Journal of Machine Learning Research 17 (2016) 1-45 Submitted 8/13; Revised 3/15; Published 4/16

On Quantile Regression in Reproducing Kernel Hilbert Spaces with the Data Sparsity Constraint

Chong Zhang CHONG.ZHANG@UWATERLOO.CA
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, ON N2L 3G1, Canada

Yufeng Liu YFLIU@EMAIL.UNC.EDU
Department of Statistics and Operations Research
Department of Genetics
Department of Biostatistics
Carolina Center for Genome Sciences
The University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA

Yichao Wu WU@STAT.NCSU.EDU
Department of Statistics
North Carolina State University
Raleigh, NC 27695, USA

Editor: Saharon Rosset

Abstract

For spline regressions, it is well known that the choice of knots is crucial for the performance of the estimator. As a general learning framework covering the smoothing splines, learning in a Reproducing Kernel Hilbert Space (RKHS) has a similar issue. However, the selection of training data points for kernel functions in the RKHS representation has not been carefully studied in the literature. In this paper we study quantile regression as an example of learning in a RKHS. In this case, the regular squared norm penalty does not perform training data selection. We propose a data sparsity constraint that imposes thresholding on the kernel function coefficients to achieve a sparse kernel function representation. We demonstrate that the proposed data sparsity method can have competitive prediction performance for certain situations, and have comparable performance in other cases compared to that of the traditional squared norm penalty. Therefore, the data sparsity method can serve as a competitive alternative to the squared norm penalty method. Some theoretical properties of our proposed method using the data sparsity constraint are obtained. Both simulated and real data sets are used to demonstrate the usefulness of our data sparsity constraint.

Keywords: kernel learning, Rademacher complexity, regression, smoothing, sparsity

1. Introduction

Regression is one of the most important and commonly used statistical tools. Given a set of data points whose predictors and responses are both available, one builds a regression model to predict the response variable for any new instance with only predictors observed. When solving a regression problem, linear regression can be insufficient. In particular, when the response has highly nonlinear dependence on the predictors, linear models can be suboptimal. To overcome this difficulty, various

nonlinear regression models such as kernel smoothers (see Hastie et al., 2009, for a review) and splines (De Boor, 1978) can be used. The main idea is to find a regression function in a nonlinear functional class that best fits the response variable.

In a typical spline regression problem with a univariate predictor, one can use a piecewise nonlinear function as the regression function. The function is smooth everywhere including at the knots, where the nonlinear pieces connect. The knots play a crucial role in spline regression. For the regular smoothing splines (see, for example, Wahba, 1990; Gu, 2002, and the references therein), the knots are located at the observed predictor values automatically. For some other types of spline regressions, one needs to determine the knots. For instance, B-splines (De Boor, 1978) commonly use a set of equally spaced knots, and certain types of P-splines (Eilers and Marx, 1996; Ruppert et al., 2003) take knots based on quantiles of the predictor variable.

For spline regression, it is known that too many knots may lead to overfitting and unnecessary fluctuation in the resulting estimator. For instance, based on Chappell (1989), Koenker et al. (1994) gave an example where the regular smoothing splines perform poorly because of too many change points, and the one-change-point method proposed by Chappell (1989) works much better. Extensive efforts have been devoted on how to choose the knots for B-spline and P-spline methods in the literature (see, for example, Friedman and Silverman, 1989; Eilers and Marx, 1996; Zhou et al., 1998; Ruppert, 2002; Hansen and Kooperberg, 2002; Mao and Zhao, 2003; Miyata and Shen, 2005; Gervini, 2006; Eilers and Marx, 2010, and the references therein).

In this paper, we consider multi-dimensional regression problems with the regression function in a Reproducing Kernel Hilbert Space (RKHS, Aronszajn, 1950; Schölkopf and Smola, 2002). This is a very general setting, which includes many well known regression techniques as special cases, for example penalized linear regressions, additive spline models with or without interactions, and the entire family of smoothing splines. Typically, the optimization of such a RKHS regression can be written in a *loss + penalty* form. Since the regression function is assumed to be in a RKHS, it is common to take the squared norm of the function as the penalty. By the well celebrated representer's theorem (Kimeldorf and Wahba, 1971), the resulting regression function can be represented as a linear combination of kernel functions determined by the training data.

Our motivation for this paper is based on the following observation. The kernel representation of the regression function is similar to the knot structure in smoothing splines, in the sense that each observation in the training data can be regarded as a "knot" in a multi-dimensional space. In particular, when we restrict the RKHS regression to the smoothing splines, the kernel function representation is equivalent to the piecewise nonlinear function representation. With the regular squared norm penalty, the resulting estimator involves all kernel functions on the training data. For large sample size problems, this estimator is known to be consistent with desirable theoretical properties. However, for problems with relatively smaller numbers of observations, using all kernel functions for the representation may introduce a similar issue as using too many knots in spline regressions. Hence it is desirable to have a regularization method that can select the kernel functions.

To this end, we propose a new penalty method to achieve a "data sparsity" model. Through simulation studies, we observe that for some cases, the data sparsity model can perform better than the regular squared norm penalty method, and for other cases, their performance is comparable. See Section 2 for more detailed discussions. Moreover, we provide some theoretical insights on the data sparsity method. In particular, we show that under very mild conditions, the asymptotic convergence rates of the estimation errors for the two methods are the same, and both are close to the "parametric rate". Furthermore, we give finite sample error bounds on the prediction errors for both

methods. We show that for a general RKHS and problems with small sample sizes, the bound for the squared norm penalty method can be large. On the other hand, the data sparsity method can enjoy a better bound because its corresponding functional space is smaller. Hence, we propose the data sparsity method as an alternative approach to the squared norm penalty for RKHS learning. Note that in the literature, Takeuchi et al. (2006) studied kernel based nonparametric quantile regression problems, and mentioned a similar method as a natural extension of their formulation. However, their work focused on nonparametric quantile regression, whereas the possible overfitting of the squared norm penalty wasn't brought to attention. Moreover, Takeuchi et al. (2006) didn't perform detailed theoretical or numerical studies on the data sparsity constraint. Our important contribution in this paper is to explore the similarity and (more importantly) differences between the data sparsity constraint and the regular squared norm penalty, through both numerical and theoretical studies.

In a regression problem, one needs to choose the loss function. The commonly used loss function is the squared error loss, which estimates the conditional mean of the response given the predictors. It is known that compared to the conditional mean estimation, the conditional median estimation is more robust against outliers. Therefore, in this paper, we consider quantile regression, and the loss function we use is the check function (Koenker and Bassett, 1978), although the idea of imposing data sparsity constraint is very general and can be applied to many other settings as well. Note that quantile regression with the check function provides the conditional median estimation as a special case. Another advantage is that it can provide more information on the conditional distribution of the response. Quantile regression has been widely used in many scientific fields, including survival analysis (Koenker and Geling, 2001), microarray study (Wang and He, 2007), economics (Koenker and Hallock, 2001), growth chart (Wei and He, 2006), and many others. Note that quantile regression in RKHS with the regular squared norm penalty was previously studied by Takeuchi et al. (2006) and Li et al. (2007).

In the machine learning literature, the Support Vector Machine (SVM, Boser et al., 1992; Cortes and Vapnik, 1995) and the Support Vector Regression (SVR, Drucker et al., 1997; Vapnik et al., 1997; Smola and Schölkopf, 1998; Stitson et al., 1999; Smola and Schölkopf, 2004) have been well studied and widely used as classification and regression tools. One attractive feature of the SVM and SVR is that even with the regular square norm penalty, due to the choice of the loss functions, the estimated classification function or regression function has a sparse representation in the dual space of the corresponding optimization problem. If the classification or regression function is in a RKHS, sparsity in the dual space representation is equivalent to sparsity in the kernel representation. However, with many other loss functions and the squared norm penalty, the advantage of a sparse representation is lost (Smola and Schölkopf, 2004). Our proposed data sparsity constraint is able to provide such a sparse representation for general loss functions. Note that in the Bayesian learning literature, Tipping (2001) proposed the relevance vector machines to obtain sparse solutions for regression and classification problems.

The rest of this article is organized as follows. In Section 2, we first discuss quantile regression problems under the RKHS learning, then introduce our data sparsity constraint. In Section 3, we derive theoretical results for both asymptotic and finite sample analysis of our data sparsity method. In Section 4, we discuss how to derive the solution path of the involved optimization problem with respect to the tuning parameter, and tackle the problem of tuning parameter selection. In Section 5, we demonstrate the performance of our data sparsity method, using both simulated and real data sets. Some discussions are provided in Section 6. All technical proofs are collected in the appendix.

2. Methodology

We are given the training data (x_i, y_i); $i = 1, \ldots, n$, which are observed according to the model $Y = f_0(X) + \varepsilon(X)$. Let D be the domain of f_0, and let the dimensionality of D be p. We assume that for any given X, $\varepsilon(X)$ has a finite mean. This assumption on ε is very general, in the sense that both the homoscedastic and heteroscedastic cases are covered, along with many commonly used distributions. To estimate the $100\tau\%$ quantile of the conditional distribution of Y given X for some quantile level $\tau \in (0,1)$, Koenker and Bassett (1978) proposed to use the check loss function, which can be written as

$$\rho_\tau(u) = \begin{cases} \tau u & \text{if } u > 0, \\ -(1-\tau)u & \text{if } u \leq 0, \end{cases}$$

where $\tau \in (0,1)$ indicates the quantile we are interested in. It is known that for a given X, the population minimizer to the check function is the $100\tau\%$ conditional quantile. For a given τ, suppose that $f_{\text{true}}(X)$ is the population minimizer to the check function. Note that in general $f_0 \neq f_{\text{true}}$. A regular quantile regression problem can be typically formulated in terms of the following optimization

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau\big(y_i - f(x_i)\big) + \lambda J(f), \tag{1}$$

where $\mathcal{F}$ is the functional class we are interested in, $J(\cdot)$ is a penalty on f to prevent overfitting, and λ is a tuning parameter that controls the magnitude of $J(\cdot)$. With $p = 1$, $J(f) = \int |\frac{d^2 f(x)}{dx^2}| dx$, and an appropriately chosen $\mathcal{F}$, Koenker et al. (1994) showed that the solution to (1) is a linear spline with knots at x_i; $i = 1, \ldots, n$.

In this article, we consider the case with $p \geq 1$ and the regression function in a RKHS, which is a more general setting than the regular smoothing splines. To begin with, we introduce some notations. A summary of important notations used in this paper can be found in Tables 11-14. Assume $\mathcal{F} = \{f = f' + b : f' \in \mathcal{H},\ b \in \mathbb{R}\}$, where $\mathcal{H}$ is a RKHS over X with the kernel function $K(\cdot,\cdot)$, and b is the intercept of the regression function. Throughout this paper, we use the notation f' for any function when it belongs to a RHKS without an extra intercept term. This definition of $\mathcal{F}$ allows a more flexible setting than $\mathcal{F} = \mathcal{H}$, because some RKHS's, for example the very popular Gaussian RKHS, do not include non-zero constant functions (Minh, 2010). In this paper, without loss of generality we assume each $f \in \mathcal{F}$ can be uniquely decomposed as $f' + b$. Let the norm in $\mathcal{H}$ be $\|\cdot\|_{\mathcal{H}}$. For more detailed discussions about $\mathcal{H}$ and $\|\cdot\|_{\mathcal{H}}$, we refer the readers to Aronszajn (1950), Wahba (1999), Schölkopf and Smola (2002), Steinwart et al. (2006), Hofmann et al. (2008), Minh (2010), and the references therein. Furthermore, we assume that the RKHS $\mathcal{H}$ is separable, and the kernel function $K(\cdot,\cdot)$ is upper bounded by 1. Our theory can be generalized to the case where $\sup_{X_1,X_2} K(X_1,X_2) < \infty$. Note that a similar assumption was previously used in Steinwart and Scovel (2007) and Blanchard et al. (2008). With a little abuse of notation, we define K to be the n by n matrix $\Big(K(x_i,x_j)\Big)$; $i, j = 1, \ldots, n$, which we call the gram matrix.

The quantile regression with the regular squared norm penalty (Takeuchi et al., 2006; Li et al., 2007) solves

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau\big(y_i - f(x_i)\big) + \lambda \|f'\|_{\mathcal{H}}^2. \tag{2}$$

By the representer's theorem, the solution to (2) can be written as

$$\tilde{f}_n(x) = \tilde{f}'_n(x) + \tilde{b} = \sum_{i=1}^{n} \tilde{\alpha}_i K(x_i, x) + \tilde{b}, \tag{3}$$

where $K(x_i, \cdot)$ is the i^{th} kernel function from the training sample, and $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n)^T$ is the estimated kernel function coefficient vector. In this paper, to clarify notation, we denote the estimated function using the squared norm penalty by $\tilde{f}_n$, and the estimated function using our proposed data sparsity constraint by $\hat{f}_n$. Because $\tilde{f}_n$ possesses such a finite form, (2) can be equivalently written as

$$\min_{\alpha, b} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \{ y_i - \big(\sum_{j=1}^{n} \alpha_j K(x_i, x_j) + b \big) \} + \lambda \alpha^T K \alpha. \tag{4}$$

Li et al. (2007) provided a solution path for (4), with respect to the tuning parameter λ.

For many commonly used kernel functions, we can assume that the gram matrix K is positive definite (Paulsen, 2009). Hence for any α_i; $i = 1, \ldots, n$, the penalty $\alpha^T K \alpha$ in (4) constrains α in an ellipsoid, which does not have any singularity at $\alpha_i = 0$. This is illustrated on the left panel of Figure 1. Note that in the linear learning literature, Fan and Li (2006) discussed the effect of singularity of penalties on variable selection. Similarly, in RKHS regression problems, because the regular squared norm penalty does not have any singularity at $\alpha_i = 0$, it does not perform "kernel function selection". As a result, the estimated $\tilde{\alpha}_i \neq 0$ for all $i = 1, \ldots, n$.

As discussed in Section 1, it is desirable to have a method that can deliver estimators with a sparse kernel function representation. To this end, we propose to penalize directly on the kernel function coefficients α such that some estimated α_i's will be set to 0. The details of our method are as follows. By the representer's theorem (Kimeldorf and Wahba, 1971), the estimated $\tilde{f}_n$ in (4) lies in a space linearly spanned by $K(x_i, \cdot)$; $i = 1, \ldots, n$, and $\tilde{\alpha}$ is constrained in an ellipsoid. To obtain a data-sparsely represented function, we propose to constrain α in an L_1 ball. In other words, we solve the following optimization problem with the data sparsity constraint

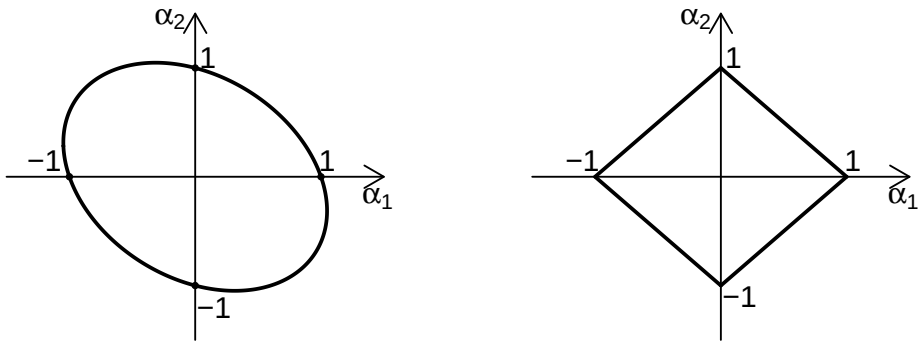
$$\min_{\alpha, b} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \big(y_i - f(x_i) \big), \text{ subject to } |b| + \sum_{i=1}^{n} |\alpha_i| \leq s, \tag{5}$$

where $s > 0$ is the tuning parameter. Note that Takeuchi et al. (2006) briefly mentioned a possible natural extension of their method that is similar to (5).

