Received: 23 April 2021 | Accepted: 14 May 2022 | Published on: 19 August 2022

DOI: 10.1111/rssb.12542

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY Series B Statistical Methodology

Universal prediction band via semi-definite programming

Tengyuan Liang

Booth School of Business, University of Chicago, Chicago, Illinois, USA

Correspondence
Tengyuan Liang, Booth School of Business, University of Chicago, Chicago, IL, USA.
Email: tengyuan.liang@chicagobooth.edu

Abstract

We propose a computationally efficient method to construct nonparametric, heteroscedastic prediction bands for uncertainty quantification, with or without any user-specified predictive model. Our approach provides an alternative to the now-standard conformal prediction for uncertainty quantification, with novel theoretical insights and computational advantages. The data-adaptive prediction band is universally applicable with minimal distributional assumptions, has strong non-asymptotic coverage properties, and is easy to implement using standard convex programs. Our approach can be viewed as a novel variance interpolation with confidence and further leverages techniques from semi-definite programming and sum-of-squares optimization. Theoretical and numerical performances for the proposed approach for uncertainty quantification are analysed.

KEYWORDS

heteroscedasticity, nonparametric prediction band, semi-definite programming, sum-of-squares, uncertainty quantification

1 | INTRODUCTION

A plausible criticism from the statistics community of modern machine learning is the lack of rigorous uncertainty quantification, with perhaps the exception in conformal prediction (Lei et al., 2018; Romano et al., 2019; Vovk et al., 2005). Instead, the machine learning community would argue that conventional uncertainty quantification based on idealized distributional assumptions may be too restrictive for real data. Nevertheless, without a doubt, uncertainty quantification for predictive modelling is essential to statistics, learning theory and econometrics. This paper will resolve the above inference dilemma by introducing a new method with provable uncertainty quantification via semi-definite programming. This paper provides an alternative approach to the now-standard conformal prediction for uncertainty quantification, with novel theoretical insights and computational advantages. The proposed method learns a data-adaptive, heteroscedastic prediction band that is: (a) universally applicable without strong distributional assumptions, (b) with desirable theoretical coverage with or without any user-specified predictive model and (c) easy to implement via standard convex programs (when used in conjunction with a wide range of positive-definite kernels).

Let $(x, y) \in \mathcal{X} \times \mathbb{R}$ be the covariates and response data pair drawn from an unknown probability distribution $\mathcal{P}$. There are plenty of regression or predictive models—denoted by $\mathsf{m}_0(x)$—that estimate $\mathsf{m}(x) := \mathbb{E}[\mathbf{y} \,|\, \mathbf{x} = x]$ sufficiently well with finite data. However, to make downstream decisions reliable, a good prediction band quantifying the uncertainty in $|\mathbf{y} - \mathsf{m}_0(\mathbf{x})|$ with provable coverage is urgently needed. The prediction band is of particular relevance to complex machine learning models that construct $\mathsf{m}_0(x)$ in a less transparent way, such as deep neural networks and boosting machines. This paper makes progress in filling in such a gap: we estimate a nonparametric, heteroscedastic prediction band $\widehat{\mathsf{PI}}(x)$ that enjoys provable coverage with minimal data assumptions for any predictive model. Our approach can be viewed as a novel variance interpolation with confidence and leverages techniques from sum-of-squares relaxations for nonparametric variance estimation. On a non-technical level, this paper enriches the toolbox of applied researchers with a theoretically justified new methodology for uncertainty quantification and visualization, as in conformal prediction.

1.1 | Semi-definite programs and prediction bands

We introduce our procedure for constructing the predictive band in this section. Let $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous symmetric and positive-definite kernel function. Given n data pairs $\{(x_i, y_i)\}_{i=1}^n$ and the corresponding kernel matrix $\mathbf{K} \in \mathbb{S}^{n \times n}$ with $\mathbf{K}_{ij} = K(x_i, x_j)$, our prediction band is constructed based on the following semi-definite program (SDP)

$$\begin{aligned} \min_{\mathbf{B}} \quad & \mathrm{Tr}(\mathbf{KB}) \\ \text{s.t.} \quad & \langle \mathsf{K}_i, \mathbf{B}\mathsf{K}_i \rangle \geq (y_i - \mathsf{m}_0(x_i))^2, \ \ i = 1, \ \dots, n \\ & \mathbf{B} \succeq 0 \end{aligned} \tag{1}$$

where the optimization variable $\mathbf{B} \in \mathbb{S}^{n \times n}$ is a symmetric positive semi-definite (PSD) matrix, $\mathsf{m}_0(\cdot) : \mathcal{X} \to \mathbb{R}$ is a given predictive model (user-specified), and $\mathsf{K}_i \in \mathbb{R}^n$ denotes the ith column of the kernel matrix $\mathbf{K}$. Given the estimated $\widehat{\mathbf{B}}$, the **prediction band**, $\widehat{\mathsf{PI}}(x)$ that maps each x to an interval, can be constructed accordingly

$$\begin{aligned}
&\widehat{\mathsf{PI}}(x,\delta) := \left[\mathsf{m}_0(x) - \sqrt{(1+\delta)\cdot\hat{\mathsf{v}}(x)}, \mathsf{m}_0(x) + \sqrt{(1+\delta)\cdot\hat{\mathsf{v}}(x)}\right], \forall x \in \mathcal{X},\\
&\text{where} \quad \hat{\mathsf{v}}(x) := \langle \mathsf{K}_x, \widehat{\mathbf{B}}\mathsf{K}_x\rangle,\\
&\text{and} \quad \mathsf{K}_x := [K(x,x_1), \ldots, K(x,x_n)]^\top \in \mathbb{R}^n,
\end{aligned} \tag{2}$$

with $\delta \in \mathbb{R}$ being a scalar quantifying confidence. δ can be later calibrated for exact coverage control, and may be set as 0 if n is large. Here $\hat{\mathsf{v}}(x)$ estimates the variability in the 'deviations' $e_i := y_i - \mathsf{m}_0(x_i)$. A few remarks on such deviations are in place.

1. First, e_i's can be computed based on any user-specified predictive model $\mathsf{m}_0(x)$ that estimates the conditional mean $\mathsf{m}_0(x) \approx \mathbb{E}[\mathbf{y}|\mathbf{x} = x]$, be it accurate or not.
2. Second, in the absence of such a predictive model for the conditional mean, one can set $\mathsf{m}_0(x) \equiv 0$ and learn a conditional second-moment function to assess uncertainty.
3. Last, as shown next in (3), in practice, one can simultaneously learn the conditional mean and variance functions using a variant of the above SDP. Therefore, a pre-specified model $\mathsf{m}_0(x)$ is not required.

Let K^{m} and K^{v} specify two kernel functions, corresponding to the conditional mean and variance functions respectively. $\mathbf{K}^{\mathsf{m}}, \mathbf{K}^{\mathsf{v}} \in \mathbb{S}^{n\times n}$ denote empirical kernel matrices on finite data with size n. For any $\gamma \geq 0$, the following convex SDP program constructs the prediction band and the conditional mean function simultaneously

$$\begin{aligned}
\min_{\alpha,\mathbf{B}} \quad & \gamma\cdot\langle\alpha, \mathbf{K}^{\mathsf{m}}\alpha\rangle + \mathrm{Tr}\,(\mathbf{K}^{\mathsf{v}}\mathbf{B})\\
\text{s.t.} \quad & \left\langle \mathsf{K}_i^{\mathsf{v}}, \mathbf{B}\mathsf{K}_i^{\mathsf{v}}\right\rangle \geq \left(y_i - \left\langle \mathsf{K}_i^{\mathsf{m}}, \alpha\right\rangle\right)^2, i = 1, \ldots, n\\
& \mathbf{B} \succeq 0
\end{aligned} \tag{3}$$

where the optimization variables are $\mathbf{B} \in \mathbb{S}^{n\times n}$ and $\alpha \in \mathbb{R}^n$. Given the solution $\widehat{\mathbf{B}}$ and $\hat{\alpha}$, the $\widehat{\mathsf{PI}}(x)$ is constructed as

$$\begin{aligned}
&\widehat{\mathsf{PI}}(x,\delta) := \left[\hat{\mathsf{m}}(x) - \sqrt{(1+\delta)\cdot\hat{\mathsf{v}}(x)}, \hat{\mathsf{m}}(x) + \sqrt{(1+\delta)\cdot\hat{\mathsf{v}}(x)}\right], \forall x \in \mathcal{X},\\
&\text{where } \hat{\mathsf{m}}(x) := \langle \mathsf{K}_i^{\mathsf{m}}, \hat{\alpha}\rangle \text{ and } \hat{\mathsf{v}}(x) := \langle \mathsf{K}_x^{\mathsf{v}}, \widehat{\mathbf{B}}\mathsf{K}_x^{\mathsf{v}}\rangle.
\end{aligned}$$

1.2 | A numerical illustration

Before diving into the motivations behind the above SDPs (Section 1.3) and corresponding theory (Section 2), let us first visually illustrate the empirical performance of the constructed prediction bands on a toy numerical example. A complete simulation study comparing our methods and conformal predictions will be deferred to Section 3.1. Details of the data generating processes will be elaborated therein as well. The quick exercise here is to showcase that convex programs (1) and (3) are easy to implement using standard optimization toolkits (say, CVX (Grant & Boyd, 2014)), and construct flexible prediction bands with desired coverage properties. As a motivating example, we try out the SDP (3), which simultaneously estimates the conditional mean and variance functions. A minimal 10-line Python implementation is provided in Listing 1.

