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)

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

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 i-th 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

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

Here $\widehat{\mathsf{v}}(x)$ estimates the variability in the "deviations" $e_i := y_i - \mathsf{m}_0(x_i)$. A few remarks on such deviations are in place.

- 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.
- 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.
- 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 witnessed by data. For any $\gamma \geq 0$, the following convex SDP program constructs the prediction band and the conditional mean function simultaneously

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

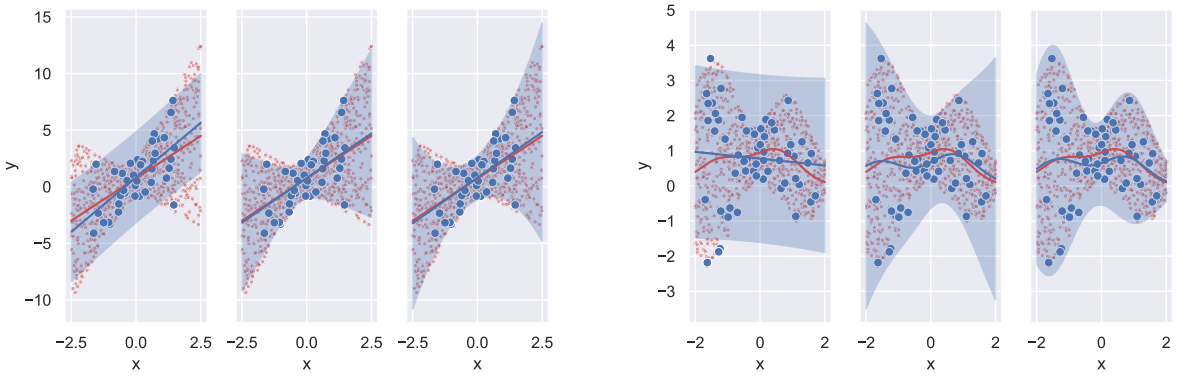
where the optimization variables are $\mathbf{B} \in \mathbb{S}^{n\times n}$ and $\alpha \in \mathbb{R}^n$. Given the solution $\widehat{\mathbf{B}}$ and $\widehat{\alpha}$, the $\widehat{\mathsf{PI}}(x)$ is constructed as

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

1.2 A Numerical Example

Before diving into the motivations behind the above SDPs (Sec. 1.3) and corresponding theory (Sec. 2), let us first illustrate the empirical performance of the constructed prediction bands on a toy numerical example. The

exercise is to showcase that convex programs (1) and (3) are easy to implement using standard optimization toolkits (say, CVX [Grant and 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.



(a) (b)

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.

The first example is a linear model with heteroskedastic error: the conditional mean $\mathsf{m}(x)$ being a linear function and variance $\mathsf{v}(x)$ being a quadratic. 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 Fig. 1a for details and Table. 1 for coverage statistics.

The second example is a non-linear, heteroskedastic 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 Fig. 1b and Table. 1. Here $\gamma = 1$.

These two numerical examples are minimal yet informative. In Fig. 1a, SLR wastes a wide prediction bandwidth on data where the conditional variances are small yet fails to capture the large conditional variance cases, resulting in an overall coverage of 86% and a median bandwidth of 8.21. SDP1/SDP2 redistributes the bandwidth budget leveraging the heteroskedastic 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 Fig. 1b. Observe first that in SDP2, the prediction band constructed by (3) almost perfectly contours the heteroskedastic 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 heteroskedasticity in data, our prediction band distributes the bandwidth in a data-adaptive way, thus improving the overall coverage.

Table 1: Simulated examples

	Coverage	Median Len	Average Len	MSE
Example 1: 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
Example 2: 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**

1.3 Sum-of-Squares and Min-Norm Interpolation

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 Nuclear-norm Minimization As shown in Prop. 2, the infinite-dimensional SDP with a nuclear-norm minimization is equivalent to (3),

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

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 $0 \leq \mathsf{v}(x) = \left(y - \mathsf{m}(x)\right)^2$, we relax the positivity constraints using a sum-of-squares form

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

It turns out that when K^{v} is universal, the above sum-of-squares function can approximate all smooth, positive functions [Fefferman and Phong, 1978, Bagnell and Farahmand, 2015, 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 and Farahmand, 2015, Marteau-Ferey et al., 2020, Curmei and Hall, 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 shatter the data, (3) aims to find the one with minimum bandwidth.