The constraint in (5) is an L_1 type regularization and imposes a soft thresholding (Tibshirani, 1996) on α. On the right panel of Figure 1, we illustrate the effect of the data sparsity constraint. For a small s, many of the estimated $\hat{\alpha}_i$ values will be set to 0. Consequently, the regression function $\hat{f}_n$ has a parsimonious representation in (3).

Through simulation studies, we demonstrate that for some settings, the data sparsity method can have a better performance, compared to the regular squared norm penalty method. For other cases, the performance difference between the two approaches is small. In particular,

- when n is small or moderate, and the underlying function can be well approximated by a sparse representation $\sum_{i=1}^{m} \gamma_i K(z_i, \cdot) + c$ for small m's, where z_i are fixed points in D and $c, \gamma_i \in \mathbb{R}$. The data sparsity can then provide a parsimonious model, and the corresponding prediction performance can be better. We demonstrate this issue using an example where $p = 1$ with the Laplacian kernel in Figure 2, and another example with $p = 2$ and the Gaussian kernel in Figure 3;



(a) The regular squared norm penalty.

(b) The proposed data sparsity constraint.

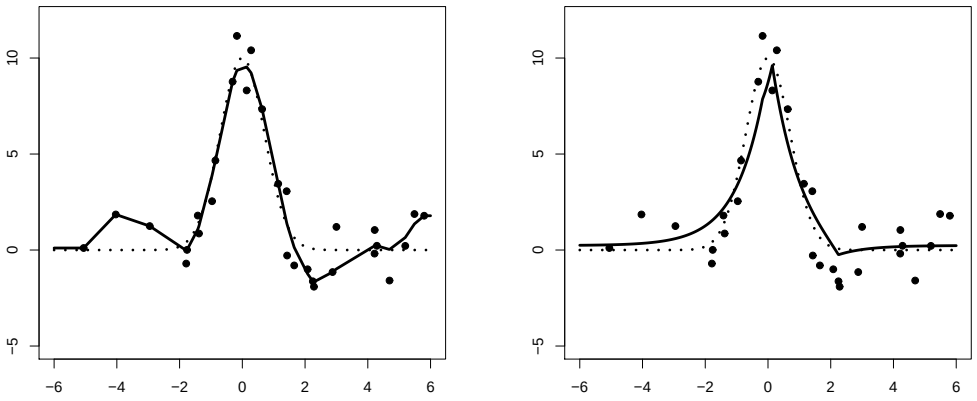
Figure 1: The left panel demonstrates the contour of the regular squared norm penalty $\alpha^T K\alpha = 1$ with $K = \left[(1,0.3)^T\ (0.3,1)^T\right]$. The right panel plots the contour of our data sparsity constraint $|\alpha_1| + |\alpha_2| = 1$. For the regular squared norm penalty, there is no singularity at the intersections of the contour and the axes ($\alpha_1 = 0$, $\alpha_2 = 0$), thus it does not encourage sparsity in the estimated kernel function coefficients. In contrast, the data sparsity penalty has singularity at the intersections and is able to achieve sparsity in the estimated kernel function coefficients.

- when n is small or moderate, and the underlying function does not possess such a sparse representation, the data sparsity method tends to choose a large s in (5). As a result, the fitted model is not sparsely represented. In this case, the prediction performance of the data sparsity method and the squared norm penalty method is often comparable. This is illustrated in Figure 4;

- and when n is large, there is enough information to estimate the underlying function accurately, and both methods can perform well in terms of prediction. In particular, we show in Section 3.1 that the estimation errors of (4) and (5) both converge at a rate very close to the "parametric rate" $O_P(n^{-1/2})$. In other words, asymptotically the data sparsity method can perform as well as the squared norm penalty. However, (5) can still provide a data sparse representation model. The advantages of such a parsimonious estimator is that the prediction for new observations can be much faster than the regular method, and a sparser model can be easier to interpret.

Therefore, the data sparsity method can be regarded as an alternative learning technique to the regular squared norm penalty method.

Notice that although we observe that the data sparsity constraint may work well under some settings when n is small or moderate, we still need a certain amount of information from the data



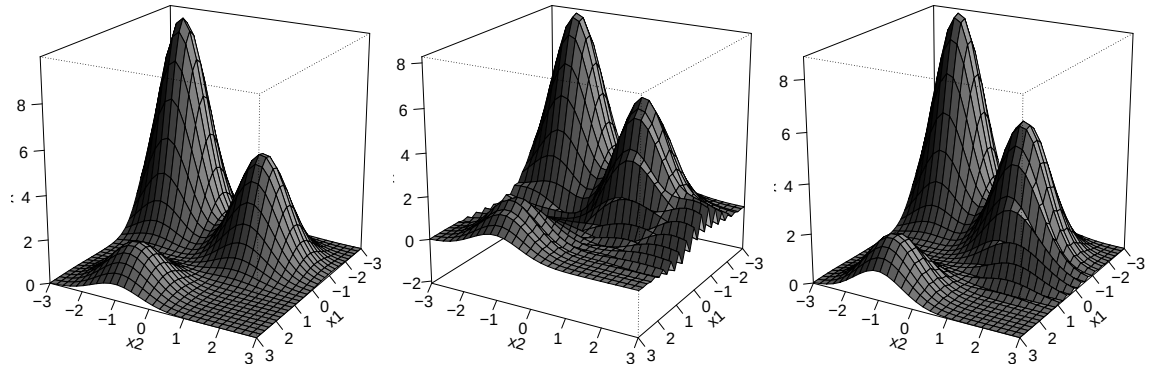
(a) The fitted $\tilde{f}_n$ from (4).

(b) The fitted $\hat{f}_n$ from (5).

Figure 2: Plot of the fitted $\tilde{f}_n$ and $\hat{f}_n$ (solid lines) in a simulated example with $n = 30$ and $\tau = 0.5$, using the regular squared norm penalty (left panel) and the proposed data sparsity constraint (right panel). The kernel used is the Laplacian kernel. The error ε follows $U(-2,2)$. The best tuning parameters (λ, s and the kernel parameter) are selected by the GACV criterion (Yuan, 2006). The dotted line is the true regression function. Note that as the regular squared norm does not have sparsity in the estimated kernel function coefficients, the estimated regression function has quite a few wiggles which degrades the prediction performance. On the other hand, our data sparsity method performs remarkably well in this example.

to estimate the underlying function reasonably well. When the dimensionality p is high and the sample size n is small, without variable selection, the curse of dimensionality would prevent most of the kernel methods from working well. Therefore, we focus on the case when p is not large in this paper.

We would like to point out that besides the data sparsity constraint on α, we also impose regularization on b in (5). Although penalizing b may not be standard, some papers, for example Fan et al. (2008), also considered penalizing the intercept. The effect of penalizing on b is two fold. Firstly, it guarantees the uniqueness of the solution path with respect to s, which is discussed in Section 4. Secondly, it prevents b from diverging too fast as $n \to \infty$. This helps to bound the complexity of $\mathcal{F}$ and the functional space we consider in (5), and consequently helps to derive the theoretical properties in Section 3. If we remove b from the constraint, more conditions are needed for the corresponding theorems to be valid. In particular, in the RKHS learning literature, many theoretical results are derived with $f = f' \in \mathcal{H}$ without the intercept term. See, for example, Bousquet and Elisseeff (2002), Chen et al. (2004), and the discussion on Page 17 of Steinwart and Christmann (2008). Our data sparsity constraint can naturally incorporate the intercept in the regularization, and consequently provide desirable theoretical properties without additional assumptions. More discus-



(a) The underlying $f_0(X)$. (b) The fitted $\tilde{f}_n$ from (4). (c) The fitted $\hat{f}_n$ from (5).

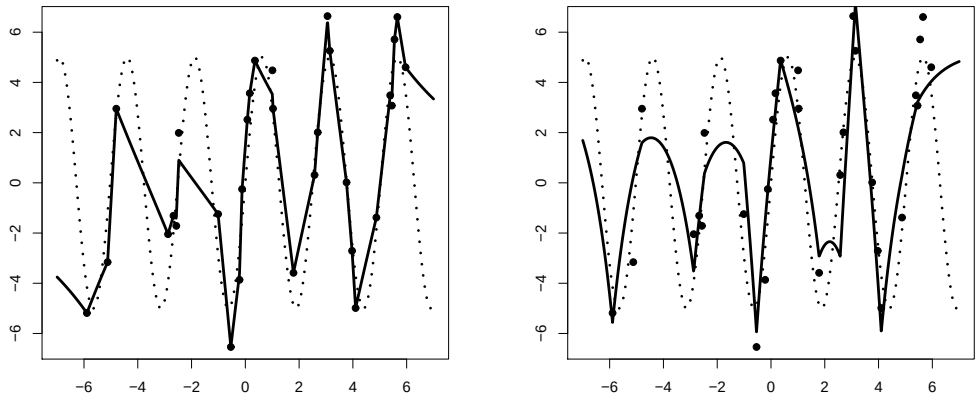
Figure 3: Panel (a) displays the underlying function $f_0(X)$, which has a sparse representation in the Gaussian RKHS. Panel (b) shows the estimated regression function $\tilde{f}_n$ from (4), using a simulated example of size 50 with the Gaussian RKHS and the regular square norm penalty. Panel (c) shows the estimated regression function $\hat{f}_n$ from the same example with our data sparsity constraint in (5). The error ε follows $N(0,1)$, and we use $\tau = 0.5$. We select the best tuning parameters (λ, s and the kernel parameter) over a grid of candidates by the GACV criterion (Yuan, 2006). On Panel (b), one can see that there are fluctuations in $\tilde{f}_n$ which degrade its prediction performance. On the other hand, $\hat{f}_n$ from our data sparsity method has less fluctuations.

sions on this issue are provided in Remark 4 and the proofs of the corresponding theorems in the appendix. In Section 5, we study the numerical performance of our method with and without $|b|$ penalized, respectively. The results demonstrate that the difference between the two settings is small, in terms of their empirical performance. In real data analysis, practitioners can choose whether to penalize b or not based on the nature of the problem and the model.

Next, we derive some theoretical properties of our data sparsity method, as well as the regular squared norm penalty method.

3. Statistical Theory

In this section, we investigate some statistical theory of our data sparsity method. In particular, we study the asymptotic behavior of our data sparsity method and the standard penalized quantile regression with the regular squared norm penalty as $n \to \infty$ in Section 3.1. An example is given in Section 3.2 to calculate the rate of convergence of the approximation error. We discuss the approximation ability of the RKHS in Section 3.3. We also derive some finite sample error bounds in Section 3.4. Note that our main results (Theorems 1-9) require only that the noise $\varepsilon(X)$ has finite mean for all X, therefore they hold in general for both homoscedastic and heteroscedastic cases.



(a) The fitted $\tilde{f}_n$ from (4).

(b) The fitted $\hat{f}_n$ from (5).

Figure 4: The left panel shows the fitted $\tilde{f}_n$ from (4) with a simulated example of size 30 and the Laplacian kernel for $\tau = 0.5$. The right panel shows the fitted $\hat{f}_n$ from (5) using the same sample data and τ. The error follows $N(0,1)$. The best tuning parameters (λ, s and the kernel parameter) are selected by minimizing the GACV criterion (Yuan, 2006). Because the underlying function is quite wiggly, a sparse representation in this case cannot perform well. However, by allowing a large s, the data sparsity constraint yields a model that is not "sparse". Therefore one can see that the two methods give comparable performance.

3.1 Asymptotic Results

Before stating our main results, we introduce some additional notations. Let $\mathcal{F}_n(s) = \{f = f' + b : f'(x) = \sum_{i=1}^n \alpha_i K(x, x_i); |b| + \sum_{i=1}^n |\alpha_i| \le s\}$ be the functional space of the optimization problem (5). Note that we can define the functional space of the regular squared norm penalized method in a similar manner. Let $\mathcal{F}(\infty) = \lim_{s\to\infty} \lim_{n\to\infty} \mathcal{F}_n(s)$. Next, we define $f_n^{(s)} = \text{arginf}_{f\in\mathcal{F}_n(s)} E\rho_\tau(Y - f(X))$ and $f^{(\infty)} = \text{arginf}_{f\in\mathcal{F}(\infty)} E\rho_\tau(Y - f(X))$. Here the expectation is taken with respect to the joint distribution of X and the noise ε. Note that the conditional $100\tau\%$ quantile function f_{true} may not belong to $\mathcal{F}(\infty)$. Now let $e(f_1, f_2) = E\rho_\tau(Y - f_1(X)) - E\rho_\tau(Y - f_2(X))$. In the following theorem, we explore the convergence rate of $e(\hat{f}_n, f^{(\infty)})$ by decomposing it into the estimation error $e(\hat{f}_n, f_n^{(s)})$ and the approximation error $e(f_n^{(s)}, f^{(\infty)})$, where $\hat{f}_n = \text{argmin}_{f\in\mathcal{F}_n(s)} \frac{1}{n}\sum_{i=1}^n \rho_\tau(y_i - f(x_i))$. We also study the convergence rate of $e(\tilde{f}_n, f^{(\infty)})$ for the regular method using the squared norm penalty.

Theorem 1 *For the data sparse L_1 method (5), we have $e(\hat{f}_n, f^{(\infty)}) = O_P(\max(sn^{-1/2}\log(n), d_{n,s}))$, where $d_{n,s} = e(f_n^{(s)}, f^{(\infty)})$ is the approximation error between $\mathcal{F}_n(s)$ and $\mathcal{F}(\infty)$.*

For the regular squared norm method with $|b|$ penalized, the estimation error of the solution $\tilde{f}_n$ to

$$\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\big(y_i - f(x_i)\big), \text{ subject to } |b|^2 + \|f'\|^2_{\mathcal{H}} \le s^2 \tag{6}$$

enjoys the same rate as above.

Remark 2 *In Theorem 1, the estimation error converges at the rate $O_P\big(sn^{-1/2}\log(n)\big)$, and $d_{n,s}$ approaches 0 as $s,n\to\infty$. Therefore the optimal value of the tuning parameter s is roughly determined by $sn^{-1/2}\log(n)\approx d_{n,s}$. It can be considered as the trade-off between the approximation error and the estimation error.*

Remark 3 *Theorem 1 is developed for a general separable RKHS such that the kernel function $K(\cdot,\cdot)$ is upper bounded. The results in Section 3 can be refined if one focuses on a smaller set of RKHS's that satisfies additional conditions. For example, the Gaussian RKHS is commonly used in the literature, and the corresponding theoretical properties are well studied (see, for example, Zhou, 2002; Keerthi and Lin, 2003; Steinwart et al., 2006; Steinwart and Scovel, 2007; Minh, 2010, and the references within). In Theorem 1, the width parameter σ of the Gaussian kernel and the dimensionality of X do not affect the convergence rate explicitly. They are both involved in the approximation error $d_{n,s}$. The choice of σ is often data dependent, as described in Section 5. The asymptotic effect of σ is studied in many papers, for example Keerthi and Lin (2003). If f is an element in a Banach space whose norm is defined to be $\|f_1 - f_2\| = |(E\rho_\tau(Y - f_1(X)) - E\rho_\tau(Y - f_2(X)))|$ with an appropriate definition of limits for Cauchy sequences, then the corresponding theory of $d_{n,s}$ can be derived as studied in Cucker and Smale (2002) and Smale and Zhou (2003). In Section 3.2, we give a simple example in which $d_{n,s}$ can be explicitly calculated and vanishes in a rate much faster than the estimation error.*

Remark 4 *The constraint on $|b|$ in (5) and (6) helps to bound the complexity of $\mathcal{F}_n(s)$ in terms of its L_2 entropy number. The definition of the L_2 entropy number is as follows. Let Q be a σ-finite measure on X. One can define the $L_2(Q)$ norm of a square integratable function f on X to be $\|f\|_{L_2(Q)} = (\int f^2 dQ)^{1/2}$. An $\eta-$net on $\mathcal{F}_n(s)$ is defined to be a set of functions $\mathcal{G} = \{g_1, g_2, \ldots\}$ such that for all $f\in\mathcal{F}_n(s)$, there exists a $g\in\mathcal{G}$ satisfying $\|f-g\|_{L_2(Q)}\le\eta$. For any fixed η, the $L_2(Q)$ entropy number of $\mathcal{F}_n(s)$ is defined as the logarithm of the cardinality of an $\eta-$net $\mathcal{G}$ on $\mathcal{F}_n(s)$ whose size is the smallest (Van der Vaart and Wellner, 2000). A bound on $|b|$ helps to control the L_2 entropy number of $\mathcal{F}_n(s)$. See Lemma 14 and its proof in the appendix for more details. Our theory can also be valid with some additional assumptions if $|b|$ is not penalized. The next corollary discusses a natural generalization of our asymptotic results without penalizing $|b|$, when we impose some assumptions on f_0 and ε. First, we define*

$$\mathcal{F}_n^*(s) = \{f = f' + b : f'(x) = \sum_{i=1}^{n}\alpha_i K(x,x_i); \sum_{i=1}^{n}|\alpha_i| \le s\},$$

and $f_n^{(*s)}$, $\mathcal{F}^*(\infty)$, and $f^{(*\infty)}$ are defined analogously as in Theorem 1. Note that if a random variable X is sub-Gaussian with the parameter S, then $pr(|X|>t)\le 2\exp(-t^2/S)$ for t large enough.