The first example is a linear model with heteroscedastic error: the conditional mean $\mathsf{m}(x)$ being a linear function and variance $\mathsf{v}(x)$ being a quadratic (with the conditional variance

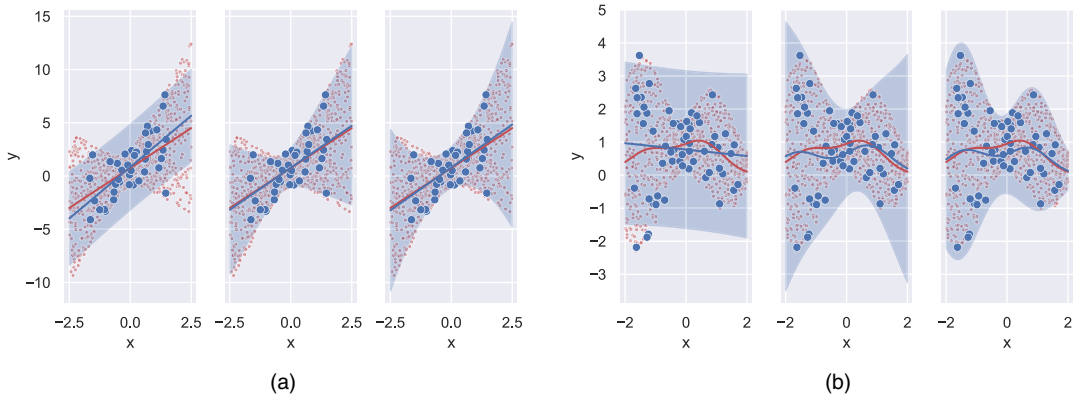


FIGURE 1 From left to right: SLR, SDP1 and SDP2. For each plot, Blue dots denote training data $\{(x_i, y_i)\}_{i=1}^n$, Blue line denotes the estimated conditional mean $\widehat{\mathsf{m}}(x)$, and Blue band denotes the estimated prediction band $\widehat{\mathsf{PI}}(x)$. Red dots represent the unknown test distribution, and Red line denotes the true conditional mean $\mathsf{m}(x) = \mathbb{E}[\mathbf{y}|\mathbf{x} = x]$. Here the training and test data share the same conditional distribution $\mathbf{y}|\mathbf{x} = x$ and thus $\mathsf{m}(x)$. The training and test data are shared in three plots. A good coverage corresponds to when Blue band covers essentially all Red dots. Statistics are summarized in Table 1 [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Simulated examples

	Coverage	Median length	Average length	MSE
Figure 1a: linear $\mathsf{m}(x)$, quadratic $\mathsf{v}(x)$				
SLR	85.88%	8.2057	8.2658	0.6294
SDP1	91.13%	7.4689	7.7173	0.1146
SDP2	94.00%	7.2962	8.3361	0.1720
Figure 1b: rbf $\mathsf{m}(x)$, rbf $\mathsf{v}(x)$				
SLR	96.13%	4.8048	4.8185	0.2556
SDP1	99.25%	4.4138	4.6196	0.1916
SDP2	99.50%	3.3488	3.7506	0.1670

generated from a uniform distribution). We generate a training dataset of size $n = 40$ and compare the coverage among three methods: (a) SLR, simple linear regression, (b) SDP1, a SDP (3) with linear kernels K^{m} and K^{v} for both mean and variance functions and (c) SDP2, a SDP (3) with a linear K^{m} and a (degree-3) polynomial kernel K^{v}. The coverage is compared on the same test dataset of size $N = 800$. Here $\gamma = 0.1$. See Figure 1a for details and Table 1 for coverage statistics.

The second example is a non-linear, heteroscedastic error model: mean $\mathsf{m}(x)$ and variance $\mathsf{v}(x)$ functions lying in a reproducing kernel Hilbert space (RKHS) with a radial basis function (rbf) kernel. Here $n = 60$ training samples and $N = 800$ test samples are generated. Three methods being compared are: (a) SLR, (b) SDP1, rbf kernel for K^{m} and linear kernel for K^{v} and (c) SDP2, rbf kernels for both K^{m} and K^{v}, summarized in Figure 1b and Table 1. Here $\gamma = 1$.

These two numerical examples are minimal yet informative. In Figure 1a, SLR misspends a wide prediction bandwidth on data where the conditional variances are small yet fails to capture the large conditional variance cases, resulting in the overall coverage of 86% and a

median bandwidth of 8.21. SDP1/SDP2 re-distributes the bandwidth budget leveraging the heteroscedastic nature and achieves an improved coverage 91%/94%, with a smaller median bandwidth of 7.47/7.30. Such an effect is even more pronounced in Figure 1b. Observe first that in SDP2, the prediction band constructed by (3) almost perfectly contours the heteroscedastic variances, thus achieving a >99% prediction coverage with a merely 3.35 median bandwidth, in contrast to SLR with a 96% coverage and a 4.80 bandwidth. Second, a better conditional variance estimate also improves performance in learning the conditional mean, as seen in the differences between Blue lines and Red lines. The errors are also numerically summarized in the column 'MSE' of Table 1. Leveraging the heteroscedasticity in data, our prediction band distributes the bandwidth in a data-adaptive way, thus improving the overall coverage.

1.3 | Sum-of-squares, interpolation and connections to literature

The SDPs proposed in (1) and (3) are inspired by recent advancements in optimization and learning theory. We will elaborate on the connections to related works and explain the innovations in our approach. We start with some basic observations about the SDPs. First, when α is not an optimization variable, (3) recovers (1). Second, the constraints set of (3) is always non-empty since $\alpha = 0, \mathbf{B} = \max_i \|\mathsf{K}_i^{\mathsf{v}}\|^{-2} y_i^2 \cdot \mathbf{I}$ is feasible.

Sum-of-squares and phase retrieval As shown in Proposition 3, the infinite-dimensional SDP with a nuclear norm minimization is equivalent to (3),

$$\begin{aligned} \min_{\beta \in \mathcal{H}^{\mathsf{m}}, \mathbf{A}: \mathcal{H}^{\mathsf{v}} \to \mathcal{H}^{\mathsf{v}}} \quad & \gamma \cdot \|\beta\|_{\mathcal{H}^{\mathsf{m}}}^2 + \|\mathbf{A}\|_\star \\ \text{s.t.} \quad & \langle \phi_{x_i}^{\mathsf{v}}, \mathbf{A} \phi_{x_i}^{\mathsf{v}} \rangle_{\mathcal{H}^{\mathsf{v}}} \geq (y_i - \langle \phi_{x_i}^{\mathsf{m}}, \beta \rangle_{\mathcal{H}^{\mathsf{m}}})^2, \quad \forall i. \\ & \mathbf{A} \succeq 0 \end{aligned}$$

Here $\mathcal{H}^{\mathsf{m}}, \mathcal{H}^{\mathsf{v}}$ denote two RKHSs where the conditional mean and variance functions reside. $\phi_{x_i} \in \mathcal{H}$ is the feature map w.r.t. the Hilbert space $\mathcal{H}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the Hilbert space inner product. We call it the infinite-dimensional SDP since the optimization variables $(\beta, \mathbf{A})$ are (function, operator) rather than finite-dimensional (vector, matrix). A few remarks are in place. First, if the kernel K^{m} is universal, $\langle \phi_x^{\mathsf{m}}, \beta \rangle_{\mathcal{H}^{\mathsf{m}}}$ is dense in L^2 and hence can universally approximate all conditional mean function $\mathsf{m}(x)$. Second, as for the conditional variance which has positivity constraints over a continuum $x \in \mathcal{X}$ with $0 \leq \mathsf{v}(x) = (y - \mathsf{m}(x))^2$, we relax the positivity constraints using a sum-of-squares form

$$0 \leq \langle \phi_x^{\mathsf{v}}, \mathbf{A} \phi_x^{\mathsf{v}} \rangle_{\mathcal{H}^{\mathsf{v}}} = (y - \mathsf{m}(x))^2, \text{ for some } \mathbf{A} \succeq 0. \tag{4}$$

It turns out that when K^{v} is universal, the above sum-of-squares function can approximate all smooth, positive functions (Bagnell & Farahmand, 2015; Fefferman & Phong, 1978; Marteau-Ferey et al., 2020), thus explaining the name 'universal' in the title. Remark that sum-of-squares optimization (Lasserre, 2001) for nonparametric estimation has recently been considered; see (Bagnell & Farahmand, 2015; Curmei & Hall, 2020; Marteau-Ferey et al., 2020). The further relaxation changing from equality in (4) to inequality will be discussed in the next paragraph. Last, the minimum nuclear norm objective translates to a particular form of "minimum bandwidth" in the prediction band as $\mathsf{v}(x) = \langle \phi_x^{\mathsf{v}}, \mathbf{A} \phi_x^{\mathsf{v}} \rangle_{\mathcal{H}^{\mathsf{v}}}$. In language, for all prediction bands that shelter the data, (3) aims to find the one with minimum bandwidth.

The curious reader may wonder where the nuclear norm $\|\mathbf{A}\|_\star$ arises from. The first reason is conceptual: the nuclear norm is a relaxation for rank, and the procedure is to minimize the number of factors (rank) that explain the variance. The second reason is a connection to phase retrieval: specify $K^{\mathsf{m}}(x,x') \equiv 0$ and $K^{\mathsf{v}}(x,x') = \langle x, x' \rangle$ (the linear kernel with $\phi_x^{\mathsf{v}} = x$), and force the inequality constraints to be equal, our SDP in (3) is equivalent to phase retrievel

$$\min_{\mathbf{A}\succeq 0} \|\mathbf{A}\|_\star, \ \text{s.t.} \quad \langle x_i, \mathbf{A} x_i \rangle = y_i^2, \quad \forall i = 1, \ \dots \ , n.$$

Conceptually, the minimum nuclear norm procedure estimates the smallest number of factors that could generate the variance.

Min-norm variance interpolation with confidence Now we discuss the tuning parameter $\gamma \in [0, \infty]$ and reveal the connection to the recent min-norm interpolation literature (Bartlett et al., 2020, 2021; Ghorbani et al., 2020; Liang & Rakhlin, 2020; Liang & Recht, 2021; Montanari et al., 2020). In the limit of $\gamma \to 0$, (3) reduces to the familiar min-norm interpolation with kernel $\mathbf{K}^{\mathsf{m}}$ (whenever it has full rank, since optimal $\mathbf{B} = 0$)

$$\begin{aligned} \min_{\alpha} \quad & \langle \alpha, \mathbf{K}^{\mathsf{m}} \alpha \rangle \\ \text{s.t.} \quad & 0 = \left(y_i - \left\langle \mathsf{K}_i^{\mathsf{m}}, \alpha \right\rangle\right)^2, \quad \forall i. \end{aligned}$$

In the limit of $\gamma \to \infty$, (3) reduces to (since optimal $\alpha = 0$)

$$\begin{aligned} \min_{\mathbf{B}} \quad & \mathrm{Tr}(\mathbf{K}^{\mathsf{v}} \mathbf{B}) \\ \text{s.t.} \quad & \langle \mathsf{K}_i^{\mathsf{v}}, \mathbf{B} \mathsf{K}_i^{\mathsf{v}} \rangle \geq y_i^2, \quad \forall i. \\ & \mathbf{B} \succeq 0 \end{aligned}$$

Now it is clear what the role of the *tuning parameter* γ is: it trades off the conditional mean $\mathsf{m}(x)$ and variance $\mathsf{v}(x)$ to explain the variability in y's witnessed on the data. A small γ aims to use a complex mean $\mathsf{m}(x)$ and a parsimonious variance $\mathsf{v}(x)$ to explain the overall variability, and vice versa.

