde{L}(\mathbf{u}) &= L(-\mathbf{u}) \\
&= \sum_{i=1}^{n}(\Gamma_{\text{low},i} - \frac{b}{n})u_{\text{low},i} + \sum_{i=1}^{n}(\Gamma_{\text{up},i} - \frac{b}{n})u_{\text{up},i} + \lambda_{\text{pen}}\sum_{i=1}^{n}(u_{\text{up},i} - u_{\text{low},i})^2 \\
&= \mathbf{a}^\top\mathbf{u} + \frac{1}{2}\mathbf{u}^\top\mathbf{M}\mathbf{u}
\end{aligned}$$

where $\mathbf{a} = [\ldots, \Gamma_{\text{low},i} - \frac{b}{n}, \ldots, \Gamma_{\text{up},i} - \frac{b}{n}, \ldots]$ and $\mathbf{M} = 2\lambda_{\text{pen}}\begin{bmatrix} I_n & -I_n \\ -I_n & I_n \end{bmatrix}$. It follows that

$$\tilde{L}^\star(\boldsymbol{\alpha}^\star) = \frac{1}{2}(\boldsymbol{\alpha}^\star - \mathbf{a})^\top\mathbf{M}^\dagger(\boldsymbol{\alpha}^\star - \mathbf{a}) \tag{46}$$

for $\boldsymbol{\alpha}^\star \in \text{Range}(\mathbf{M}) + \mathbf{a}$ (Becker et al., 2019) and where $\mathbf{M}^\dagger$ is the pseudo-inverse of $\mathbf{M}$. It remains to compute $\mathbf{M}^\dagger$ and $\text{Range}(\mathbf{M}) + \mathbf{a}$.

First, observe that the pseudo-inverse of the matrix $\begin{bmatrix} I_n & -I_n \\ -I_n & I_n \end{bmatrix}$ is given by $\frac{1}{4}\begin{bmatrix} I_n & -I_n \\ -I_n & I_n \end{bmatrix}$, which leads to:

$$\mathbf{M}^\dagger = \left(2\lambda_{\text{pen}}\begin{bmatrix} I_n & -I_n \\ -I_n & I_n \end{bmatrix}\right)^\dagger = \frac{1}{8\lambda_{\text{pen}}}\begin{bmatrix} I_n & -I_n \\ -I_n & I_n \end{bmatrix} = \frac{1}{16\lambda^2_{\text{pen}}}\mathbf{M}.$$

For identifying $\text{Range}(\mathbf{M}) + \mathbf{a}$, let $\mathbf{v} \in \mathbb{R}^{2n}$. Then,

$$\mathbf{M}\mathbf{v} = 2\lambda_{\text{pen}}\begin{bmatrix} I_n & -I_n \\ -I_n & I_n \end{bmatrix}\begin{bmatrix} v_{1:n} \\ v_{(n+1):2n} \end{bmatrix} = 2\lambda_{\text{pen}}\begin{bmatrix} v_{1:n} - v_{(n+1):2n} \\ v_{(n+1):2n} - v_{1:n} \end{bmatrix}$$

and it follows that

$$\text{Range}(\mathbf{M}) = \left\{\mathbf{v} \in \mathbb{R}^{2n}:\ v_{1:n} = -v_{(n+1):2n}\right\}.$$

Denoting $\boldsymbol{\alpha}_0 \in \mathbb{R}^n$ the vector which consists of the first n components of $\mathbf{v}$, then $\boldsymbol{\alpha}^\star$ in Equation (46) can be parameterized as

$$\boldsymbol{\alpha}^\star = (\boldsymbol{\alpha}^\star_{\text{low}}, \boldsymbol{\alpha}^\star_{\text{up}}) = [\boldsymbol{\alpha}_0 + \mathbf{a}_{1:n},\ -\boldsymbol{\alpha}_0 + \mathbf{a}_{(n+1):2n}] \in \text{Range}(\mathbf{M}) + \mathbf{a}, \tag{47}$$

such that Equation (46) becomes

$$L^\star(-\boldsymbol{\alpha}^\star) = \frac{1}{4\lambda_{\text{pen}}}\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^\top.$$

Next, since f is defined as a separable sum, its conjugate is the sum of the conjugates: $f^\star(\mathbf{A}^\star_{\text{low}}, \mathbf{A}^\star_{\text{low}}) = \Omega^\star_{+,\text{low}}(\mathbf{A}^\star_{\text{low}}) + \Omega^\star_{+,\text{up}}(\mathbf{A}^\star_{\text{up}})$. Finally, to compute $R^\star\boldsymbol{\alpha}^\star$ notice that

$$\begin{aligned}
R\mathbf{x} &= (\ldots, \tilde{f}_{\mathbf{A}_{\text{low}}}(X_i), \ldots, \tilde{f}_{\mathbf{A}_{\text{up}}}(X_i), \ldots) \\
&= \Big(\text{Diag}(\mathbf{V}_{\text{low}}\mathbf{A}_{\text{low}}\mathbf{V}_{\text{low}}^\top), \text{Diag}(\mathbf{V}_{\text{up}}\mathbf{A}_{\text{up}}\mathbf{V}_{\text{up}}^\top)\Big).
\end{aligned}$$

By considering the Frobenius scalar product on $(\mathbb{S}^n)^2$, we thus have

$$\begin{aligned} R^\star \boldsymbol{\alpha}^\star &= \left(\mathbf{V}_{\text{low}}\text{Diag}(\boldsymbol{\alpha}^\star_{\text{low}})\mathbf{V}_{\text{low}}^\top, \mathbf{V}_{\text{up}}\text{Diag}(\boldsymbol{\alpha}^\star_{\text{up}})\mathbf{V}_{\text{up}}^\top\right) \\ &= (R^\star_{\text{low}}\boldsymbol{\alpha}^\star_{\text{low}}, R^\star_{\text{up}}\boldsymbol{\alpha}^\star_{\text{up}}). \end{aligned}$$

where $R^\star_{(\cdot)}\boldsymbol{\alpha}^\star_{(\cdot)} = \mathbf{V}_{(\cdot)}\text{Diag}(\boldsymbol{\alpha}^\star_{(\cdot)})\mathbf{V}_{(\cdot)}^\top$. The value of $f^\star$ at $R^\star\boldsymbol{\alpha}^\star$ is finally given by

$$\begin{aligned} f^\star(R^\star\boldsymbol{\alpha}^\star) &= \Omega^\star_{+,\text{low}}(\mathbf{V}_{\text{low}}\text{Diag}(\boldsymbol{\alpha}^\star_{\text{low}})\mathbf{V}_{\text{low}}^\top) + \Omega^\star_{+,\text{up}}(\mathbf{V}_{\text{up}}\text{Diag}(\boldsymbol{\alpha}^\star_{\text{up}})\mathbf{V}_{\text{up}}^\top) \\ &= \Omega^\star_{+,\text{low}}(\mathbf{V}_{\text{low}}\text{Diag}(\boldsymbol{\alpha}_0 + \mathbf{a})\mathbf{V}_{\text{low}}^\top) + \Omega^\star_{+,\text{up}}(\mathbf{V}_{\text{up}}\text{Diag}(-\boldsymbol{\alpha}_0 + \mathbf{a})\mathbf{V}_{\text{up}}^\top), \end{aligned}$$

which concludes the proof of Lemma A.17. □

Putting things together, using Lemma A.17 and Equation (41), our main dual problem writes

$$\begin{aligned} d = &\sup_{\substack{\boldsymbol{\Gamma}_{\text{low}}\in\mathbb{R}^n_+, \boldsymbol{\Gamma}_{\text{up}}\in\mathbb{R}^n_+ \\ \boldsymbol{\alpha}_0\in\mathbb{R}^n}} (\boldsymbol{\Gamma}_{\text{up}} - \boldsymbol{\Gamma}_{\text{low}})\mathbf{r}^\top - \frac{1}{4\lambda_{\text{pen}}}\boldsymbol{\alpha}_0\boldsymbol{\alpha}_0^\top \\ &\qquad - \Omega^\star_{+,\text{low}}(\mathbf{V}_{\text{low}}\text{Diag}((\boldsymbol{\Gamma}_{\text{low}} + \boldsymbol{\alpha}_0)_{-\mathbf{b}})\mathbf{V}_{\text{low}}^\top) - \Omega^\star_{+,\text{up}}(\mathbf{V}_{\text{up}}\text{Diag}((\boldsymbol{\Gamma}_{\text{up}} - \boldsymbol{\alpha}_0)_{-\mathbf{b}})\mathbf{V}_{\text{up}}^\top) \\ = &\sup_{\substack{\boldsymbol{\Gamma}_{\text{low}}\in\mathbb{R}^n_+, \boldsymbol{\Gamma}_{\text{up}}\in\mathbb{R}^n_+ \\ \boldsymbol{\alpha}_0\in\mathbb{R}^n}} g(\boldsymbol{\Gamma}_{\text{low}}, \boldsymbol{\Gamma}_{\text{up}}, \boldsymbol{\alpha}_0). \end{aligned}$$

Gradient computation.

$$\frac{\partial g}{\partial \boldsymbol{\Gamma}_{\text{low}}} = -\mathbf{r} - \text{Diag}\left(\mathbf{V}_{\text{low}}^\top\left[\nabla\Omega^\star_{+,\text{low}}\left(\mathbf{V}_{\text{low}}\text{Diag}((\boldsymbol{\Gamma}_{\text{low}} + \boldsymbol{\alpha}_0)_{-\mathbf{b}})\mathbf{V}_{\text{low}}^\top\right)\right]^\top\mathbf{V}_{\text{low}}\right)$$

$$\frac{\partial g}{\partial \boldsymbol{\Gamma}_{\text{up}}} = \mathbf{r} - \text{Diag}\left(\mathbf{V}_{\text{up}}^\top\left[\nabla\Omega^\star_{+,\text{up}}\left(\mathbf{V}_{\text{up}}\text{Diag}((\boldsymbol{\Gamma}_{\text{up}} - \boldsymbol{\alpha}_0)_{-\mathbf{b}})\mathbf{V}_{\text{up}}^\top\right)\right]^\top\mathbf{V}_{\text{up}}\right)$$

$$\begin{aligned} \frac{\partial g}{\partial \boldsymbol{\alpha}_0} = &-\frac{1}{2\lambda_{\text{pen}}}\boldsymbol{\alpha}_0 - \text{Diag}\left(\mathbf{V}_{\text{low}}^\top\left[\nabla\Omega^\star_{+,\text{low}}\left(\mathbf{V}_{\text{low}}\text{Diag}((\boldsymbol{\Gamma}_{\text{low}} + \boldsymbol{\alpha}_0)_{-\mathbf{b}})\mathbf{V}_{\text{low}}^\top\right)\right]^\top\mathbf{V}_{\text{low}}\right) \\ &+ \text{Diag}\left(\mathbf{V}_{\text{up}}^\top\left[\nabla\Omega^\star_{+,\text{up}}\left(\mathbf{V}_{\text{up}}\text{Diag}((\boldsymbol{\Gamma}_{\text{up}} - \boldsymbol{\alpha}_0)_{-\mathbf{b}})\mathbf{V}_{\text{up}}^\top\right)\right]^\top\mathbf{V}_{\text{up}}\right) \end{aligned}$$

These gradients come from elementary computations following Allain et al. (2026).

Recovering the solution from optimal Lagrange multipliers. Similarly to Proposition A.8, by denoting $(\widehat{\boldsymbol{\Gamma}}_{\text{low}}, \widehat{\boldsymbol{\Gamma}}_{\text{up}}, \widehat{\boldsymbol{\alpha}}_0) \in \mathbb{R}^n_+ \times \mathbb{R}^n_+ \times \mathbb{R}^n$ the optimal variables of the dual problem, to reconstruct the matrices $\widehat{\mathbf{A}}_{\text{low}}, \widehat{\mathbf{A}}_{\text{up}}$ we have:

$$\begin{aligned} \widehat{\mathbf{A}}_{\text{low}} &= \nabla\Omega^\star_{+,\text{low}}\left(\mathbf{V}_{\text{low}}\text{Diag}((\widehat{\boldsymbol{\Gamma}}_{\text{low}} + \widehat{\boldsymbol{\alpha}}_0)_{-\mathbf{b}})\mathbf{V}_{\text{low}}^\top\right) \\ &= \frac{1}{2\lambda_{\text{low2}}}\left[\mathbf{V}_{\text{low}}\text{Diag}((\widehat{\boldsymbol{\Gamma}}_{\text{low}} + \widehat{\boldsymbol{\alpha}}_0)_{-\mathbf{b}})\mathbf{V}_{\text{low}}^\top - \lambda_{\text{low1}}\mathbf{I}_n\right]_+ \\ \widehat{\mathbf{A}}_{\text{up}} &= \nabla\Omega^\star_{+,\text{up}}\left(\mathbf{V}_{\text{up}}\text{Diag}((\widehat{\boldsymbol{\Gamma}}_{\text{up}} - \widehat{\boldsymbol{\alpha}}_0)_{-\mathbf{b}})\mathbf{V}_{\text{up}}^\top\right) \\ &= \frac{1}{2\lambda_{\text{up2}}}\left[\mathbf{V}_{\text{up}}\text{Diag}((\widehat{\boldsymbol{\Gamma}}_{\text{up}} - \widehat{\boldsymbol{\alpha}}_0)_{-\mathbf{b}})\mathbf{V}_{\text{up}}^\top - \lambda_{\text{up1}}\mathbf{I}_n\right]_+ \end{aligned}$$

see Theorem 8 in Marteau-Ferey et al. (2020) and Appendix A.2 in Allain et al. (2026). □

A.4. Error bounds

To derive the error bounds for both penalties, we make the following assumption on the space $\mathcal{H}$ (Assumption 2(a) in Rudi et al. (2025)).

Assumption A.18. For a bounded open set $\mathcal{X} \in \mathbb{R}^d$, the RKHS $\mathcal{H}$ of functions on $\mathcal{X}$ with norm $\|\cdot\|_{\mathcal{H}}$ satisfies $f|\mathcal{X} \in \mathcal{H}$, $\forall f \in C^\infty(\mathbb{R}^d)$. Moreover $\forall u, v \in \mathcal{H}$, $u \cdot v \in \mathcal{H}$ and $\exists M \geq 1$ such that

$$\|u \cdot v\|_{\mathcal{H}} \leq \mathrm{M}\|u\|_{\mathcal{H}}\|v\|_{\mathcal{H}}.$$

This assumption has an important implication: if it holds, kernel SoS functions will live in the RKHS $\mathcal{H}$, and their norm can be controlled by the nuclear norm of their operator, as elaborated in the following lemma.

Lemma A.19 (Lemma 9 in Rudi et al. (2025).)**.** *Let $\mathcal{X}$ and $\mathcal{H}$ satisfy Assumption A.18. If $\mathcal{A}$ is a trace-class operator, then $f_{\mathcal{A}}(X) = \langle \phi(X), \mathcal{A}\phi(X)\rangle \in \mathcal{H}$ and*

$$\|f_{\mathcal{A}}\|_{\mathcal{H}} \leq \mathrm{M}\|\mathcal{A}\|_\star.$$

A.4.1. Operator penalty.

We can now state the error bound with the operator penalty, which controls the difference between two kernel SoS functions with the nuclear norm of their operator difference.

Proposition A.20 (Error bound with operator penalty)**.** *Let $\mathcal{H}$ be a reproducing kernel Hilbert space with associated kernel k and feature map ϕ which satisfies Assumption A.18. If k is bounded such that $\|k\|_\infty := \sup_{x\in\mathcal{X}} \sqrt{k(X,X)} < \infty$, then for any two PSD operators $\mathcal{A}_1, \mathcal{A}_2 \in \mathbb{S}_+(\mathcal{H})$, we have*

$$\sup_{x\in\mathcal{X}} |f_{\mathcal{A}_1}(x) - f_{\mathcal{A}_2}(x)| \leq \mathrm{M}\|k\|_\infty \, \|\mathcal{A}_1 - \mathcal{A}_2\|_\star,$$

where $f_{\mathcal{A}_1}(X) = \langle \phi(X), \mathcal{A}_1\phi(X)\rangle \in \mathcal{H}$ and $f_{\mathcal{A}_2}(X) = \langle \phi(X), \mathcal{A}_2\phi(X)\rangle \in \mathcal{H}$ are the kernel sum-of-squares functions associated to $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively.

Proof. Writing $\Delta f = f_1 - f_2$, if $\mathcal{H}$ satisfies Assumption A.18 then $\Delta f \in \mathcal{H}$, such that

$$\Delta f(X) = \langle \Delta f, k(X, \cdot)\rangle_{\mathcal{H}} \leq \|\Delta f\|_{\mathcal{H}}\|k\|_\infty,$$

see Steinwart & Christmann (2008) Lemma 4.23. We conclude the proof by applying Lemma A.19 to $\Delta f = \langle \phi(X), (\mathcal{A}_1 - \mathcal{A}_2)\phi(X)\rangle$ with $\mathcal{A} = \mathcal{A}_1 - \mathcal{A}_2$. □

A.4.2. Training set penalty.

While the operator penalty controls the difference of kSoS functions through the difference of their associated operators, the training set penalty relies instead on the difference of the functions on the training points.
For this penalty, we first need a new assumption on the space $\mathcal{H}$ (Assumption 2(d) in Rudi et al. (2025) with $m = 1$).