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 [Liang and Rakhlin, 2020,

Bartlett et al., 2020, 2021, Ghorbani et al., 2020, Montanari et al., 2020, Liang and Recht, 2021]. 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 $\mathbf{B} = 0$)

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

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

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

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 and Sur, 2020].

1.4 Other Related Works

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. The elegant theory of conformal prediction, motivated by online learning and sequential prediction, to some extent 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 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 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" vs. "probability" is discussed extensively in Section 2.2 of Shafer and Vovk [2008].

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 neighborhood 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 away from each other in high dimensions, posing challenges to pool the residuals. One can also use either the obtained residuals or the original responses y to fit a conditional quantile regression model [Koenker and Bassett Jr, 1978, Koenker and Hallock, 2001, Chernozhukov and Hansen, 2005], $\widehat{\xi}^{\tau}(\cdot) := \arg\min_{\xi} \frac{1}{n}\sum_{i=1}^{n} \rho_{\tau}\big(y_i - \xi(x_i)\big)$ 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.

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 $\delta \in (0,1]$

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

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 $\sup_{x \in \mathcal{X}} K(x,x) \leq \mathrm{C}$. In addition, eigenvalues of the associated integral operator $\mathcal{T} : \mathcal{H} \to \mathcal{H}$ satisfy $\lambda_j(\mathcal{T}) \leq \mathrm{C} j^{-\tau}, j \in \mathbb{N}$ for some constant $\tau > 1$.
- [S2] (Non-trivial uncertainty) There exist constants $\eta \in (0,1), \xi > 0$ such that $\mathbb{P}\left[\mathbf{y}^2 > \xi \cdot K(\mathbf{x},\mathbf{x}) \mid \mathbf{x} = x\right] > \eta$ holds for all $x \in \mathcal{X}$.
- [S3] (Non-wild uncertainty) There exists a constant $\omega > 0$ such that $\mathbb{P}\left[\mathbf{y}^2 > t \cdot K(\mathbf{x},\mathbf{x})\right] < \exp(-\mathrm{C}t^{\omega})$ 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 $\mathbf{y}|\mathbf{x} = x$. [S3] is the most stringent one, which requires the variability of $\mathbf{y}$ to exhibit a certain tail-decay. Bounded or Gaussian $\mathbf{y}|\mathbf{x} = x$ 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 $\mathbb{P}\left[\mathbf{y}^2 > 0 \mid \mathbf{x} = x\right] = 1$.

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

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

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

$$\mathop{\mathbb{P}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\left[\mathbf{y} \notin \widehat{\mathsf{PI}}(\mathbf{x}, \delta)\right] \leq \delta^{-1}(\widehat{\mathsf{Opt}}_n \vee 1)\sqrt{\tfrac{C_{\tau,\xi,\eta,\omega} \cdot \log(n)}{n}} \ ,$$

$$\textit{and} \ \ \widehat{\mathsf{Opt}}_n \leq \big[\log(n)\big]^{c_{\omega}} \ ,$$

with probability at least $1 - n^{-10}$ on $\{(x_i, y_i)\}_{i=1}^n$. Here the constants $C_{\tau,\xi,\eta,\omega}, c_{\omega}$ 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 $\mathbf{x}, \mathbf{y}$)

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

With $\delta = 0.5$, the bandwidth $\mathrm{Length}\big[\widehat{\mathsf{PI}}(x)\big] = 2.45\sqrt{\widehat{\mathsf{v}}(x)}$ is at a heteroskedastic level adaptive to x with corresponding coverage probability at least $1 - O^\star(\sqrt{\frac{1}{n}})$. 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 $\{(x_i, y_i\}_{i=1}^n$.