From the above discussion, it is also clear that the SDP (3) can be viewed as a min-norm variance interpolation with confidence. Instead of having the typical equality constraints in interpolation

$$\left\langle \mathsf{K}_i^{\mathsf{v}}, \mathbf{B} \mathsf{K}_i^{\mathsf{v}} \right\rangle = \left(y_i - \left\langle \mathsf{K}_i^{\mathsf{m}}, \alpha \right\rangle\right)^2,$$

which violates the disciplined convex programming ruleset (due to the quadratic form on the RHS), we further relax to inequality constraints to incorporate additional 'confidence' (and to make the problem convex at the same time)

$$\left\langle \mathsf{K}_i^{\mathsf{v}}, \mathbf{B} \mathsf{K}_i^{\mathsf{v}} \right\rangle \geq \left(y_i - \left\langle \mathsf{K}_i^{\mathsf{m}}, \alpha \right\rangle\right)^2.$$

As we shall see, the notion of confidence in this variance interpolation is closely related to the notion of margin in classification (Bartlett et al., 1998; Liang & Sur, 2020).

Support vector regression We illustrate that minor modifications to our SDP formulation lead to other problems, including support vector regression and kernel ridge regression.

Specify the variance kernel as the trivial one $K^{\mathrm{v}}(x,x') = \mathbb{1}(x = x')$, then the decision variable $\mathbf{B}$ only matters in its diagonal component, and our SDP (3) reduces to the kernel ridge regression

$$\min_{\alpha} \quad \gamma \cdot \langle \alpha, \mathbf{K}^{\mathrm{m}} \alpha \rangle + \sum_{i=1}^{n} \left(y_i - \left\langle \mathsf{K}_i^{\mathrm{m}}, \alpha \right\rangle \right)^2.$$

Moreover, a slight modification of (3) is to use the absolute deviation rather than the squared deviation in the constraints, namely

$$\left\langle \mathsf{K}_i^{\mathrm{v}}, \mathbf{B}\mathsf{K}_i^{\mathrm{v}} \right\rangle \geq \left| y_i - \left\langle \mathsf{K}_i^{\mathrm{m}}, \alpha \right\rangle \right|.$$

In this case support vector regression is exactly our procedure with the specification $K^{\mathrm{v}}(x,x') = \mathbb{1}(x = x')$,

$$\min_{\alpha} \quad \gamma \cdot \langle \alpha, \mathbf{K}^{\mathrm{m}} \alpha \rangle + \sum_{i=1}^{n} \left| y_i - \left\langle \mathsf{K}_i^{\mathrm{m}}, \alpha \right\rangle \right|.$$

In summary, our SDP generalize beyond support vector machines, with the new non-trivial variance component for rigorous uncertainty quantification for heteroscedastic data.

1.4 | Literature review

There are increasingly many approaches proposed to address the uncertainty quantification dilemma in machine learning due to its significance and centrality. However, very few methods are theoretically grounded and universally applicable to the best of our knowledge. Many approaches are merely heuristics or data visualization tools. This section divides related theoretical studies in the literature into two categories and discusses how our method significantly differs from them and could potentially lead to a stronger theory.

Conformal prediction Based on the exchangeability of data and a user-specified nonconformity measure, Vovk et al. (2005); Shafer and Vovk (2008) pioneered the field of conformal prediction, which uses past data to determine precise levels of confidence in new predictions. To some extent, the elegant theory of conformal prediction, motivated by online learning and sequential prediction, resolved the uncertainty quantification dilemma. The conformal prediction algorithm (see, for instance Shafer and Vovk, 2008, section 4.3) usually requires to enumerate over all the possibilities of $z = (x,y) \in \mathcal{X} \times \mathcal{Y}$, and for each possibility, calculate n nonconformity measures via the leave-one-out method. Therefore, the total budget is $n \times |\mathcal{Y}| \times |\mathcal{X}|$, which can be expensive for continuous y and multi-dimensional x. Much of the above computation can be saved if additional information about the metric structure in $x \in \mathcal{X}$ can be leveraged. In contrast, our SDP approach constructs the prediction band over all the x's at once, leverages the metric structure in $\mathcal{X}$, and suffers at most a computational budget of n^2. An additional key feature in our approach is in the coverage theory established in Theorem 1: the prediction band has coverage probability >95% on a new data point $(\mathbf{x}, \mathbf{y})$, for 99.9999% dataset $\{(x_i, y_i)\}_{i=1}^n$ of size n drawn from the same distribution. Such a distinction on 'confidence' versus 'probability' is discussed extensively in section 2.2 of Shafer and Vovk (2008).

There has been a vast line of recent work on extending the conformal prediction idea further to address the bottlenecks above in the regression setting. The body of work proliferates, and we certainly cannot do justice here. Lei et al. (2018) alleviates the computational burden of the conformal prediction by introducing the sample-splitting technique. Remarkably, theory on the bandwidth is also studied in Lei et al. (2018), thus providing an angle to probe the statistical efficiency. Romano et al. (2019) studies the problem that existing conformal methods can form nearly constant or weakly varying bandwidth and provide conservative results. Romano et al. (2019) proposes conformalized quantile regression to address this issue. One shared feature of our SDP approach and the method in Romano et al. (2019) is that the prediction band is fully adaptive to heteroscedasticity. Finally, we would like to emphasize that conformal prediction constructs prediction bands in a numerical, black-box fashion without a structural understanding of the variance function. In contrast, our SDP approach provides a transparent and efficient way of learning the variance function, a complementary contribution to the conformal literature.

Residual subsampling and quantile regression An alternative approach for uncertainty quantification that leverages the metric structure in $x \in \mathcal{X}$ is to resample the residuals locally. Typically, this is done by first fitting a predictive model $\mathsf{m}_0(x)$, and defining a local neighbourhood around a new data $\mathbf{x}$, then subsampling the residuals for uncertainty quantification via (conditional) quantiles. The validity of the above approach crucially depends on how many 'similar residuals' to pool information from. However, the curse of dimensionality comes in since data points are far from each other in high dimensions, posing challenges in pooling the residuals. One can also use either the obtained residuals or the original responses y to fit a conditional quantile regression model (Belloni & Chernozhukov, 2011; Belloni et al., 2019; Koenker & Bassett Jr, 1978; Koenker & Hallock, 2001), $\widehat{\xi}^{\tau}(\cdot) := \arg\min_{\xi} \frac{1}{n}\sum_{i=1}^{n}\rho_{\tau}(y_i - \xi(x_i))$ where $\tau \in (0,1)$ is a quantile parameter, $\rho_{\tau}(\cdot) : \mathbb{R} \to \mathbb{R}_+$ is the tilted absolute value function, and $\widehat{\xi}^{\tau}(\cdot) : \mathcal{X} \to \mathbb{R}$ is the estimated conditional quantile function. However, it is not guaranteed that over all $x \in \mathcal{X}$, the estimated conditional quantile function satisfies $\widehat{\xi}^{\tau_1}(x) < \widehat{\xi}^{\tau_2}(x)$ for two quantiles $\tau_1 < \tau_2$. In other words, it is entirely possible that for several x's, the conditional prediction intervals are empty (Chernozhukov et al., 2010).

2 | THEORY FOR UNCERTAINTY QUANTIFICATION

In this section, we develop a theory for the coverage property of the prediction band constructed above, under the mild assumption that the data are i.i.d. drawn with $(\mathbf{x}, \mathbf{y}) \sim \mathcal{P}$. To highlight the main arguments in a simple form, let us consider the setting $\mathsf{m}_0(x) \equiv 0$. Otherwise, the same proof follows by replacing $\mathbf{y}$ with $\mathbf{y} - \mathsf{m}_0(\mathbf{x})$. Define the corresponding prediction band, with a confidence parameter $\delta \in (0, 1]$

$$\widehat{\mathsf{PI}}(x, \delta) = \left[\pm\sqrt{(1+\delta)\cdot\widehat{\mathsf{v}}(x)}\right]. \tag{5}$$

We need the following assumptions before stating the theorem, where $(\mathbf{x}, \mathbf{y}) \sim \mathcal{P}$ and $\mathrm{C} > 0$ denotes a universal constant.

[S1] (Kernel and RKHS) The continuous symmetric kernel K is positive definite and satisfies $\sup_{x\in\mathcal{X}} K(x,x) \le \mathrm{C}$. In addition, eigenvalues of the associated integral operator $\mathcal{T} : \mathcal{H} \to \mathcal{H}$ satisfy $\lambda_j(\mathcal{T}) \le \mathrm{C}j^{-\tau}, j \in \mathbb{N}$ for some constant $\tau > 1$.

[S2] (Non-trivial uncertainty) There exist constants $\eta \in (0,1), \xi > 0$ such that $\mathbb{P}[\mathbf{y}^2 > \xi \cdot K(\mathbf{x},\mathbf{x}) \mid \mathbf{x} = x] > \eta$ holds for all $x \in \mathcal{X}$.

[S3] (Non-wild uncertainty) There exists a constant $\omega > 0$ such that $\mathbb{P}[\mathbf{y}^2 > t \cdot K(\mathbf{x},\mathbf{x})] < \exp(-Ct^{\omega})$ for all $t \geq 1$.

Discussion of Assumptions All the above assumptions are mild. The eigenvalue decay in [S1] is almost identical to $\mathrm{Tr}(\mathcal{T}) < \infty$ (bounded trace integral operator). [S2] is also minimal, since it is only not true when there is no variability in $\mathbf{y} \mid \mathbf{x} = x$. [S3] is the most stringent one, which requires the variability of $\mathbf{y}$ to exhibit a certain tail-decay. For small $\omega \in (0, 1)$, [S3] can be much milder than exponential tail-decay. Bounded or Gaussian $\mathbf{y} \mid \mathbf{x} = x$ satisfies [S3] with arbitrarily large ω or $\omega = 2$ respectively. With some extra work, [S3] can be relaxed to the case of a sufficiently rapid polynomial tail-decay. [S2] can be relaxed to restricting only to x with $\mathbb{P}[\mathbf{y}^2 > 0 \mid \mathbf{x} = x] = 1$.

Theorem 1. *Define the objective value of the SDP in* (1)

$$\begin{aligned} \widehat{\mathsf{Opt}}_n := \min_{\mathbf{B}} \quad & \mathrm{Tr}(\mathbf{KB}) \\ s.t. \quad & \langle \mathsf{K}_i, \mathbf{B}\mathsf{K}_i \rangle \geq y_i^2, i = 1, \ldots, n. \\ & \mathbf{B} \succeq 0 \end{aligned}$$

Assume that [S1]–[S3] *hold. For any* $\delta \in (0,1]$, *the following non-asymptotic, data-dependent prediction band coverage guarantee holds,*

$$\begin{aligned} \mathop{\mathbb{P}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}} \left[\mathbf{y} \in \widehat{\mathsf{PI}}(\mathbf{x},\delta) \right] \geq 1 - \delta^{-1}(\widehat{\mathsf{Opt}}_n \vee 1) \sqrt{\frac{\mathrm{C}_{\tau,\xi,\eta,\omega} \cdot \log(n)}{n}}, \\ \text{and } \widehat{\mathsf{Opt}}_n \leq \left[\log(n)\right]^{\mathrm{c}_{\omega}}, \end{aligned}$$

with probability at least $1 - n^{-10}$ on $\{(x_i, y_i)\}_{i=1}^n$. *Here the constants* $\mathrm{C}_{\tau,\xi,\eta,\omega}, \mathrm{c}_{\omega}$ *only depend on parameters in* [S1]–[S3].