Assumption A.21. For a bounded open set $\mathcal{X} \in \mathbb{R}^d$, the RKHS $\mathcal{H}$ of functions on $\mathcal{X}$ with associated kernel k satisfies

$$\max_{|\alpha|=1} \sup_{x,y\in\mathcal{X}} |\partial_x^\alpha \partial_y^\alpha k(x,y)| \leq \mathrm{D}^2 < \infty$$

for some $\mathrm{D} \geq 1$.

Assumption A.21 implies that $\mathcal{H} \subseteq C^1(\mathcal{X})$ and that $\forall u \in \mathcal{H}$,

$$\max_{|\alpha|=1} \sup_{x\in\mathcal{X}} |\partial_x^\alpha u(x)| \leq \mathrm{D} \tag{48}$$

see Remark 2 in Rudi et al. (2025).
We also require the following geometric property on the domain $\mathcal{X}$:

Assumption A.22. $\mathcal{X}$ writes as $\mathcal{X} = \cup_{x\in S} \mathrm{B}_r(x)$, where S is a bounded subset of $\mathbb{R}^d$ and $\mathrm{B}_r(x)$ is the ball of center x and radius r.

In order to show that controlling the difference of functions on a finite subset $\widehat{X} = \{X_1, \dots, X_n\}$ allows to control the difference of functions on their whole domain $\mathcal{X}$, we rely on traditional scattered data approximation techniques. In particular, they bring into play the distance of any data point x to $\widehat{X}$, known as the *fill-in distance* defined by:

$$\rho_{\widehat{X},\mathcal{X}} = \sup_{x\in\mathcal{X}} \min_{X_i\in\widehat{X}} \|x - X_i\|.$$

Proposition A.23 (Error bound with training set penalty)**.** *Let $\mathcal{X}$ satisfying Assumption A.22 and $\widehat{X} = \{X_1, \dots, X_n\}$ be a finite subset of $\mathcal{X}$. Let $\mathcal{H}$ be a RKHS of functions defined on $\mathcal{X}$ with associated kernel k satisfying Assumptions A.18 and A.21. Then for any two PSD operators $\mathcal{A}_1, \mathcal{A}_2 \in \mathcal{S}_+(\mathcal{H})$, it holds*

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

where $f_{\mathcal{A}_1}(X) = \langle\phi(X), \mathcal{A}_1\phi(X)\rangle \in \mathcal{H}$ and $f_{\mathcal{A}_2}(X) = \langle\phi(X), \mathcal{A}_2\phi(X)\rangle \in \mathcal{H}$ are the kernel sum-of-squares functions associated to $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively, and $C_{f_1,f_2} = 2d\mathrm{DM}\|\mathcal{A}_1 - \mathcal{A}_2\|_\star$ with D and M constants depending on kernel k.

Proof. The proof relies on the following general result (Wendland & Rieger (2005), Rudi et al. (2025) Theorem 13 with $k = m = 0$) which states that for any function $f\colon \mathcal{X} \to \mathbb{R}$, if $\mathcal{X}$ satisfies Assumption A.22 and f is at least $C^1(\mathcal{X})$ then

$$\sup_{x\in\mathcal{X}} |f(x)| \leq 2\, C_f\rho_{\widehat{X},\mathcal{X}} + \max_{X_i\in\widehat{X}} |f(X_i)|$$

with $C_f = \sum_{|\alpha|=1} \frac{1}{\alpha!} \sup_{x\in\mathcal{X}} |\partial^\alpha f|$.
The core idea of the proof is to apply this result to the difference of kSoS functions. First, notice that $\Delta f = f_1 - f_2 \in C^1$ as a difference of C^1 functions. Then, we need to explicit the bound C_f. By Assumption A.21, applying Equation (48) to Δf we have:

$$\max_{|\alpha|=1} \sup_{x\in\mathcal{X}} |\partial^\alpha \Delta f(x)| \leq \mathrm{D}\|\Delta f\|_{\mathcal{H}}.$$

Next, applying Lemma A.19 to Δf we obtain $\|\Delta f\|_{\mathcal{H}} \leq \mathrm{M}\|\mathcal{A}_1 - \mathcal{A}_2\|_\star$ where M is the constant from Assumption A.18. Putting things together, we have:

$$C_{\Delta f} = \sum_{|\alpha|=1} \frac{1}{\alpha!} \sup_{x\in\mathcal{X}} |\partial^\alpha \Delta f(x)| \leq d\mathrm{DM}\|\mathcal{A}_1 - \mathcal{A}_2\|_\star.$$

To conclude the proof, we use

$$\max_{X_i\in\widehat{X}} |\Delta f(X_i)| \leq \sqrt{\sum_{i=1}^{n}(\Delta f(X_i))^2}.$$

□

In our work, we focus on the Matérn $5/2$ kernel for its superior empirical performance. Crucially, the following remark shows that this kernel actually satisfies all the assumptions needed for Proposition A.23 and Proposition A.20.

Remark A.24. *Following Proposition 1 in Rudi et al. (2025), the Matérn $5/2$ kernel with lengthscale θ is a Sobolev kernel with $\nu = s - d/2$ which satisfies Assumptions A.18 and A.21 with constants*

$$\mathrm{M} = 2^{3+d/2}(2\pi)^{d/2} \quad \text{and} \quad \mathrm{D} = \frac{1}{\theta}(2\pi)^{d/4}\sqrt{d/3}.$$

A.4.3. Post-optimization error bounds.

Proposition A.20 and Proposition A.23 give error bounds for the infinite dimensional operators. We need to show that the finite representation of those operators still controls the difference between the finite representation of the functions. This is done using the following lemma.

Lemma A.25. *Let $\mathcal{H}$ be a reproducing kernel Hilbert space with associated kernel k and feature map ϕ. For any matrix $\mathbf{A} \in \mathbb{S}^n$, any partial isometry $\mathrm{U}\colon \mathcal{H} \to \mathbb{R}^n$ and any $p \geq 1$, we have*

$$\|\mathbf{A}\|_p = \|\mathrm{U}^\star \mathbf{A} \mathrm{U}\|_p,$$

where $\|\cdot\|_p$ is the p-Schatten norm.

Proof. Let $\mathrm{U}\colon \mathcal{H} \to \mathbb{R}^n$ be a partial isometry (i.e. $\mathrm{U}\mathrm{U}^\star$ is the identity over $\mathbb{R}^n$).

By first noticing that $(\mathrm{U}^\star \mathbf{A} \mathrm{U})^\star (\mathrm{U}^\star \mathbf{A} \mathrm{U}) = \mathrm{U}^\star \mathbf{A}^\star \mathbf{A} \mathrm{U}$, non-zero singular values of $\mathrm{U}^\star \mathbf{A} \mathrm{U}$ are given by

$$\begin{aligned}
\sigma(\mathrm{U}^\star \mathbf{A} \mathrm{U}) &= \sqrt{\lambda((\mathrm{U}^\star \mathbf{A} \mathrm{U})^\star (\mathrm{U}^\star \mathbf{A} \mathrm{U}))} \\
&= \sqrt{\lambda(\mathrm{U}^\star \mathbf{A}^\star \mathbf{A} \mathrm{U})} \\
&= \sqrt{\lambda(\mathbf{A}^\star \mathbf{A})} \\
&= \sigma(\mathbf{A}),
\end{aligned}$$

where the third equality comes from Pedersen (2012) Exercise 4.1.3, which states that $\lambda(\mathcal{A}\mathcal{B}) \setminus \{0\} = \lambda(\mathcal{B}\mathcal{A}) \setminus \{0\}$ for any two bounded operators, applied to $\mathrm{U}^\star \mathbf{A}^\star$ and $\mathbf{A}\mathrm{U}$ (which are bounded because $\mathrm{U}^\star \mathbf{A}^\star$ is defined on $\mathbb{R}^n$ and $\mathbf{A}\mathrm{U}$ is its adjoint). This concludes the proof because p-Schatten norms are solely defined by the singular values. □

Proposition A.26. *Under the same hypothesis as Proposition A.20, if $(\mathbf{A}_{\mathrm{low}}, \mathbf{A}_{\mathrm{up}})$ are solutions of Equation* (31) *then*

$$\sup_{x \in \mathcal{X}} |\tilde{f}_{\mathbf{A}_{\mathrm{low}}}(x) - \tilde{f}_{\mathbf{A}_{\mathrm{up}}}(x)| \leq \mathrm{M}\|k\|_\infty \, \|\mathbf{A}_{\mathrm{low}} - \mathbf{A}_{\mathrm{low}}\|_\star$$

where $\tilde{f}_{\mathbf{A}_{\mathrm{low}}}(X) = \mathbf{\Phi}(X)^\top \mathbf{A}_{\mathrm{low}} \mathbf{\Phi}(X) \in \mathcal{H}$ and $\tilde{f}_{\mathbf{A}_{\mathrm{up}}}(X) = \mathbf{\Phi}(X)^\top \mathbf{A}_{\mathrm{up}} \mathbf{\Phi}(X) \in \mathcal{H}$ are the kernel sum-of-squares functions associated to $\mathbf{A}_{\mathrm{low}}$ and $\mathbf{A}_{\mathrm{up}}$, respectively.

Proof. Let $(\mathbf{A}_{\mathrm{low}}, \mathbf{A}_{\mathrm{up}})$ be solutions of Equation (31) and write $(\mathcal{A}_{\mathrm{low}}, \mathcal{A}_{\mathrm{up}})$ their associated operators. Let us consider $\mathrm{U}\colon \mathcal{H} \to \mathbb{R}^n$, the following partial isometry $f \longmapsto \mathbf{V}^{-\top} (\langle \phi(X_1), f \rangle, \ldots, \langle \phi(X_n), f \rangle)$.

First, we show that $f_{\mathcal{A}_{\mathrm{low}}}(X) = \tilde{f}_{\mathbf{A}_{\mathrm{low}}}(X)$ for any $X \in \mathcal{X}$:

$$\begin{aligned}
\tilde{f}_{\mathbf{A}_{\mathrm{low}}}(X) &= \langle \mathbf{\Phi}(X), \mathbf{A}_{\mathrm{low}} \mathbf{\Phi}(X) \rangle \\
&= \langle \mathrm{U}\phi(X), \mathbf{A}_{\mathrm{low}} \mathrm{U}\phi(X) \rangle \\
&= \langle \phi(X), \mathrm{U}^\star \mathbf{A}_{\mathrm{low}} \mathrm{U}\phi(X) \rangle \\
&= \langle \phi(X), \mathcal{A}_{\mathrm{low}} \phi(X) \rangle \\
&= f_{\mathcal{A}_{\mathrm{low}}}(X).
\end{aligned}$$

The same results holds for $_{\mathrm{up}}$: $f_{\mathcal{A}_{\mathrm{up}}}(X) = \tilde{f}_{\mathbf{A}_{\mathrm{up}}}(X)$ for any $X \in \mathcal{X}$.

By applying Proposition A.20 to $(\mathcal{A}_{\mathrm{low}}, \mathcal{A}_{\mathrm{up}})$, we have

$$\sup_{x \in \mathcal{X}} |f_{\mathcal{A}_{\mathrm{low}}}(x) - f_{\mathcal{A}_{\mathrm{up}}}(x)| \leq \mathrm{M}\|k\|_\infty \, \|\mathcal{A}_{\mathrm{low}} - \mathcal{A}_{\mathrm{up}}\|_\star,$$

and

$$\sup_{x \in \mathcal{X}} |\tilde{f}_{\mathbf{A}_{\mathrm{low}}}(x) - \tilde{f}_{\mathbf{A}_{\mathrm{up}}}(x)| \leq \mathrm{M}\|k\|_\infty \, \|\mathcal{A}_{\mathrm{low}} - \mathcal{A}_{\mathrm{up}}\|_\star.$$

Finally, by applying Lemma A.25 with $p = 1$ to $\mathbf{A}_{\mathrm{low}} - \mathbf{A}_{\mathrm{up}}$ and U, we have that $\|\mathbf{A}_{\mathrm{low}} - \mathbf{A}_{\mathrm{up}}\|_\star = \|\mathrm{U}^\star(\mathbf{A}_{\mathrm{low}} - \mathbf{A}_{\mathrm{up}})\mathrm{U}\|_\star = \|\mathcal{A}_{\mathrm{low}} - \mathcal{A}_{\mathrm{up}}\|_\star$, which concludes the proof. □

Proposition A.27. *Under the same hypothesis as Proposition A.23, if $(\mathbf{A}_{\mathrm{low}}, \mathbf{A}_{\mathrm{up}})$ are solutions of Equation* (39) *then*

$$\sup_{x \in \mathcal{X}} |\tilde{f}_{\mathbf{A}_{\mathrm{low}}}(x) - \tilde{f}_{\mathbf{A}_{\mathrm{up}}}(x)| \leq 2\, C_{\tilde{f}_{\mathbf{A}_{\mathrm{low}}}, \tilde{f}_{\mathbf{A}_{\mathrm{up}}}} \rho_{\widehat{X}, \mathcal{X}} + \sqrt{\sum_{i=1}^n (\tilde{f}_{\mathbf{A}_{\mathrm{low}}}(X_i) - \tilde{f}_{\mathbf{A}_{\mathrm{up}}}(X_i))^2}$$

where $\tilde{f}_{\mathbf{A}_{\mathrm{low}}}(X) = \mathbf{\Phi}(X)^\top \mathbf{A}_{\mathrm{low}} \mathbf{\Phi}(X) \in \mathcal{H}$ *and* $\tilde{f}_{\mathbf{A}_{\mathrm{up}}}(X) = \mathbf{\Phi}(X)^\top \mathbf{A}_{\mathrm{up}} \mathbf{\Phi}(X) \in \mathcal{H}$ *are the kernel sum-of-squares functions associated to* $\mathbf{A}_{\mathrm{low}}$ *and* $\mathbf{A}_{\mathrm{up}}$*, respectively, and with* $C_{\tilde{f}_{\mathbf{A}_{\mathrm{low}}}, \tilde{f}_{\mathbf{A}_{\mathrm{up}}}} = 2d\mathrm{DM}\|\mathbf{A}_{\mathrm{low}} - \mathbf{A}_{\mathrm{up}}\|_\star$.

Proof. Let $(\mathbf{A}_{\mathrm{low}}, \mathbf{A}_{\mathrm{up}})$ be solutions of Equation (39) and write $(\mathcal{A}_{\mathrm{low}}, \mathcal{A}_{\mathrm{up}})$ their associated operators. As in the proof of Proposition A.26, we have $f_{\mathcal{A}_{\mathrm{low}}}(X) = \tilde{f}_{\mathbf{A}_{\mathrm{low}}}(X)$ and $f_{\mathcal{A}_{\mathrm{up}}}(X) = \tilde{f}_{\mathbf{A}_{\mathrm{up}}}(X)$ for any $X \in \mathcal{X}$.

We now apply Proposition A.23 to $(\mathcal{A}_{\mathrm{low}}, \mathcal{A}_{\mathrm{up}})$ to get

$$\sup_{x \in \mathcal{X}} |f_{\mathcal{A}_{\mathrm{low}}}(x) - f_{\mathcal{A}_{\mathrm{up}}}(x)| \leq 2\, C_{f_{\mathcal{A}_{\mathrm{low}}}, f_{\mathcal{A}_{\mathrm{up}}}} \rho_{\widehat{X}, \mathcal{X}} + \sqrt{\sum_{i=1}^{n} (f_{\mathcal{A}_{\mathrm{low}}}(X_i) - f_{\mathcal{A}_{\mathrm{up}}}(X_i))^2}$$

where $C_{f_{\mathcal{A}_{\mathrm{low}}}, f_{\mathcal{A}_{\mathrm{up}}}} = 2d\mathrm{DM}\|\mathcal{A}_{\mathrm{low}} - \mathcal{A}_{\mathrm{up}}\|_\star$ and

$$\sup_{x \in \mathcal{X}} |\tilde{f}_{\mathbf{A}_{\mathrm{low}}}(x) - \tilde{f}_{\mathbf{A}_{\mathrm{up}}}(x)| \leq 2\, C_{f_{\mathcal{A}_{\mathrm{low}}}, f_{\mathcal{A}_{\mathrm{up}}}} \rho_{\widehat{X}, \mathcal{X}} + \sqrt{\sum_{i=1}^{n} (\tilde{f}_{\mathbf{A}_{\mathrm{low}}}(X_i) - \tilde{f}_{\mathbf{A}_{\mathrm{up}}}(X_i))^2}$$

Finally, by applying Lemma A.25 with $p = 1$ to $\mathbf{A}_{\mathrm{low}} - \mathbf{A}_{\mathrm{up}}$ and U, we have that $\|\mathbf{A}_{\mathrm{low}} - \mathbf{A}_{\mathrm{up}}\|_\star = \|\mathrm{U}^\star(\mathbf{A}_{\mathrm{low}} - \mathbf{A}_{\mathrm{up}})\mathrm{U}\|_\star = \|\mathcal{A}_{\mathrm{low}} - \mathcal{A}_{\mathrm{up}}\|_\star$, which gives $C_{f_{\mathcal{A}_{\mathrm{low}}}, f_{\mathcal{A}_{\mathrm{up}}}} = C_{\tilde{f}_{\mathbf{A}_{\mathrm{low}}}, \tilde{f}_{\mathbf{A}_{\mathrm{up}}}}$ and concludes the proof. □

A.5. Bounds on 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.