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

$$\mathrm{Length}\big[\widehat{\mathsf{PI}}(x)\big] = \big(1 + O^\star(\sqrt{\tfrac{1}{n}})\big) \cdot \sqrt{\widehat{\mathsf{v}}(x)}\,. \tag{6}$$

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

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

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 enables us to 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. Remark that $\widehat{\mathsf{Opt}}_n = \|\widehat{\mathsf{v}}(\cdot)\|_\star^2$ is also a particular norm of the heteroskedastic variance function, quantified by the nuclear norm of the associated PSD operator $\widehat{\mathbf{A}} \succeq 0$ with $\widehat{\mathsf{v}}(x) = \langle \phi_x, \mathbf{A}\phi_x \rangle_{\mathcal{H}}$. Curiously, a simpler variance function $\widehat{\mathsf{v}}(x)$ (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 Proof Sketch

In this section, we sketch the proof of Theorem 1. Observe that by definition

$$(\mathrm{LHS}) := \mathop{\mathbb{P}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\big[\mathbf{y} \notin \widehat{\mathsf{PI}}(\mathbf{x}, \delta)\big] = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}}\big[\mathbb{1}(\mathbf{y}^{-2}\widehat{\mathsf{v}}(\mathbf{x}) < \tfrac{1}{1+\delta})\big]\,.$$

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

$$\mathbb{1}(t < \tfrac{1}{1+\delta}) \le h_\delta(t),\ \forall t \in \mathbb{R}\,,$$

and thus

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

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 Prop. 2 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 $y^{-2}\widehat{\mathsf{v}}(x) = f_{\widehat{\mathbf{A}}}(z)$. There exists a $k\in\mathbb{N}$ such that $f_{\widehat{\mathbf{A}}}\in\mathcal{F}_k$ with $2^{k-1}\le \widehat{\mathsf{Opt}}_n < 2^k$, and thus we continue to bound

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

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 Prop. 3. 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 $2^{k_0} = [\log(n)]^{\mathrm{c}_\omega}$, where we use the estimate on $\widehat{\mathsf{Opt}}_n$ obtained in Prop. 4, which is implied by Assumption [S3]. With probability at leat $1-2\exp(-t)$ on the data $\{z_i\}_1^n$, uniformly over all $k\le k_0$

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

where the last step follows from the duality between the nuclear norm and operator norm. Before getting into the deviation term (iii), first recall $2^k \le 2(\widehat{\mathsf{Opt}}_n \vee 1)$, we know

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

Similarly, by Prop. 3, the deviation term (iii) can be bounded by $6(\widehat{\mathsf{Opt}}_n\vee 1)\cdot\sup_{x,y}\|\frac{\phi_x}{y}\|_{\mathcal{H}}^2\cdot\sqrt{\frac{k_0+t}{n}}$ 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

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

and that

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

Naively applying the matrix Bernstein inequality, one would expect the term (iv) to behave like $\sup_{x,y} \| \frac{\phi_x}{y} \|_{\mathcal{H}}^2 \cdot \left(\sqrt{\frac{t}{n}} \vee \frac{t}{n} \right)$ 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 $\phi_x^{\leq m}$ 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 Prop. 5, we derive that

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

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

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

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

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

This is based on two facts. First,

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

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 whole analyses upper bounding (LHS) from (8)-(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 $Q_{\mathbf{y}^2|\mathbf{x}=x}(1-\eta) \geq \xi \cdot K(x,x)$, and thus

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

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

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

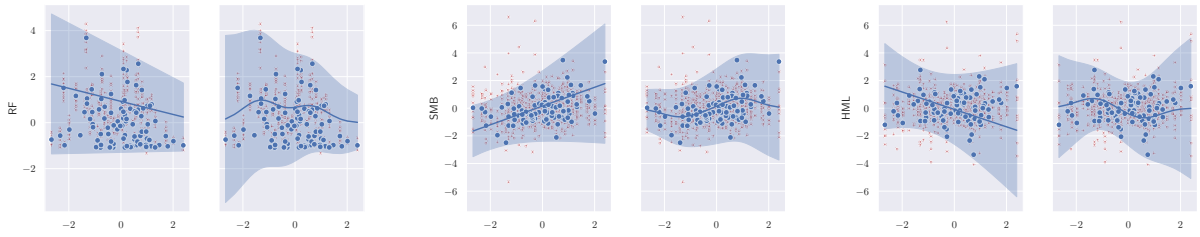
with probability at least $1 - \exp(-\mathrm{c}_\eta \cdot n)$ on $\{z_i\}_1^n$.