Corollary 5 *Suppose that f_0 is uniformly bounded, and the error $\varepsilon(X)$ follows a sub-Gaussian distribution with a common finite parameter for any X. Then the solution $\hat{f}_n^*$ to the following optimization*

$$\min_{\alpha,b} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\big(y_i - f(x_i)\big), \text{ subject to } \sum_{i=1}^{n} |\alpha_i| \le s,$$

satisfies that $e(\hat{f}_n^*, f^{(*\infty)}) = O_P(\max(sn^{-1/2}\log(n), d_{n,s}^*))$, where $d_{n,s}^* = e(f_n^{(*s)}, f^{(*\infty)})$ is the approximation error between $\mathcal{F}_n^*(s)$ and $\mathcal{F}^*(\infty)$.

In Corollary 5, we impose the assumption that the distributions of error terms ε_i; $i = 1, \ldots, n$, are all sub-Gaussian, which covers many commonly used distributions. Consequently, the probability that any observed y_i being significantly away from $f_0(x_i)$ can be well controlled. If the distribution of ε_i has a heavier tail than sub-Gaussian, and we do not penalize b, the convergence rate of the estimation error can be slower than that in Theorem 1. See the proof in the appendix for more discussions. Compared to Corollary 5, our asymptotic theory with $|b|$ in the constraint only requires the noise ε being integrable, hence is more general.

Remark 6 *Note that our results in Theorem 1 also apply to the regular squared norm penalty method. This suggests that asymptotically, the two methods can both perform well. However, for problems with a moderate or small n, asymptotic results may be less useful. In Section 3.4, we give bounds on the prediction errors $E\rho_\tau(Y - \hat{f}_n)$ and $E\rho_\tau(Y - \tilde{f}_n)$. In particular, we give two such bounds. The first bound works for both the regular method with the squared norm penalty and the proposed data sparsity method. The second bound is only for the data sparsity method. We show that for a small n, the second bound can be better than the first one. Therefore, when the true function can be well approximated by functions in $\mathcal{F}_m(s_0)$ for small m and s_0, the data sparsity method can enjoy a smaller prediction error bound.*

In the next section, we give an example where $f_0 \in \mathcal{F}_m(s_0)$ for some fixed s_0 and m, and the approximation error $d_{n,s}$ converges to 0 in a speed much faster than $O_P(n^{-1/2}\log(n))$. In that case, $e(\hat{f}_n, f^{(\infty)}) = O_P(n^{-1/2}\log(n))$.

In the literature, Takeuchi et al. (2006) derived finite sample bounds on the estimation error for general quantile regression problems, using the Rademacher complexity technique (Mohri et al., 2012). They showed that the estimation error can be upper bounded by the Rademacher complexity of the corresponding functional space (plus a small penalty which exists only in finite sample problems). Under various settings where the Rademacher complexity is well studied, one can obtain the asymptotic convergence rate of the estimation error accordingly. For example, when we perform learning with a radial basis function kernel such that $K(\cdot,\cdot)$ is upper bounded, or when the functional space has finite VC dimensions, one can verify that the corresponding Rademacher complexity converges to zero at a rate close or equal to $O_P(n^{-1/2})$. Li et al. (2007) also studied the asymptotic convergence rate of $e(\tilde{f}_n, f^{(\infty)})$ under some assumptions. For example, a bound on the complexity of $\mathcal{F}$ in terms of the L_2 metric entropy was assumed. Our asymptotic theory for quantile regression with the data sparsity constraint is more general, as we only use the assumption that the RKHS is separable and bounded. This is a very weak assumption and can be satisfied by most commonly used kernel spaces. Furthermore, in Section 3.4, we obtain finite sample error bounds on the prediction error for our proposed method with the data sparsity constraint. Our bounds can be directly calculated using the training data and the corresponding tuning parameter.

3.2 An Illustrative Example on the Approximation Error

In this section we give an example to calculate $d_{n,s}$, where we know f_0 and the distribution of X. The Gaussian RKHS is considered.

For simplicity, let $p = 1$, $\tau = 1/2$, $\sigma = 1$ and X be uniformly distributed on $[0,1]$. Moreover, assume that ε is symmetric with respect to 0 for all X. Suppose the underlying model is $f_{\text{true}} = f_0(x) = \exp(-x^2)$. One can verify that when s is fixed at 1, $d_{n,s} \leq \frac{1}{2}E(|Y - f_0(X)|) - E(|Y - f_{(1)}(X)|)$, where $f_{(1)}(x) = \exp(-(1-x_{(1)})^2)$ with $x_{(1)}$ being the smallest order statistic of the sample $x = (x_1, \ldots, x_n)$. Note that the probability density function of $x_{(1)}$ is $n(1-x)^{n-1}I_{[0,1]}$ and $d_{n,s} \leq \frac{1}{2}E(|f_0(X) - f_{(1)}(X)|)$. Since the largest difference between $f_0(x)$ and $f_{(1)}(x)$ occurs at $x = 0$, $d_{n,s} \leq \frac{1}{2}\int_0^1 (1 - \exp(-x^2))n(1-x)^{n-1}dx$. By the Taylor's expansion, one can verify that $(1 - \exp(-x^2)) \leq 2x^2$ for all $x \in [0,1]$. Thus, $d_{n,s} \leq \frac{1}{2}\int_0^1 2nx^2(1-x)^{n-1}dx = \frac{2}{(n+1)(n+2)}$. Hence in this example, $d_{n,s} = O_P(n^{-2})$, which converges to 0 at a much faster rate than $O_P\big(n^{-1/2}\log(n)\big)$.

In general, when $f_0(x) = \sum_{j=1}^m \gamma_j K(x, z_j) + c$, where $c \in \mathbb{R}$, $\gamma_j \in \mathbb{R}$, and $z_j \in \mathbb{R}^p$ are fixed points (not the observed data points), we can have s fixed and the approximation error vanishes quickly as $n \to \infty$. This is because with growing n, there will be some observed x_i's that are close to z_j; $j = 1, \ldots, m$, and the approximation error $d_{n,s}$ may converge at a rate faster than $O_P(n^{-1/2}\log(n))$ with $s = |c| + \sum_{j=1}^m |\gamma_j|$.

3.3 Approximation Ability of $\mathcal{F}(\infty)$

We have explored the convergence rate of the estimation errors $e(\hat{f}_n, f_n^{(s)})$ and $e(\tilde{f}_n, f_n^{(s)})$ in Section 3.1, and in Section 3.2 we have given an example to illustrate the convergence rate of the approximation error $d_{n,s} = e(f_n^{(s)}, f^{(\infty)})$. For real applications, it is desirable to study the approximation ability of $\mathcal{F}(\infty)$. In other words, how well $f^{(\infty)}$ can approximate f_{true}. However, this approximation ability depends on the properties of f_{true} (i.e., the smoothness, etc.), the richness of the RKHS, and the underlying marginal distribution of X. In the literature of RKHS learning, Steinwart and Scovel (2007) studied the approximation ability of the Gaussian RKHS for support vector machines. In this section, we provide a discussion on this issue for quantile regression with a general RKHS. We measure such approximation ability by $A(\infty) := E\big(\rho_\tau(Y - f^{(\infty)})\big) - E\big(\rho_\tau(Y - f_{\text{true}})\big)$. We first show that if f_{true} is a bounded piecewise step function, an upper bound on $A(\infty)$ for the Gaussian kernel learning can be obtained. This upper bound depends on the marginal distribution of X. In particular, when the marginal distribution of X is absolutely continuous with respect to the Lebesgue measure, the Gaussian RKHS can approximate f_{true} arbitrary well. Then, we extend to the case where f_{true} is a Lipschitz function. Finally, we note that this upper bound can be generalized to other kernels satisfying certain conditions.

We begin with the description of f_{true} and some further notations. Recall the definition of D, and without loss of generality assume $D = \mathbb{R}^p$. As mentioned above, we assume that f_{true} is a bounded step function. In particular, assume $f_{\text{true}} = a_i$ on D_i, where a_i is a constant, D_i is a measurable set in D, and $D = \bigcup D_i$. Let $a > 0$ be the upper bound of $|f_{\text{true}}|$. Next, for any $x \in D_i$, define the distance of x to other D_j; $j \neq i$ as $\psi_x = \min_{j \neq i} \text{dis}(x, D_j)$, where $\text{dis}(x, D_j) = \inf_{x' \in D_j} \|x - x'\|$ and $\|\cdot\|$ is the usual Euclidean norm in D. By this definition of ψ_x, one can verify that $B(x, \psi_x) \in D_i$ for all x, where $B(x, \psi_x)$ is the ball centered at x with the radius ψ_x. Note that $B(x, \psi_x)$ is well defined for all $x \in D$. Recall that σ is the kernel parameter of the Gaussian RKHS.

The next theorem gives an upper bound on $A(\infty)$.

Theorem 7 *Suppose f_{true} is a piecewise step function with $|f_{true}| \leq a$ for some $a > 0$. Define ψ_x and $A(\infty)$ as above. One has*

$$A(\infty) \leq \max(\tau, 1-\tau) E_{P_X}\left\{8a \exp\left(-\psi_x^2/(2p\sigma^2)\right)\right\}, \tag{7}$$

where P_X is the marginal distribution of X, p is the dimensionality of X and σ is the kernel parameter of the Gaussian RKHS.

For a fixed p, the upper bound in Theorem 7 depends on ψ_x, σ and the distribution P_X. If P_X is absolutely continuous with respect to the Lebesgue measure, one can verify that $A(\infty) \to 0$ with $\sigma \to 0$. This means that $f^{(\infty)}$ can approximate f_{true} arbitrarily well almost everywhere.

Next we consider the case where f_{true} is a general bounded measurable function. All bounded measurable functions can be approximated arbitrarily well by step functions. However, if f_{true} is too wiggly or is discontinuous at too many points in D, it cannot be well approximated by the Gaussian RKHS functions. For example, when f_{true} is discontinuous on a dense subset of D, $\psi_x = 0$ for all x if the step function is close enough to f_{true}. This leads to the right hand side of (7) being large. Therefore, we need some smoothness condition on f_{true}. The next corollary shows that when f_{true} is Lipschitz, $A(\infty) \to 0$ as $\sigma \to 0$.

Corollary 8 *Assume that f_{true} is a bounded Lipschitz function, and P_X is absolutely continuous with respect to the Lebesgue measure on D. Then $A(\infty) \to 0$ as $\sigma \to 0$.*

The discussions above have focused on the Gaussian RKHS. We note that it is possible to generalize the obtained results to more general RKHS's. For example, using similar techniques, one can verify that similar results as in Theorem 7 and Corollary 8 still hold if we consider many other radial kernels such as the Laplacian kernel. Other kernels, for example the polynomial kernel, may not have such guarantee that functions in the kernel space can approximate the underlying function arbitrary well. See the proof of Theorem 7 in the appendix for more discussions.

3.4 Finite Sample Error Bounds

The theory in Section 3.1 gives the asymptotic convergence rate of the estimation error. It is useful when the sample size is large. In this section, we derive some finite sample bounds on the prediction errors $E\rho_\tau(Y - \hat{f}_n)$ and $E\rho_\tau(Y - \tilde{f}_n)$, which can be used to assess the goodness of fit of the resulting model, when n is not large. In this section, we focus on the comparison between $E\rho_\tau(Y - \hat{f}_n)$ and $E\rho_\tau(Y - \tilde{f}_n)$. In particular, we show that the Rademacher complexity for $\hat{f}_n$ can be smaller compared to that of $\tilde{f}_n$. Hence, the prediction error bound for the data sparsity method can be better, and this demonstrates the usefulness of our proposed method. Note that Takeuchi et al. (2006) also used the Rademacher complexity to bound the estimation error. In this paper, we further consider how to bound the Rademacher complexity, especially for the data sparsity method.

To begin with, we introduce the following assumption.

Assumption A: The noise ε is bounded such that $|\varepsilon| \leq t$ for some positive t.

Theorem 9 *Suppose Assumption A holds. Then the solution $\hat{f}_n$ to (5) satisfies that, with probability at least $1-\delta$ with a small and positive δ,*

$$E\rho_\tau(Y - \hat{f}_n) \leq \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\big(y_i - \hat{f}_n(x_i)\big) + Z_n + \max(\tau, 1-\tau)\mu,$$

where $Z_n = 3\max(\tau, 1-\tau)\Big(n^{-1}(2s^2+2t^2)\log(2/\delta)\Big)^{1/2}$ *and*

$$\mu = s\sqrt{\frac{2\log(2n+2)}{n}}.$$

Moreover, the solution $\tilde{f}_n$ *to (6) satisfies that, with probability at least* $1-\delta$*,*

$$E\rho_\tau(Y-\tilde{f}_n) \le \frac{1}{n}\sum_{i=1}^{n}\rho_\tau\big(y_i-\tilde{f}_n(x_i)\big) + Z_n + \max(\tau,1-\tau)\mu,$$

where Z_n *is defined as above and*

$$\mu = 2sn^{-1/4} + \sqrt{n^{-1}\big(2^{11}n^{1/4}+2\log(5)+0.5\log(n)\big)}.$$

One can see from Theorem 9 that μ for $\hat{f}_n$ is much smaller than $\tilde{f}_n$. This is because the functional space of (5) is smaller than that of (4) (see Lemma 13 in the appendix). The key to the proof for $\tilde{f}_n$ is to control the covering number of the functional space in (4), as in Lemma 14. In particular, we study the covering number of a unit ball in $\mathcal{H}$, i.e., $\{f' : \|f'\|_{\mathcal{H}} \le 1\}$. On the other hand, the key to the proof for $\hat{f}_n$ is that the Rademacher complexity of the functional space of (5) is equivalent to the Rademacher complexity of a convex hull of functions with $2n+2$ vertices, and the latter enjoys a much better bound compared to the Rademacher complexity in (4).

From Theorem 9, one can conclude that when the underlying function can be well approximated by a function that has a sparse representation (in other words, the term $\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\big(y_i-\hat{f}_n(x_i)\big)$ and s are both small), the prediction error bound for the data sparsity method can be better. This observation provides some insight on the usefulness of the data sparsity method, which is illustrated by our numerical examples in Section 5, and discussed in Section 2.