Definition A.28 (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 $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a kernel function and let $\mathcal{H}$ be the associated reproducing kernel Hilbert space. The maximum mean discrepancy between P_X and P_Y is defined as

$$\begin{aligned} \mathrm{MMD}(P_X, P_Y) &= \sup_{\|f\|_{\mathcal{H}} \leq 1} |\mathbb{E}_{X \sim P_X}[f(X)] - \mathbb{E}_{Y \sim P_Y}[f(Y)]| \\ &= \|\mu_{P_X} - \mu_{P_Y}\|_{\mathcal{H}} \end{aligned}$$

where $\mu_{P_X} = \int k(x, \cdot) dP_X(x)$ and $\mu_{P_Y} = \int k(x, \cdot) dP_Y(x)$ are the *kernel mean embeddings* of P_X and P_Y, respectively.

The squared MMD admits the following closed-form expression:

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

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 $k((u, v), (u', v')) = k_{\mathcal{X}}(u, u') k_{\mathcal{Y}}(v, v')$. A measure of the dependence between U and V can then be defined as the distance between the mean embedding $\mu_{P_{UV}}$ of P_{UV} and $\mu_{P_U} \otimes \mu_{P_V}$ of $P_U \otimes P_V$, the joint distribution with independent marginals P_U and P_V:

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

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

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

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. Let us now state our proposition again, before giving a detailed proof.

Proposition A.29. *Let $\widehat{C}_{\mathcal{D}_N}$ be the prediction intervals built from a score function $S(X,Y) = \max\big(l(X) - Y, Y - u(X)\big)$ through split CP with $\mathcal{D}_N = \mathcal{D}_n \cup \mathcal{D}_m$. Then for any ω_X in $\mathcal{F}_X$ such that $\mathbb{P}(X \in \omega_X) \geq \delta$, denoting $p_{\mathcal{D}_N} = \mathbb{P}(Y_{N+1} \in \widehat{C}_{\mathcal{D}_N}(X_{N+1}) | \mathcal{D}_N, X_{N+1} \in \omega_X)$ we have*

$$p_{\mathcal{D}_N} \geq 1 - \alpha - \frac{1}{\delta}\sqrt{1 - \alpha_1 \exp(\mathrm{MI}(\tilde{r}_{\mathcal{D}_n}(X,Y), W_{\mathcal{D}_n}(X)))} \tag{49}$$

and

$$p_{\mathcal{D}_N} \geq 1 - \alpha - \frac{1}{\delta}\sqrt{1 - \frac{\alpha_1}{1 - \alpha_2 \mathrm{HSIC}(\tilde{r}_{\mathcal{D}_n}(X,Y), W_{\mathcal{D}_n}(X))}}, \tag{50}$$

where $\tilde{r}_{\mathcal{D}_n}(X,Y) = |Y - (\widehat{u}_{\mathcal{D}_n}(X) + \widehat{l}_{\mathcal{D}_n}(X))/2|$ are the absolute centered residuals and $W_{\mathcal{D}_n}(X) = (\widehat{u}_{\mathcal{D}_n}(X) - \widehat{l}_{\mathcal{D}_n}(X))/2$ is the width of the prediction bands. α_1 is a constant and α_2 only depends on the kernel used for HSIC*, which must be characteristic.*

Proof. First, let us rewrite the score as a function of the width and the absolute centered residuals.

$$\begin{aligned}
S(X,Y) &= \max\big(l(X) - Y, Y - u(X)\big) \\
&= -\frac{u(X) - l(X)}{2} + \max\left(l(X) - Y + \frac{u(X) - l(X)}{2}, Y - u(X) + \frac{u(X) - l(X)}{2}\right) \\
&= -\frac{u(X) - l(X)}{2} + \max\left(\frac{u(X) + l(X)}{2} - Y, Y - \frac{u(X) + l(X)}{2}\right) \\
&= -\frac{u(X) - l(X)}{2} + \left|Y - \frac{u(X) + l(X)}{2}\right| \\
&= -\frac{W(X)}{2} + |Y - \widetilde{m}(X)| \\
&= -\frac{W(X)}{2} + R(X,Y) \\
&= f(W(X), R(X,Y)).
\end{aligned}$$

Then, placing ourselves in the context of split CP, we are working conditionally on $\mathcal{D}_N$, thus $\widehat{u}_{\mathcal{D}_n}(\cdot)$ and $\widehat{l}_{\mathcal{D}_n}(\cdot)$ are deterministic functions and to lighten notations we will write $W = W(X)$ and $R = R(X,Y)$. We now have

$$\begin{aligned}
\mathrm{MI}(X,S) &= \mathrm{MI}(X, f(W,R)) & \\
&\leq \mathrm{MI}(X,(R,W)) & \forall g,\ \mathrm{MI}(g(X),Y) \leq \mathrm{MI}(X,Y) \\
&\leq \mathrm{MI}((X,Y),(R,W)) & \mathrm{MI}((X_1,X_2),Y) \geq \mathrm{MI}(X_1,Y) \\
&= H(R,W) - H(R,W|X,Y) & \mathrm{MI}(X,Y) = H(X) - H(X|Y) \\
&= H(R,W) & H(Y|X) = 0 \Leftrightarrow \exists g,\ Y = g(X) \\
&= H(R) + H(W) - \mathrm{MI}(W,R) & \mathrm{MI}(X,Y) = H(X) + H(Y) - H(X,Y) \\
&\leq H(X,Y) + H(X) - \mathrm{MI}(W,R) & \forall g,\ H(g(X)) \leq H(X)
\end{aligned}$$

and we can observe that only $\mathrm{MI}(W,R)$ depends on $\widehat{u}_{\mathcal{D}_n}(\cdot)$ and $\widehat{l}_{\mathcal{D}_n}(\cdot)$. We thus deduce that

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

where $\alpha_1 = \exp(-H(X,Y) - H(X))$ is independent from $\widehat{u}_{\mathcal{D}_n}(\cdot)$ and $\widehat{l}_{\mathcal{D}_n}(\cdot)$. Deutschmann et al. (2024) showed that

$$p_{\mathcal{D}_N} \geq 1 - \alpha - \frac{1}{\delta}\sqrt{1 - \exp(-\mathrm{MI}(X,S))}$$

and we obtain Equation (49) by using the previous bound.

For the second part of the proposition, from Wang & Tay (2023, Equation 15), we have the bound

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

where $\mathrm{TV}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|$ for $\mathbb{P}$, $\mathbb{Q}$ defined on a measurable space $(\Omega, \mathcal{F})$ and $\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q})$ 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 $\mathbb{P} = P_{WR}$ the joint distribution of (W, R) and $\mathbb{Q} = P_W \otimes P_R$ the joint distribution with independent marginals P_W and P_R, to get

$$1 - \exp(-\mathrm{MI}(W, R)) \geq \mathrm{TV}^2(P_{WR}, P_W \otimes P_R) \geq \alpha_2 \mathrm{HSIC}(W, R),$$

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

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

and Equation (50) follows.

□

B. Additional experiments and details

B.1. Cross-validation for kernel hyperparameter estimation and Kruskal-Wallis rank test

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 $\hat{f}_{\text{low},-k}, \hat{f}_{\text{up},-k}$ the lower and upper bands trained on $\mathcal{D}_{-k}$. Define two sets,

$$R_K = \bigcup_{k=1}^{K} \left\{ \left| Y_i - \widehat{m}_n(X_i) - \frac{\hat{f}_{\text{up},-k}(X_i) - \hat{f}_{\text{low},-k}(X_i)}{2} \right| \right\}_{i \in \mathcal{D}_k} \quad \text{and} \quad W_K = \bigcup_{k=1}^{K} \{\hat{f}_{\text{up},-k}(X_i) + \hat{f}_{\text{low},-k}(X_i)\}_{i \in \mathcal{D}_k}.$$

We seek

$$\max_{(\theta_{\text{low}}, \theta_{\text{up}}) \in \mathbb{R}^d \times \mathbb{R}^d} \quad \widehat{\text{HSIC}}\,(W, R)\,, \tag{51}$$

where $\widehat{\text{HSIC}}(W, R)$ is estimated with samples W_K and R_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).

Kruskal-Wallis rank test. To assess whether HSIC varies significantly across different λ_{pen}, we treat each value of λ_{pen} as a "group" and the HSIC values computed from each cross-validation replicate as repeated observations. However, because cross-validation replicates are not independent and identically distributed across groups (they overlap, share data, and violate the independent-samples assumption of classical parametric tests), we use a permutation-based nonparametric framework rather than standard ANOVA or rank-based tests with asymptotic null distributions. Specifically, we first compute the classical Kruskal–Wallis test statistic H_{obs} on the pooled data (after rank-transforming all HSIC values), but derive its null distribution via permuting the λ_{pen} labels across all replicates. We generate $B = 2000$ random label permutations and compute the permuted H-statistics, producing an empirical null distribution. The p-value is the fraction of permutations whose permuted H exceeds or equals the observed H_{obs}. This yields a valid significance test under the null hypothesis that the distributions of HSIC are identical across λ_{pen}, without requiring independence of replicates or parametric assumptions.

B.2. Implementation details

Optimization. For the primal problems with the SDP formulation, we use the SCS algorithm (O'Donoghue, 2021; O'Donoghue et al., 2023) available in the convex optimization software `CVXPY` (Diamond & Boyd, 2016; Agrawal et al., 2018), with a maximum number of iterations equal to 10000. An example of script that implements the SDP problems defined by Equations (27), (31) and (39) is shown below. Here, `matrix_variable_low` and `matrix_variable_upp` denote the unknowns $\mathbf{A}_{\text{low}} \in \mathbb{S}_+^n$ and $\mathbf{A}_{\text{up}} \in \mathbb{S}_+^n$. The remaining variables, such as the feature vectors `phi_matrix_low` ($\mathbf{\Phi}_{\text{low}}(\mathbf{X})$) and `phi_matrix_upp` ($\mathbf{\Phi}_{\text{up}}(\mathbf{X})$) have to be computed by the user using the training set and the chosen kernels.

```
import numpy as np
import cvxpy as cp
import scs

penalty = "Asymmetric"/"Operator"/"Training set"

# Variables
matrix_variable_low = cp.Variable((n, n), symmetric=True)
matrix_variable_upp = cp.Variable((n, n), symmetric=True)

# Expressions
f_low = cp.reshape(
    cp.diag(phi_matrix_low.T @ matrix_variable_low @ phi_matrix_low),
    (-1, 1),
    order="C",
)
f_upp = cp.reshape(
    cp.diag(phi_matrix_upp.T @ matrix_variable_upp @ phi_matrix_upp),
    (-1, 1),
    order="C",
)

# Objective
```

```
match penalty:
    case "Asymmetric":
        objective = cp.Minimize(
            + (b/n) * cp.sum(f_low+f_upp)
            + lambda1_low * cp.trace(matrix_variable_low) + lambda2_low * cp.square(cp.
                norm(matrix_variable_low, "fro"))
            + lambda1_upp * cp.trace(matrix_variable_upp) + lambda2_upp * cp.square(cp.
                norm(matrix_variable_upp, "fro"))
        )
    case "Training set":
        objective = cp.Minimize(
            + (b/n) * cp.sum(f_low+f_upp)
            + lambda1_low * cp.trace(matrix_variable_low) + lambda2_low * cp.square(cp.
                norm(matrix_variable_low, "fro"))
            + lambda1_upp * cp.trace(matrix_variable_upp) + lambda2_upp * cp.square(cp.
                norm(matrix_variable_upp, "fro"))
            + lambda1_pen * cp.sum(cp.square(f_low-f_upp))
        )
    case "Operator":
        objective = cp.Minimize(
            + (b/n) * cp.sum(f_low+f_upp)
            + lambda1_low * cp.trace(matrix_variable_low) + lambda2_low * cp.square(cp.
                norm(matrix_variable_low, "fro"))
            + lambda1_upp * cp.trace(matrix_variable_upp) + lambda2_upp * cp.square(cp.
                norm(matrix_variable_upp, "fro"))
            + lambda1_pen * cp.norm((matrix_variable_low-matrix_variable_upp), "nuc")
            + lambda2_pen * cp.square(cp.norm((matrix_variable_low-matrix_variable_upp), "
                fro"))
        )

# Positive semi-definite constraints
constraints = [matrix_variable_low >> 0]
constraints += [matrix_variable_upp >> 0]

# 100% coverage constraints
constraints += [y_train <= mean_predictions + f_upp]
constraints += [y_train >= mean_predictions - f_low]

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

# Solve
prob.solve(solver=cp.SCS)

# Retrieve final matrices
Ahat_low = matrix_variable_low.value
Ahat_upp = matrix_variable_upp.value
```

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

For dual problems, unlike Allain et al. (2026) who developed a projected gradient method with Nesterov acceleration, we prefer to use the L-BFGS algorithm (Liu & Nocedal, 1989) for its robustness and faster convergence, since we only have at most a few thousands optimization variables in our experiments. We use `SciPy`'s minimize algorithm (Virtanen et al., 2020), with a maximum number of iterations equal to 10000, tolerance equal to 10^{-2} and initialize all Lagrange multipliers to 0 (except when using the warm-start strategy detailed below).

Hyperparameter and training pipeline. In Algorithm 1 we give details about the steps to follow in order to train the penalized kSoS model for both penalties with the Kruskal-Wallis rank test and the HSIC independence test. The conformal step is performed after the final stage when $(\widehat{\mathbf{A}}_{\text{low}}, \widehat{\mathbf{A}}_{\text{up}})$ are computed.

Algorithm 1 Hyperparameter selection and training for the penalized kSoS model

Require: Training set $\mathcal{D}_n$; mean model $\widehat{m}_n$; candidate sets Θ (pairs $(\theta_{\text{low}}, \theta_{\text{up}})$) and Λ (values of λ_{pen}); number of CV folds K; bootstrap HSIC repetitions B; validation HSIC test repetitions R; p-value thresholds $\alpha_{\text{KW}}, \alpha_{\text{HSIC}}$; default lengthscales θ_{hom} for homoscedastic noise (typically 20) and default penalization λ_{sym} for symmetric noise (typically 100).

Ensure: Estimator $(\widehat{\mathbf{A}}_{\text{low}}, \widehat{\mathbf{A}}_{\text{up}})$.

// Stage 1 — Cross-validation grid search
// Can be coupled with warm-start
Partition $\mathcal{D}_n$ into K folds $\{\mathcal{D}_k\}_{k=1}^K$
for each $(\theta, \lambda) \in \Theta \times \Lambda$ **do**
  for $k = 1, \ldots, K$ **do**
    Fit Pen. kSoS on $\mathcal{D}_{-k}$ with (θ, λ) and predict $(\widehat{f}_{\text{low}}, \widehat{f}_{\text{up}})$ on $\mathcal{D}_k$
  end for
  Aggregate out-of-fold residuals R_K and widths W_K
  Compute observed $\text{HSIC}(\theta, \lambda)$ and bootstrap replicates $\{\text{HSIC}^{(b)}(\theta, \lambda)\}_{b=1}^B$
end for

// Stage 2 — Best $\theta^\star$ per λ
for each $\lambda \in \Lambda$ **do**
  $\theta^\star(\lambda) \leftarrow \arg\max_{\theta \in \Theta} \text{HSIC}(\theta, \lambda)$
  $\widetilde{\text{HSIC}}^{(b)}(\lambda) \leftarrow \max_{\theta \in \Theta} \text{HSIC}^{(b)}(\theta, \lambda)$, for each $b = 1, \ldots, B$
end for

// Stage 3 — Kruskal–Wallis rank test with bootstrap
$p_{\text{KW}} \leftarrow \text{TEST}_{\text{KW}}\big(\{\widetilde{\text{HSIC}}^{(b)}(\lambda)\}_{b \in [B], \lambda \in \Lambda}\big)$
if $p_{\text{KW}} > \alpha_{\text{KW}}$ **then**
  $\lambda^\star \leftarrow \lambda_{\text{sym}}$ ▷ *default to symmetric model*
else
  $\lambda^\star \leftarrow \arg\max_{\lambda \in \Lambda} \text{HSIC}(\theta^\star(\lambda), \lambda)$
end if
$\theta^\star \leftarrow \theta^\star(\lambda^\star)$

// Stage 4 — HSIC test of independence
for $j = 1, \ldots, R$ **do**
  Partition $\mathcal{D}_n$ into K folds and run CV with $(\theta^\star, \lambda^\star)$ to obtain $R_K^{(j)}$ and $W_K^{(j)}$
  $p_j \leftarrow \text{TEST}_{\text{HSIC}}\left(R_K^{(j)}, W_K^{(j)}\right)$
end for
if $\text{median}(p_1, \ldots, p_R) \geq \alpha_{\text{HSIC}}$ **then**
  $\theta^\star \leftarrow \theta_{\text{hom}},\ \lambda^\star \leftarrow \lambda_{\text{sym}}$ ▷ *default to homoscedastic model*
end if

// Stage 5 — Final fit
$(\widehat{\mathbf{A}}_{\text{low}}, \widehat{\mathbf{A}}_{\text{up}}) \leftarrow$ Pen. kSoS trained on full $\mathcal{D}_n$ with $(\theta^\star, \lambda^\star)$
return $(\widehat{\mathbf{A}}_{\text{low}}, \widehat{\mathbf{A}}_{\text{up}})$

Scaling with respect to n. To illustrate the advantage of our dual formulation over the primal one, in Figure 7 we compare the computation time for both of them when the number of samples n increases. The primal formulation can only handle up to $n = 200$ samples, while the dual solver easily scales to $n = 1000$.