2.1 | What does the theorem entail

A few remarks are in order before we sketch the proof of Theorem 1.

Coverage First, the above theorem says that the prediction band constructed using the SDP based on a dataset of size n, will correctly cover a fresh data point $(\mathbf{x}, \mathbf{y}) \sim \mathcal{P}$ drawn from the same distribution, with a non-asymptotic coverage probability (on the new data $\mathbf{x}, \mathbf{y}$)

$$1 - \delta^{-1} \sqrt{\frac{\log^3(n)}{n}}.$$

With $\delta = 0.5$, the bandwidth $\mathrm{Length}[\widehat{\mathsf{PI}}(x)] = 2.45\sqrt{\hat{\mathrm{v}}(x)}$ is at a heteroscedastic level adaptive to x with corresponding coverage probability at least $1 - O^{\star}\left(\sqrt{\frac{1}{n}}\right)$. Here $O^{\star}$ hides polylog factors. The coverage can be arbitrary close to 1 with large n without the need of increasing δ, which is in clear distinction to the conventional wisdom that coverage 1 can only be possible with an

increasing δ regardless of n. Again, we emphasize that the above coverage guarantee holds essentially on $99.9999\% \ll 1 - n^{-10}$ of the datasets $\{x_i, y_i\}_{i=1}^n$. In Section 2.2, we propose a rigorous calibration algorithm to choose $\delta^\star(\alpha)$ to achieve a constant coverage level $1-\alpha \in (0, 1)$.

Optimality If one wishes to obtain the classic 95% coverage probability, then choosing $\delta = O^\star\left(\sqrt{\frac{1}{n}}\right)$ suffices, which translates to

$$\text{Length } \left[\widehat{\mathsf{PI}}(x)\right] = \left(1 + O^\star\left(\sqrt{\frac{1}{n}}\right)\right) \cdot \sqrt{\hat{\mathsf{v}}(x)}. \tag{6}$$

Recall that in classic simple linear regression, the prediction interval is of length

$$\left(1 + \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma_i (x_i - \bar{x})^2}}\right) \cdot 3.92\hat{s} \tag{7}$$

with $\hat{s} = \sqrt{\frac{\Sigma_i \hat{e}_i^2}{n-2}}$ being the estimated residual standard error. The fact that (6) and (7) share the $\sqrt{\frac{1}{n}}$ fluctuation seems to indicate the optimality of our Theorem (in terms of the dependence on n).

Data adaptivity Curiously, the objective value of the convex optimization program quantifies the uncertainty of the prediction band: a smaller $\widehat{\mathsf{Opt}}_n$ implies (a) a better confidence/coverage guarantee and (b) a narrower prediction band overall. More importantly, the $\widehat{\mathsf{Opt}}_n$ can be calculated directly from data! We find such an optimization/inference interface exciting: the data-adaptive bound lets us know the coverage guarantee specific to the current dataset. Put differently, the convex program constructs the prediction band via its solution and at the same time, reveals the confidence via its objective value. Since $\widehat{\mathsf{Opt}}_n$ is a function of the dataset, our Theorem reveals which dataset allows for a better prediction band. Remark that $\widehat{\mathsf{Opt}}_n = \|\hat{\mathsf{v}}(\cdot)\|_\star^2$ is also a particular norm of the heteroscedastic variance function, quantified by the nuclear norm of the associated PSD operator $\widehat{\mathbf{A}} \succeq 0$ with $\hat{\mathsf{v}}(x) = \langle \phi_x, \mathbf{A}\phi_x \rangle_{\mathcal{H}}$. Curiously, a simpler variance function $\hat{\mathsf{v}}(x)$ (with a small norm) will simultaneously result in a narrower band and better coverage. We emphasize that the above discussion is in sharp contrast to the conventional wisdom that a narrow band usually leads to poor coverage guarantees.

2.2 | Calibration and coverage control

One nice feature about conformal prediction is that it directly operates on a user-specified coverage level (e.g. 95%), albeit the resulting procedure only achieves some form of coverage guarantee in the marginal sense. In contrast, the coverage level guarantee of the current SDP approach in Theorem 1 is in an inequality form with a mild dependence on the non-explicit universal constant $C_{\gamma,\xi,\eta,\omega}$; however, the coverage is in a stronger conditional sense conditioned on $\{(x_i, y_i)\}_{i=1}^n$. The constant will not significantly affect the coverage guarantee in the large n setting; nevertheless, curious readers may wonder if more transparent control could be achieved by tuning δ. This section provides a theoretically justified calibration procedure in choosing δ (in the SDP band) to control coverage at a user-specified level $1-\alpha \in (0,1)$, as in conformal

prediction. Such fine calibration on δ can be helpful numerically in the moderate n and constant α setting (e.g. 5%).

The calibration idea is based on sample-splitting. Split the samples into two parts, the training set $\{(x_i, y_i)\}_{i=1}^n$ and the calibration set $\{x'_j, y'_j\}_{j=1}^m$, with in total $n+m$ data points drawn i.i.d. from $\mathcal{P}$. The training set will be used to construct prediction band $\widehat{\mathsf{PI}}(\cdot, \delta)$. The calibration dataset will be used to choose δ to calibrate coverage.

Calibration procedure Suppose that there exists large enough constants $\Delta, L > 0$ such that

$$\mathop{\mathbb{P}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\left[\mathbf{y} \in \widehat{\mathsf{PI}}(\mathbf{x}, \Delta)\right] = 1, \left|\frac{d}{d\delta}\mathop{\mathbb{P}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\left[\mathbf{y} \notin \widehat{\mathsf{PI}}(\mathbf{x}, \delta)\right]\right| < L.$$

The calibration uses a dyadic search to select $\delta^\star(\alpha) \in [-1, \Delta]$ with the set $\{x'_j, y'_j\}_{j=1}^m$. The goal of the calibration is to ensure $\mathbb{P}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\left[\mathbf{y} \in \widehat{\mathsf{PI}}(\mathbf{x}, \delta^\star(\alpha))\right] \geq 1-\alpha$.

Algorithm 1: Calibration for Coverage Control

Data: Calibration set $\{x'_j, y'_j\}_{j=1}^m$, coverage level $1-\alpha$;
Result: Calibrated $\delta^\star(\alpha)$;
Initialize $\delta = -1$;
while $\frac{1}{m}\sum_{j=1}^m \mathbb{1}[y'_j \notin \widehat{\mathsf{PI}}(x'_j, \delta)] > \frac{3}{4}\alpha$, **do**
  $\delta \leftarrow \frac{\delta+\Delta}{2}$;
end
return $\delta^\star(\alpha) \leftarrow \delta$.

The following result derives the theoretical ground for the calibration procedure. We defer the proof to Section A.2.

Lemma 2 (Calibration). *Consider the calibration procedure in Algorithm* 1 *and* L, Δ, α *specified therein. If the size of the calibration set* m *is large enough such that*

$$\sqrt{\frac{\log\left(\lceil\log_2\left(\frac{2L(\Delta+1)}{\alpha}\right)\rceil + 1\right) + 10\log(m)}{m}} \leq \frac{1}{4}\alpha,$$

then the calibrated $\delta^\star(\alpha)$ *satisfies the coverage control*

$$\mathop{\mathbb{P}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\left[\mathbf{y} \in \widehat{\mathsf{PI}}(\mathbf{x}, \delta^\star(\alpha))\right] \geq 1-\alpha,$$

with probability at least $1 - 2m^{-10}$ *on the calibration set* $\{x'_j, y'_j\}_{j=1}^m$.

2.3 | Intuition and proof sketch

We first explain the key intuition before presenting the details of the proof sketch. The proof first leverages a representation theorem that relates the finite-dimensional (kernelized) SDP to

an infinite-dimensional SDP to decouple the dependencies among x_i's. Next, we propose to use empirical process theory to analyse the prediction coverage, inspired by the margin-based analysis originally done in analysing classification. Finally, in controlling the uniform deviations between empirical and population coverage, we use properties of the PSD operators and conditional quantile functions. The additional calibration procedure in fine-tuning the confidence parameter $\delta^\star(\alpha)$ for a fixed coverage level $1-\alpha \in (0,1)$ also hinges on uniform deviation arguments. Our analysis fundamentally relies on empirical process theory and is crucially different from conformal prediction analysis (based on exchangeability). Since the SDP provides a rigorous coverage guarantee as conformal prediction, we hope the new proof idea opens new doors to study uncertainty quantification.

Now we sketch the proof of Theorem 1. Observe that by definition

$$(\text{LHS}) := \mathop{\mathbb{P}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\left[\mathbf{y} \notin \widehat{\mathsf{Pl}}(\mathbf{x},\delta)\right] = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\left[\mathbb{1}\left(\mathbf{y}^{-2}\hat{\mathsf{v}}(\mathbf{x}) < \frac{1}{1+\delta}\right)\right].$$

Define a hinge function $h_\delta(t) : t \mapsto \max\left\{\frac{1+\delta}{\delta}(1-t), 0\right\}$, we have

$$\mathbb{1}\left(t < \frac{1}{1+\delta}\right) \le h_\delta(t), \quad \forall t \in \mathbb{R},$$

and thus

$$(\text{LHS}) \le \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\left[h_\delta(\mathbf{y}^{-2}\hat{\mathsf{v}}(\mathbf{x}))\right]. \tag{8}$$

Define a real positive function (indexed by $\mathbf{A}$) on the data $z=(x,y)$, $f_{\mathbf{A}}(z) : z \mapsto \left\langle \frac{\phi_x}{y}, \mathbf{A}\frac{\phi_x}{y}\right\rangle_{\mathcal{H}}$. Here $\frac{\phi_x}{y} \in \mathcal{H}$ lies in the RKHS, and $\mathbf{A} : \mathcal{H} \to \mathcal{H}$ is a PSD operator. Define a sequence of function spaces according to its nuclear norm radius $\mathcal{F}_k := \{f_{\mathbf{A}} : 2^{k-1} < \|\mathbf{A}\|_\star \le 2^k\}$ for all $k \in \mathbb{N}$ and $\mathcal{F}_0 := \{f_{\mathbf{A}} : \|\mathbf{A}\|_\star \le 1\}$.