Notice that our theory is for a general RKHS with very weak assumptions. Thus, the first choice of μ, $\mu = 2sn^{-1/4} + \sqrt{n^{-1}\big(2^{11}n^{1/4}+2\log(5)+0.5\log(n)\big)}$, can be refined if one considers specific kernels with additional assumptions, such as the decay rate of the eigenvalues of the corresponding Hilbert-Schmidt integral operator. See, for example, Zhou (2002) and Steinwart and Scovel (2007). This can lead to a better bound in Theorem 9. For example, with Gaussian kernel learning (including intercepts), an application of the result in Zhou (2002) gives $\mu = 2sn^{-1/3} + 2^{p+2}p^{1/2}n^{-1/3}\log^{p/2}(n)$, where p is the dimensionality of X. Hence for a small p, our μ on $\tilde{f}_n$ in Theorem 9 can be loose in this case. Nevertheless, one can verify that the bound on the data sparsity method is still better than $\mu = 2sn^{-1/3} + 2^{p+2}p^{1/2}n^{-1/3}\log^{p/2}(n)$. If we use the Gaussian kernel without an intercept, then one can verify that we have $\mu = O_P(n^{-1/2})$ for $\tilde{f}_n$ (Mendelson, 2003). However, without the intercept, the empirical prediction error term might be large for some problems, which can lead to suboptimal results.

As a remark, we note that Assumption A ensures that the response variable y is bounded if we restrict our consideration to the space $\mathcal{F}_n(s)$. Because we want to bound a finite sample error, any large noise in the response data, that is, an observed y_i being significantly away from its expected value given the predictors, can result in the failure of our bound derived in Theorem 9. Assumption A excludes the possibility that this large noise happens. This assumption can be removed if one assumes that the tail of the distribution of $|\varepsilon|$ satisfies certain properties, and similar results can

be obtained by modifying the proof of Theorem 9 accordingly. Corollary 10 gives one possible generalization of the bound in Theorem 9, where we make an assumption of the distribution of the error ε. Note that the results in Theorem 9 and Corollary 10 can be calculated directly from the data and the tuning parameter s we use.

Corollary 10 *Suppose that the errors ε_i; $i = 1, \ldots, n$ follow a common distribution with a continuous cumulative distribution function Φ_ε. For simplicity, assume that the distribution of ε is symmetric with respect to 0. Then $\hat{f}_n$ and $\tilde{f}_n$ are controlled by the same finite sample bounds as in Theorem 9 except that $Z_n = 3\max(\tau, 1-\tau)\Big(n^{-1}(2s^2+2t^2)\log(4/\delta)\Big)^{1/2}$ and $t = \Phi_\varepsilon^{-1}\big(0.5+0.5(1-\delta/2)^{1/n}\big)$.*

If ε follows a Gaussian distribution, one can verify that for a fixed δ, $t = O_P\big(\log(n)^{1/2}\big)$ as $n \to \infty$. Hence, t diverges in a slow rate. For other error distributions, one can obtain similar results by studying the corresponding Φ_ε, and we omit the details here.

4. Optimization and Tuning Procedure

The numerical optimization of (5) for fixed s and τ can be done by a simple linear programming. However, it is often desirable to have the entire solution path of $\hat{\alpha}$ and $\hat{b}$ with respect to s. For example, when we need to perform a comprehensive tuning procedure for choosing the optimal s, the solution path can significantly reduce the computational cost. In the literature of penalized quantile regression, Li et al. (2007) developed the solution path for (4) with respect to λ, Li and Zhu (2008) derived the solution path for L_1 penalized quantile regression with linear learning, and Rosset (2009) studied the solution surface with respect to both λ and τ in (4). In this section, we first briefly discuss how to derive the corresponding solution path with respect to s, then consider how to select the tuning parameter s.

Let $\tilde{K}$ be the n by $(n+1)$ matrix $(1\ K)$, where 1 is a vector of 1 of length n, and let $\bar{\alpha} = (b, \alpha_1, \ldots, \alpha_n)$. With b penalized, one can verify that the optimization (5) is equivalent to

$$\min_{\bar{\alpha}} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\big(y_i - f(x_i)\big), \text{ subject to } |\bar{\alpha}| \le s,$$

where $\big(f(x_1), \ldots, f(x_n)\big) = \tilde{K}\bar{\alpha}$. We note that the solution path of this optimization problem can be obtained in a similar manner as in Li and Zhu (2008) without an intercept in their notation, despite that they only considered linear learning problems. We omit the details here.

To illustrate the algorithm, using the data considered in Figure 2, we plot the piecewise linear solution path $\{b(s), \alpha(s)\}$ in Figure 5. Because the entire set $\{b(s), \alpha(s)\}$ consists of 31 piecewise linear functions, we only report a subset of $\{b(s), \alpha(s)\}$ in Figure 5 to make the plot clear. Moreover, if we plot the solution path on $[0, s_1]$ for large s_1, the lines become less clear on $[0, s_2]$ with $s_2 \ll s_1$. Hence, we only plot the solution path on $[0, 20]$ for a demonstration.

Remark 11 *As mentioned in Section 2, penalizing $|b|$ helps to guarantee the uniqueness of the solution path. In particular, if the intercept is not regularized, when s is large (or equivalently, the model is mildly penalized), there exist cases where b is not uniquely determined. See Li et al. (2007) and Li and Zhu (2008) for detailed discussions on this issue.*

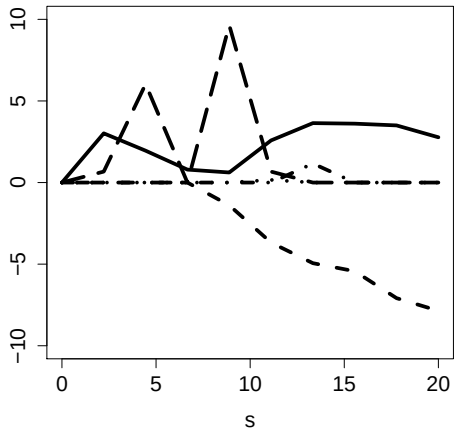


Figure 5: A subset of the solution path $\{b(s), \alpha(s)\}$ as a function of s. Notice that for brevity, we only plot five $\hat{\alpha}_i(s)$'s and the intercept b, and we restrict s such that $s \in [0, 20]$ to illustrate the solution path. The solid line corresponds to $b(s)$.

Next, we briefly discuss how to select the optimal tuning parameter s in (5) for a given regression problem. Similar to many other penalized techniques, a proper choice of s is crucial in practice. In particular, s being too large can lead to an overfitted model, and s being too small can lead to an underfitted model. For either cases, the prediction accuracy can be low. Here we discuss two commonly used criteria for kernel quantile regression. The first criterion is the Schwarz Information Criterion (SIC, Schwarz, 1978; Koenker et al., 1994), which can be written as

$$\mathrm{SIC}(s) = \log\left\{\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\big(y_i - \hat{f}(x_i)\big)\right\} + \frac{\log(n)df}{n}.$$

Here df measures the dimensionality of the model, and $\frac{\log(n)df}{n}$ balances the model complexity and goodness-of-fit. The second criterion is the Generalized Approximate Cross-Validation criterion (GACV, Yuan, 2006), defined as

$$\mathrm{GACV}(s) = \frac{1}{n - df}\left\{\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\big(y_i - \hat{f}(x_i)\big)\right\}.$$

Note that GACV was originally proposed as a stable estimator of the generalized comparative Kullback-Leibler distance for the model.

To estimate df, it has been proposed to use the divergence (Nychka et al., 1995; Yuan, 2006)

$$\mathrm{div}(\hat{f}) = \sum_{i=1}^{n}\frac{\partial\hat{f}(x_i)}{\partial y_i}.$$

Li et al. (2007) showed that for the kernel quantile regression with the regular squared norm penalty, $\mathrm{div}(\hat{f})$ coincides with the number of interpolated y_i's, and this makes the estimation of df convenient. The next proposition shows that for the proposed quantile regression with the data sparsity constraint, we have the same estimation formula, given the "one at a time" condition (Efron et al., 2004).

Proposition 12 *Assume the "one at a time" condition holds. For* $(y_i;\ i = 1,\ldots,n) \in \mathbb{R}^n$ *except on a set of Lebesgue measure* 0 *and any fixed s, we have*

$$\sum_{i=1}^{n} \frac{\partial \hat{f}(x_i)}{\partial y_i} = |\mathcal{E}|,$$

where $\mathcal{E}$ *is the set of interpolated* y_i*'s.*

5. Numerical Examples

In this section, we examine the performance of our proposed method with the data sparsity constraint using both simulated and real data sets. We demonstrate the effect of the SIC and GACV criteria. As a comparison, we also apply quantile regression using RKHS learning with the regular squared norm penalty, which was previously studied in Takeuchi et al. (2006) and Li et al. (2007).

5.1 Simulated Examples

We use three simulated examples to compare the performance of the proposed method with the data sparsity constraint and the standard method using the regular squared norm penalty. In particular, we study the three examples given in Section 2, which cover two cases. The first case (Examples 1 and 2) considers the situation when $f_0(x)$ can be well approximated by functions of the form $\sum_{j=1}^{m} \gamma_j K(x, z_j)$ for some fixed z_j and γ_j, where m is a small positive integer. For the second case (Example 3), f_0 is constructed to have many fluctuations. In this situation, f_0 cannot be well approximated by functions of the form $\sum_{j=1}^{m} \gamma_j K(x, z_j)$ with a small m.

For each example, we consider two choices of n: moderate ($n \in [30, 50]$) and very large ($n = 1000$). We show that for Examples 1-2 with moderate n's, the sparsely represented model from our method using the data sparsity constraint can have better prediction performance. On the other hand, for the case with a moderately large n but with the true function being quite wiggly (Example 3), or with a very large n, the performance of the two methods is comparable in terms of prediction accuracy. This is consistent with our theoretical insights.

We explore the performance under various settings of the noise. In particular, for the homoscedastic case, we let ε follow the standard normal distribution $N(0,1)$ (homoscedastic normal distribution, ho-norm. in Tables 2-7), and the t distribution with degrees of freedom 3 (t3). For the case when the noise is heteroscedastic, we let $\varepsilon(x) \sim U[-\|x\|_2/1.5, \|x\|_2/1.5]$ (heteroscedastic uniform distribution, unif. in Tables 2-7), and $\varepsilon(x) \sim N(0, \|x\|_2/3)$ (heteroscedastic normal distribution, he-norm. in Tables 2-7), where $\|\cdot\|_2$ is the usual Euclidian norm. Then we generate a training data set. Prediction models that correspond to different tuning parameters are built on the training data. We select the best tuning parameters by minimizing the SIC and GACV criteria on the training data respectively. The kernel parameters are also selected via the tuning procedure. We predict on a separate testing data set with the size 10000, to compare the performances of the two criteria. For

the choice of quantiles, since ε follows symmetric distributions with respect to 0, we choose $\tau = 0.1$, 0.3 and 0.5.

As discussed in Section 3, in this section, we will demonstrate that empirically, the effect of whether to penalize the intercept b or not is not large. In particular, for all the examples, we fit the models of the two compared methods with or without $|b|$ penalized. Note that for the regular squared norm method, we include $|b|$ in the penalty as in (6).

To measure the goodness of fit of the model, we calculate the prediction error

$$\frac{1}{10000} \sum_{i \in \text{test set}} \rho_\tau(y_i - \hat{f}_n(x_i)),$$

and the L_1 norm of $\hat{f}_n - f_{\text{true}}$. This procedure is repeated 1000 times and we report the average prediction error, average L_1 norm and average model complexity in terms of df. For our data sparsity method, we also report the percentage of non-zero α_i's as a measurement of how sparse the model is. Note that this percentage for the regular squared norm method is always 100%.

The simulation results are reported in Tables 2-7. For brevity, we only report the results for certain settings listed in Table 1. The other results are omitted because the general pattern is similar. Notice that for the cases where we compare moderate to large n's, we penalize the intercept of the data sparsity method, but not the intercept of the regular squared norm method, because the numerical difference is small.

Example	τ	Comparison
Ex 1	$\tau = 0.1$	Penalize b or not
	$\tau = 0.3$	Moderate or large n
Ex 2	$\tau = 0.3$	Penalize b or not
	$\tau = 0.5$	Moderate or large n
Ex 3	$\tau = 0.5$	Penalize b or not
	$\tau = 0.1$	Moderate or large n

Table 1: Summary of the settings for the reported simulation results.

Example 1: We generate the data in the same way as in Figure 2. In particular, x is one dimensional and follows the uniform distribution between -6 and 6. The underlying $f_0(x) = 10\exp(-x^2)$. We use $n = 30$ for the training data. The Laplacian kernel is used.

Example 2: The data are generated in the same way as in Figure 3. In particular, we let x be uniformly distributed in $[-3,3] \times [-3,3]$. The underlying true model is given by $f_0(x) = 10\exp(\|x - (-2,-2)^T\|_2^2) + 5\exp(\|x - (-1,1)^T\|_2^2) + 2\exp(\|x - (2,-1)^T\|_2^2)$. There are 50 observations in the training data set. We apply the Gaussian kernel for this example.

Example 3: The data are generated similarly as in Figure 4. To be specific, x is one dimensional, uniformly distributed in $[-7,7]$. We have the underlying function $f_0(x) = 5\sin(2.5x)$. The training data consist of 30 observations, and the Laplacian kernel is employed.

We summarize our findings from the simulation results as follows.

- For Examples 1 and 2 with a moderate n, the data sparsity method tends to choose a simpler model than that of the square norm penalty, with either the GACV or the SIC criterion. Furthermore, the data sparsity constraint performs better than the regular squared norm penalty.

This implies that functions estimated by the squared norm penalty can have potential overfitting because of too many kernel functions, as indicated by Figures 2 and 3. For a large n, the prediction performance of the two methods is comparable, because asymptotically both methods can perform well.

- For Example 3 with a moderate n, the data sparsity method tends to choose a model with a large number of non-zero α_i's. In this case, the model from (5) is not sparse, and the two methods perform similarly.
- For a large n in all the examples, the data sparsity method can choose models with simpler representation than the regular method, while keeping similar prediction performance. Consequently, the data sparsity method yields a parsimonious model that has advantages in terms of computational efficiency and interpretability.
- The SIC always tends to choose a simpler model than the GACV criterion. In these examples, GACV overall works slightly better than SIC, for both the data sparsity constraint and the squared norm penalty.
- As τ gets closer to 0.5, the performances of the two penalties and the two criteria become better, in terms of the L_1 norm.

		$\|b\|$ penalized				$\|b\|$ not penalized			
		Squared norm		Data sparsity		Squared norm		Data sparsity	
	Dist.	GACV	SIC	GACV	SIC	GACV	SIC	GACV	SIC
Pred	ho-norm.	0.760	0.910	0.517	0.596	0.752	0.909	0.523	0.579
	t3	0.893	1.013	0.677	0.699	0.849	1.075	0.711	0.726
	unif.	0.869	0.995	0.636	0.701	0.933	0.959	0.598	0.700
	he-norm.	0.913	0.925	0.708	0.734	0.958	1.033	0.722	0.748
L_1 Norm	ho-norm.	2.061	2.291	0.733	0.759	2.062	2.294	0.715	0.753
	t3	2.233	2.426	1.239	1.297	2.236	2.429	1.244	1.280
	unif.	2.362	2.217	0.665	0.687	2.246	2.199	0.626	0.699
	he-norm.	2.343	2.366	1.290	1.375	2.109	2.256	1.276	1.391
df	ho-norm.	13.24	12.55	6.939	6.703	13.56	12.33	6.899	6.716
	t3	13.19	11.87	6.771	6.529	13.48	11.44	6.784	6.545
	unif.	13.28	11.13	7.044	6.512	13.64	11.23	7.033	6.529
	he-norm.	15.16	14.57	8.182	7.903	15.22	14.91	8.452	8.005
Percent.	ho-norm.	-	-	17.33	16.25	-	-	17.49	16.88
	t3	-	-	17.18	16.10	-	-	17.36	16.42
	unif.	-	-	17.58	16.69	-	-	17.42	17.01
	he-norm.	-	-	19.21	18.34	-	-	18.87	18.55

Table 2: Results of the simulation Example 1 with $\tau = 0.1$ and $|b|$ penalized (left) or not (right). The best mean prediction errors are 0.40 (ho-norm., homoscedastic normal distribution), 0.56 (t3), 0.59 (unif., heteroscedastic uniform distribution), and 0.60 (he-norm., heteroscedastic normal distribution). The standard errors of the prediction errors range from 0.0040 to 0.0053. The standard errors of the L_1 norms range from 0.0068 to 0.0095. The standard errors of df range from 0.050 to 0.066. The standard errors of the percentage of non-zero α_i's for the data sparsity method range from 0.032 to 0.051.