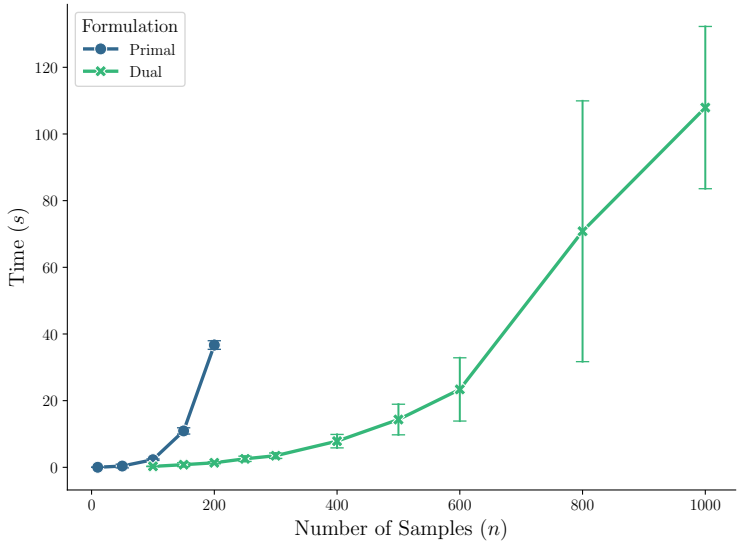


Figure 7. Dataset 1: computing time for solving SDP and dual formulation as a function of n (penalty 1, $b = 0$, $\theta_{\text{low}} = \theta_{\text{up}} = 0.3$, $\lambda_{\text{pen}} = 1$, max iter = 10^4), mean±sd over 20 repetitions.

The simulations were performed on an AMD Ryzen 7 9700X 8-Core Processor (3.80 GHz) with four max threads.

Interaction between b and $(\theta_{\text{low}}, \theta_{\text{up}})$. In Allain et al. (2026), it was observed that the penalty intensity b on the one hand, and the kernel lengthscales θ (for the symmetric kSoS model) on the other hand, compensate each other through an interaction. Indeed, both influence the optimal kSoS function complexity: for a given (b, θ), similar bands with equivalent adaptivity can be obtained for higher values of both hyperparameters, as long as b is sufficiently large (typically 10 or larger).

In order to corroborate this assertion for our new penalized kSoS, we focus on one of our analytical test cases and compute the HSIC criterion for a grid of $(\lambda_{\text{pen}}, \theta)$ values (we fix $\theta_{\text{low}} = \theta_{\text{up}} = \theta$) and four different values of $b = 1,\ 10,\ 100,\ 1000$. Figure 8 displays these contour plots for three different random seeds (one row corresponds to one seed).

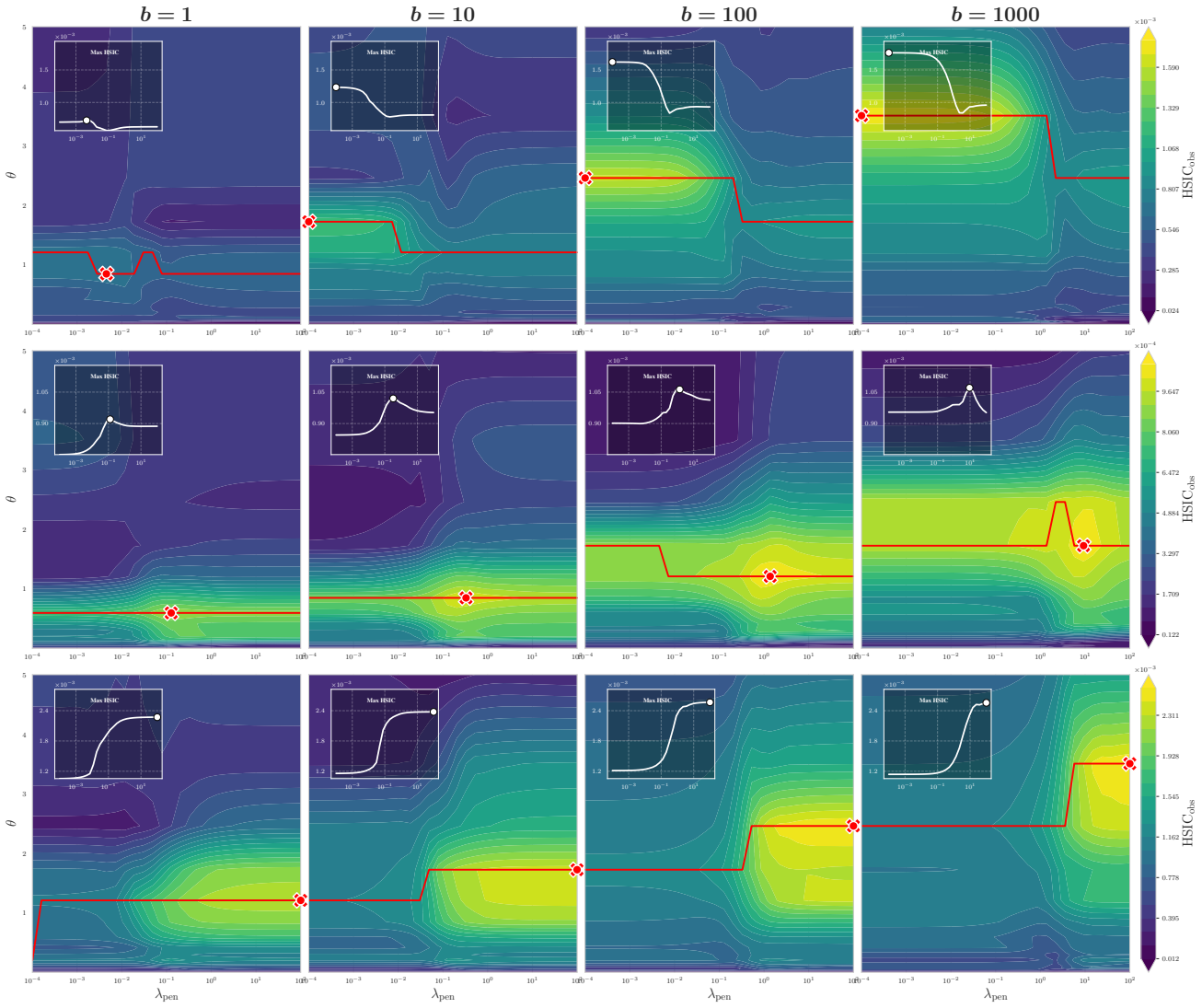


Figure 8. Contour plots of the HSIC criterion for a grid of $(\lambda_{\text{pen}}, \theta)$ and $b = 1,\ 10,\ 100,\ 1000$, case 2, $n = 100$.

For all seeds, we recover the expected behavior: when b increases, the optimal HSIC is reached for larger values of θ, and we observe that the value of the optimal HSIC is attained as soon as $b = 10$ or $b = 100$ in average. Interestingly, we can deepen the analysis by having a closer look at the differences between the seeds. At the top, we see that our model systematically chooses the smallest value for λ_{pen}, thus indicating that asymmetry is strongly favored for this seed, and only θ increases to reach the optimal HSIC. Similarly, we notice the exact same phenomenon, although mirrored, in the bottom contour plots. This time, a strong symmetry is identified for the random samples, with only large selected values of λ_{pen}. In between, the seed points towards including a small amount of symmetry for better adaptivity, where intermediate penalty values are preferred. But the optimal λ_{pen} increases with b and θ: to preserve the same amount of beneficial symmetry, the model necessitates a stronger constraint to compensate for the effect of increasing both b and θ.

Warm-start optimization strategy. To search for the best $(\theta_{\text{low}}, \theta_{\text{up}}, \lambda_{\text{pen}})$ that maximizes HSIC, a brute-force and naive approach consists in solving the dual problem separately for all hyperparameter values to test. However, for fixed kernel lengthscales $(\theta_{\text{low}}, \theta_{\text{up}})$, it is reasonable and intuitive to expect that the solution of the dual will only slightly vary between two close values of λ_{pen}. This means that once a penalized kSoS is trained for a given value of λ_{pen}, the corresponding optimal Lagrange multipliers can be used as initialization for training a subsequent penalized kSoS with a new value of λ_{pen}. If both are close, we anticipate a drastic reduction in the number of iterations required for convergence.

In order to validate this intuition, we consider one of our analytical test case and record the number of iterations at convergence for two strategies: a) the brute-force approach where all dual problems are solved separately and are initialized at the same value (the *cold-start approach*) and b) an iterative approach where we first solve the dual for $\lambda_{\text{pen}} = 10^{-4}$ and gradually increase its value while using the previous optimum as the initial point for the next problem (the *warm-start approach*). Results are reported in Figure 9.

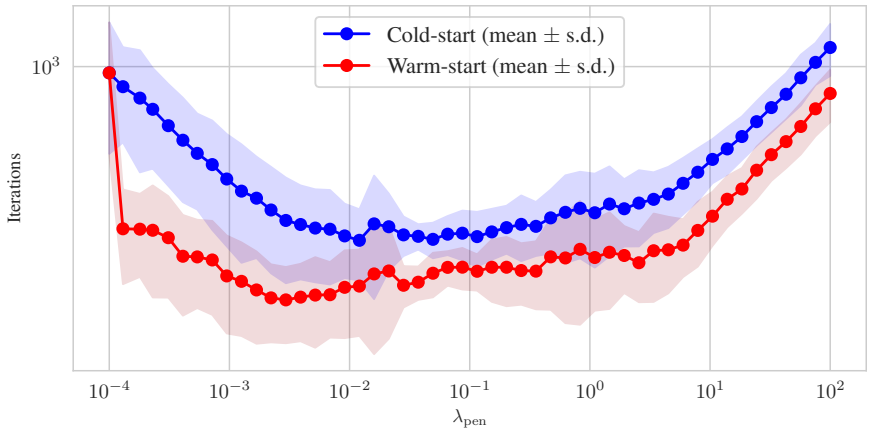


Figure 9. Dataset 1. Number of iterations at convergence for the cold-start (blue) and warm-start (red) approach with the dual formulation and penalty on the training set, $n = 100$, $b = 10$ (mean±sd on 10 repetitions).

We observe first that for both approaches, the number of iterations can substantially vary with the value of λ_{pen}: this means that some dual problems are harder to solve than others, typically those associated to very low or very high values of λ_{pen}. But crucially, we also clearly see that the warm-start approach yields considerable computational savings, with a reduction factor that can reach 3 or 4. In average, warm-start is 65% cheaper than cold-start: this means that in practice, we can investigate 10 different values of λ_{pen} for maximizing HSIC at the cost of 3 single kSoS trainings only.

B.3. Additional numerical experiments

B.3.1. Illustration of asymmetry and adaptivity of noise distribution

The most commonly pursued feature in prediction bands is adaptivity, also known as conditional coverage, and is a needed property when the noise exhibits heteroscedastic behavior, as showcased in Figure 10, bottom row, where the noise is a function of the input X. On the other hand, both homoscedastic (top row) and heteroscedastic noise (bottom row) can exhibit either symmetric (left column) or asymmetric noise (right column). **Adaptivity** refers to whether the noise distribution is independent of X or not, while **asymmetry** refers to whether the noise distribution is symmetric around the mean or not.

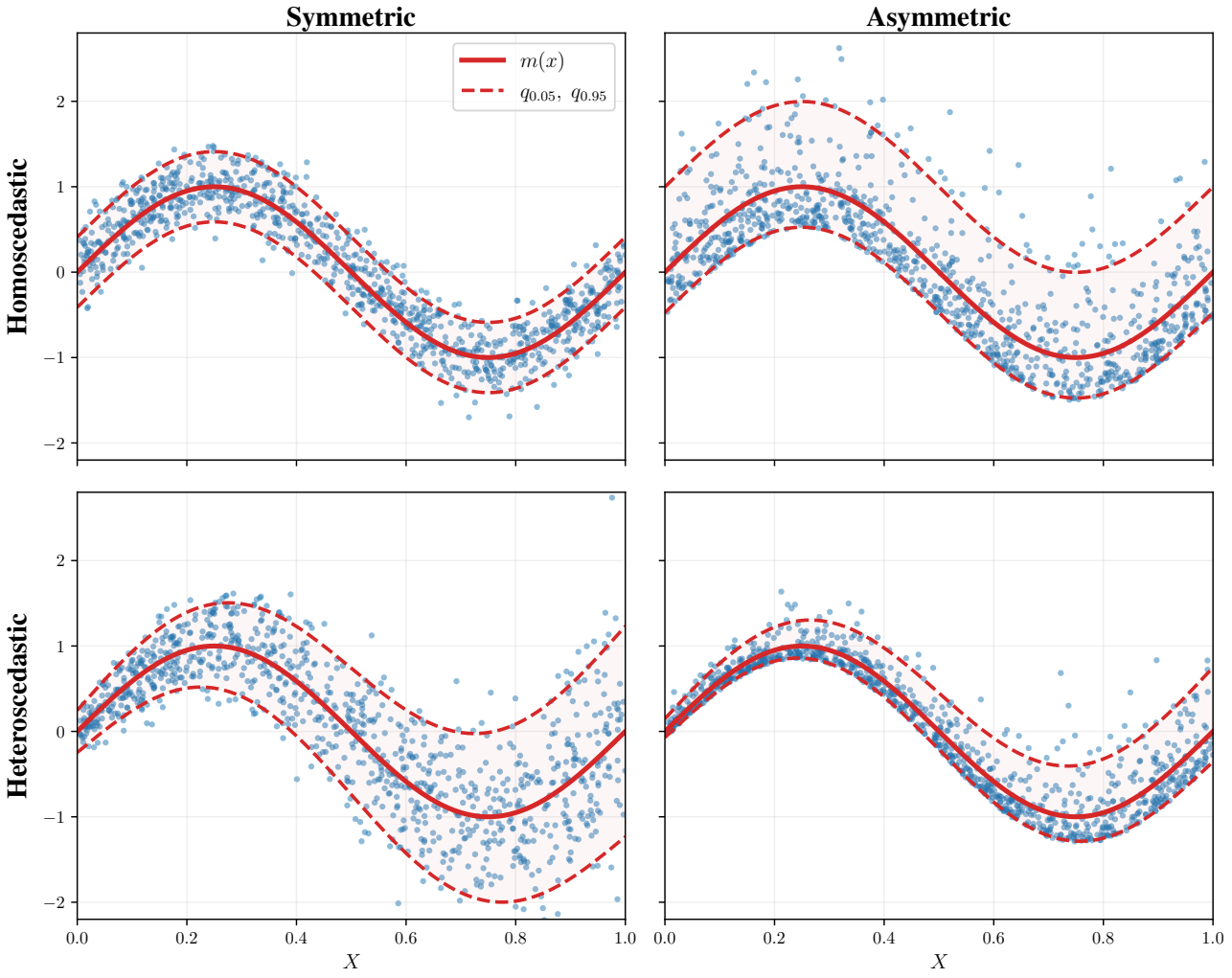


Figure 10. Noise taxonomy between two properties: adaptivity (homoscedastic versus heteroscedastic, rows) and asymmetry (columns).

In our numerical experiments on synthetic datasets, we thus consider several of these variants: datasets 1–5 are heteroscedastic while dataset 6 is homoscedastic, and asymmetric noise appears in datasets 3–5. Dataset 4 is special because it has homoscedastic lower noise and heteroscedastic upper noise, resulting in asymmetric noise.

B.3.2. Discussion on evaluation metrics

As is usually done in the literature to compare interval adaptivity, we compute the mean width of the prediction intervals on the test set. Unfortunately, when considered alone, this performance metric is of limited relevance. Indeed it is very common to have intervals with similar mean width but with highly different local coverage quality (we illustrate this phenomenon below). As such, to better measure adaptivity, we also consider additional metrics.

For our experiments on analytical test cases, we consider the local coverage, obtained by approximating $\mathbb{P}(Y_{N+1} \in \widehat{C}_{\mathcal{D}_N}(X_{N+1}) \mid X_{N+1} = x)$ by its empirical counterpart with samples from Y (of size n_Y) at different random locations X_i (of size n_X), denoted $\hat{p}(X_i)$. We then compute a global measure of local coverage quality, the *absolute coverage gap*, by considering the average distance to the target level $1 - \alpha$:

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

For real-world datasets, since such conditional samples are not available, we rely instead on 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 $\mathbb{P}(Y_{N+1} \in \widehat{C}_{\mathcal{D}_N}(X_{N+1}) | X_{N+1} \in \mathcal{R}_l)$. The worst-set coverage WSC is defined as the minimum of all these coverages: the closer it is to the target $(1 - \alpha)$, 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 identity the 100-th closest neighbors in the feature space to estimate the marginal coverage.