With the Proposition 3 establishing the equivalence between the kernelized SDP and the infinite-dimensional SDP, $\widehat{\mathbf{A}} = \sum_{i,j}\widehat{\mathbf{B}}_{ij}\phi_{x_i}\otimes\phi_{x_j}$, we know that $y^{-2}\hat{\mathsf{v}}(x) = f_{\widehat{\mathbf{A}}}(z)$. There exists a $k \in \mathbb{N}$ such that $f_{\widehat{\mathbf{A}}} \in \mathcal{F}_k$ with $2^{k-1} \le \widehat{Opt}_n < 2^k$, and thus we continue to bound

$$\begin{aligned}(\text{LHS}) &\le \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\left[h_\delta \circ f_{\widehat{\mathbf{A}}}(\mathbf{z})\right] \\ &\le \underbrace{\widehat{\mathbb{E}}\left[h_\delta \circ f_{\widehat{\mathbf{A}}}(\mathbf{z})\right]}_{\text{(i)}} + \underbrace{\sup_{f\in\mathcal{F}_k}(\mathbb{E}-\widehat{\mathbb{E}})[h_\delta\circ f]}_{\text{(ii)}}.\end{aligned}$$

For term (i), recall the optimality condition of (1),

$$\langle \mathsf{K}_i, \widehat{\mathbf{B}}\mathsf{K}_i\rangle \ge y_i^2 \Leftrightarrow f_{\widehat{\mathbf{A}}}(z_i) \ge 1$$

which further implies $h_\delta \circ f_{\widehat{\mathbf{A}}}(z_i) = 0$ for all $i = 1,\cdots,n$. Therefore term (i) is zero.

For term (ii), we will use the high probability symmetrization in Proposition 4. Introduce i.i.d. Rademacher variables $\{\epsilon_i\}_1^n$ independent of the data. Note that we only need to consider $k \le k_0$ such that $2^{k_0} = [\log(n)]^{c_\omega}$, where we use the upper estimate on $\widehat{Opt}_n$ obtained in Proposition 5,

which is implied by Assumption [S3]. With probability at least $1-2\exp(-t)$ on the data $\{z_i\}_1^n$, uniformly over all $k \le k_0$

$$\begin{aligned}
\text{(ii)} &\le 2 \cdot \mathop{\mathbb{E}}_{\{\epsilon_i\}_1^n} \sup_{f \in \mathcal{F}_k} \frac{1}{n}\sum_{i=1}^n \epsilon_i \left(h_\delta \circ f\right)(z_i) + \text{(iii)} \\
&\le 2 \cdot \mathrm{Lip}(h_\delta) \cdot \mathop{\mathbb{E}}_{\{\epsilon_i\}_1^n} \sup_{f \in \mathcal{F}_k} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(z_i) + \text{(iii)} \\
&= \frac{2(1+\delta)}{\delta} \mathop{\mathbb{E}}_{\{\epsilon_i\}_1^n} \sup_{\|\mathbf{A}\|_\star \le 2^k} \left\langle \frac{1}{n}\sum_{i=1}^n \epsilon_i \frac{\phi_{x_i}}{y_i} \otimes \frac{\phi_{x_i}}{y_i}, \mathbf{A} \right\rangle + \text{(iii)} \\
&\le \frac{2(1+\delta)}{\delta} 2^k \underbrace{\mathop{\mathbb{E}}_{\{\epsilon_i\}_1^n} \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i \frac{\phi_{x_i}}{y_i} \otimes \frac{\phi_{x_i}}{y_i} \right\|_{\mathrm{op}}}_{\text{(iv)}} + \text{(iii)}
\end{aligned}$$

where the last step follows from the duality between the nuclear norm and operator norm. Before getting into the deviation term (iii) (originated by Proposition 4, formally upper bounded in (15)), first recall $2^k \le 2(\widehat{\mathrm{Opt}}_n \vee 1)$, we know

$$\text{(ii)} \le \frac{4(1+\delta)}{\delta}(\widehat{\mathrm{Opt}}_n \vee 1) \cdot \text{(iv)} + \text{(iii)}. \tag{9}$$

Similarly by Proposition 4, the deviation term (iii) can be bounded by $6(\widehat{\mathrm{Opt}}_n \vee 1) \cdot \sup_{x,y} \left\| \frac{\phi_x}{y} \right\|_{\mathcal{H}}^2 \cdot \sqrt{\frac{k_0+t}{n}}$ with $k_0 = \mathrm{c}_\omega \log\log(n)$.

To bound the expected operator norm for the above random matrix, namely term (iv), we rely on matrix Bernstein's inequality plus a truncation technique. Observe that

$$\mathop{\mathbb{E}}_{\epsilon}\left[\epsilon \frac{\phi_x}{y} \otimes \frac{\phi_x}{y}\right] = 0, \text{ and } \left\|\epsilon \frac{\phi_x}{y} \otimes \frac{\phi_x}{y}\right\|_{\mathrm{op}} \le \sup_{x,y}\left\|\frac{\phi_x}{y}\right\|_{\mathcal{H}}^2 \text{ a.s.},$$

and that

$$\left\| \sum_{i=1}^n \mathop{\mathbb{E}}_{\epsilon}\left[\left(\epsilon_i \frac{\phi_{x_i}}{y_i} \otimes \frac{\phi_{x_i}}{y_i}\right)^2\right] \right\|_{\mathrm{op}} \le \left(\sup_{x,y}\left\|\frac{\phi_x}{y}\right\|_{\mathcal{H}}^2\right)^2 \cdot n.$$

Naively applying the matrix Bernstein inequality, one would expect the term (iv) to behave like $\sup_{x,y}\left\|\frac{\phi_x}{y}\right\|_{\mathcal{H}}^2 \cdot \left(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\right)$ with probability $1 - \dim(\phi_x)\cdot\exp(-t)$ on $\{\epsilon_i\}_1^n$. This is educative yet wrong, since $\dim(\phi)$ is infinity. To make things rigorous, we rely on a truncation technique to look at a finite-dimensional version $\phi_x^{\le m}$ truncated at a level $m = \mathrm{poly}(n)$ to apply matrix Bernstein, and then estimate the remaining contribution from $\phi_x^{>m}$ by the eigenvalue decay in Assumption [S1]. With details given in Proposition 6, we derive that

$$\text{(iv)} \le \mathrm{C}_\tau \cdot \sup_{x,y}\left\|\frac{\phi_x}{y}\right\|_{\mathcal{H}}^2 \cdot \left(\sqrt{\frac{\log(n)}{n}} \vee \frac{\log(n)}{n}\right). \tag{10}$$

The final piece of the puzzle lies in the term $\sup_{(x,y)\in\mathrm{dom}(\mathcal{P})}\|\frac{\phi_x}{y}\|_{\mathcal{H}}^2$, which appears in both the main term (iv) and deviation term (iii). It is not true that a.s. for all x, y, the above term is bounded. To resolve this issue, we rely on a conditional quantile technique. Introduce the conditional quantile function $Q_{\mathbf{y}^2|\mathbf{x}=x}(\cdot) : [0,1] \to \mathbb{R}_+$ for the conditional random variable $\mathbf{y}^2 \,|\, \mathbf{x} = x$. Let's only look at data (x_i, y_i)'s lying in the region

$$\Omega := \{(x,y)|y^2 > Q_{\mathbf{y}^2|\mathbf{x}=x}(1-\eta)\},$$

and denote $\mathcal{P}|_\Omega$ as the conditional distribution of data $(\mathbf{x}, \mathbf{y})$ conditioning on the region Ω. Claim that for any $\widehat{\mathsf{PI}}(\mathbf{x}, \delta)$

$$\mathop{\mathbb{P}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}[\mathbf{y} \notin \widehat{\mathsf{PI}}(\mathbf{x},\delta)] \le \mathop{\mathbb{P}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}|_\Omega}[\mathbf{y} \notin \widehat{\mathsf{PI}}(\mathbf{x},\delta)]. \tag{11}$$

This is based on two facts. First,

$$\begin{aligned}
&\mathbb{P}\left[\mathbf{y}^2 > (1+\delta)\widehat{\mathsf{v}}(x) \,|\, \mathbf{x} = x\right] \\
&\le \frac{\mathbb{P}\left[\mathbf{y}^2 > (1+\delta)\widehat{\mathsf{v}}(x) \vee Q_{\mathbf{y}^2|\mathbf{x}=x}(1-\eta) | \mathbf{x} = x\right]}{\mathbb{P}\left[\mathbf{y}^2 > Q_{\mathbf{y}^2|\mathbf{x}=x}(1-\eta) | \mathbf{x} = x\right]} \\
&= \mathbb{P}\left[\mathbf{y}^2 > (1+\delta)\widehat{\mathsf{v}}(x) \,|\, \mathbf{x} = x, (\mathbf{x},\mathbf{y}) \in \Omega\right].
\end{aligned} \tag{12}$$

regardless of the ordering of $Q_{\mathbf{y}^2|\mathbf{x}=x}(1-\eta)$ and $(1+\delta)\widehat{\mathsf{v}}(x)$. Second, conditioning on Ω does not change the marginal distribution of $\mathbf{x}$ due to the quantile construction, namely $\mathcal{P}_\mathbf{x}|_\Omega \equiv \mathcal{P}_\mathbf{x}$. Marginalizing (12) over x proves the above claim.

The inequality (11) makes the analyses upper bounding (LHS) from (8) and (9) applicable, with the changes: (a) $\mathcal{P}|_\Omega$ replacing $\mathcal{P}$, and (b) $\widehat{\mathbb{E}}$ denoting average over data points inside Ω rather than the whole dataset. With the conditioning on Ω, Assumption [S2] implies $Q_{\mathbf{y}^2|\mathbf{x}=x}(1-\eta) \ge \xi \cdot K(x,x)$, and thus

$$\sup_{(x,y)\in\mathrm{dom}(\mathcal{P}|_\Omega)} \left\|\frac{\phi_x}{y}\right\|_{\mathcal{H}}^2 \le \frac{K(x,x)}{Q_{\mathbf{y}^2|\mathbf{x}=x}(1-\eta)} \le \xi^{-1}. \tag{13}$$

Now, we only need to estimate the effective sample size inside Ω to complete the analyses. By the quantile construction, $\mathbb{P}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}[(\mathbf{x},\mathbf{y}) \in \Omega] = \eta$, a simple Bernstein's inequality asserts that

$$|\{i : (x_i, y_i) \in \Omega\}| > \frac{\eta}{2} \cdot n$$

with probability at least $1 - \exp(-\mathrm{c}_\eta \cdot n)$ on $\{z_i\}_1^n$.