		$n = 30$				$n = 1000$			
		Squared norm		Data sparsity		Squared norm		Data sparsity	
	Dist.	GACV	SIC	GACV	SIC	GACV	SIC	GACV	SIC
Pred	ho-norm.	0.847	0.903	0.711	0.716	0.514	0.510	0.512	0.512
	t3	1.044	1.070	0.928	0.946	0.688	0.691	0.685	0.693
	unif.	0.914	0.916	0.833	0.820	0.655	0.667	0.659	0.671
	he-norm.	0.966	0.989	0.831	0.845	0.711	0.709	0.705	0.713
L_1 Norm	ho-norm.	1.462	1.841	0.517	0.552	0.657	0.644	0.659	0.650
	t3	1.533	1.897	0.583	0.694	1.024	1.058	1.062	0.989
	unif.	1.541	1.882	0.755	0.795	0.578	0.560	0.556	0.581
	he-norm.	1.515	1.798	0.689	0.726	0.913	0.926	0.912	0.904
df	ho-norm.	15.67	12.77	6.523	6.040	20.44	20.16	18.16	18.05
	t3	13.97	12.05	6.164	6.070	21.52	21.09	19.13	18.07
	unif.	15.85	13.71	6.775	6.433	19.89	19.15	18.29	18.14
	he-norm.	16.77	15.84	7.782	7.503	22.10	22.28	21.85	22.34
Percent.	ho-norm.	-	-	17.33	16.25	-	-	0.614	0.603
	t3	-	-	17.18	16.10	-	-	0.610	0.606
	unif.	-	-	17.44	16.96	-	-	0.612	0.606
	he-norm.	-	-	18.92	17.77	-	-	0.693	0.687

Table 3: Results of the simulation Example 1 with $\tau = 0.3$ and moderate n (left) or large (right). The best mean prediction errors are 0.40 (ho-norm., homoscedastic normal distribution), 0.56 (t3), 0.59 (unif., heteroscedastic uniform distribution), and 0.60 (he-norm., heteroscedastic normal distribution). The standard errors of the prediction errors range from 0.0030 to 0.0046. The standard errors of the L_1 norms range from 0.0055 to 0.0079. The standard errors of df range from 0.048 to 0.059. The standard errors of the percentage of non-zero α_i's for the data sparsity method range from 0.031 to 0.055.

5.2 Real Data Analysis

In this section, we apply our proposed method (5) to several real data sets. In particular, we consider 20 data sets studied in Section 5 of Takeuchi et al. (2006), and the well known annual salary of baseball players data studied in He et al. (1998), Yuan (2006) and Li et al. (2007). The description and a summary table of the first 20 data sets can be found in Takeuchi et al. (2006), and we do not repeat it here. For the baseball data, it consists of statistics for 263 North American major league baseball players in the year 1986. The original data set has 22 predictors and the players' 1987 annual salary as the response, and we use the whole data for our analysis reported in Tables 8-10. Furthermore, following He et al. (1998), Yuan (2006) and Li et al. (2007), we use two representative predictors from the baseball data to perform an illustrative analysis, which is reported in Figure 7. In particular, we measure players' performance by the number of home runs in the latest year, and measure player's seniority by the number of years played.

For all real data analysis, the Gaussian kernel is used. For the results reported in Tables 8-10, we first standardize the predictors and response to make the results comparable. We split each data set into 10 parts of roughly the same size. We choose 1 part as the testing data set, and the remaining as the training data. Then we select the best tuning parameter and kernel parameter on the training data, and predict on the testing data. We continue to the next random split once all the 10 parts have served as the testing data. This random split is repeated 100 times for each data set, and we

		$\|b\|$ penalized				$\|b\|$ not penalized			
		Squared norm		Data sparsity		Squared norm		Data sparsity	
	Dist.	GACV	SIC	GACV	SIC	GACV	SIC	GACV	SIC
Pred	ho-norm.	0.885	1.075	0.527	0.673	0.847	1.031	0.557	0.624
	t3	1.068	1.118	0.677	0.894	1.124	1.341	0.710	0.909
	unif.	1.057	1.243	0.644	0.692	1.147	1.286	0.613	0.629
	he-norm.	1.021	1.226	0.697	0.801	1.089	1.146	0.703	0.811
L_1 Norm	ho-norm.	1.826	2.078	0.760	0.911	1.864	2.271	0.796	0.894
	t3	1.873	2.213	0.875	1.244	1.905	2.400	0.892	1.303
	unif.	2.002	2.199	0.841	1.199	1.953	2.168	0.872	1.240
	he-norm.	2.112	2.351	0.873	0.992	1.913	2.325	0.846	1.023
df	ho-norm.	18.16	16.22	12.33	9.885	18.83	16.06	12.51	10.21
	t3	17.91	16.22	11.97	8.678	18.05	15.80	12.03	9.146
	unif.	18.44	15.78	12.91	9.913	18.31	16.03	12.50	10.12
	he-norm.	19.05	18.79	14.41	12.90	19.33	19.03	14.29	12.46
Percent.	ho-norm.	-	-	13.24	12.90	-	-	13.44	13.00
	t3	-	-	14.26	12.92	-	-	14.55	13.22
	unif.	-	-	16.64	15.68	-	-	17.02	15.94
	he-norm.	-	-	17.22	16.78	-	-	16.89	16.59

Table 4: Results of the simulation Example 2 with $\tau = 0.3$ and $|b|$ penalized (left) or not (right). The best mean prediction errors are 0.40 (ho-norm., homoscedastic normal distribution), 0.56 (t3), 0.59 (unif., heteroscedastic uniform distribution), and 0.60 (he-norm., heteroscedastic normal distribution). The standard errors of the prediction errors range from 0.0055 to 0.0068. The standard errors of the L_1 norms range from 0.0071 to 0.0133. The standard errors of df range from 0.066 to 0.072. The standard errors of the percentage of non-zero α_i's for the data sparsity method range from 0.044 to 0.057.

report the average prediction error (Pred) and its sample standard deviation (SSD) in Tables 8-10 for $\tau = 0.1, 0.5, 0.9$. We only report the results where the intercept is penalized for the data sparsity method, but not for the standard method with the squared norm penalty. Similar to the results in Section 5.1, the numerical difference of whether the intercept is penalized or not is not large. For the results reported in Figure 7, we train the model using the entire data set. We then plot the predicted values against the two dimensional input space.

Similar to Takeuchi et al. (2006), we perform a two-sided paired-sample t-test to compare the prediction performance of the two methods. In Tables 8-10, we can see that for the caution, sniffer, GAGurine, topo, CobarOre, and baseball data sets, the performance of the data sparsity method is overall better than that of the squared norm penalty method. For birthwt, engel, gilgais, and mcycle, the data sparsity method is slightly better. For BostonHousing and cpus, the squared norm penalty is slightly better. For the other data sets, their performance is comparable. This demonstrates the usefulness of the data sparsity method. Moreover, we plot the fitted functions $\hat{f}_n$ and $\tilde{f}_n$ with $\tau = 0.5$ for the mcycle data on the left panel of Figure 6. Compared to Figure 3 in Takeuchi et al. (2006), one can see that the data sparsity model has less wiggles, and yields a more interpretable result. We also plot the fitted functions with $\tau = 0.1$ on the right panel of Figure 6. In this case, one can see that $\tilde{f}_n$ is quite wiggly compared to $\hat{f}_n$.

For the results on the illustrative analysis of the baseball data, from the right panels of Figure 7 (our data sparsity constraint), we can see that for all players, the income increases with their perfor-

		$n = 30$				$n = 1000$			
		Squared norm		Data sparsity		Squared norm		Data sparsity	
	Dist.	GACV	SIC	GACV	SIC	GACV	SIC	GACV	SIC
Pred	ho-norm.	0.922	0.984	0.546	0.683	0.487	0.466	0.482	0.478
	t3	1.243	1.309	0.798	0.913	0.634	0.633	0.612	0.629
	unif.	1.155	1.272	0.877	0.916	0.637	0.640	0.642	0.638
	he-norm.	1.133	1.287	0.764	0.787	0.688	0.685	0.681	0.684
L_1 Norm	ho-norm.	1.416	1.813	0.653	0.772	0.714	0.720	0.711	0.715
	t3	1.679	2.002	0.718	1.060	0.922	0.918	0.918	0.927
	unif.	1.744	2.102	0.899	1.132	0.559	0.576	0.565	0.563
	he-norm.	1.553	1.842	0.957	1.086	1.010	1.004	0.989	0.995
df	ho-norm.	18.40	15.66	12.17	10.11	16.14	16.01	15.98	15.79
	t3	18.54	16.11	11.10	9.264	16.48	16.22	16.14	15.88
	unif.	18.25	16.94	11.46	10.72	15.66	15.41	15.36	15.20
	he-norm.	19.91	17.27	14.62	13.78	18.89	18.43	17.67	17.40
Percent.	ho-norm.	-	-	12.28	11.90	-	-	0.591	0.582
	t3	-	-	13.42	13.00	-	-	0.611	0.590
	unif.	-	-	16.54	14.30	-	-	0.622	0.605
	he-norm.	-	-	16.88	15.53	-	-	0.630	0.616

Table 5: Results of the simulation Example 2 with $\tau = 0.5$ and moderate n (left) or large (right). The best mean prediction errors are 0.40 (ho-norm., homoscedastic normal distribution), 0.56 (t3), 0.59 (unif., heteroscedastic uniform distribution), and 0.60 (he-norm., heteroscedastic normal distribution). The standard errors of the prediction errors range from 0.0052 to 0.0068. The standard errors of the L_1 norms range from 0.0083 to 0.0114. The standard errors of df range from 0.044 to 0.078. The standard errors of the percentage of non-zero α_i's for the data sparsity method range from 0.065 to 0.097.

mances. On the other hand, the salary does not necessarily increase with the seniority. For many players, especially the high income ones ($\tau = 0.75$), the salary increases with the seniority until a golden age, then it decreases. This is consistent with our intuition. For the results with the squared norm penalty (the left panels), we can see the same trend. However, for the very senior players, because the estimated salary function has fluctuations from kernel functions, the salary decreases if their performances increase from 20 to 30. This is against our intuition. Therefore, in this data set, our data sparsity constraint performs well and gives a good interpretation of the data.

6. Discussion

In this paper, we study the learning problem in a RKHS. In particular, we propose a data sparsity constraint that can achieve a parsimonious representation of the resulting learning function. Using quantile regression as an example, we numerically show that when the underlying function can be well approximated by functions that have a sparse representation in the corresponding RKHS and n is not large, the data sparsity method can perform better than the regular squared norm penalty method. For other cases, such as when the true function is relatively difficult to be approximated by

		$\|b\|$ penalized				$\|b\|$ not penalized			
		Squared norm		Data sparsity		Squared norm		Data sparsity	
	Dist.	GACV	SIC	GACV	SIC	GACV	SIC	GACV	SIC
Pred	ho-norm.	1.770	1.812	1.749	1.796	1.742	1.814	1.759	1.803
	t3	2.119	2.230	2.138	2.206	2,178	2.267	2.104	2.257
	unif.	1.997	1.989	1.846	2.014	1.884	1.892	1.865	1.911
	he-norm.	1.849	1.897	1.833	1.878	1.845	1.904	1.890	1.926
L_1 Norm	ho-norm.	1.657	1.690	1.664	1.711	1.649	1.698	1.625	1.709
	t3	2.058	2.224	2.060	2.215	2.121	2.269	2.150	2.298
	unif.	1.887	1.845	1.869	1.893	1.853	1.829	1.850	1.837
	he-norm.	1.848	1.820	1.851	1.827	1.857	1.841	1.866	1.836
df	ho-norm.	24.66	24.03	13.76	13.09	24.48	23.19	14.90	13.55
	t3	25.53	24.69	13.96	13.14	25.56	23.92	14.14	14.00
	unif.	25.12	23.55	15.47	14.16	25.09	24.61	14.98	13.58
	he-norm.	25.79	24.15	14.16	13.90	25.12	24.07	13.92	13.23
Percent.	ho-norm.	-	-	66.24	61.32	-	-	68.45	65.37
	t3	-	-	70.18	66.26	-	-	68.91	65.28
	unif.	-	-	68.93	65.21	-	-	70.52	66.42
	he-norm.	-	-	71.29	67.84	-	-	68.35	66.91

Table 6: Results of the simulation Example 3 with $\tau = 0.5$ and $|b|$ penalized (left) or not (right). The best mean prediction errors are 0.40 (ho-norm., homoscedastic normal distribution), 0.56 (t3), 0.59 (unif., heteroscedastic uniform distribution), and 0.60 (he-norm., heteroscedastic normal distribution). The standard errors of the prediction errors range from 0.0091 to 0.0143. The standard errors of the L_1 norms range from 0.0128 to 0.0175. The standard errors of df range from 0.088 to 0.144. The standard errors of the percentage of non-zero α_i's for the data sparsity method range from 0.244 to 0.351.

functions in the RKHS, or when n is large, the data sparsity method can have comparable performance as the regular method. Therefore, the data sparsity method can be regarded as an alternative penalization method to solve learning problems with RKHS learning. Moreover, because of the sparsity in the kernel representation, the prediction for new data sets can be computationally faster. Through theoretical comparisons, we demonstrate that the data sparsity method can achieve the same convergence rate of the estimation error, compared with the squared norm penalty method. Furthermore, we show that for certain cases, the data sparsity method can enjoy a smaller bound on the finite sample prediction error. This helps to shed some light on the usefulness of the data sparsity constraint. We also discuss how to obtain a solution path with respect to the tuning parameter s.

We would like to point out several open problems for the theory developed in Section 3. The technique used to prove Theorems 1 and 9 there does not take into account the fact that the "active functional space" of the estimated function is often smaller than the entire $\mathcal{F}_n(s)$. Therefore, one possible way to obtain better results is to consider the "localized" covering number of the active functional space of (5). See the discussion of localization idea in, for example, Bartlett et al. (2005). In that case, we can expect a faster convergence rate and a tighter bound on the prediction error. Another open problem is to consider a combination of the L_2 and L_1 penalties, which can be a more general form than the pure L_1 or L_2 penalty. In the literature of linear learning, Zou and Hastie (2005) proposed the elastic net penalty as a convex combination of the L_2 and L_1 penalties. In

		$n = 30$				$n = 1000$			
		Squared norm		Data sparsity		Squared norm		Data sparsity	
	Dist.	GACV	SIC	GACV	SIC	GACV	SIC	GACV	SIC
Pred	ho-norm.	1.569	1.658	1.554	1.772	0.745	0.761	0.746	0.753
	t3	2.044	2.168	2.015	2.247	0.810	0.813	0.825	0.839
	unif.	1.717	1.826	1.679	1.774	0.876	0.891	0.859	0.880
	he-norm.	1.946	2.091	1.925	2.083	0.845	0.886	0.839	0.891
L_1 Norm	ho-norm.	1.746	1.766	1.753	1.760	0.924	0.995	0.957	0.986
	t3	2.193	2.324	2.301	2.332	1.103	1.166	1.098	1.156
	unif.	1.986	2.009	2.054	2.013	0.883	0.916	0.849	0.872
	he-norm.	2.094	2.129	1.954	2.058	1.124	1.196	1.049	1.087
df	ho-norm.	25.65	24.88	15.52	14.49	33.14	33.09	32.45	32.16
	t3	25.97	25.12	15.66	14.79	30.26	30.28	31.06	30.73
	unif.	26.36	25.77	16.28	16.11	32.28	32.06	32.44	31.98
	he-norm.	27.48	27.01	16.94	16.59	33.04	32.85	33.30	32.19
Percent.	ho-norm.	-	-	50.53	46.68	-	-	4.221	4.057
	t3	-	-	55.16	53.25	-	-	4.528	4.247
	unif.	-	-	55.23	54.71	-	-	4.567	4.059
	he-norm.	-	-	56.14	54.90	-	-	4.778	4.670

Table 7: Results of the simulation Example 3 with $\tau = 0.1$ and moderate n (left) or large (right). The best mean prediction errors are 0.40 (ho-norm., homoscedastic normal distribution), 0.56 (t3), 0.59 (unif., heteroscedastic uniform distribution), and 0.60 (he-norm., heteroscedastic normal distribution). The standard errors of the prediction errors range from 0.0107 to 0.0177. The standard errors of the L_1 norms range from 0.0109 to 0.0186. The standard errors of df range from 0.099 to 0.158. The standard errors of the percentage of non-zero α_i's for the data sparsity method range from 0.414 to 0.572.

kernel learning, how to perform such a generalization effectively can be an interesting problem to pursue.