These two measures indicate how well prediction intervals capture the noise distribution and are powerful in practice to compare CP procedures. However, they only focus on the central part of the intervals: for asymmetric noise distributions, they may fail to detect if a prediction interval is better than another one. Instead, we can consider their asymmetric variants, where we evaluate local coverage below or above the prediction interval, i.e. $p^{\mathrm{low}}(X_{N+1}) = \mathbb{P}(Y_{N+1} > l(X_{N+1}) \mid X_{N+1} = x)$ and $p^{\mathrm{upp}}(X_{N+1}) = \mathbb{P}(Y_{N+1} < u(X_{N+1}) \mid X_{N+1} = x)$ for an interval $\widehat{C}_{\mathcal{D}_N}(X_{N+1}) = [l(X_{N+1}), u(X_{N+1})]$, see Linusson et al. (2014) and Romano et al. (2019). Ideally, we would like $p^{\mathrm{low}}(X_{N+1})$ and $p^{\mathrm{upp}}(X_{N+1})$ to be greater than $1 - \alpha/2$ for all X_{N+1}, which would mean the prediction interval also captures well the lower and upper tails of the distribution. If this is the case, traditional local coverage also holds since

$$\begin{aligned}
\mathbb{P}(Y_{N+1} \notin \widehat{C}_{\mathcal{D}_N}(X_{N+1}) \mid X_{N+1} = x) &= \mathbb{P}(Y_{N+1} < l(X_{N+1}) \cup Y_{N+1} > u(X_{N+1} \mid X_{N+1} = x) \\
&\leq \mathbb{P}(Y_{N+1} < l(X_{N+1}) \mid X_{N+1} = x) + \mathbb{P}(Y_{N+1} > u(X_{N+1}) \mid X_{N+1} = x) \\
&= (1 - p^{\mathrm{low}}(X_{N+1})) + (1 - p^{\mathrm{upp}}(X_{N+1})) \\
&\leq \alpha/2 + \alpha/2 = \alpha.
\end{aligned}$$

Following this idea, we can define the lower and upper absolute coverage gaps

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

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

Similarly, we consider the lower and upper worst-set coverage $\mathrm{WSC}^{\mathrm{low}}$ and $\mathrm{WSC}^{\mathrm{upp}}$. When aggregating such lower and upper adaptivity measures over several experimental replications, we will concatenate lower and upper indicators to draw boxplots or compute averages: we will thus refer to them as "combined", with notations ACG^c and WSC^c.

Now, we illustrate numerically two important facts which motivate the evaluation metrics discussed so far:

1. Mean width alone is not sufficient to measure adaptivity, since two intervals can have similar mean width but one can be locally adaptive while the other one is not.
2. For asymmetric noise distribution, local coverage is not sufficient because an interval can have equivalent or better local coverage than another one, while failing at capturing lower and upper tails.

We focus on the analytical test case 4 (described later on), and compare the intervals produced by heteroscedastic GP and our penalized kSoS (we consider both methods after split CP calibration which guarantees marginal coverage). For two replications, Figure 11 shows the prediction intervals from the two procedures as well as their mean width and ACG variants. Visually, it is clear that heteroscedastic GP is not adaptive, since it does not capture the noise distribution with an interval composed of "holes" on the right. On the contrary, kSoS learns the shape of the noise and is much more adaptive. But when

looking at the metrics, we observe that heteroscedastic GP has equivalent or better mean width and ACG than kSoS, which clearly shows that they are not sufficient to properly evaluate adaptivity. On the other side, both $\mathrm{ACG}^{\mathrm{low}}$ and $\mathrm{ACG}^{\mathrm{upp}}$ indicate that kSoS certainly outperforms heteroscedastic GP.

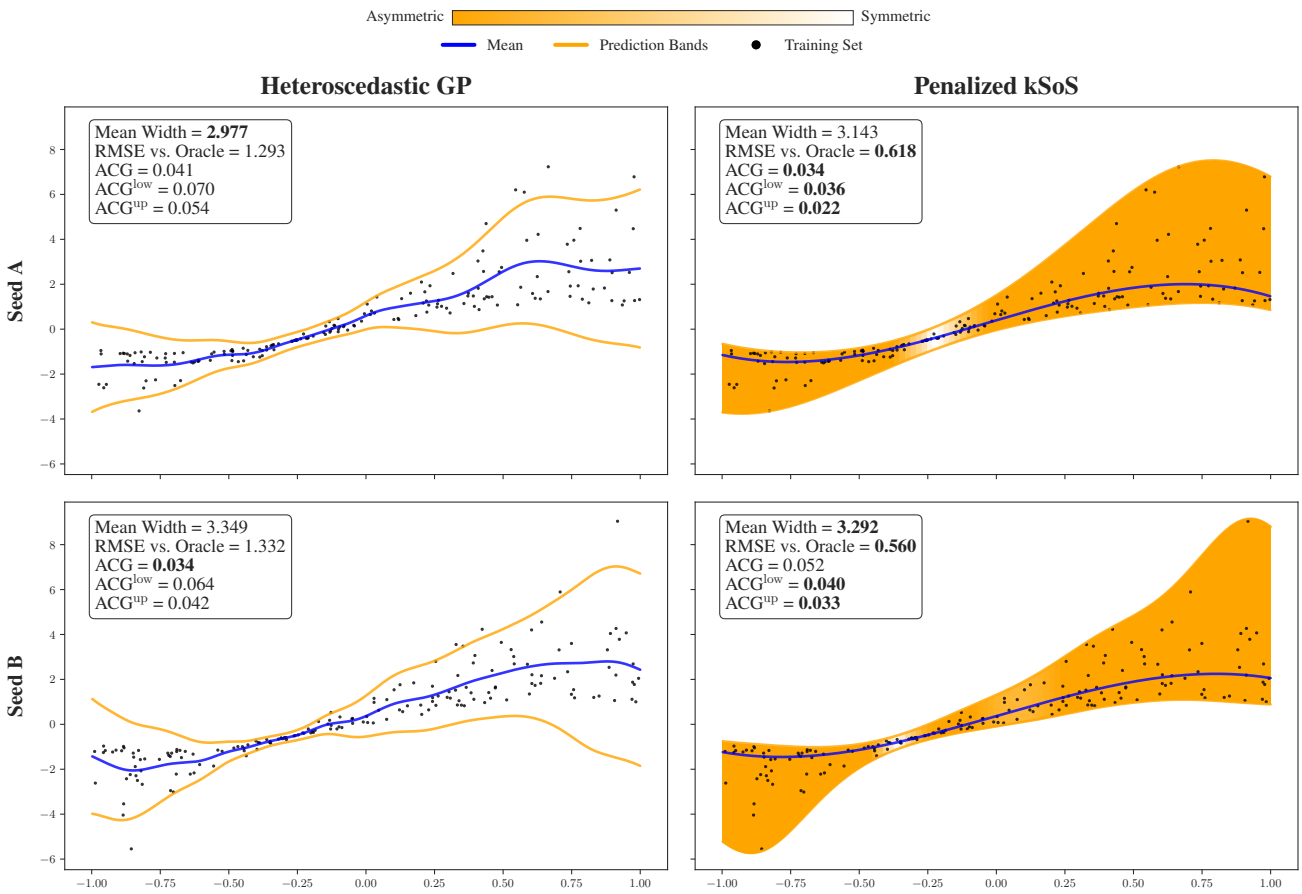
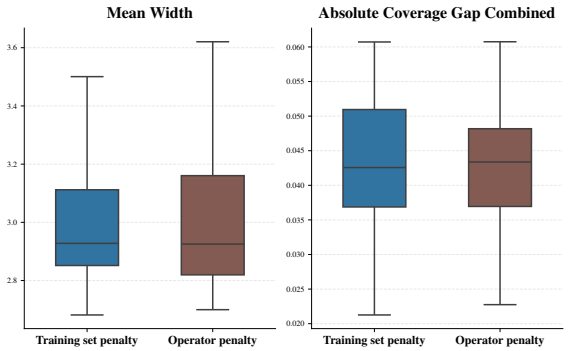


Figure 11. Prediction intervals for heteroscedastic GP and penalized kSoS for two random seeds on test case 5.

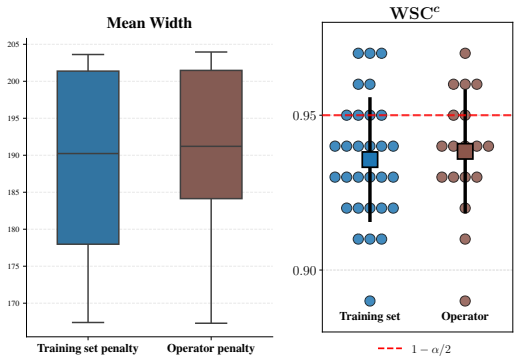
All in one, we advocate the comparison of ACG^c or WSC^c first (since in practice we do not know beforehand if the noise distribution is symmetric or not, it may be misleading to use ACG or WSC), and for methods that have similar lower and upper adaptivity, to compare as a second step their mean width to potentially break the ties. This is similar in spirit to the assertion "maximize the sharpness of the predictive distributions subject to calibration" of Gneiting & Raftery (2007), which was later emphasized again in Chung et al. (2021).

B.3.3. Operator penalty versus training set penalty

Both penalties achieve similar goals but differ in implementation and theoretical properties. The operator penalty provides a continuous functional view of the problem, and consequently inherits tighter bound independent of the fill-in distance. But in terms of computational complexity, the operator penalty involves $\mathcal{O}(n^2)$ dual variables, as opposed to $\mathcal{O}(n)$ for the training set one, and is also less flexible since the exact same kernels (and lengthscales) must be used for both bands. In this particular case, we illustrate in Figure 12 that the operator penalty does not yield improvement in either mean width or coverage over the training set penalty for two test cases, but we observe this phenomenon for all the datasets we investigated.



(a) Mean width and absolute coverage gap combined for both penalties on dataset 1.



(b) Mean width and worst-set coverage combined for both penalties on dataset Diabetes.

Figure 12. Comparison of operator and training set penalties on an analytical test case (left) and a real-world dataset (right).

We thus recommend to use the operator penalty for small datasets where strongest theoretical control is desired and computation is not limiting, but use the training set one in all other instances.

B.3.4. Difference with symmetric kSoS

When the predictive model $\widehat{m}_n(\cdot)$ is fixed, the symmetric kSoS problem introduced in Allain et al. (2026) writes

$$\begin{aligned} \inf_{\mathcal{A}\in\mathcal{S}_+(\mathcal{H})} \quad & \frac{b}{n}\sum_{i=1}^{n} f_{\mathcal{A}}(X_i) + \Omega(\mathcal{A}) \\ \text{s.t.} \quad & \left(Y_i - \widehat{m}_n(X_i)\right)^2 \leq f_{\mathcal{A}}(X_i),\ i \in [n] \end{aligned}$$

while our penalty operator formulation is

$$\begin{aligned} \inf_{\substack{\mathcal{A}_{\text{low}}\in\mathcal{S}_+(\mathcal{H}) \\ \mathcal{A}_{\text{up}}\in\mathcal{S}_+(\mathcal{H})}} \quad & \frac{b}{n}\sum_{i=1}^{n} \left(f_{\mathcal{A}_{\text{low}}}(X_i) + f_{\mathcal{A}_{\text{up}}}(X_i)\right) + \Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}}) + \Omega_{\text{pen}}(\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}}) \\ \text{s.t.} \quad & \widehat{m}_n(X_i) - Y_i - f_{\mathcal{A}_{\text{low}}}(X_i) \leq 0,\ i \in [n] \\ & Y_i - \widehat{m}_n(X_i) - f_{\mathcal{A}_{\text{up}}} \leq 0,\ i \in [n] \end{aligned}$$

where $\Omega_{\text{pen}}(\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}}) = \lambda_{\text{pen1}}\|\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}}\|_\star + \lambda_{\text{pen2}}\|\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}}\|_F^2$.
When $\lambda_{\text{pen1}}, \lambda_{\text{pen2}} \to +\infty$, $\Omega_{\text{pen}}(\mathcal{A}_{\text{low}} - \mathcal{A}_{\text{up}}) \to \iota_{\mathcal{A}_{\text{low}}=\mathcal{A}_{\text{up}}}$ with ι_C the indicator function of the set C. The penalized formulation thus tends to

$$\begin{aligned} \inf_{\substack{\mathcal{A}_{\text{low}}\in\mathcal{S}_+(\mathcal{H}) \\ \mathcal{A}_{\text{up}}\in\mathcal{S}_+(\mathcal{H}) \\ \mathcal{A}_{\text{low}}=\mathcal{A}_{\text{up}}}} \quad & \frac{b}{n}\sum_{i=1}^{n} \left(f_{\mathcal{A}_{\text{low}}}(X_i) + f_{\mathcal{A}_{\text{up}}}(X_i)\right) + \Omega_{\text{low}}(\mathcal{A}_{\text{low}}) + \Omega_{\text{up}}(\mathcal{A}_{\text{up}}) \\ \text{s.t.} \quad & \widehat{m}_n(X_i) - Y_i - f_{\mathcal{A}_{\text{low}}}(X_i) \leq 0,\ i \in [n] \\ & Y_i - \widehat{m}_n(X_i) - f_{\mathcal{A}_{\text{up}}}(X_i) \leq 0,\ i \in [n]\,. \end{aligned}$$

Now, since by definition $\mathcal{A}_{\text{low}} = \mathcal{A}_{\text{up}}$ implies $\forall X,\ f_{\mathcal{A}_{\text{low}}}(X) = f_{\mathcal{A}_{\text{up}}}(X)$, the last problem is equivalent to

$$\begin{aligned} \inf_{\mathcal{A}\in\mathcal{S}_+(\mathcal{H})} \quad & \frac{b}{n}\sum_{i=1}^{n}\left(f_{\mathcal{A}}(X_i) + f_{\mathcal{A}}(X_i)\right) + \Omega_{\text{low}}(\mathcal{A}) + \Omega_{\text{up}}(\mathcal{A}) \\ \text{s.t.} \quad & \widehat{m}_n(X_i) - Y_i - f_{\mathcal{A}}(X_i) \leq 0,\ i \in [n] \\ & Y_i - \widehat{m}_n(X_i) - f_{\mathcal{A}}(X_i) \leq 0,\ i \in [n]\,. \end{aligned}$$

If we also take $\lambda_{\text{low}1} = \lambda_{\text{up}1}$ and $\lambda_{\text{low}2} = \lambda_{\text{up}2}$ such that $\Omega_{\text{low}}(\mathcal{A}) = \Omega_{\text{up}}(\mathcal{A}) = \Omega(\mathcal{A})$, this boils down to

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

with a score function which becomes $S(X,Y) = \max\left(\widehat{m}_n(X) - Y - f_{\mathcal{A}}(X), Y - \widehat{m}_n(X) - f_{\mathcal{A}}(X)\right) = |Y - \widehat{m}_n(X)| - f_{\mathcal{A}}(X)$. At first sight, this closely resembles the symmetric kSoS of Allain et al. (2026), where the constraints write $\left(Y_i - \widehat{m}_n(X_i)\right)^2 \leq f_{\mathcal{A}}(X_i) \Leftrightarrow |Y_i - \widehat{m}_n(X_i)| \leq \sqrt{f_{\mathcal{A}}(X_i)}$. From this perspective, $f_{\mathcal{A}}(X)$ in our work plays the role of a standard deviation function, while in Allain et al. (2026) it relates to a variance function. However, there are important differences:

- To get a strict equivalence of the problems, we should use $\sum_{i=1}^{n}\sqrt{f_{\mathcal{A}}(X_i)}$ instead of $\sum_{i=1}^{n} f_{\mathcal{A}}(X_i)$. This means that the optimal solutions are not the same
- More importantly, the score functions are fundamentally different and the final intervals differ:

$$\begin{aligned} C^{\text{Allain}}(X) &= \widehat{m}_n(X) \pm q_\alpha \left(\frac{|Y_j - \widehat{m}_n(X_j)|}{\sqrt{f_{\mathcal{A}}(X_j)}}\right)\sqrt{f_{\mathcal{A}}(X)} \\ C^{\text{ours}}(X) &= \widehat{m}_n(X) \pm \left(q_\alpha\left(|Y_j - \widehat{m}_n(X_j)| - f_{\mathcal{A}}(X_j)\right) + f_{\mathcal{A}}(X)\right) \end{aligned}$$

Ultimately, the method by Allain et al. (2026) is not a special case of our penalty formulation, because of both the objective function and the CP score functions. However, we insist that when the penalty is large, our problem still produces symmetric prediction bands, although different from Allain et al. (2026).

B.3.5. SYMMETRIC VERSUS ASYMMETRIC CALIBRATION

As discussed in Romano et al. (2019), score functions of the form $S(X,Y) = \max\big(l(X) - Y, Y - u(X)\big)$, as in CQR or penalized kSoS, can easily be used in an asymmetric calibration procedure. Denoting $\widehat{q}_{\alpha_{\text{low}}}$ and $\widehat{q}_{\alpha_{\text{upp}}}$ the adjusted quantiles of the sets $\{l(X_i) - Y_i,\ i \in \mathcal{D}_m\}$ and $\{Y_i - l(X_i),\ i \in \mathcal{D}_m\}$, respectively, then

$$\widehat{C}_{\mathcal{D}_N}(X_{N+1}) = \big[l(X_{N+1}) - \widehat{q}_{\alpha_{\text{low}}}, u(X_{N+1}) + \widehat{q}_{\alpha_{\text{upp}}}\big]$$

has $1-\alpha$ marginal coverage as long as $\alpha_{\text{low}} + \alpha_{\text{upp}} = \alpha$. But Romano et al. (2019) also mentions that the stronger coverage guarantee (marginal coverage below and above the interval) comes at the cost of slightly longer intervals. We also investigate this behavior on twelve real-world datasets (detailed later on). For both CQR and penalized kSoS, we examine how often asymmetric calibration produces longer intervals, and how close to local coverage intervals after either calibration are, when measured with WSC^{c}. Results are given in Table 3, averaged over 10 repetitions.

Table 3. Percentage of times asymmetric calibration yields larger intervals than symmetric calibration, and mean distance between worst-set coverage (lower and upper) and target level $1-\alpha/2$, over 10 repetitions and for both CQR and penalized kSoS.