Finally, plug (13) into upper bounds on terms (iii) and (iv), with $\frac{\eta}{2} \cdot n$ replacing n in (10) and (9), we have proved that

$$\text{(LHS)} \le \delta^{-1}(\widehat{\mathsf{Opt}}_n \vee 1)\sqrt{\frac{\mathrm{C}_{\tau,\xi,\eta}\log(n)}{n}} \quad \text{(main term)} \tag{14}$$

$$+ (\widehat{\mathsf{Opt}}_n \vee 1)\sqrt{\frac{\mathrm{C}_{\xi,\eta,\omega}(\log\log(n)+t)}{n}} \quad \text{(deviation term)} \tag{15}$$

with probability at least $1 - \exp(-\mathrm{c}_\eta \cdot n) - 2\exp(-t)$ on $\{z_i\}_1^n$.

3 | NUMERICAL STUDIES

We now study the numerical performance of our procedure.

3.1 | Empirical example: Comparison to conformal prediction

This section compares our SDP methods to the conformal prediction methods, measuring the coverage and statistical efficiency of the prediction bands. We compare five methods on two simulated datasets, including (a) standard prediction band using simple linear regression (SLR), without accounting for heteroscedasticity, as a baseline; (b) SDP prediction bands proposed in this paper, with two specifications of the kernels, denoted as SDP1 and SDP2; (c) conformal prediction bands, including full conformal prediction (CF) and split conformal prediction (SplitCF), see (Lei et al., 2018, Algorithms 1 and 2 respectively). For each method, we report the coverage probability, efficiency statistics (including median length and average length) and finally, the mean squared error (MSE) of the estimated conditional mean function.

We first explain the two simulated datasets. Here the $\mathbf{x} \sim \text{Unif}([-\sqrt{3}, \sqrt{3}])$, and the conditional distribution $\mathbf{y} \,|\, \mathbf{x} = x \sim \epsilon \cdot \sqrt{\mathsf{v}(x)}$ where the independent error ϵ is either drawn from a standard normal $\text{N}(0,1)$ (Figure 2b) or a uniform distribution $\text{Unif}([-\sqrt{3}, \sqrt{3}])$ (Figure 2b). The conditional mean function $\mathsf{m}(x) = 0$. The heteroscedastic variance function scales as $\mathsf{v}(x) = 1 + x + 4x^2$, depending on x. For each simulated dataset, we generate $\{(x_i, y_i)\}_{i=1}^n$ i.i.d. from the above data generating process (DGP), with $n = 50$ as the training data, marked by Blue dots. The test dataset are drawn from the same DGP, with $N = 500$ marked by Red dots. For the SDP1/SDP2, and SplitCF, an independent calibration dataset with $n = 50$ is used. The calibration set is used to choose δ in SDPs as in Algorithm 1, and the homoscedastic conformal bandwidth as in SplitCF. We compare five methods that construct the prediction bands, illustrated by the Blue band, on the Gaussian error dataset in Figure 2a, and Uniform error dataset in Figure 2b. For all methods, the desired coverage is set at $1 - \alpha = 95\%$. For the SDPs in (3) $\gamma = 10$. Table 2 summarizes the coverage, efficiency and estimation error.

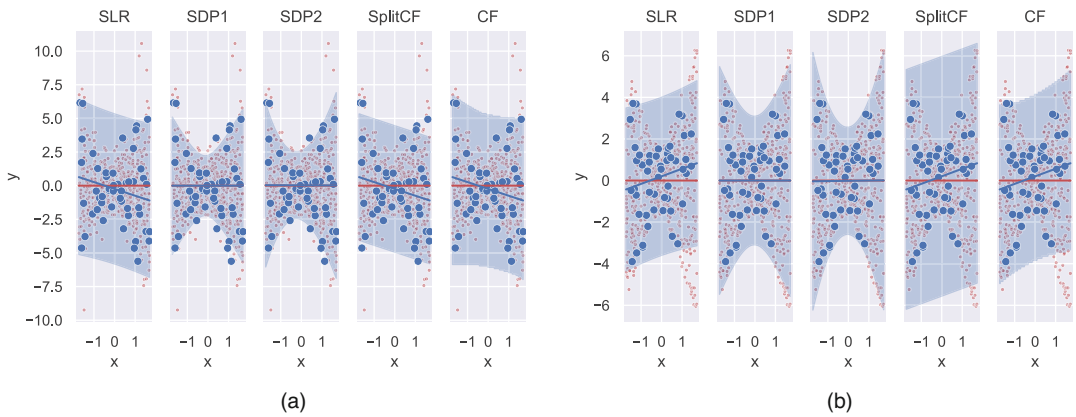


FIGURE 2 From left to right: SLR, SDP1, SDP2, split conformal (SplitCF), and full conformal (CF). Same style as in Figure 1a-b. Statistics are summarized in Table 2 [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Simulated examples

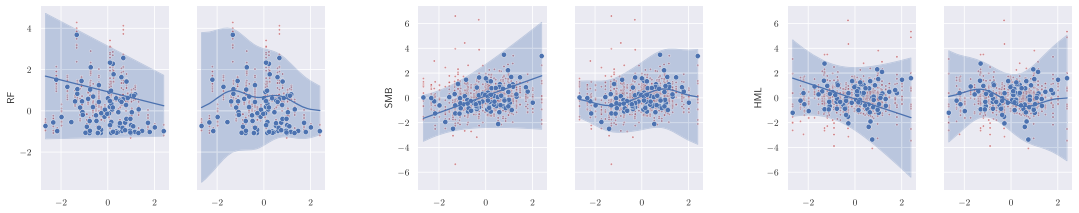
	Coverage	Median length	Average length	MSE
Figure 2a: $m(x) = 0$, quadratic $v(x)$, $\mathbf{y}|\mathbf{x} \sim$ Gaussian				
SLR	97.40%	11.2272	11.2537	0.5554
SDP1	92.80%	6.5172	6.9019	0.0002
SDP2	94.20%	7.0025	7.6900	0.0272
SplitConformal	96.00%	9.3960	9.3960	0.5554
Conformal	97.40%	11.5152	11.5911	0.5554
Figure 2b: $m(x) = 0$, quadratic $v(x)$, $\mathbf{y}|\mathbf{x} \sim$ Uniform				
SLR	89.60%	7.6882	7.7077	0.4405
SDP1	95.40%	7.9161	8.1540	0.0000
SDP2	96.60%	7.3064	7.8175	0.0183
SplitConformal	97.80%	11.5199	11.5199	0.4405
Conformal	92.80%	8.0808	8.1651	0.4405

Nearly all methods achieve 95% desired coverage, with the only exception of SLR. The focus will be on comparing efficiencies, namely, which method estimates a smaller, truly heteroscedastic band in achieving the desired coverage. As seen visually in Figure 2a and b and numerically in Table 2, the two conformal methods, SplitCF and CF, estimates a conservative, wide prediction band that is almost homoscedastic. In contrast, both SDP approaches estimate desirable heteroscedastic bands that are on average much shorter, with the closest (94.20% and 95.40%) to the desired 95% coverage. Finally, we would like to remark that all four methods SLR, SDP1, SDP2 and SplitCF are efficient to compute. Yet, the full conformal method CF involves discretizing input space, which is computationally intensive. Our empirical results show that the two conformal methods can be unnecessarily conservative and form bands of nearly constant width across x (Romano et al., 2019).

3.2 | Real data example: Fama-French factors

In this section, we apply our method of constructing the prediction band to the celebrated three-factor dataset created by Fama and French (1993). We choose this dataset for three reasons: (a) financial data are known to suffer severe heteroscedasticity, (b) the factors are believed to be different sources explaining returns of diversified portfolios, thus when conditioned on one factor, the other factors should have large, heteroscedastic conditional variability and (c) the factors—Market, Size and Value—correspond nicely to our common sense about the financial market for exploratory data analysis.

Let us first explain the data in plain language. The dataset consists of yearly and monthly observations of four variables from July 1926 to December 2020. The four variables are (a) Risk-free return rate (RF), the 1-month Treasury bill rate (i.e. interest rate), (b) Market factor (MKT), the excess return on the market (i.e. market return minus interest rate), (c) Size factor (SMB, Small Minus Big), the average difference in returns between small and big portfolios according to the market capitalization and (d) Value factor (HML, High Minus Low), the difference



(a) RF (b) SMB (c) HML

FIGURE 3 From left to right: response variable y corresponds to RF, SMB and HML, with x being MKT. Blue dots denote $n = 94$ training data $\{(x_i, y_i)\}_{i=1}^n$, Blue line denotes the estimated conditional mean $\hat{\mathsf{m}}(x)$, and Blue band denotes the estimated prediction band $\hat{\mathsf{PI}}(x)$. Red dots represent $N = 1134$ test data points [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Real data: Fama-French

	Kernel	Coverage	Median length	Average length
RF	lin m(x), quad v(x)	98.68%	4.3616	4.4358
RF	rbf m(x), quad v(x)	98.59%	4.5693	4.6847
SMB	lin m(x), quad v(x)	95.77%	5.2560	5.2798
SMB	rbf m(x), quad v(x)	97.53%	5.5407	5.4290
HML	lin m(x), quad v(x)	96.56%	5.2822	5.5556
HML	rbf m(x), quad v(x)	97.27%	4.9180	5.3640

in returns between value and growth portfolios. We design two experiments, one focusing on the prediction coverage and bandwidth, and the other on exploring the role of the tuning parameter γ in trading off mean and variance.

The first experiment aims to access the prediction coverage in the SDP (3), using MKT (as x) to predict other variables (as y): RF and two other factors SMB and HML. Here we use yearly data ($n = 94$ from 1927 to 2020, shown as Blue dots) to construct the prediction bands, each illustrated in Figure 3a–c. As for the test data, we use the standardized monthly data ($N = 1134$, normalized to zero mean and unit standard deviation, shown as Red dots) as a surrogate for test (x,y) pairs. Namely, we match 12 test data to each training data. We verified that after standardization, the histograms of yearly and monthly data match nicely for all four variables. For each type of response variable, we run two SDPs with different kernels. The summary statistics about the coverage probability, median and mean bandwidth are given in Table 3.

We note a few observations regarding the empirical results. First, all models achieve desirable coverage (all >95%). Second, controlling for MKT, all other factors have significant heteroscedastic error left unexplained. For the RF, a high MKT return implies a low expected RF interest, and more importantly, a small variability, compared to the low MKT return case. For the size factor SMB, the conditional variability is much larger when the MKT is high versus low, so does the conditional expectation. The conditional variability in SMB is roughly minimized when the market is significantly below average. While for the value factor HML, conditional variability is minimized when the market is slightly below its average.