Acknowledgments

The authors would like to thank the Action Editor Professor Saharon Rosset and three reviewers for their constructive comments and suggestions, which led to substantial improvements of the presentation of this paper. The authors are supported in part by National Science and Engineering Research Council of Canada (NSERC), National Natural Science Foundation of China (NSFC 61472475), NSF grant DMS-1407241, NSF DMS-1055210, NIH/NCI grant R01 CA-149569, and NIH/NCI P01 CA-142538.

Appendix A. Proof of Theorem 1

In the following proofs, when the technique can be applied to both the proposed data sparsity constraint and the regular squared norm penalty, we omit the difference between $\hat{f}_n$ and $\tilde{f}_n$ for brevity.

Before giving the proof of Theorem 1, we first introduce a lemma.

	Square norm				Data sparsity			
	GACV		SIC		GACV		SIC	
Data	Pred	SSD	Pred	SSD	Pred	SSD	Pred	SSD
baseball	10.09	0.67	10.14	0.52	**9.814**	0.43	10.33	0.49
BigMac2003	6.442	0.64	6.414	0.52	**6.297**	0.57	6.371	0.77
birthwt	18.44	0.84	18.80	0.72	**18.38**	0.69	18.92	0.74
BostonHousing	**5.543**	0.35	5.569	0.29	5.677	0.52	5.712	0.39
*caution	9.782	0.74	9.891	0.56	**8.433**	0.61	8.598	0.59
*CobarOre	15.74	1.16	16.12	1.28	12.19	0.95	**12.10**	0.84
*cpus	4.692	0.16	**4.575**	0.20	5.132	0.13	5.242	0.17
crabs	3.952	0.05	4.127	0.12	**3.893**	0.09	4.016	0.10
engel	5.490	0.42	**5.336**	0.39	5.356	0.51	5.561	0.35
ftcollinssnow	15.99	3.36	16.43	3.28	**15.62**	3.03	15.88	3.00
GAGurine	8.224	0.32	**8.138**	0.24	8.261	0.35	8.210	0.38
geyser	**8.440**	0.49	9.158	0.58	8.923	0.49	8.682	0.65
gilgais	6.731	0.32	6.849	0.29	**6.680**	0.29	7.003	0.33
heights	14.91	0.38	15.56	0.35	**14.69**	0.37	15.27	0.41
highway	8.964	0.57	9.059	0.63	9.112	0.71	**8.877**	0.56
*mcycle	7.732	0.25	8.105	0.31	**7.012**	0.21	7.033	0.29
*sniffer	6.372	0.33	6.457	0.30	**5.416**	0.28	5.608	0.28
snowgeese	5.788	1.01	5.892	0.84	**5.694**	0.70	6.010	0.69
topo	6.514	0.48	6.449	0.42	**6.268**	0.34	6.435	0.35
ufc	**10.11**	1.10	11.06	0.64	10.49	0.89	10.83	0.93
UN3	12.29	1.11	12.04	1.24	**11.92**	0.92	12.09	1.05

Table 8: Results of the real data analysis for $\tau = 0.1$. We reported $100 \times$ prediction error for Pred. Here $*$ means the difference of prediction error between the two methods, both GACV vs. GACV and SIC vs. SIC, is statistically significant at level 0.05 using two-sided paired-sample t-test.

Lemma 13 *Suppose the RKHS is separable and* $\sup_{X_1,X_2} K(X_1,X_2) = 1$. *Then* $\sum_{i=1}^n |\alpha_i| \le s$ *implies* $\|f'\|_{\mathcal{H}}^2 \le s^2$.

Lemma 13 indicates that a bound on $\sum_{i=1}^n |\alpha_i|$ is a stronger constraint than the usual squared norm constraint. Hence the effect of our data sparsity penalty is two fold: control the complexity of f and impose a soft threshold to gain data sparsity. Lemma 13 helps to bound the covering number of the functional class in Lemma 14.

Proof of Lemma 13: For any $f'(x) = \sum_{i=1}^n \alpha_i K(x_i,x)$ with $\sum_{i=1}^n |\alpha_i| \le s$, we have that $\|f'\|_{\mathcal{H}}^2 = \alpha^T K \alpha = \sum_{i=1}^n \sum_{j=1}^n K(x_i,x_j)\alpha_i\alpha_j \le \sum_{i=1}^n |\alpha_i| \cdot \sum_{j=1}^n |\alpha_j| \le s^2$, because $K(\cdot,\cdot) \le 1$. □

To prove Theorem 1, note that $\rho_\tau(y-f) \le |y-f|$. Hence in the following arguments, we can consider the loss $L(a,b) = |a-b|$ instead of the check function. Recall that the definition of $f_n^{(s)}$ is $f_n^{(s)} = \text{argmin}_{f \in \mathcal{F}_n(s)} L(Y,f)$. Define $g_f(\cdot) = (2s)^{-1}(L(\cdot,f) - L(\cdot,f_n^{(s)}))$, and $\mathcal{G} = \{g_f : f \in \mathcal{F}_n(s)\}$.

First we provide a lemma that controls the complexity of $\mathcal{G}$ in terms of its covering number. Notice that this technique can also be applied to the regular squared norm method, therefore, the result in Theorem 1 is also valid for the regular method that penalizes $|b|$ (or make additional assumptions to avoid this penalty on $|b|$). Before that we introduce some further notation. Let T_X be the empirical measure of a training set $((x_1,y_1),\ldots,(x_n,y_n))$, and the L_2 norm be defined as

	Square norm				Data sparsity			
	GACV		SIC		GACV		SIC	
Data	Pred	SSD	Pred	SSD	Pred	SSD	Pred	SSD
*baseball	24.47	0.87	25.61	0.78	22.16	0.69	**21.58**	0.70
BigMac2003	18.81	2.24	19.00	2.51	**18.42**	1.86	19.13	1.72
*birthwt	36.44	1.31	37.68	1.09	**33.20**	1.21	33.97	1.10
*BostonHousing	**10.92**	0.55	11.43	0.59	13.10	0.47	13.55	0.61
*caution	23.17	1.23	23.20	1.08	20.19	0.99	**20.01**	1.04
*CobarOre	41.26	1.84	40.07	1.45	**35.98**	1.76	36.62	1.64
cpus	**3.128**	0.23	3.269	0.30	3.202	0.22	3.294	0.25
crabs	**5.133**	0.24	5.249	0.22	5.227	0.31	5.158	0.28
engel	**14.18**	0.82	14.76	0.86	14.25	0.68	14.40	0.91
ftcollinssnow	40.58	5.14	41.89	5.46	41.43	4.91	**40.29**	3.88
*GAGurine	14.98	0.56	15.25	0.45	**13.07**	0.43	13.56	0.42
geyser	**29.16**	1.57	32.45	1.29	30.08	1.06	31.66	1.11
*gilgais	13.15	0.51	13.59	0.44	**11.02**	0.47	11.27	0.39
heights	36.72	1.13	**35.48**	0.78	36.14	0.81	35.55	0.92
highway	27.21	2.17	27.49	2.24	**26.79**	1.82	27.34	1.70
mcycle	18.24	0.77	19.17	0.61	**17.53**	0.71	18.32	0.64
sniffer	**10.17**	0.67	10.35	0.71	10.68	0.55	10.51	0.60
snowgeese	17.76	1.91	18.14	1.54	**17.42**	1.64	18.08	1.73
*topo	15.28	0.55	16.01	0.62	14.01	0.43	**13.76**	0.50
ufc	22.78	1.27	**21.75**	1.01	23.04	1.21	22.11	1.24
UN3	21.71	1.55	22.94	1.42	**21.25**	1.62	22.18	1.39

Table 9: Results of the real data analysis for $\tau = 0.5$. We reported $100\times$ prediction error for Pred. Here $*$ means the difference of prediction error between the two methods, both GACV vs. GACV and SIC vs. SIC, is statistically significant at level 0.05 using two-sided paired-sample t-test.

$\|f\|_{L_2(T_X)} = \left(\frac{1}{n}\sum_{i=1}^n |f(x_i,y_i)|^2\right)^{1/2}$. For any $\eta > 0$, define $\mathcal{M}$ to be a η-net of a class of function $\mathcal{F}$ if, for any $f \in \mathcal{F}$, there exists $m \in \mathcal{M}$ such that $\|m - f\|_{L_2(T_X)} \le \eta$. Now let the $L_2(T_X)$ covering number $N(\eta, \mathcal{F}, L_2(T_X))$ be the minimal size of all possible η-nets.

Lemma 14 *For $\eta > 0$ small enough and $C_0 = 2^{10}$, we have that*

$$\sup_{T_X} N\big(\eta, \mathcal{G}, L_2(T_X)\big) \le \frac{5\exp(C_0\eta^{-2})}{\eta}.$$

Proof of Lemma 14: The proof consists of two steps. The first step is to bound the entropy number when there is no intercept in the regression function. The second step is to add the intercept into consideration, and bound the corresponding entropy based on the results obtained in the first step.

Here we focus on the covering number of $\mathcal{G}_{\mathcal{H},b} := \{(2s)^{-1}L(\cdot,f) : f \in \mathcal{F}_n(s)\}$, because $\mathcal{G}_{\mathcal{H},b}$ has the same covering number as $\mathcal{G}$. To that end, we first calculate the covering number of $\mathcal{G}_{\mathcal{H}} := \{(2s)^{-1}L(\cdot,f') : f' = \sum_{i=1}^n \alpha_i K(x_i,\cdot), \sum_{i=1}^n |\alpha_i| \le s\}$. Define

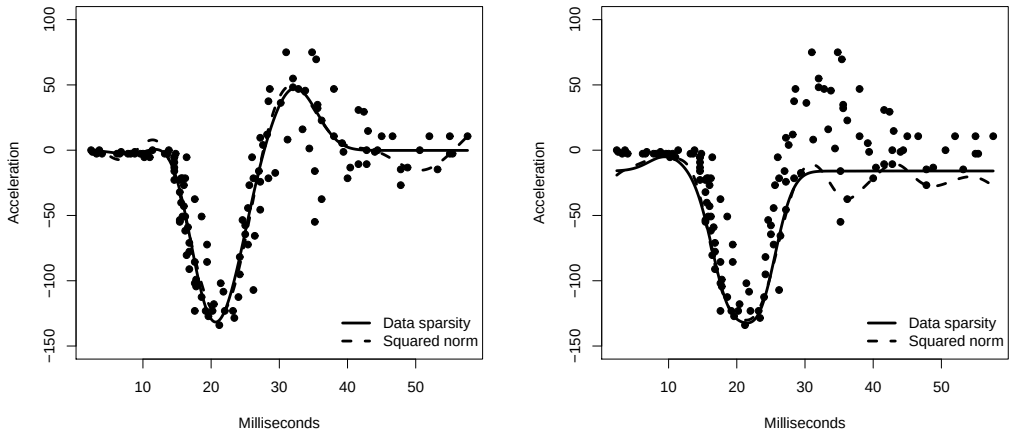
$$\mathcal{G}'_{\mathcal{H}} := \{(2s)^{-1}f' : f' = \sum_{i=1}^n \alpha_i K(x_i,\cdot), \sum_{i=1}^n |\alpha_i| \le s\}.$$

	Square norm				Data sparsity			
	GACV		SIC		GACV		SIC	
Data	Pred	SSD	Pred	SSD	Pred	SSD	Pred	SSD
*baseball	19.61	0.77	19.22	0.64	**15.52**	0.69	15.99	0.59
BigMac2003	11.33	0.92	11.84	0.74	10.92	0.81	**10.44**	0.99
birthwt	16.44	0.90	**15.89**	0.74	16.27	0.72	16.03	0.65
BostonHousing	**7.775**	0.36	8.104	0.40	8.158	0.34	7.960	0.39
*caution	16.76	0.56	16.33	0.49	**12.51**	0.51	13.09	0.62
*CobarOre	14.43	0.88	15.16	0.82	**12.55**	0.75	13.08	0.80
cpus	1.551	0.28	**1.479**	0.24	1.492	0.20	1.505	0.20
crabs	2.461	0.18	**2.447**	0.20	2.569	0.15	2.450	0.14
*engel	6.168	0.42	6.284	0.35	**5.041**	0.38	5.297	0.40
ftcollinssnow	20.18	4.14	22.46	4.51	**19.59**	3.97	22.13	4.22
*GAGurine	10.01	0.42	10.57	0.44	**8.558**	0.53	9.101	0.46
geyser	**12.63**	0.66	13.10	0.71	12.76	0.59	12.91	0.73
gilgais	6.167	0.40	6.228	0.29	**6.109**	0.31	6.334	0.34
heights	15.21	0.58	15.42	0.64	15.63	0.69	**15.02**	0.56
highway	16.96	1.42	16.27	1.59	**15.88**	1.49	16.59	1.34
mcycle	7.162	0.62	7.246	0.41	**6.197**	0.53	6.331	0.49
*sniffer	5.691	0.31	5.743	0.39	**4.409**	0.24	4.553	0.28
snowgeese	8.166	0.74	8.259	0.88	**7.919**	0.80	8.260	0.92
*topo	11.57	0.40	12.18	0.35	**10.51**	0.40	10.86	0.38
ufc	10.52	0.82	11.77	0.62	**10.24**	0.77	10.31	0.65
UN3	**7.774**	0.87	8.126	1.00	7.910	0.94	7.885	0.83

Table 10: Results of the real data analysis for $\tau = 0.9$. We reported $100 \times$ prediction error for Pred. Here $*$ means the difference of prediction error between the two methods, both GACV vs. GACV and SIC vs. SIC, is statistically significant at level 0.05 using two-sided paired-sample t-test.

Define T_X' to be the empirical measure of the set $(x_1, \ldots, x_n)$. For any $g_1 := (2s)^{-1}L(\cdot, f_1') \in \mathcal{G}_{\mathcal{H}}$, $g_2 := (2s)^{-1}L(\cdot, f_2') \in \mathcal{G}_{\mathcal{H}}$, $|g_1 - g_2| \le (2s)^{-1}|f_1' - f_2'|$. Hence, an $L_2(T_X')$ net on $\mathcal{G}_{\mathcal{H}}'$ naturally introduces an $L_2(T_X)$ net on $\mathcal{G}_{\mathcal{H}}$, and furthermore the $L_2(T_X)$ covering number of $\mathcal{G}_{\mathcal{H}}$ is upper bounded by the $L_2(T_X')$ covering number of $\mathcal{G}_{\mathcal{H}}'$. Moreover, by Lemma 13, $\|(2s)^{-1}f'\|_{\mathcal{H}} \le 1$. Thus, $\mathcal{G}_{\mathcal{H}}' \subset B_{\mathcal{H}}$, where $B_{\mathcal{H}}$ is the unit ball in $\mathcal{H}$. Hence, we only need to bound $N(\eta, B_{\mathcal{H}}, L_2(T_X'))$. This can be done by a similar argument as in Theorem 2.1 of Steinwart and Scovel (2007). In particular, from analogous arguments as those that lead to (21) in Steinwart and Scovel (2007), we have that $\sup_{T_X'} N(\eta, B_{\mathcal{H}}, L_2(T_X')) \le \exp(\frac{C_0\eta^{-2}}{4})$, where one can choose $C_0 = 2^{10}$ (Carl and Stephani, 1990). In one words, we have that $\sup_{T_X} N(\eta, \mathcal{G}_{\mathcal{H}}, L_2(T_X)) \le \sup_{T_X'} N(\eta, \mathcal{G}_{\mathcal{H}}', L_2(T_X')) \le \sup_{T_X'} N(\eta, B_{\mathcal{H}}, L_2(T_X')) \le \exp(\frac{C_0\eta^{-2}}{4})$. Note that Theorem 2.1 of Steinwart and Scovel (2007) considered only the Gaussian RKHS, however the proof of the entropy bound for $p = 2$ in their notation only requires that the RKHS is separable.