Dataset	$\text{MW}_{\text{asym}} > \text{MW}_{\text{sym}}$	$\lvert\text{WSC}^{\text{c}}_{\text{asym}} - (1-\alpha/2)\rvert$	$\lvert\text{WSC}^{\text{c}}_{\text{sym}} - (1-\alpha/2)\rvert$
All datasets	69.17%	0.0322	0.0315
Concrete	65%	0.01700	0.01425
Bike	75%	0.05400	0.05775
Diabetes	70%	0.02100	0.02350
Housing log	65%	0.02975	0.02725
Housing	55%	0.02900	0.02675
MPG	95%	0.02375	0.02375
Boston	60%	0.01700	0.01025
Energy	85%	0.02500	0.02375
Miami	40%	0.04500	0.04225
Sulfur	75%	0.04175	0.04425
Power	75%	0.06600	0.06525
Yacht	70%	0.01725	0.01950

Table 4. Mean width relative increase for intervals with asymmetric calibration vs intervals with symmetric calibration, averaged over 10 repetitions.

Dataset	$\frac{\text{MW}_{\text{asym}} - \text{MW}_{\text{sym}}}{\text{MW}_{\text{sym}}}$
All datasets	1.48%
Concrete CQR	1.47%
Bike CQR	1.20%
Bike kSoS	1.30%
Boston CQR	2.96%
Diabetes CQR	2.02%
Energy kSoS	3.06%
Housing kSoS	1.45%
MPG CQR	8.32%
MPG kSoS	5.68%
Yacht CQR	2.46%
Yacht kSoS	3.52%

We observe first that both calibration methods have equivalent lower and upper worst-set coverage, meaning that asymmetric calibration does not improve local coverage in these examples. However, except for Miami, asymmetric calibration predominantly produces longer intervals, in 69.17% of the cases in average. To go further, we study in Table 4 the relative increase of mean width induced by asymmetric calibration on some of these datasets. We observe that in average the increase is limited, but it can be very large in specific instances. Since asymmetric calibration was observed to come with no benefits on lower and upper local coverage in these experiments, we only consider symmetric calibration in all the following experiments.

B.3.6. Additional analytical test cases and results

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}. Mean width is estimated with a test set of size $n_{\text{test}} = 1000$. We consider CQR with random forests and both homoscedastic and heteroscedastic GPs with Matérn $5/2$ kernel. As for kSoS, we also use a Matérn $5/2$ kernel with a predictive model given by a homoscedastic GP for symmetric cases and a cubic spline for asymmetric ones. We evaluate the symmetric variant of Allain et al. (2026) and our penalized version with the penalty on the training set only, since we did not observe differences with the operator penalty in our experiments. For both kSoS methods, we train first an initial model with $\theta_{\text{low}} = \theta_{\text{up}}$ equal to the median of the feature distances (a usual rule-of-thumb for kernel methods) and extract the values of the mean-width and the norms, which serve as a normalization before setting the hyperparameter values $\lambda_{(\cdot)1} = \lambda_{(\cdot)2} = 1$ and b (depending on the test case). For all methods, we use a calibration set of size $m = 2000$, and also compute the root mean-squared error (RMSE) with respect to the oracle prediction bands, as well as plot the density of local coverage.

Case 1. Inspired from Gramacy & Lee (2009).

$$
\begin{aligned}
&d = 1, \quad X \sim \mathcal{U}[-1, 1], \quad Y = m(X) + \sigma(X)\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \\
&m(X) = \begin{cases} \sin(\pi(2X + 1/5)) + 0.2\cos(4\pi(2X + 1/5)) & \text{if } 10X + 1 \leq 9.6 \\ X - 9/10 & \text{otherwise} \end{cases} \\
&\sigma(X) = \sqrt{0.1 + 2X^2}
\end{aligned}
$$

We display the adaptivity metrics in Figure 13. CQR and homoscedastic GP tend to produce intervals that overcover and are too large. Heteroscedastic GP, symmetric and penalized kSoS yield better ACG^c (with a slight advantage for penalized kSoS), but both kSoS models have smaller mean width among all competitors. This is confirmed by the density plot of local coverage, where heteroscedastic GP, symmetric and penalized kSoS have local coverage density closer to the target $1 - \alpha/2$ than homoscedastic GP and CQR. This illustrates that the automatic choice of the penalty allows to recover the best performing results of symmetric kSoS, which is also confirmed with both of them reaching the smallest RMSE.

Figure 14 gives the optimal solution of our dual formulation for $n = 1000$.

Case 2. Corresponds to setting 1 in Hore & Barber (2024).

$$
\begin{aligned}
&X \sim \mathcal{N}_d(0, I_d), \quad Y = m(X) + \sigma(X)\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \\
&m(X) = 0.5 \sum_{i=1}^{d} X^{(i)} \\
&\sigma(X) = \sum_{i=1}^{d} |\sin(X^{(i)})|
\end{aligned}
$$

Similarly to the previous case, penalized kSoS and heteroscedastic GP have better ACG^c than other competitors and equivalent to symmetric kSoS, as can be seen in Figure 15. This is again confirmed by the local coverage densities: they show that homoscedastic GP tends to undercover and overcover more than other methods, which can be explained by the constant width of the produced bands. But penalized kSoS has much smaller mean width and RMSE, although it does not reach the mean width or RMSE of symmetric kSoS.

With $n = 1000$, we obtain in Figure 16 the following optimal solution of the dual formulation.

Case 3. Inspired from Braun et al. (2025), with a lognormal noise:

$$
\begin{aligned}
&d = 1, \quad X \sim \mathcal{U}[-1, 1], \quad Y = m(X) + \sigma(X)\epsilon, \quad \epsilon \sim \text{Lognormal}(0, 1) \\
&m(X) = \sin(5X) \\
&\sigma(X) = X
\end{aligned}
$$

Figure 17 demonstrates once again that on such asymmetric case, penalized kSoS produces intervals with much better ACG^c and smaller mean width than all competitors, except for homoscedastic GP in terms of mean width. But as already discussed,

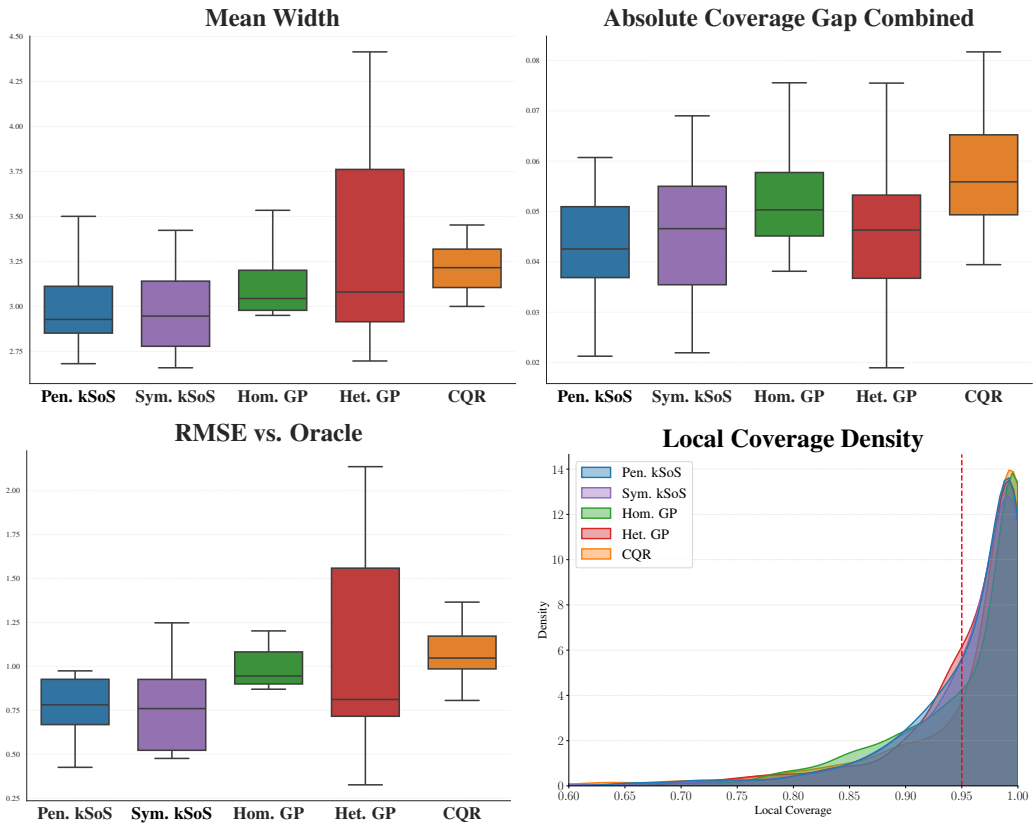


Figure 13. Test case 1 with $d = 1$ and $n = 100$. Mean width, Absolute Coverage Gap combined, RMSE versus oracle and density of local coverage lower and upper combined, $b = 10$.

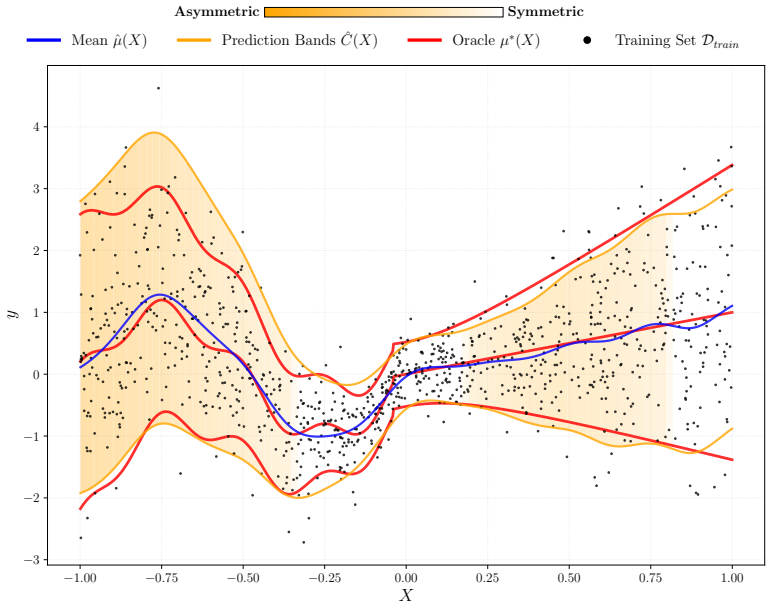


Figure 14. Test case 1 with $d = 1$ and $n = 1000$. Optimal solution of dual formulation with penalty 1.

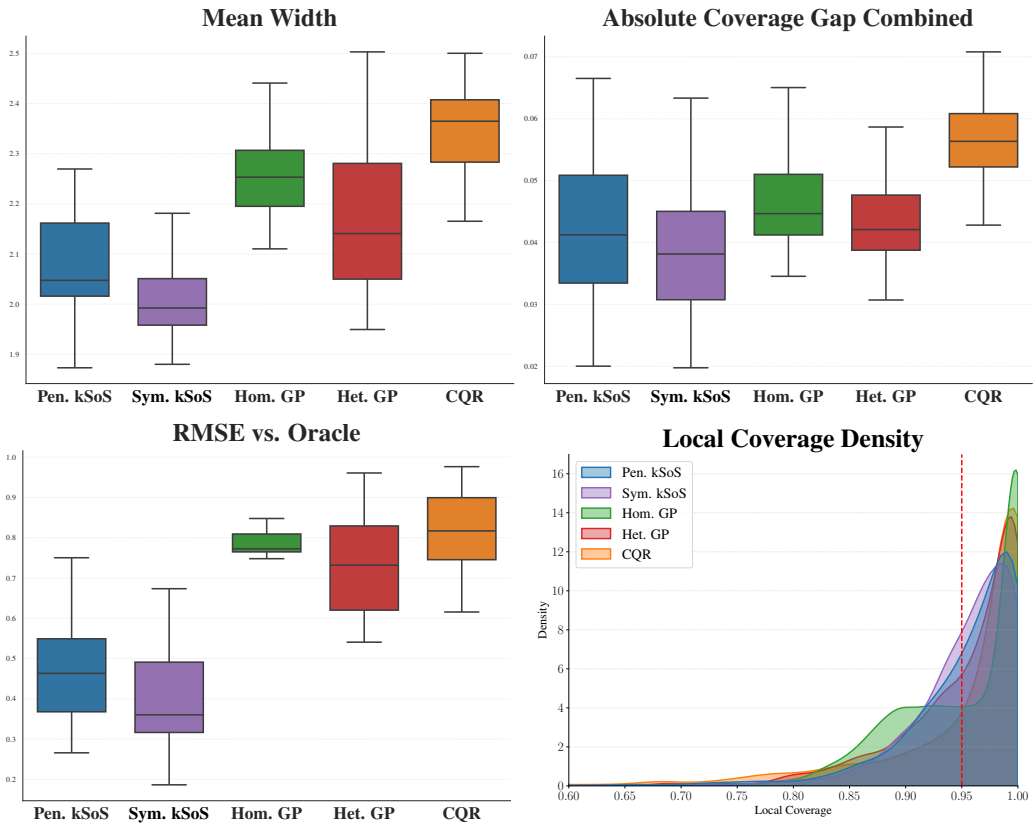


Figure 15. Test case 2 with $d = 1$ and $n = 100$. Mean width, Absolute Coverage Gap combined, RMSE versus oracle and density of local coverage lower and upper combined, $b = 100$.

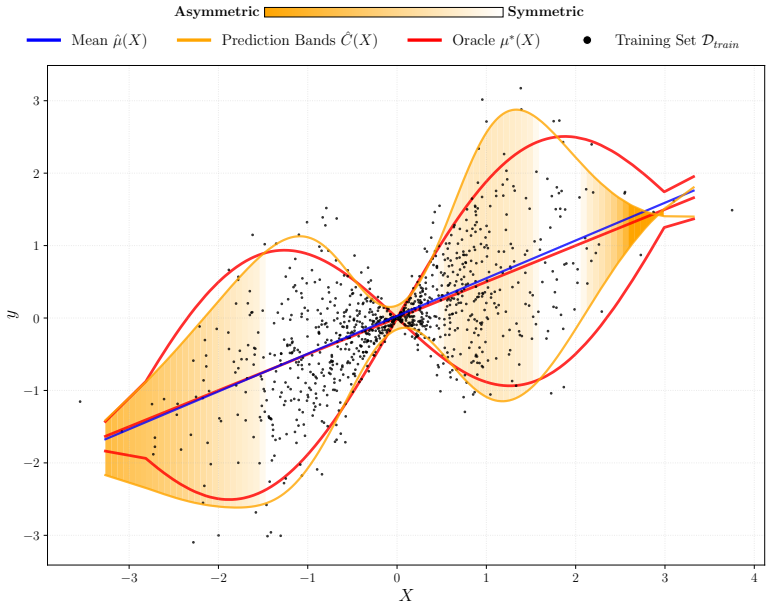


Figure 16. Test case 2 with $d = 1$ and $n = 1000$. Optimal solution of dual formulation with penalty 1.

using this indicator only can be misleading. As in the previous test case, penalized kSoS also approximates the oracle prediction intervals with a much higher accuracy, with a very small RMSE. The local coverage density plot clearly shows that penalized kSoS has the best local coverage. It also showcases a similar behavior for all symmetric methods (symmetric kSoS, homoscedastic and heteroscedastic GP) with a tendency to overcover in some input regions and undercover in others.

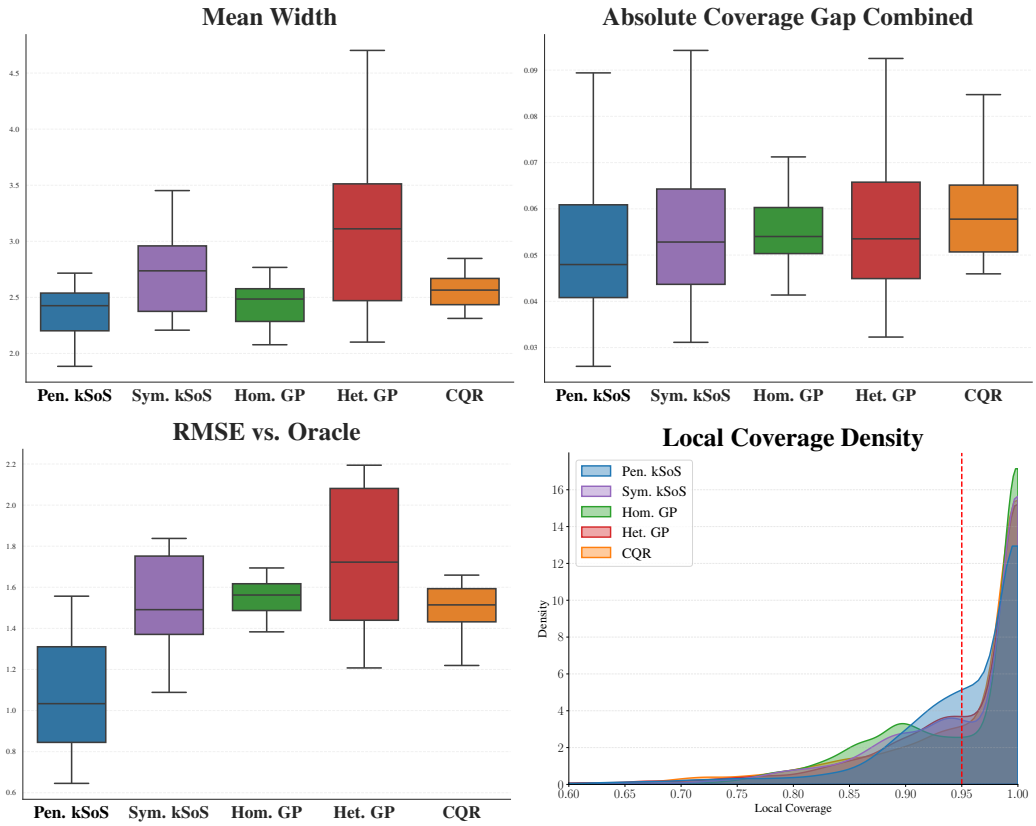


Figure 17. Test case 3 with $d = 1$ and $n = 100$. Mean width, Absolute Coverage Gap combined, RMSE versus oracle and density of local coverage lower and upper combined, $b = 10$.