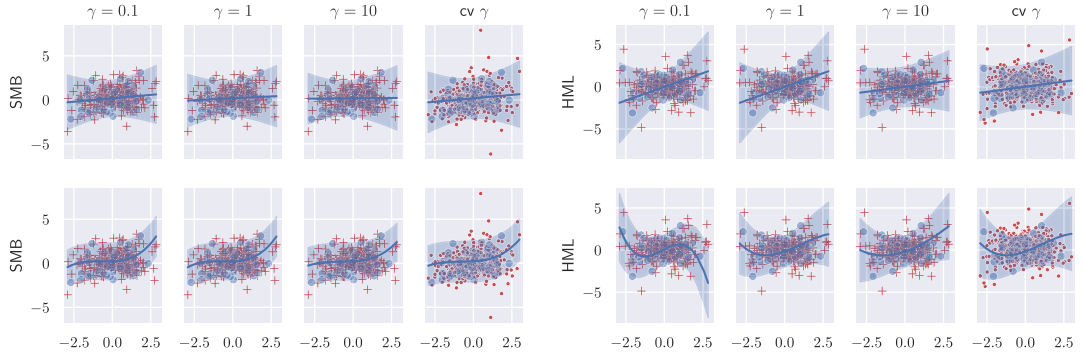


FIGURE 4 Cross-validate γ experiment. On the left is SMB as the response and on the right is HML. For each response variable, we run two sub-experiments: the top row corresponds to a linear $\mathsf{m}(x)$ and a quadratic $\mathsf{v}(x)$, and the bottom row corresponds to a degree-3 polynomial $\mathsf{m}(x)$ and a quadratic $\mathsf{v}(x)$. Each sub-figure corresponds to a specific γ, noted in its title. The cv γ denotes the cross-validated optimal γ using the validation dataset. Here the (train, valid, test) dataset has size proportional to 1:3:9, denoted by Blue dots, Red pluses, and Red dots respectively [Colour figure can be viewed at wileyonlinelibrary.com]

The second experiment aims to verify the mean and variance trade-offs by tuning the parameter γ, discussed in Section 1.3. Here we use the monthly return data, and for each sub-experiment, we split the data into (train, valid, test) parts. We train models with different $\gamma = 0.1, 1, 10$ on the training data, then valid their performances on the validation data. We finally evaluate the performance using the test data with the cross-validated optimal γ (based on the validation data). A nice feature about this experiment is that, one can visualize how the SDPs trade a complex/large conditional variance $\mathsf{v}(x)$ for a simple/small conditional mean $\mathsf{m}(x)$ in explaining $\mathbf{y}|\mathbf{x} = x$ as γ increases, illustrated by Figure 4.

4 | SUMMARY

The current paper progresses to resolve the uncertainty quantification dilemma faced by modern machine learning models. There are two innovative viewpoints we are taking. First, rather than relying on idealized parametric distributional assumptions on error $\mathbf{y} - \mathsf{m}(\mathbf{x})$, we make minimal assumptions. Both the conditional mean and variance functions are modelled nonparametrically and can universally approximate all functions. It is worth noting that such flexibility does not hinder computational feasibility due to the sum-of-squares and convex relaxations. The computational complexity and statistical guarantee scale favourably with high-dimensional covariates x. Second, rather than modelling the conditional mean only and giving up the variance (Frequentist justification, the conditional mean is assumed inside an RKHS, see Caponnetto and De Vito (2007), Liang and Rakhlin (2020), Liang et al. (2020)), or modelling the conditional variance function only (Bayesian justification of kriging/Gaussian processes regression, the covariance function is specified by a kernel, see Handcock and Stein (1993), Stein (2005, 2012)) for the variability in data, we model both the mean and the variance and prove strong, non-asymptotic Frequentist coverage guarantees. Such a modelling advantage enables the uncertainty quantification with or without any black-box predictive model, whether accurate or not.

To conclude, our Theorem 1 established a strong, non-asymptotic coverage guarantee in the language of Neyman, yet with two distinct new features. First, the coverage probability can go

to 1 with a fixed confidence parameter δ as long as the sample size n is large enough. Second, the data-adaptive quantity $\widehat{\mathsf{Opt}}_n$ controls both the average bandwidth and the coverage guarantee of the prediction band $\widehat{\mathsf{PI}}(x)$. A small objective value of the SDP makes the prediction band accurate and narrow simultaneously. Finally, our procedure for constructing prediction bands can be viewed as a novel variance interpolation with confidence and further leverages techniques from semi-definite programming and sum-of-squares optimization. We conducted simulated and real data experiments to validate the prediction interval's numerical performance for uncertainty quantification. A minimal 10-line Python implementation is provided in Listing 1 for interested readers.

ACKNOWLEDGEMENTS

Liang acknowledges the generous support from the NSF Career award (DMS-2042473), the George C. Tiao Fellowship and the William S. Fishman faculty fellowship. Liang thanks Michael Stein, Ruey Tsay and Vladimir Vovk for constructive comments.

ORCID

Tengyuan Liang https://orcid.org/0000-0002-6202-9605

REFERENCES

Bagnell, J.A. & Farahmand, A.-M. (2015) Learning positive functions in a Hilbert space. In *NIPS Workshop on Optimization (OPT2015)*.

Bartlett, P., Freund, Y., Lee, W.S. & Schapire, R.E. (1998) Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5), 1651–1686.

Bartlett, P.L., Long, P.M., Lugosi, G. & Tsigler, A. (2020) Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48), 30063–30070.

Bartlett, P.L., Montanari, A. & Rakhlin, A. (2021) Deep learning: a statistical viewpoint. *arXiv:2103.09177*.

Belloni, A. & Chernozhukov, V. (2011) 1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1), 82–130.

Belloni, A., Chernozhukov, V., Chetverikov, D. & Fernández-Val, I. (2019) Conditional quantile processes based on series or many regressors. *Journal of Econometrics*, 213(1), 4–29.

Caponnetto, A. & De Vito, E. (2007) Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3), 331–368.

Chernozhukov, V., Fernández-Val, I. & Galichon, A. (2010) Quantile and probability curves without crossing. *Econometrica*, 78(3), 1093–1125.

Curmei, M. & Hall, G. (2020) Shape-constrained regression using sum of squares polynomials. *arXiv preprint arXiv:2004.03853*.

Fama, E.F. & French, K.R. (1993) Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56.

Fefferman, C. & Phong, D.H. (1978) On positivity of pseudo-differential operators. *Proceedings of the National Academy of Sciences of the United States of America*, 75(10), 4673.

Ghorbani, B., Mei, S., Misiakiewicz, T. & Montanari, A. (2020) Linearized two-layers neural networks in high dimension. *arXiv:1904.12191*.

Grant, M. & Boyd, S. (2014) CVX: Matlab software for disciplined convex programming, version 2.1. Available from: http://cvxr.com/cvx

Handcock, M.S. & Stein, M.L. (1993) A Bayesian analysis of kriging. *Technometrics*, 35(4), 403–410.

Koenker, R. & Bassett, G. (1978) Regression quantiles. *Econometrica*, 46(1), 33–50. Available from: https://doi.org/10.2307/1913643

Koenker, R. & Hallock, K.F. (2001) Quantile regression. *Journal of Economic Perspectives*, 15(4), 143–156.

Lasserre, J.B. (2001) Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3), 796–817.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J. & Wasserman, L. (2018) Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523), 1094–1111.

Liang, T. & Rakhlin, A. (2020) Just interpolate: Kernel "Ridgeless" regression can generalize. *The Annals of Statistics*, 48(3), 1329–1347.

Liang, T. & Recht, B. (2021) Interpolating classifiers make few mistakes. *arXiv preprint arXiv:2101.11815*.

Liang, T. & Sur, P. (2020) A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586, The Annals of Statistics*, forthcoming.

Liang, T., Rakhlin, A. & Zhai, X. (2020) On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In: *Proceedings of 33rd Conference on Learning Theory*, Volume 125, PMLR, pp. 2683–2711.

Marteau-Ferey, U., Bach, F. & Rudi, A. (2020) Non-parametric models for non-negative functions. *arXiv:2007.03926*.

Montanari, A., Ruan, F., Sohn, Y. & Yan, J. (2020) The generalization error of max-margin linear classifiers: high-dimensional asymptotics in the overparametrized regime. *arXiv:1911.01544*.

Romano, Y., Patterson, E. & Candes, E. (2019) Conformalized quantile regression. In: *Advances in neural information processing systems*, Volume 32, Curran Associates, Inc.

Shafer, G. & Vovk, V. (2008) A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12), 371–421.

Stein, M.L. (2005) Space–time covariance functions. *Journal of the American Statistical Association*, 100(469), 310–321.

Stein, M.L. (2012) *Interpolation of spatial data: some theory for kriging*. Berlin: Springer Science & Business Media.

Vovk, V., Gammerman, A. & Shafer, G. (2005) *Algorithmic learning in a random world*. Berlin: Springer Science & Business Media.

How to cite this article: Liang, T. (2022) Universal prediction band via semi-definite programming. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(4), 1558–1580. Available from: https://doi.org/10.1111/rssb.12542

APPENDIX A

A.1. Remaining propositions

In this section, we collect the remaining propositions.

Proposition 3 (Representation). *The kernelized version of the SDP as in* (3) *is equivalent to the following infinite-dimensional SDP*

$$\begin{aligned} \min_{\beta\in\mathcal{H}^{\mathrm{m}},\,\mathbf{A}:\mathcal{H}^{\mathrm{v}}\to\mathcal{H}^{\mathrm{v}}} \quad & \gamma\cdot\|\beta\|_{\mathcal{H}^{\mathrm{m}}}^2+\|\mathbf{A}\|_\star \\ s.t. \quad & \langle\phi_{x_i}^{\mathrm{v}},\mathbf{A}\phi_{x_i}^{\mathrm{v}}\rangle_{\mathcal{H}^{\mathrm{v}}}\geq(y_i-\langle\phi_{x_i}^{\mathrm{m}},\beta\rangle_{\mathcal{H}^{\mathrm{m}}})^2, \quad \forall i. \\ & \mathbf{A}\succeq 0 \end{aligned}$$

Proof. Noticing that the solution to the infinite-dimensional problem must lie in the span of data, namely $\mathbf{A}=\sum_{i,j}\mathbf{B}_{ij}\phi_{x_i}^{\mathrm{v}}\otimes\phi_{x_j}^{\mathrm{v}}$ with some PSD $\mathbf{B}\in\mathbb{S}^{n\times n}$, and $\beta=\sum_i\alpha_i\phi_{x_i}^{\mathrm{m}}$ with some $\alpha\in\mathbb{R}^n$. With the above representation, plug in the infinite-dimensional SDP and recall $\mathrm{Tr}(\mathbf{A})=\|\mathbf{A}\|_\star$, we can derive (3). When β is not a decision variable, this representation theorem applies to (1). ∎