We proceed to bound the entropy number of $\mathcal{G}_{\mathcal{H},b}$. Define $\mathcal{G}_{\mathcal{H},b}' = \{(2s)^{-1}f : f \in \mathcal{F}_n(s)\}$. By similar arguments as above, $N(\eta, \mathcal{G}_{\mathcal{H},b}, L_2(T_X)) \le N(\eta, \mathcal{G}_{\mathcal{H},b}', L_2(T_X'))$. Suppose $\mathcal{G}''$ is a minimal $\frac{\eta}{2}$-net of $\mathcal{G}_{\mathcal{H}}'$. One can verify that the union $\bigcup_{i=-S}^{S}\{\mathcal{G}'' + is\frac{\eta}{2}\}$ is an η-net of $\mathcal{G}_{\mathcal{H},b}'$, where S is the smallest integer that is larger than $\frac{2}{\eta}$. To see this, let $g_1^b = (2s)^{-1}(f_1' + b_1)$ be an arbitrary point



(a) The fitted regression functions with $\tau = 0.5$ in the mcycle data.

(b) The fitted regression functions with $\tau = 0.1$ in the mcycle data.

Figure 6: The estimated functions for the mcycle data with $\tau = 0.5$ and $\tau = 0.1$. The dashed lines correspond to $\tilde{f}_n$ using the squared norm penalty method, and the solid lines correspond to $\hat{f}_n$ using the data sparsity method. Note that Takeuchi et al. (2006) also plotted the estimator for the squared norm penalty in their Figure 3. Compared to the dashed lines, the solid lines on both panels have less fluctuations especially when the predictor value is large, and are more interpretable.

in $\mathcal{G}'_{\mathcal{H},b}$, and $g_2^b = (2s)^{-1}(f_2' + b_2)$, with $(2s)^{-1}f_2'$ being the corresponding point in $\mathcal{G}''$ that has an $L_2(T_X')$ distance to $(2s)^{-1}f_1'$ smaller than $\frac{\eta}{2}$, and $b_2 = is\frac{\eta}{2}$ for some integer $i \in [-S, S]$, such that $|b_2 - b_1| \le s\frac{\eta}{2}$. Now the $L_2(T_X')$ distance between g_1^b and g_2^b is

$$\begin{aligned}
&\Big(\int (g_1^b - g_2^b)^2\Big)^{1/2} \\
&\le (2s)^{-1}\Big(\int (2(f_1' - f_2')^2 + 2(b_1 - b_2)^2)^2\Big)^{1/2} \\
&\le \eta,
\end{aligned}$$

where the integral is taken with respect to the counting measure on T_X'. Therefore, the covering number of $\mathcal{G}'_{\mathcal{H},b}$ is less than $(\frac{4}{\eta} + 2)\exp(C_0\eta^{-2})$. Consequently, the covering number of $\mathcal{G}_{\mathcal{H},b}$ is upper bounded by $(\frac{4}{\eta} + 2)\exp(C_0\eta^{-2})$. The desired result follows when η is small enough. □

Proof of Theorem 1: The outline of the proof is as follows. First, we define $M = \sqrt{2}n^{-1/2}\log(n)$. Notice the difference between M (a number) and $\mathcal{M}$ (a functional space used for the definition of the entropy number introduced just before Lemma 14). Then we bound the probability $P(e(\hat{f}_n, f^{(\infty)}) \ge 8sM + d_{n,s})$. In particular, we show that $P(e(\hat{f}_n, f^{(\infty)}) \ge 8sM + d_{n,s}) \le 6(1 - \frac{1}{16nM^2})^{-1}\exp(-nM^2)$, then apply the Borel-Cantelli Lemma to obtain the result.

For $M = \sqrt{2}n^{-1/2}\log(n)$, we first verify that for a large n, this M satisfies

$$(\log_2 \frac{16\sqrt{6}\eta_{n,0}}{M} + 1)^2 (\frac{256C_0}{n}) \leq \frac{M^2}{256}, \tag{8}$$

where $\eta_{n,0} > 0$ is chosen to satisfy

$$\frac{C_0}{\eta_{n,0}^2} + \log \frac{5}{\eta_{n,0}} = \frac{1}{4} n M^2, \tag{9}$$

and $C_0 = 2^{10}$ is a constant as in Lemma 14. From (9), one can verify that $\eta_{n,0}$ goes to 0. Now (8) is equivalent to $(\log_2 \frac{16\sqrt{6}\eta_{n,0}}{M} + 1)^2 \leq \frac{nM^2}{2^{16}C_0}$. Note that $\frac{nM^2}{2^{16}C_0}$ is of the order $O_P(\log(n))^2$. The order of $(\log_2 \frac{16\sqrt{6}\eta_{n,0}}{M} + 1)^2$ is less than that of $(\log \frac{1}{M})^2$, which is $O_P(\log \frac{n^{1/2}}{\log(n)})^2$, and $O_P(\log \frac{n^{1/2}}{\log(n)})^2 < O_P(\log(n))^2$. Thus with n large enough, (8) holds.

Now we prove $P(e(\hat{f}_n, f^{(\infty)}) \geq 8sM + d_{n,s}) \leq 6(1 - \frac{1}{16nM^2})^{-1} \exp(-nM^2)$. We have, by definition of $d_{n,s}$,

$$P(e(\hat{f}_n, f^{(\infty)}) > 8sM + d_{n,s}) \leq P(e(\hat{f}_n, f_n^{(s)})(2s)^{-1} > 4M).$$

Then because $\frac{1}{n}\sum_{i=1}^n (L(f_n^{(s)}, y_i) - L(\hat{f}, y_i)) > 0$, we have

$$\begin{aligned}
&P(e(\hat{f}_n, f^{(\infty)}) > 8sM + d_{n,s}) \\
&\leq P^*(\sup_{f \in \mathcal{F}_n(s): e(f, f_n^{(s)})(2s)^{-1} > 4M} \frac{1}{n}\sum_{i=1}^n (L(f_n^{(s)}, y_i) - L(f, y_i)) > 0) \\
&\leq P^*(\sup_{f \in \mathcal{F}_n(s): e(f, f_n^{(s)})(2s)^{-1} > 4M} -\frac{1}{n}\sum_{i=1}^n [g_f(y_i) - E g_f(y)] > (2s)^{-1} E(L(f, Y) - L(f_n^{(s)}, Y))) \\
&\leq P^*(\sup_{g_f \in \mathcal{G}} |P_n g_f - P g_f| > 4M).
\end{aligned}$$

Here P^* denotes the outer probability, and the expectation is taken jointly with respect to the distribution of X and the noise. In the last inequality, for brevity we have the empirical process $g_f \to P_n g_f - P g_f$, where $g_f \in \mathcal{G}$, $P g_f = \int g_f$ and $P_n g_f = \frac{1}{n}\sum_{i=1}^n g_f(y_i)$.

Now we show that $P^*(\sup_{g_f \in \mathcal{G}} |P_n g_f - P g_f| > 4M) \leq 6(1 - \frac{1}{16nM^2})^{-1} \exp(-nM^2)$.

This part of proof follows a similar line as Theorem A.2 in Wang and Shen (2007). There are two steps involved. The first step is to sample n observations without replacement from $N = 2n$ instances, which are *i.i.d.* samples from P, and let $(W_1, \ldots, W_N)$ be uniformly distributed on the set of all permutations of $1, \ldots, N$. Define $\tilde{P}_{n,N} = \frac{1}{n}\sum_{i=1}^n \delta_{(X)_{W_i}}$, and $P_N = \frac{1}{N}\sum_{i=1}^N \delta_{(X)_i}$, where $\delta_{(X)_i}$ is the Dirac measure at the observation X_i. Then we can bound the LHS of the required inequality by Lemma 2.14.18 in Van der Vaart and Wellner (2000),

$$P^*(\sup_{g_f \in \mathcal{G}} |P_n g_f - P g_f| > 4M) \leq (1 - \frac{1}{16nM^2})^{-1} P^*_{|N}(\sup_{\mathcal{G}} |\tilde{P}_{n,N} g_f - P_N g_f| > M), \tag{10}$$

where $P_{|N}$ is the conditional probability given N observations.

The second step is to bound $P^*_{|N}(\sup_{\mathcal{G}} |\tilde{P}_{n,N} g_f - P_N g_f| > M)$. We apply the chaining technique here. Let $\eta_{n,0} > \eta_1 > \cdots > \eta_T > 0$ be a sequence of positive numbers to be determined later

on. Let $\mathcal{G}_q$ be the minimal η_q-net for $\mathcal{G}$ with respect to the $L_2(T_X)$ norm. For each q, let $\pi_q g = \text{argmin}_{h\in\mathcal{G}_q}\|g-h\|_{L_2(T_X)}$. That means, $\pi_q g$ is the closest point to g within $\mathcal{G}_q$. By definition, $|\mathcal{G}_q| = N(\eta_q,\mathcal{G},L_2(T_X))$, and $\|\pi_q g - g\|_{L_2(T_X)} \le \eta_q$. Hence, decompose $P^*_{|N}(\sup_{\mathcal{G}}|\tilde{P}_{n,N}g_f - P_N g_f| > M)$ into

$$\begin{aligned}
&P^*_{|N}(\sup_{\mathcal{G}}|\tilde{P}_{n,N}g_f - P_N g_f| > M)\\
\le\ & P^*_{|N}(\sup_{\mathcal{G}}|(\tilde{P}_{n,N}-P_N)(\pi_0 g_f)| > 7M/8)\\
&+P^*_{|N}(\sup_{\mathcal{G}}|(\tilde{P}_{n,N}-P_N)(\pi_0 g_f - \pi_T g_f)| > M/16)\\
&+P^*_{|N}(\sup_{\mathcal{G}}|(\tilde{P}_{n,N}-P_N)(\pi_T g_f - g_f)| > M/16)\\
\le\ & |\mathcal{G}_0|\sup_{\mathcal{G}}P^*_{|N}(|(\tilde{P}_{n,N}-P_N)(\pi_0 g_f)| > 7M/8)\\
&+\sum_{q=1}^{T}|\mathcal{G}_q||\mathcal{G}_{q-1}|\sup_{\mathcal{G}}P^*_{|N}(|(\tilde{P}_{n,N}-P_N)(\pi_0 g_f - \pi_T g_f)| > \chi)\\
&+P^*_{|N}(\sup_{\mathcal{G}}|(\tilde{P}_{n,N}-P_N)(\pi_T g_f - g_f)| > M/16)\\
:=\ & P_1+P_2+P_3,
\end{aligned}$$

where $T\chi \le M/16$. Next, we bound P_1, P_2 and P_3 individually.

For P_3, one can verify that $\tilde{P}_{n,N}f \le 2P_N f$ for any non-negative f. Note that we have $((\tilde{P}_{n,N}-P_N)z)^2 \le 2(\tilde{P}_{n,N}z^2+P_N z^2)$ for any z. Thus, $|(\tilde{P}_{n,N}-P_N)(\pi_T g_f - g_f)|^2 \le 2(\tilde{P}_{n,N}+P_N)(\pi_T g_f - g_f)^2 \le 6\eta_T^2$. This yields $P_3 = 0$ if we choose $\eta_T = \frac{M}{16\sqrt{6}}$.

For P_1, note that $0 \le \pi_0 g_f \le 1$ for any $g_f \in \mathcal{G}$ because $\mathcal{G}$ is scaled. By Hoeffding's inequalities for sums of bounded random variables (Hoeffding, 1963), $P^*_{|N}(|(\tilde{P}_{n,N}-P_N)(\pi_0 g_f)| > 7M/8) \le 2\exp(-2n(7/8)^2M^2)$. Thus by assumption (9) and Lemma 14,

$$P_1 \le 2N(\eta_{n,0},\mathcal{G},L_2(T_X))\exp(-2n(7/8)^2M^2) \le 2\exp(-nM^2).$$

For P_2, if $\eta_{n,0} \le \frac{M}{16\sqrt{6}}$, then let $\eta_q = \eta_{n,0}; q = 1,\ldots,T$, and we have $P_2 = 0$ by a similar argument as in the P_3 part discussed above. So suppose $\eta_{n,0} > \frac{M}{16\sqrt{6}} > \eta_T$. Note that $P_N(\pi_q g_f - \pi_{q-1}g_f)^2 \le 2(P_N(\pi_q g_f - g_f)^2 + P_N(\pi_{q-1}g_f - g_f)^2) \le 4\eta_{q-1}^2$. By Massart's inequality from Lemma 2.14.19 in Van der Vaart and Wellner (2000), we have $P^*_{|N}(|(\tilde{P}_{n,N}-P_N)(\pi_0 g_f - \pi_T g_f)| > \chi) \le 2\exp(-\frac{n\chi^2}{2\sigma_N^2})$ with $\sigma_N^2 = P_N(\pi_q g_f - \pi_{q-1}g_f)^2 \le 4\eta_{q-1}^2$. So,

$$\begin{aligned}
P_2 &\le 2\sum_{q=1}^{T}|\mathcal{G}_q|^2\exp(-\frac{n\chi^2}{2\sigma_N^2})\\
&\le 2\sum_{q=1}^{T}\exp(2C_0\eta_q^{-2}+2\log\frac{5}{\eta_q}-\frac{n\chi^2}{8\eta_{q-1}^2}).
\end{aligned}$$

Let $\eta_q = 2^{-q}\eta; q = 1,\ldots,T$, and let T be the greatest integer that does not exceed $\log_2\frac{16\sqrt{6}\eta_{n,0}}{M}$. Then let $\chi = (\frac{256C_0}{n})^{1/2}$. By assumption (8), we can verify that $T\chi \le M/16$ as satisfied. Because

when η is small enough, $\frac{C_0}{\eta^2} > \log\frac{5}{\eta}$, and this leads to

$$\begin{aligned} P_2 &\le 2\sum_{q=1}^{T}\exp(\frac{-4}{\eta_q^2}C_0) \\ &\le 2\sum_{q=1}^{T}\exp(-2^{2q-1}(nM^2)) \\ &\le 4\exp(-nM^2). \end{aligned}$$

Thus, $P^*_{|N}(\sup_{\mathcal{G}}|\tilde{P}_{n,N}g_f - P_N g_f| > M) < 6\exp(-nM^2)$. The desired result follows after we take expectation with respect to the distribution of N observations. We have proved $P^*(\sup_{g_f\in\mathcal{G}}|P_n g_f - Pg_f| > 4M) \le 6(1-\frac{1}{16nM^2})^{-1}\exp(-nM^2)$.

Finally, observe that $nM^2 = 2(\log(n))^2 > 2\log(n)$. We have $\exp(-nM^2) \le \exp(-2\log(n)) = \frac{1}{n^2}$. The desired result in Theorem 1 then follows from Borel-Cantelli Lemma. ■

Appendix B. Proof of Corollary 5

The key to the proof is to show that with a high probability, the estimated b would be bounded in a range. Then we can apply the same technique as that in the proof of Theorem 1 to prove the desired result.