In Figure 18 we display the optimal solution of the dual formulation for $n = 1000$.

Case 4. Asymmetric test case with split-normal noise.

$$
\begin{aligned}
&d = 1, \quad X \sim \mathcal{U}[0, 4\pi], \quad Y = m(X) + \epsilon\Big[\sigma_-(X)\,\mathbf{1}_{\{\epsilon<0\}} + \sigma_+(X)\,\mathbf{1}_{\{\epsilon\geq 0\}}\Big], \quad \epsilon \sim \mathcal{N}(0,1)\\
&m(X) = \sin(X)\\
&\sigma_-(X) = 0.2\\
&\sigma_+(X) = 0.4(\sin(X) + 1) + 0.1
\end{aligned}
$$

Case 4 involves asymmetric bands with a heteroscedastic upper band and a homoscedastic lower one. The flexibility of our framework, which allows to use different kernels for the upper and lower bands, is particularly appropriate here. Figure 19 demonstrates that penalized kSoS is much more adaptive than any other method, when measured with the ACG^c or when looking at the densities of local coverage. Penalized kSoS also produces much narrower prediction bands and achieves much higher accuracy, with a very small RMSE.

We give in Figure 20 the optimal solution of the dual formulation obtained with $n = 1000$.

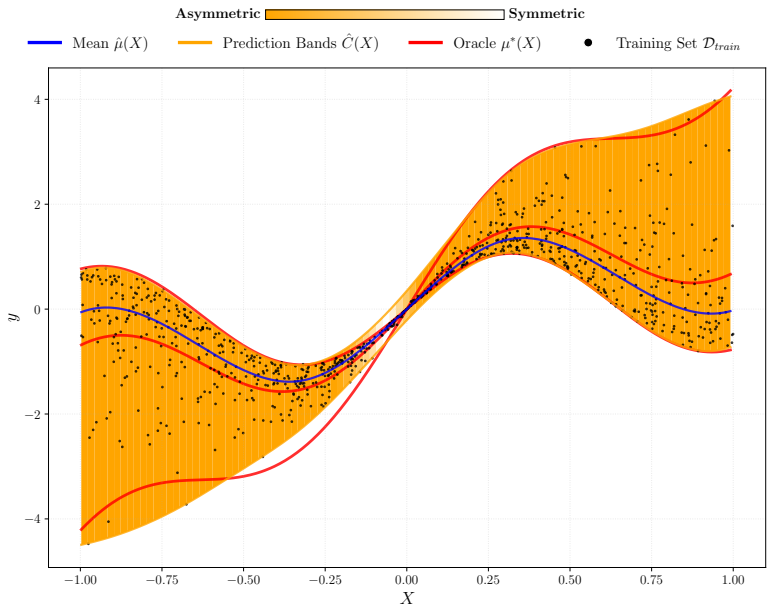


Figure 18. Test case 3 with $d = 1$ and $n = 1000$. Optimal solution of dual formulation with penalty 1.

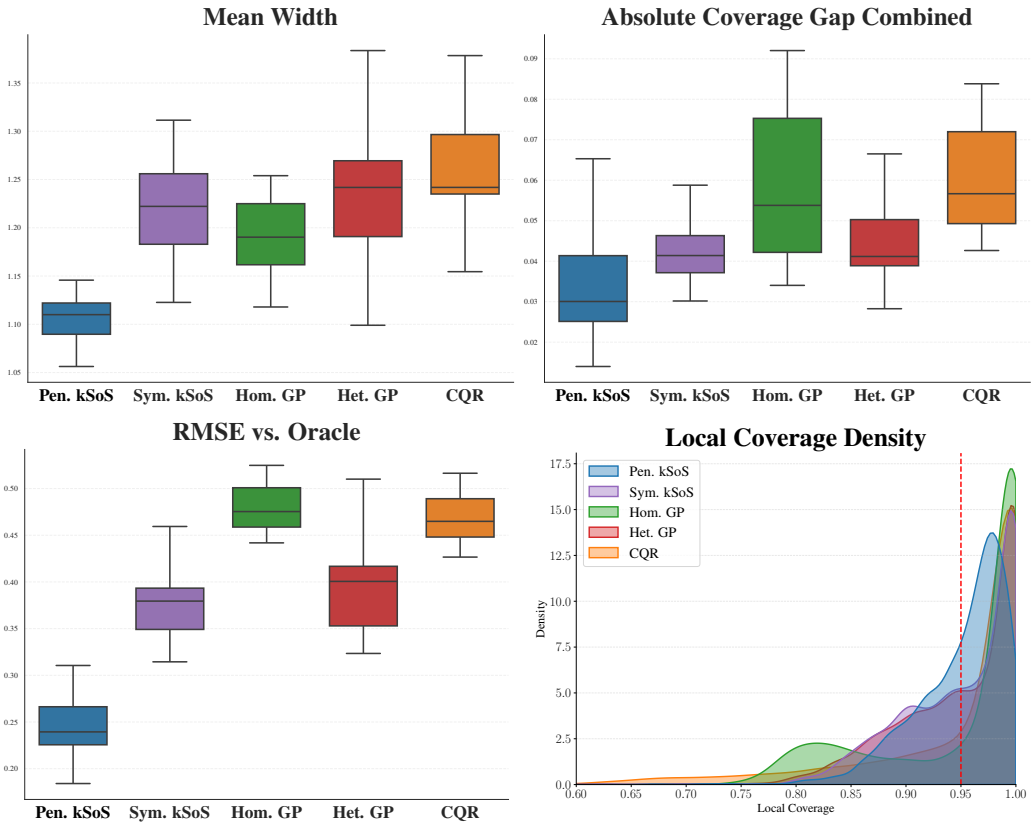


Figure 19. Test case 4 with $d = 1$ and $n = 500$. Mean width, Absolute Coverage Gap combined, RMSE versus oracle and density of local coverage lower and upper combined, $b = 10$.

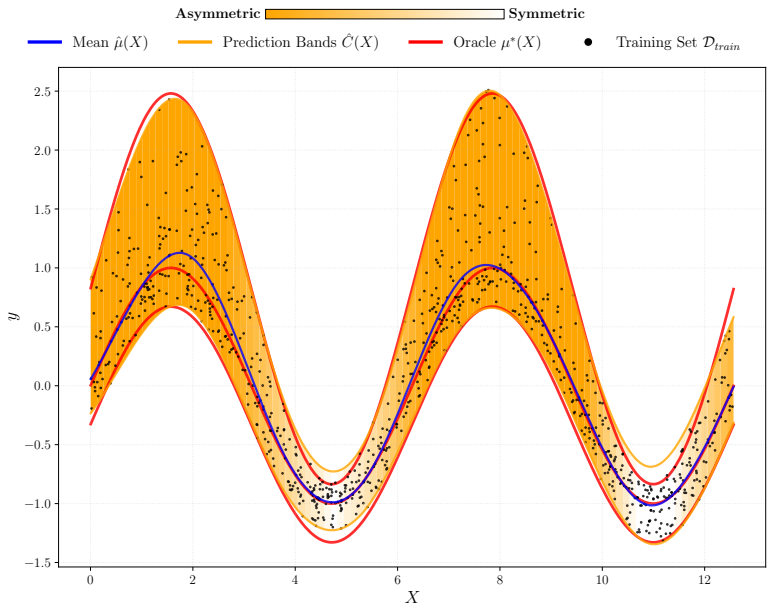


Figure 20. Test case 4 with $d = 1$ and $n = 1000$. Optimal solution of dual formulation with penalty 1.

Case 5. Test case from Braun et al. (2025), which involves an exponentially distributed noise:

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

For this asymmetric dataset, Figure 21 shows that penalized kSoS has much better local coverage (as measured by the ACG^c and in the density of local coverage plot) than all competitors, where all symmetric procedures have poor adaptivity. In addition, penalized kSoS also yields the smallest mean width, with a much smaller RMSE with respect to oracle prediction bands.

The optimal solution of the dual formulation obtained for $n = 1000$ is given in Figure 22.

Case 6. Inspired from Kivaranovic et al. (2020).

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

The oracle prediction bands are close to be constant for this test case: we expect the automatic HSIC independence test to activate. In dimension $d = 1$ we set $\beta = 1$ and obtain the remaining adaptivity metrics given in Figure 23.

As expected, homoscedastic GP, which produces almost constant intervals, performs the best in this setting. Symmetric kSoS activates the HSIC test of independence on most random seeds, and thus yields similar performance. Interestingly, our penalized kSoS almost always selects a symmetric model, thus allowing to get metrics close to the best ones.

Figure 24 shows the optimal solution of the dual formulation obtained for $n = 1000$.

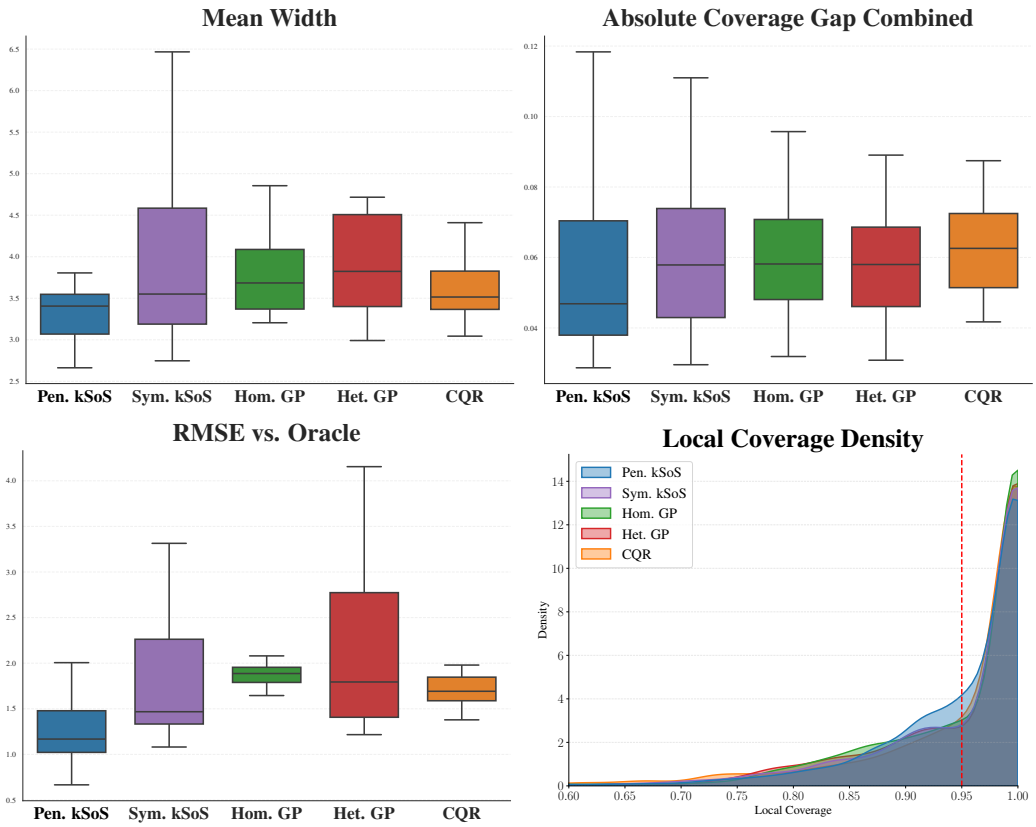


Figure 21. Test case 5 with $d = 1$ and $n = 100$. Mean width, Absolute Coverage Gap combined, RMSE versus oracle and density of local coverage lower and upper combined, $b = 10$.

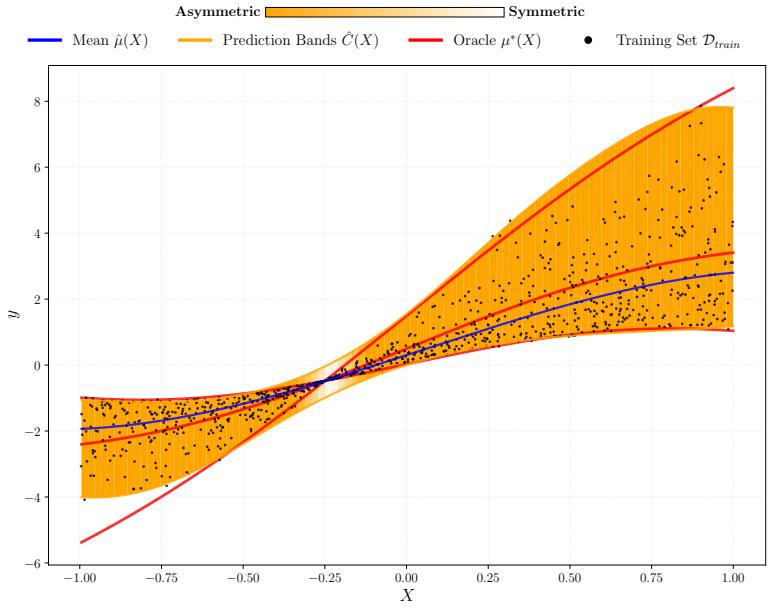


Figure 22. Test case 5 with $d = 1$ and $n = 1000$. Optimal solution of dual formulation with penalty 1.

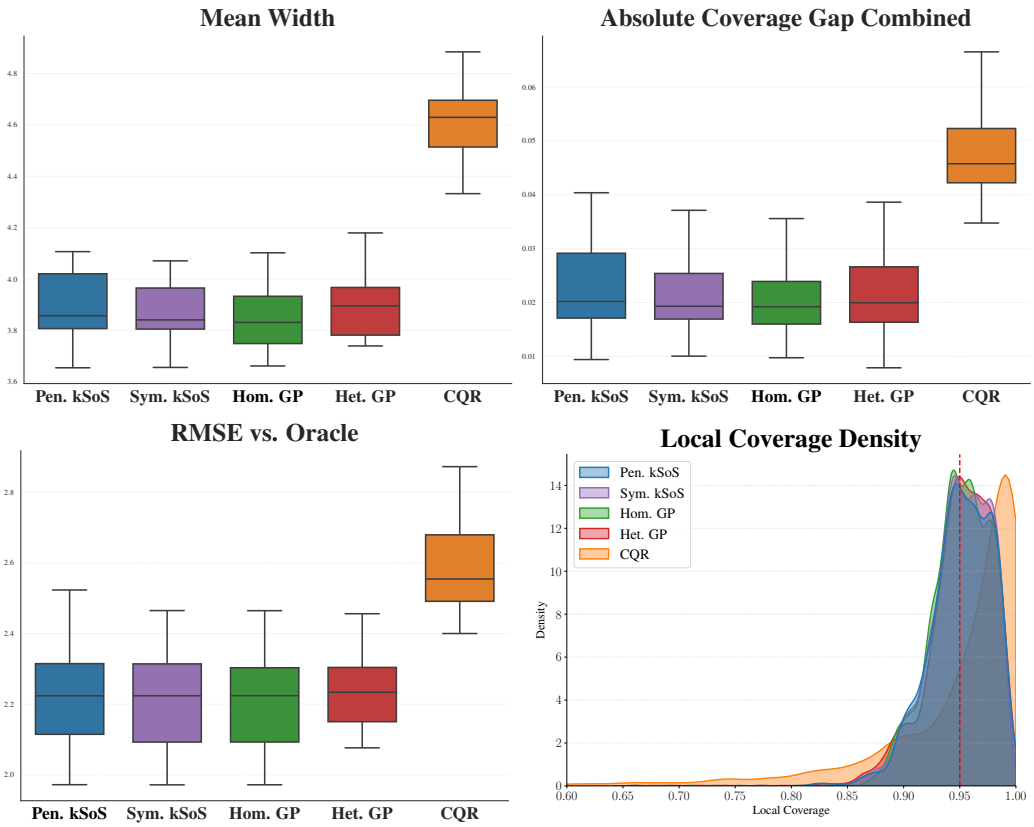


Figure 23. Test case 6 with $d = 1$ and $n = 100$. Mean width, Absolute Coverage Gap combined, RMSE versus oracle and density of local coverage lower and upper combined, $b = 0$.

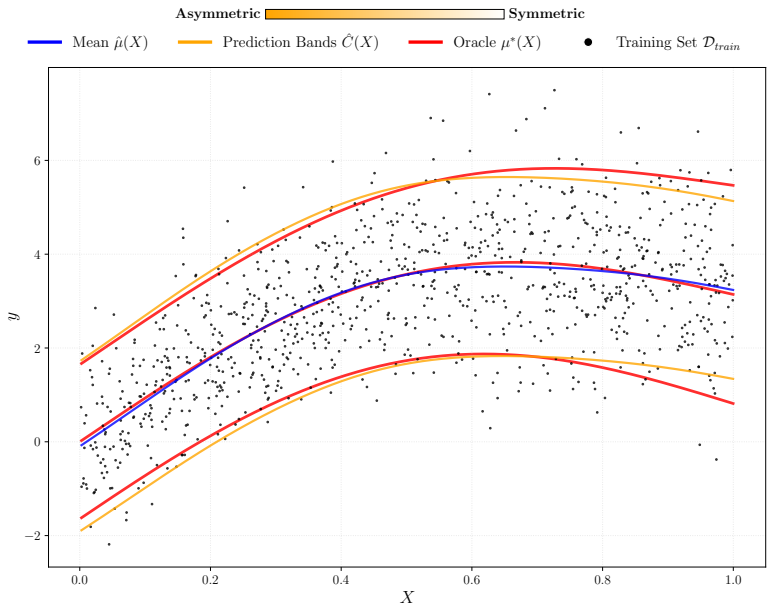


Figure 24. Test case 6 with $d = 1$ and $n = 1000$. Optimal solution of dual formulation with penalty 1.