Proposition 4 (Symmetrization). *Let $\mathcal{F}$ be a class of functions $f : \mathcal{Z} \to \mathbb{R}$, with $\sup_{x\in\mathcal{Z}}|f(z)| \le M$. Then with probability at least $1-2\exp(-t)$ on $\{z_i\}_{i=1}^n$ i.i.d. drawn from a distribution, we have*

$$\sup_{f\in\mathcal{F}}|\mathbb{E}[f(\mathbf{z})] - \widehat{\mathbb{E}}[f(\mathbf{z})]| \le 2\cdot \mathop{\mathbb{E}}_{\epsilon}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(z_i) + 3M\sqrt{\frac{t}{2n}}.$$

Proof. First, with McDiarmid's inequality, we know w.h.p.

$$\sup_{f\in\mathcal{F}}|\mathbb{E}[f(\mathbf{z})] - \widehat{\mathbb{E}}[f(\mathbf{z})]| \le \mathop{\mathbb{E}}_{\{z_i\}_1^n}\sup_{f\in\mathcal{F}}|\mathbb{E}[f(\mathbf{z})] - \widehat{\mathbb{E}}[f(\mathbf{z})]| + M\sqrt{\frac{t}{2n}}.$$

Apply Giné-Zinn symmetrization to the first term on the RHS, then apply McDiarmid's inequality again, we can establish the claim. See Liang and Rakhlin (2020) for details. ■

Proposition 5 (Objective value estimate). *Under* [S3], *the following holds with probability at least* $1-n^{-10}$,

$$\widehat{\mathrm{O}pt}_n \le [\log(n)]^{\mathrm{c}_\omega}.$$

Proof. Apply union bound on the tails given by [S3], with the choice $t_0 = [\log(n)]^{\mathrm{c}_\omega}$ we know

$$\mathbb{P}\left[y_i^2 \le t_0\cdot K(x_i,x_i), \forall 1\le i\le n\right] \ge 1 - n\cdot\exp(-\mathrm{C}t_0^\omega) \ge 1-n^{-10}.$$

In view of Proposition 3, the above certifies that $\mathbf{A} := t_0\cdot\mathbf{I}$ lies in the feasibility set $\langle\phi_{x_i}, \mathbf{A}\phi_{x_i}\rangle_{\mathcal{H}} = t_0\cdot K(x_i,x_i) \ge y_i^2$, which implies t_0 being an upper bound on $\widehat{\mathrm{O}pt}_n$. ■

Proposition 6 (Operator-norm estimate). *Under* [S1], *for any $\{x_i\}_{i=1}^n$, the following holds*

$$\mathop{\mathbb{E}}_{\{\epsilon_i\}_1^n}\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i\phi_{x_i}\otimes\phi_{x_i}\right\|_{\mathrm{op}} \le \mathrm{C}_\tau\cdot\sup_x\|\phi_x\|_{\mathcal{H}}^2\cdot\left(\sqrt{\frac{\log(n)}{n}}\vee\frac{\log(n)}{n}\right).$$

Proof. Recall [S1], due to the Mercer's theorem, one can represent ϕ_x as an infinite-dimensional vector, with each coordinate of ϕ_x^j corresponding to the eigenfunction of the integral operator, with $j=1,\dots,\infty$ and $\lambda_j \le \mathrm{C}j^{-\tau}$. To bound the operator norm, recall the Rayleigh quotient form, for any $h\in\mathcal{H}$ with $\|h\|_{\mathcal{H}}^2 = 1$

$$\left\langle h, \left(\frac{1}{n}\sum_i \epsilon_i\phi_{x_i}\otimes\phi_{x_i}\right)h\right\rangle_{\mathcal{H}} = \frac{1}{n}\sum_i \epsilon_i\langle\phi_{x_i},h\rangle_{\mathcal{H}}^2. \tag{A1}$$

Note $\langle\phi_x,h\rangle_{\mathcal{H}} = \langle\phi_x^{\le m}, h^{\le m}\rangle_{\mathcal{H}} + \langle\phi_x^{>m}, h^{>m}\rangle_{\mathcal{H}}$ where the superscript indicates a truncation on the coordinates of ϕ_x. By Cauchy–Schwarz

$$\langle\phi_{x_i},h\rangle_{\mathcal{H}}^2 \le 2\langle\phi_{x_i}^{\le m}, h^{\le m}\rangle_{\mathcal{H}}^2 + 2\langle\phi_{x_i}^{>m}, h^{>m}\rangle_{\mathcal{H}}^2.$$

Therefore LHS in (A1) can be upper bounded by

$$2\left\|\frac{1}{n}\sum_{i=1}^n\epsilon_i\phi_{x_i}^{\le m}\otimes\phi_{x_i}^{\le m}\right\|_{\mathrm{op}} + 2\sup_i\|\phi_{x_i}^{>m}\|_{\mathcal{H}}^2.$$

For the first term, now we can apply the matrix Bernstein inequality. With probability $1-2\exp(-t)$ the following upper bound on the first term holds

$$\sup_x \|\phi_x\|_{\mathcal{H}}^2 \cdot \left(\sqrt{\frac{\log(m)+t}{n}} \vee \frac{\log(m)+t}{n}\right).$$

For the second term, recall the eigenvalue decay $\lambda_j \leq Cj^{-\tau}$ with $\tau>1$, we know it is upper bounded by $Cm^{-(\tau-1)}$. Choosing $\log(m) = C_\tau \log(n)$ with a constant large enough, we know the second term is dominated by the first term. By integrating the tail bound to obtain a bound on the expected value, we complete the proof. ■

A.2. Proof of the calibration lemma

Proof of Lemma 2. Define the Bernoulli random variables $z_j(\delta) := \mathbb{1}[y'_j \notin \widehat{\mathsf{PI}}(x'_j, \delta)]$, then for any fixed $\delta \in [-1, \Delta]$, we know by Hoeffding's inequality that

$$\left|\mathop{\mathbb{P}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}[\mathbf{y} \notin \widehat{\mathsf{PI}}(\mathbf{x}, \delta)] - \frac{1}{m}\sum_{j=1}^{m}\mathbb{1}[y'_j \notin \widehat{\mathsf{PI}}(x'_j, \delta)]\right| \leq \sqrt{\frac{t}{2m}} \tag{A2}$$

with probability $1-2\exp(-t)$ on $\{x'_j, y'_j\}_{j=1}^m$. Therefore, if we can identify a δ such that

$$\frac{1}{m}\sum_{j=1}^{m}\mathbb{1}[y'_j \notin \widehat{\mathsf{PI}}(x'_j, \delta)] \leq \frac{3}{4}\alpha, \tag{A3}$$

and for some later specified choice of t that

$$\sqrt{\frac{t}{2m}} \leq \frac{1}{4}\alpha, \tag{A4}$$

the proof will complete. Observe that both $\mathbb{P}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}[\mathbf{y} \notin \widehat{\mathsf{PI}}(\mathbf{x}, \delta)]$ and its empirical counterparts $\frac{1}{m}\sum_{j=1}^m \mathbb{1}[y'_j \notin \widehat{\mathsf{PI}}(x'_j, \delta)]$ are monotonic in δ. We claim that the Algorithm 1 will terminate with at most $\log_2\left(\frac{2L(\Delta+1)}{\alpha}\right)$ iterations (namely, disjoint choices of δ in the while loop). To prove this claim, we note that the algorithm must terminate in the interval $\delta \in [\Delta - \frac{\alpha}{2L}, \Delta]$ since we know

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m}\mathbb{1}\left[y'_j \notin \widehat{\mathsf{PI}}\left(x'_j, \Delta - \frac{\alpha}{2L}\right)\right] &\leq \mathop{\mathbb{P}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\left[\mathbf{y} \notin \widehat{\mathsf{PI}}\left(\mathbf{x}, \Delta - \frac{\alpha}{2L}\right)\right] \\ &\quad + \sqrt{\frac{t}{2m}} \leq L\frac{\alpha}{2L} + \frac{1}{4}\alpha \leq \frac{3}{4}\alpha,\end{aligned} \tag{A5}$$

by the mean value theorem and the upper bound on the Lipschitz constant. With the dyadic search, the algorithm will terminate after at most $\left\lceil \log_2\left(\frac{2L(\Delta+1)}{\alpha}\right)\right\rceil$ pre-determined dyadic grids of δ's with the form $G_{\text{dyadic}} := \left\{\left(1-\frac{1}{2^k}\right)\Delta - \frac{1}{2^k} | k = 0, 1, \ldots, \left\lceil \log_2\left(\frac{2L(\Delta+1)}{\alpha}\right)\right\rceil\right\}$.

Therefore, take $t = \log\left(\left\lceil \log_2\left(\frac{2L(\Delta+1)}{\alpha}\right)\right\rceil + 1\right) + 10\log(m)$ and recall that m is large enough such that

$$\sqrt{\frac{\log\left(\left\lceil \log_2\left(\frac{2L(\Delta+1)}{\alpha}\right)\right\rceil + 1\right) + 10\log(m)}{m}} \leq \frac{1}{4}\alpha,$$

then uniformly over the fixed dyadic grid $\delta \in G_{\text{dyadic}}$,

$$\sup_{\delta \in G_{\text{dyadic}}} \mathbb{P}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}[\mathbf{y} \notin \widehat{\mathsf{PI}}(\mathbf{x},\delta)] - \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}[y'_j \notin \widehat{\mathsf{PI}}(x'_j,\delta)] \leq \frac{1}{4}\alpha \tag{A6}$$

with probability at least $1 - 2m^{-10}$. It is easy to see that $\delta^\star(\alpha) \in G_{\text{dyadic}}$, and thus on the same event,

$$\mathbb{P}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\left[\mathbf{y} \notin \widehat{\mathsf{PI}}(\mathbf{x},\delta^\star(\alpha))\right] \leq \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}[y'_j \notin \widehat{\mathsf{PI}}(x'_j,\delta^\star(\alpha))] + \frac{1}{4}\alpha \leq \alpha. \tag{A7}$$

■

A.3. Remaining experimental details

All experiments are conducted using the Python language. The minimal implementation is provided below

Listing 1: Minimal python code

```
import cvxpy as cp

def sdpDual(K1, K2, Y, n, gamma = 1e1):
# K1 kernel for conditional mean, 1st moment
# K2 kernel for conditional variance, 2nd moment
# Define and solve the CVXPY problem.
    # Create a symmetric matrix variable \hat{B}
    hB = cp.Variable((n,n), symmetric=True)
    # Create a vector variable \hat{a}
    ha = cp.Variable(n)

  # PSD and inequality constraints
    constraints = [hB >> 0]
    constraints += [
        K2[i,:]@hB@K2[i,:] >=
        cp.square(Y[i] - K1[i,:]@ha) for i in range(n)
    ]
    prob = cp.Problem(cp.Minimize(
        gamma*cp.quad_form(ha, K1) + cp.trace(K2@hB)
    ), constraints)

    # Solve the SDP
    prob.solve()
    print("Optimal_Value", prob.value)

    return [ha.value, hB.value]
```