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

$$\begin{aligned}(\text{LHS}) \leq \delta^{-1}(\widehat{\mathsf{Opt}}_n \vee 1)\sqrt{\tfrac{\mathrm{C}_{\tau,\xi,\eta}\log(n)}{n}} \quad &\text{(main term)} \\ + (\widehat{\mathsf{Opt}}_n \vee 1)\sqrt{\tfrac{\mathrm{C}_{\xi,\eta,\omega}(\log\log(n)+t)}{n}} \quad &\text{(deviation term)}\end{aligned}$$

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

3 Real Data Example: Fama-French Factors



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

Figure 2: 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 $\widehat{\mathsf{m}}(x)$, and Blue band denotes the estimated prediction band $\widehat{\mathsf{PI}}(x)$. Red dots represent $N = 1134$ test data points.

In this section, we apply our method of constructing the prediction band to the celebrated three-factor dataset constructed by Fama and French [1993]. We choose this dataset for three reasons: (a) financial data are known to suffer severe heteroskedasticity, (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, heteroskedastic 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 one-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 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 1927-2020, shown as Blue dots) to construct the prediction bands, each illustrated in Fig. 2a-2c. 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 2.

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 heteroskedastic error left unexplained. For the RF, a high MKT return implies a low expected RF interest, and more importantly, a small

Table 2: Real data: Fama-French

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

variability, compared to the low MKT return case. For the size factor SMB, the conditional variability is much larger when the MKT is high vs. 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.

The second experiment aims to verify the mean and variance trade-offs by tuning the parameter γ, discussed in Sec. 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. With the cross-validated optimal γ (based on the validation data), we finally evaluate the performance on the test 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 Fig. 3.

4 Summary

The current paper makes progress in resolving 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 $y - f(x)$, we make minimal assumptions. Both the conditional mean and variance functions are modeled in a nonparametric way and 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 scales nicely with high-dimensional covariates x. Second, rather than modeling the conditional mean only and giving up the variance (Frequentist justification, the conditional mean is assumed inside a RKHS, see [Caponnetto and De Vito, 2007, Liang and Rakhlin, 2020, Liang et al., 2020]), or modeling 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 modeling advantage enables the uncertainty quantification with or without any black-box predictive model, be it 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 both 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.

References

Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

C Fefferman and Duong Hong Phong. On positivity of pseudo-differential operators. *Proceedings of the National Academy of Sciences of the United States of America*, 75(10):4673, 1978.

J Andrew Bagnell and Amir-massoud Farahmand. Learning positive functions in a hilbert space. In *NIPS Workshop on Optimization (OPT2015)*, 2015.

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Non-parametric Models for Non-negative Functions. *arXiv:2007.03926*, 2020.

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

Mihaela Curmei and Georgina Hall. Shape-constrained regression using sum of squares polynomials. *arXiv preprint arXiv:2004.03853*, 2020.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "Ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, June 2020. doi: 10.1214/19-AOS1849.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, December 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1907378117.

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: A statistical viewpoint. *arXiv:2103.09177*, 2021.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *arXiv:1904.12191*, February 2020.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv:1911.01544*, 2020.

Tengyuan Liang and Benjamin Recht. Interpolating classifiers make few mistakes. *arXiv preprint arXiv:2101.11815*, 2021.

Peter Bartlett, Yoav Freund, Wee Sun Lee, and Robert E. Schapire. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, October 1998. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1024691352.

Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*, 2020.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Glenn Shafer and Vladimir Vovk. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. ISSN 1533-7928. URL http://jmlr.org/papers/v9/shafer08a.html.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

Victor Chernozhukov and Christian Hansen. An iv model of quantile treatment effects. *Econometrica*, 73 (1):245–261, 2005.

Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, February 1993. ISSN 0304-405X. doi: 10.1016/0304-405X(93)90023-5.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Proceedings of 33rd Conference on Learning Theory*, volume 125, pages 2683–2711. PMLR, July 2020.

Mark S Handcock and Michael L Stein. A bayesian analysis of kriging. *Technometrics*, 35(4):403–410, 1993.