B.3.7. ADDITIONAL REAL-WORLD DATASETS AND RESULTS

The 13 real-world datasets we consider are the following:

1. Concrete (Yeh, 1998) and Bike (Fanaee-T, 2013) are taken from Romano et al. (2019).
2. Diabetes (Efron et al., 2004), Housing (Pace & Barry, 1997), MPG (Quinlan, 1993), Boston (Harrison Jr & Rubinfeld, 1978), Energy (Tsanas & Xifara, 2012), Miami (Mayer et al., 2022), Sulfur (Fortuna et al., 2007), Power (Tfekci & Kaya, 2014), Yacht (Gerritsma et al., 1981) and Kin8nm (Ghahramani, 1996) are standard regression datasets.

Table 5 provides a detailed description of each of them in terms of number of features and total sample size. Note that for some of them, we perform a preliminary preprocessing step, by removing outliers (with a homoscedastic GP model) and removing inactive features (following previous literature where they were investigated for CP or regression tasks). For Housing specifically, we remove censored data (target ≥ 5) and consider two variants with or without logarithm transformation of the target. We also apply such transformation for Yacht.

Table 5. Description of thirteen 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.

Dataset	Nb features	Total sample size	n_{train}	n_{cal}	n_{test}	Outliers
Concrete	8	1030	412	412	206	30%
Bike	13 (18)	10886	1000	1000	1000	20%
Diabetes	10	442	101	170	171	NA
Housing log	8	20640	1000	1000	1000	NA
Housing	8	20640	1000	1000	1000	NA
MPG	7	398	100	146	146	NA
Boston	10 (13)	490	100	200	190	NA
Energy	6 (8)	768	300	234	234	NA
Miami	7 (16)	13932	1000	1000	1000	NA
Sulfur	6	10081	1000	1000	1000	NA
Power	4	9568	1000	1000	1000	30%
Yacht	1 (6)	308	100	108	100	20%
Kin8nm	6 (8)	8192	1000	1000	1000	NA

Each experiment is repeated 10 times, where we randomly sample the training, calibration and test datasets. We compare our penalized kSoS (predictive model obtained with a homoscedastic GP and hyperparameters estimated with our HSIC criterion, isotropic Matérn 5/2 kernel) with CQR, homoscedastic GP and heteroscedastic GP (both also with Matérn 5/2 kernel, but anisotropic) and symmetric kSoS from Allain et al. (2026). For kSoS, we apply the same normalizing preprocessing step as for analytical test cases, and only consider the training set penalty which scales linearly with the number of training samples. For completeness, we first provide in Table 6 the estimated marginal coverage on the test set: as expected from theory, all methods achieve the target coverage ($\alpha = 0.1$ here).

The mean width achieved by each method is given in Table 7. First observe that homoscedastic GP very often yields the smaller intervals among all methods, in particular for Concrete, Housing, Boston and Kin8nm. But we will see below that precisely for these datasets, it comes at the price of a poorer local coverage, thus confirming our assertion that mean width alone is not sufficient to differentiate competitors. Our penalized kSoS model also exhibits the smaller mean width in most instances, but with much better local coverage, as elaborated in what follows.

To investigate adaptivity, we thus also analyze the worst-set coverage (where lower and upper variants are combined) in Figure 25. In average, CQR and kSoS (penalized and symmetric) have better worst-set coverage than both GPs: this confirms

Table 6. Estimated marginal coverage on the test test for thirteen real-world datasets (mean±sd on 10 repetitions).

Dataset	CQR	Het. GP	Hom. GP	Sym. kSoS	Pen. kSoS
Concrete	0.906 ± 0.023	0.900 ± 0.027	0.903 ± 0.027	0.904 ± 0.018	0.916 ± 0.026
Bike	0.897 ± 0.017	0.897 ± 0.020	0.902 ± 0.018	0.895 ± 0.016	0.894 ± 0.014
Diabetes	0.881 ± 0.044	0.888 ± 0.024	0.907 ± 0.023	0.902 ± 0.027	0.899 ± 0.023
Housing log	0.900 ± 0.009	0.900 ± 0.016	0.897 ± 0.016	0.895 ± 0.013	0.899 ± 0.011
Housing	0.902 ± 0.007	0.903 ± 0.012	0.899 ± 0.014	0.892 ± 0.011	0.888 ± 0.011
MPG	0.901 ± 0.048	0.905 ± 0.025	0.889 ± 0.048	0.886 ± 0.042	0.894 ± 0.049
Boston	0.910 ± 0.021	0.900 ± 0.032	0.902 ± 0.022	0.904 ± 0.017	0.912 ± 0.016
Energy	0.902 ± 0.024	0.896 ± 0.021	0.897 ± 0.026	0.904 ± 0.015	0.903 ± 0.016
Miami	0.899 ± 0.016	0.899 ± 0.016	0.903 ± 0.016	0.900 ± 0.017	0.900 ± 0.011
Sulfur	0.897 ± 0.015	0.900 ± 0.014	0.891 ± 0.014	0.901 ± 0.014	0.902 ± 0.011
Power	0.895 ± 0.007	0.894 ± 0.010	0.898 ± 0.010	0.896 ± 0.007	0.897 ± 0.008
Yacht	0.895 ± 0.055	0.901 ± 0.034	0.901 ± 0.036	0.909 ± 0.040	0.907 ± 0.041
Kin8nm	0.902 ± 0.017	0.900 ± 0.013	0.895 ± 0.014	0.898 ± 0.011	0.894 ± 0.014

Table 7. Mean width of prediction intervals on the test test for thirteen real-world datasets (**mean**±sd on 10 repetitions). Mean width values within 1% of the minimum are displayed in bold.

Dataset	CQR	Het. GP	Hom. GP	Sym. kSoS	Pen. kSoS
Concrete	22.66 ± 1.06	21.53 ± 1.58	$\mathbf{21.01} \pm 1.58$	28.00 ± 2.78	21.74 ± 1.19
Bike	217.98 ± 6.54	199.23 ± 14.27	169.57 ± 7.83	177.53 ± 11.82	$\mathbf{162.47} \pm 5.86$
Diabetes	$\mathbf{188.65} \pm 12.59$	196.25 ± 16.01	193.21 ± 15.80	195.80 ± 15.81	$\mathbf{190.46} \pm 14.63$
Housing log	0.967 ± 0.039	0.855 ± 0.041	$\mathbf{0.768} \pm 0.040$	0.967 ± 0.140	0.836 ± 0.037
Housing	1.77 ± 0.055	1.70 ± 0.14	$\mathbf{1.60} \pm 0.082$	2.02 ± 0.38	1.80 ± 0.14
MPG	9.82 ± 1.06	9.76 ± 1.31	$\mathbf{9.33} \pm 1.02$	9.68 ± 1.35	$\mathbf{9.32} \pm 1.14$
Boston	12.16 ± 1.18	10.85 ± 0.92	$\mathbf{9.76} \pm 0.80$	17.93 ± 18.06	11.05 ± 1.23
Energy	1.50 ± 0.16	1.46 ± 0.098	1.72 ± 0.079	1.79 ± 0.42	$\mathbf{1.39} \pm 0.12$
Miami	$31.4e^4 \pm 1.17e^4$	$36.4e^4 \pm 8.23e^4$	$29.5e^4 \pm 2.29e^4$	$28.3e^4 \pm 1.53e^4$	$\mathbf{26.9e^4} \pm 1.45e^4$
Sulfur	$\mathbf{0.051} \pm 0.003$	0.055 ± 0.009	$\mathbf{0.050} \pm 0.003$	0.064 ± 0.010	$\mathbf{0.050} \pm 0.002$
Power	13.29 ± 0.30	$\mathbf{12.88} \pm 0.40$	13.05 ± 0.36	13.62 ± 0.45	$\mathbf{12.99} \pm 0.29$
Yacht	0.645 ± 0.085	$\mathbf{0.595} \pm 0.086$	0.609 ± 0.082	0.616 ± 0.097	$\mathbf{0.599} \pm 0.105$
Kin8nm	0.607 ± 0.020	0.425 ± 0.023	$\mathbf{0.386} \pm 0.009$	0.453 ± 0.043	0.393 ± 0.011

the initial findings of Allain et al. (2026), and shows that our penalized kSoS has the desired adaptive property. Now, if we compare exhaustively penalized kSoS against competitors, we notice that:

1. Penalized kSoS almost always outperforms CQR in terms of worst-set coverage, while the latter systematically yields intervals with much larger mean width
2. Penalized and symmetric kSoS have equivalent worst-set coverage in all cases, but symmetric kSoS always produces bands with larger mean width
3. For the datasets where homoscedastic GP had smaller mean width than penalized kSoS (Concrete, Housing, Boston and Kin8nm), it is significantly behind in terms of worst-set coverage in most instances. For the other datasets, where penalized kSoS and GPs have similar worst-set coverage:
    - Bike: homoscedastic GP and penalized kSoS also have similar mean width, which hints towards a symmetric dataset. In Figure 26, selected λ_{pen} values for kSoS indicate that pure asymmetry is never considered for our model.
    - Diabetes: both kSoS have mean width equivalent to homoscedastic GP, once again suggesting this is a symmetric dataset. This intuition is confirmed by Figure 26, where we automatically select a symmetric model half of the time.

- Energy: penalized kSoS clearly outperforms both GPs in terms of mean width.
- Miami: same behavior as for Energy.
- Sulfur: mean widths for penalized kSoS and homoscedastic GP are similar, from which we can infer underlying symmetry. Figure 26 shows that our model indeed heavily favors high values of λ_{pen}.
- Power: heteroscedastic GP produces intervals with the same mean width as penalized kSoS. As can be seen in Figure 26, our model also selects larger penalty values in average, which corroborates the hypothesis of a symmetric noise distribution.
- Yacht: same behavior as for Power.

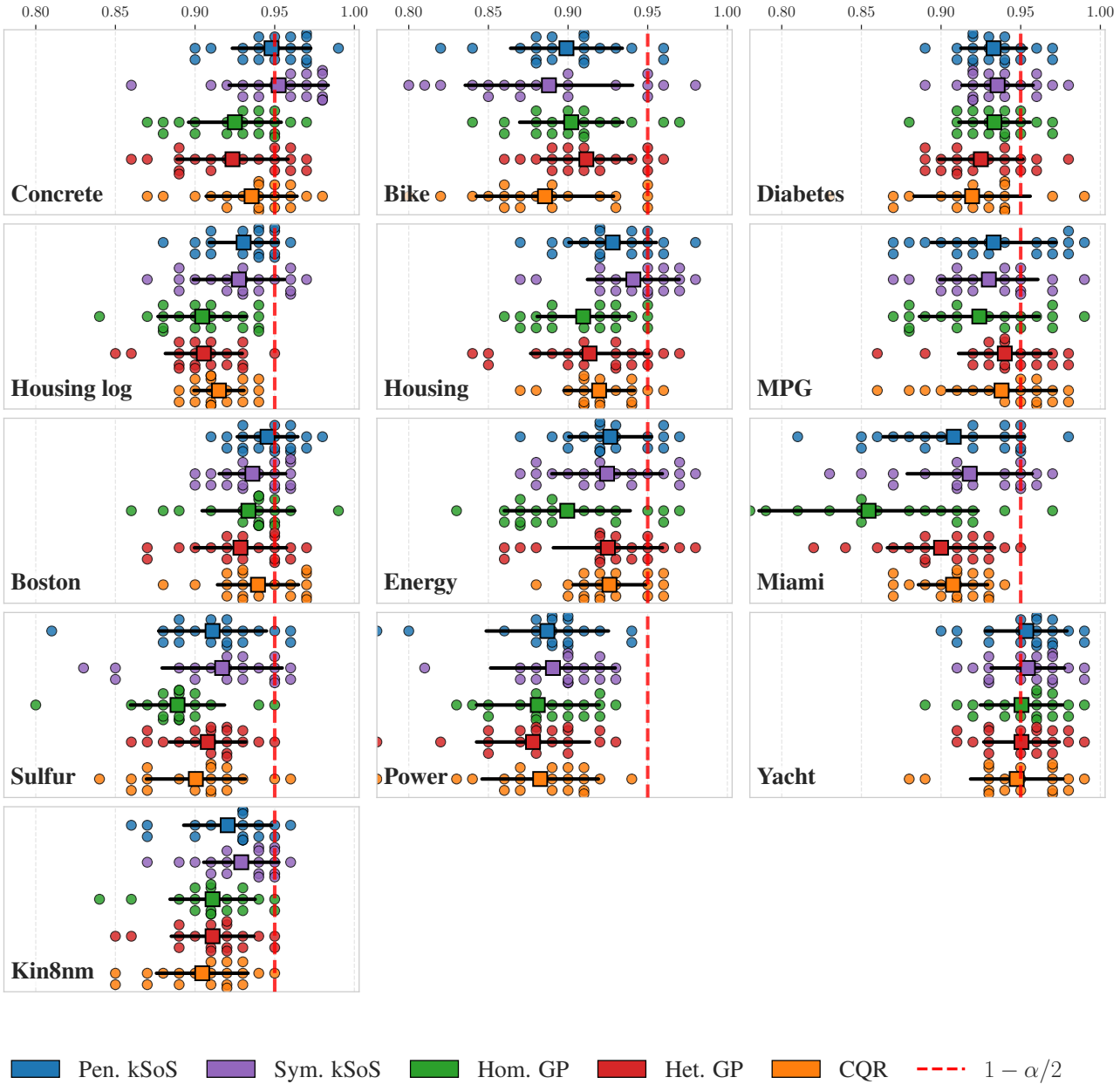


Figure 25. Mean and standard deviation of worst-set coverage lower and upper combined, 10 repetitions.

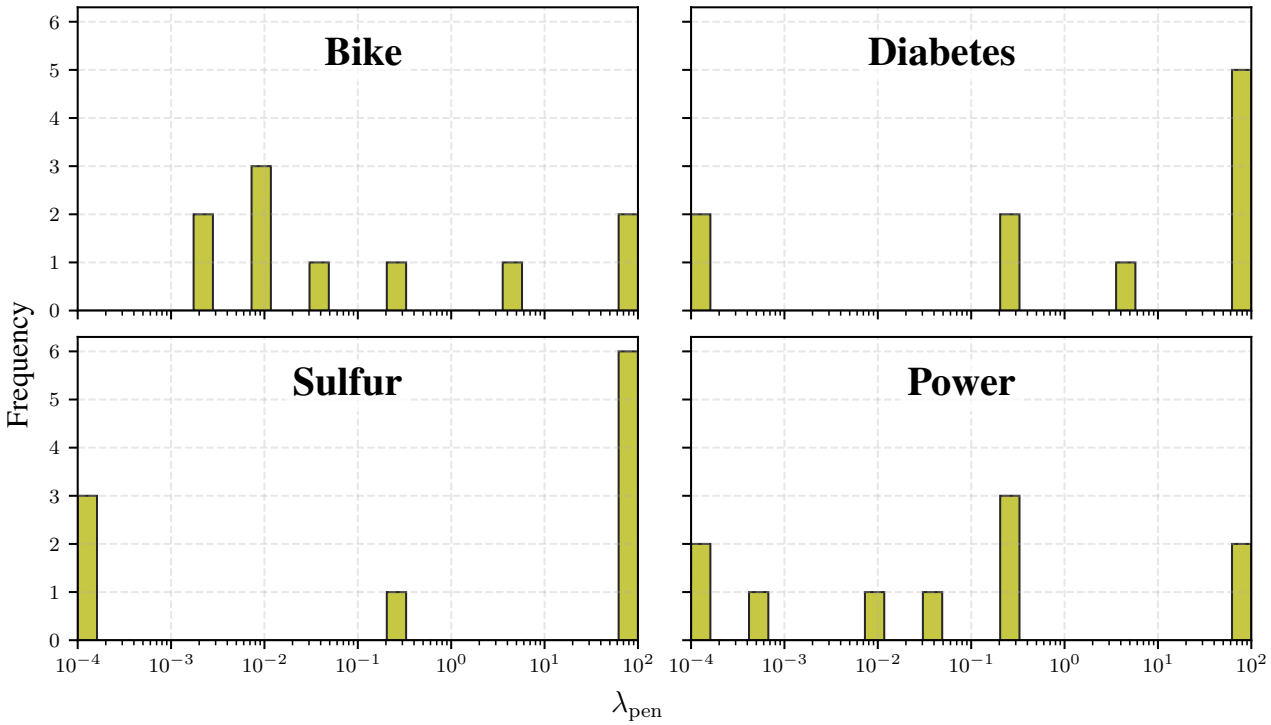


Figure 26. Histogram of automatically selected value of λ_{pen} with HSIC and Kruskal-Wallis test, 10 repetitions.