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

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

A Appendix

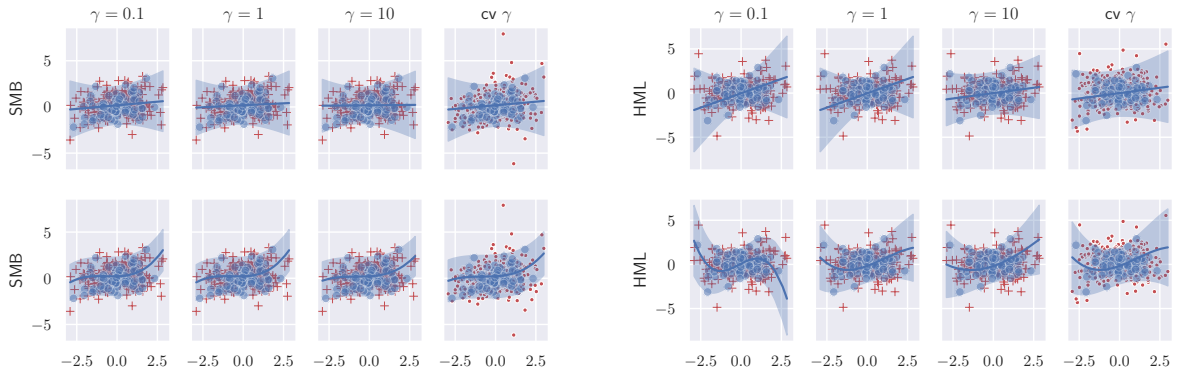


Figure 3: Cross-validate γ experiment. On the left is SMB as the response and on the right is HML. For each response variable, we run two sub-experiments: the top row corresponds to a linear $\mathsf{m}(x)$ and a quadratic $\mathsf{v}(x)$, and the bottom row corresponds to a degree-3 polynomial $\mathsf{m}(x)$ 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.

A.1 Remaining Propositions

In this section, we collect the remaining propositions.

Proposition 2 (Representation). *The kernelized version of the SDP as in (3) is equivalent to the following infinite-dimensional SDP*

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

Proof. Noticing that the solution to the infinite-dimensional problem must lie in the span of data, namely $\mathbf{A}=\sum_{i,j}\mathbf{B}_{ij}\phi_{x_i}^{\mathsf{v}}\otimes\phi_{x_j}^{\mathsf{v}}$ with some PSD $\mathbf{B}\in\mathbb{S}^{n\times n}$, and $\beta=\sum_i\alpha_i\phi_{x_i}^{\mathsf{m}}$ 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 3 (Symmetrization). *Let $\mathcal{F}$ be a class of functions $f:\mathcal{Z}\to\mathbb{R}$, with $\sup_{x\in\mathcal{Z}}|f(z)|\leq M$. Then with probability at least $1-2\exp(-t)$ on $\{z_i\}_{i=1}^n$ i.i.d. drawn from a distribution, we have*

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

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

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

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

$$\widehat{\mathsf{Opt}}_n\leq[\log(n)]^{c_\omega}\ .$$

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

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

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

Proposition 5 (Operator-norm estimate). *Under [S1], for any $\{x_i\}_{i=1}^n$, the following holds*

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

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 $j=1,\ldots,\infty$ and $\lambda_j\leq\mathrm{C}j^{-\tau}$. To bound the operator norm, recall the Rayleigh quotient form, for any $h\in\mathcal{H}$ with $\|h\|_{\mathcal{H}}^2=1$

$$\Big\langle h,\big(\tfrac{1}{n}\sum_i\epsilon_i\phi_{x_i}\otimes\phi_{x_i}\big)h\Big\rangle_{\mathcal{H}}=\tfrac{1}{n}\sum_i\epsilon_i\langle\phi_{x_i},h\rangle_{\mathcal{H}}^2\ . \tag{14}$$

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

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

Therefore LHS in (14) can be upper bounded by

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

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

$$\sup_x\|\phi_x\|_{\mathcal{H}}^2\cdot\Big(\sqrt{\tfrac{\log(m)+t}{n}}\vee\tfrac{\log(m)+t}{n}\Big)\,.$$

For the second term, recall the eigenvalue decay $\lambda_j\le \mathrm{C}j^{-\tau}$ with $\tau>1$, we know it is upper bounded by $\mathrm{C}m^{-(\tau-1)}$. Choosing $\log(m)=\mathrm{C}_\tau\log(n)$ 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 Remaining Experimental Details

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

```
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]
```

Listing 1: Minimal python code