Without loss of generality, assume $|f_0(X)| < \zeta$ for a fixed $\zeta > 0$. Moreover, for simplicity, we assume that $\varepsilon(X)$ follows a common sub-Gaussian distribution with c.d.f. Φ_ε. The generalization to heteroscedastic cases is straightforward, because we are only concerned with the tail probability $\mathrm{pr}(|\varepsilon(X)| > t)$. Next, for a small positive number δ, define $t^* = \Phi_\varepsilon^{-1}\big(0.5+0.5(1-\delta/2)^{1/n}\big)$. One can verify that with probability at least $1-\delta/2$, all the errors $\varepsilon_i;\ i = 1,\ldots,n$ are in $[-t^*, t^*]$. Therefore, for any τ, we have that with probability at least $1-\delta/2$, $|b| \le \zeta + t^*$. This is because the estimated function cannot be smaller (or larger) than all the observations. Hence, letting $s^* = s+\zeta+t^*$, $M^* = \sqrt{2}n^{-1/2}\log(n)/t^*$ and using similar techniques as that in the proof of Theorem 1, we have $P(e(\hat{f}_n^*, f^{(\infty)}) > 8s^*M^* + d_{n,s^*}) \le 6(1-\frac{1}{16nM^{*2}})^{-1}\exp(-nM^{*2}) + \delta/2$. Let δ converge to zero at the rate $O_P(n^{-2}\log(n))$. The last step is to check that (8) and (9) are both true with our new choice of M^*. Because we assume that Φ_ε is the c.d.f. of a sub-Gaussian distribution with a fixed parameter, one can verify that t^* diverges at a rate slower than $O_P(\log(n))$. Hence, (8) and (9) remain valid, and the Borel-Cantelli Lemma as in the final step of the proof of Theorem 1 holds. This completes the proof. ■

Appendix C. Proof of Theorem 7

The proof uses a similar technique as Theorem 2.7 in Steinwart and Scovel (2007). Consider the function

$$V(x) = C\int_D \exp(\frac{|x-x'|^2}{2\sigma^2})f_{\text{true}}(x')dx', \tag{11}$$

where C is a constant that depends only on σ and p and is to be determined later on. One can verify that $V(x) \in \mathcal{F}(\infty)$ (Steinwart et al., 2006). Hence,

$$\begin{aligned} A(\infty) &\leq E\big(\rho_\tau(Y-V)\big) - E\big(\rho_\tau(Y-f_{\text{true}})\big) \\ &\leq E_{P_X}\big(\rho_\tau(V-f_{\text{true}})\big). \end{aligned}$$

Therefore, one only needs to bound $E_{P_X}\big(\rho_\tau(V-f_{\text{true}})\big)$. We have

$$\begin{aligned} V(x) &= C\int_D \exp\Big(\frac{|x-x'|^2}{2\sigma^2}\Big)\big(f_{\text{true}}(x')+a\big)dx' - a \\ &\geq C\int_{B(x,\psi_x)} \exp\Big(\frac{|x-x'|^2}{2\sigma^2}\Big)\big(f_{\text{true}}(x')+a\big)dx' - a \\ &= a_i - (a_i+a)P'(|u|\geq \psi_x), \end{aligned}$$

where the first inequality is because $f_{\text{true}}(x')+a$ is lower bounded by 0, and on $B(x,\psi_x)$ the function f_{true} is constant. Here one chooses C such that P' is the measure of a spherical Gaussian in D (see the definition of spherical Gaussian in, for example, Steinwart, 2002). By the inequality (3.5) on page 59 of Ledoux and Talagrand (1991), $P'(|u|\geq\psi_x) \leq 4\exp\big(-\psi_x^2/(2p\sigma^2)\big)$, and consequently we have that on D_i,

$$V - f_{\text{true}} \geq -8a\exp\big(-\psi_x^2/(2p\sigma^2)\big).$$

An analogous derivation on $-f_{\text{true}}$ and $-V$ gives

$$V - f_{\text{true}} \leq 8a\exp\big(-\psi_x^2/(2p\sigma^2)\big).$$

Therefore

$$E\big(\rho_\tau(V-f_{\text{true}})\big) \leq \max(\tau,1-\tau)E_{P_X}\big\{8a\exp\big(-\psi_x^2/(2p\sigma^2)\big)\big\}.$$

This completes the proof.

Notice that the core part of the proof is at the inequality (3.5) on page 59 of Ledoux and Talagrand (1991). For the Gaussian kernel (and other radial kernels), the probability $P'(|u|\geq\psi_x)$ vanishes as $\sigma\to 0$. However, for other kernels this may not be true. Take the polynomial kernel as an example. One can verify that when $|x_1-x_2|$ is large, $K(x_1,x_2)$ is large, and this leads to $P'(|u|\geq\psi_x)$ being large. Therefore, the result here does not hold true for general RKHS's. ■

Appendix D. Proof of Corollary 8

Without loss of generality, let the Lipschitz constant of f_{true} be 1. In this proof, let ε be a small positive number, instead of the noise as $Y = f_0+\varepsilon$. We first consider the approximation of V to f_{true} on $[0,1]^p$, where V is defined as in the proof of Theorem 7. Because $[0,1]^p$ is a compact set, there exists a finite set of $B(x_j,\varepsilon);\ j=1,\ldots,J$ that covers $[0,1]^p$. Here $B(x_j,\varepsilon)$ is a ball with center at x_j and radius ε, and J is a positive integer. Based on $B(x_j,\varepsilon);\ j=1,\ldots,J$, one can construct sets $S_{x_j}\subset B(x_j,\varepsilon);\ j=1,\ldots,J$ such that S_{x_i} and S_{x_j} are non-overlapping for $i\neq j$, and $\bigcup_{j=1}^J S_{x_j} = [0,1]^p$. On each S_{x_j}, one can verify that

$$f_{\text{true}}(x_j) - \varepsilon \leq f_{\text{true}} \leq f_{\text{true}}(x_j) + \varepsilon.$$

Now on S_{x_j}, define $\bar{f}_{\text{true}} = f_{\text{true}}(x_j)$. We have that

$$E_{[0,1]^p}\rho_\tau(V - f_{\text{true}}) \leq E_{[0,1]^p}\rho_\tau(V - \bar{f}_{\text{true}}) + E_{[0,1]^p}\rho_\tau(\bar{f}_{\text{true}} - f_{\text{true}}). \tag{12}$$

By Theorem 7 and the discussion thereafter, the first part on the right hand side of (12) goes to 0 with $\sigma \to 0$, and the second part is upper bounded by ε. Let $\varepsilon \to 0$ and this proves that V can approximate f_{true} arbitrarily well on $[0,1]^p$. For the approximation on D, notice that D can be decomposed into countably many sets such as $[0,1]^p$, and a similar argument as above proves the corollary. ■

Appendix E. Proof of Theorem 9

To prove Theorem 9, we need to introduce the Rademacher variables (see, for example, Bartlett and Mendelson (2002), Koltchinskii and Panchenko (2002), Shawe-Taylor and Cristianini (2004), Bartlett et al. (2005), Koltchinskii (2006), Mohri et al. (2012) and the references therein). With a little abusing of notations, let σ_i; $i = 1,\ldots,n$ be *i.i.d.* random variables that take 1 with probability $1/2$, and -1 with probability $1/2$. Denote by S a sample of (x_i, y_i); $i = 1,\ldots,n$, *i.i.d.* from the joint distribution of X and Y. Recall the definition of $\mathcal{F}_n(s)$ in Section 3.1. With S fixed, we define the empirical Rademacher complexity of the function class $\mathcal{F}_n(s)$ as

$$\hat{R}_n(\mathcal{F}_n(s)) = E_\sigma[\sup_{f \in \mathcal{F}_n(s)} \frac{1}{n}\sum_{i=1}^n \sigma_i \rho_\tau\big(y_i - f(x_i)\big)],$$

where E_σ represents the expectation with respect to $\sigma = (\sigma_1,\ldots,\sigma_n)$. Moreover, let the Rademacher complexity of $\mathcal{F}_n(s)$ be

$$R_n(\mathcal{F}_n(s)) = E_S \hat{R}_n(\mathcal{F}_n(s)),$$

where E_S is the expectation with respect to the distribution of S.

The proof of Theorem 9 follows directly from Lemmas 15, 17 and 18. Lemma 15 bounds $E\rho_\tau(Y - \hat{f}_n)$ or $E\rho_\tau(Y - \tilde{f}_n)$ in terms of the sum of its empirical measurement, the Rademacher complexity of the function class $\mathcal{F}_n(s)$, and a penalty term on δ, where δ is the small probability that the bound fails. Lemma 17 and Lemma 18 bound the Rademacher complexity. In particular, Lemma 17 provides the bound with $\mu = \sqrt{n^{-1}\big(2^{11}n^{1/4} + 2\log(5) + 0.5\log(n)\big)}$ that works for both the regular squared norm method and the data sparsity method. This is because the complexity bound of $\mathcal{F}_n(s)$ is from Lemma 14, which holds for both methods. As discussed in the main text, we provide another bound that only works for the data sparsity method in Lemma 18, which leads to $\mu = s\sqrt{\frac{2\log(2n+2)}{n}}$.

Lemma 15 *Define $R_n(\mathcal{F}_n(s))$ and $\hat{R}_n(\mathcal{F}_n(s))$ as above. Suppose Assumption A holds. With probability at least $1-\delta$,*

$$E\rho_\tau(Y - \hat{f}_n) \leq \frac{1}{n}\sum_{i=1}^n \rho_\tau(y_i - \hat{f}_n(x_i)) + 2R_n(\mathcal{F}_n(s)) + T_n(\delta), \tag{13}$$

where $T_n(\delta) = \max(\tau, 1-\tau)\Big(n^{-1}(2s^2 + 2t^2)\log(1/\delta)\Big)^{1/2}$. In addition, with probability at least $1-\delta$,

$$E\rho_\tau(Y - \hat{f}_n) \leq \frac{1}{n}\sum_{i=1}^n \rho_\tau(y_i - \hat{f}_n(x_i)) + 2\hat{R}_n(\mathcal{F}_n(s)) + 3T_n(\delta/2). \tag{14}$$

Moreover, (13) and (14) hold for $\tilde{f}_n$.

Proof of Lemma 15: We divide the proof into three parts. In the first part, we bound the left hand side of (13) in terms of its empirical estimation and the expectation of their supremum difference, by the McDiarmid inequality (McDiarmid, 1989). The second part bounds the expectation of the supremum difference from the first step by the previously defined Rademacher complexity with a symmetrization technique. In the third part, we bound the Rademacher complexity by its empirical version. In this proof, we focus on $\hat{f}_n$, as the proof for $\tilde{f}_n$ is the same.

We begin the proof by introducing some notation. For a given sample S, let

$$\phi(S) = \sup_{f \in \mathcal{F}_n(s)} [E\rho_\tau(Y - f(X)) - \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - f(x_i))].$$

Define $S^{(i,x)} = \{(x_1, y_1), \ldots, (x_i', y_i), \ldots, (x_n, y_n)\}$ to be another sample from the joint distribution of X and Y. Notice that the difference between S and $S^{(i,x)}$ is only on the x value of their i^{th} pair. Similarly, define $S^{(i,y)} = \{(x_1, y_1), \ldots, (x_i, y_i'), \ldots, (x_n, y_n)\}$ with the y values in the i^{th} pair being different. Then we have

$$\begin{aligned} |\phi(S) - \phi(S^{(i,x)})| = | &\sup_{f \in \mathcal{F}_n(s)} [E\rho_\tau(Y - f(X)) - \frac{1}{n}\sum_{S} \rho_\tau(y_i - f(x_i))] \\ &- \sup_{f \in \mathcal{F}_n(s)} [E\rho_\tau(Y - f(X)) - \frac{1}{n}\sum_{S^{(i,x)}} \rho_\tau(y_i - f(x_i))]|. \end{aligned} \tag{15}$$

For simplicity, we consider only the case where there exists a measurable function $f^S \in \mathcal{F}_n(s)$ that achieves the supremum of $\phi(S)$. Note that the case of no function achieving the supremum can be treated similarly, with only minor modification on the proof, and the details are omitted. Substitute f^S in (15) and after some calculation, we have that

$$|\phi(S) - \phi(S^{(i,x)})| \leq \frac{2}{n}\max(\tau, 1-\tau)s.$$

Similarly, one can verify that $|\phi(S) - \phi(S^{(i,y)})| \leq \frac{2}{n}\max(\tau, 1-\tau)t$. Hence, by the McDiarmid inequality, for any $z > 0$, $P(\phi(S) - E\phi(S) \geq z) \leq \exp\Big(-\frac{2z^2}{\frac{1}{n}\Big(\max(\tau,1-\tau)\Big)^2(4s^2+4t^2)}\Big)$. Therefore, with probability at least $1-\delta$, $\phi(S) - E\phi(S) \leq T_n(\delta)$. This proves the first part of the lemma.

In the second step, we bound $E\{\phi(S)\}$ by the Rademacher complexity $R_n(\mathcal{F}_n(s))$ using a symmetrization technique. To this end, define $S' = \{(x_i', y_i')\ (i = 1, \ldots, n)\}$ as a duplicate sample of S with size n, and assume the distribution of S' is the same as S. Recall the definition of E_S. Moreover, notice that

$$E_{S'}\big[\frac{1}{n}\sum_{S}\big\{\rho_\tau\big(y_i' - \hat{f}_n(x_i')\big) \mid S\big\}\big] = \frac{1}{n}\sum_{S}\big\{\rho_\tau\big(y_i' - \hat{f}_n(x_i')\big),$$

and

$$E_{S'}\big[\frac{1}{n}\sum_{S'}\big\{\rho_\tau\big(y_i' - \hat{f}_n(x_i')\big) \mid S\big\}\big] = E\rho_\tau\big(Y - \hat{f}_n(X)\big).$$

Hence, with the Jensen's inequality and the definition of σ, we have

$$\begin{aligned}
E\{\phi(S)\} &= E_S\big(\sup_{f\in\mathcal{F}_n(s)} E_{S'}\big[\frac{1}{n}\sum_{S'}\{\rho_\tau(y_i'-\hat{f}_n(x_i'))\}-\frac{1}{n}\sum_{S}\{\rho_\tau(y_i-\hat{f}_n(x_i))\}\big]\mid S\big)\\
&\le E_{S,S'}\big[\frac{1}{n}\sum_{S'}\{\rho_\tau(y_i'-\hat{f}_n(x_i'))\}-\frac{1}{n}\sum_{S}\{\rho_\tau(y_i-\hat{f}_n(x_i))\}\big]\\
&= E_{S,S',\sigma}\big[\sup_{f\in\mathcal{F}_n(s)}\frac{1}{n}\sum_{S'}\{\rho_\tau(y_i'-\hat{f}_n(x_i'))\}-\frac{1}{n}\sum_{S}\{\rho_\tau(y_i-\hat{f}_n(x_i))\}\big]\\
&\le 2R_n(\mathcal{F}_n(s)).
\end{aligned}$$

This completes the proof of the second step.

The third step bounds $R_n(\mathcal{F}_n(s))$ by the empirical counterpart $\hat{R}_n(\mathcal{F}_n(s))$. This part of the proof is similar to that of the first part, in the sense that we apply the McDiarmid inequality on $\hat{R}_n(\mathcal{F}_n(s))$ and its expectation $R_n(\mathcal{F}_n(s))$. One can then verify that with probability at least $1-\delta$, $R_n(\mathcal{F}_n(s)) \le \hat{R}_n(\mathcal{F}_n(s))+T_n(\delta)$.

The proof of Lemma 15 is thus completed, after combining the results in Steps 1-3 and replacing δ by $\delta/2$. □

The next lemma, Lemma 17, bounds $\hat{R}_n(\mathcal{F}_n(s))$ with the data and tuning parameter that we use. It employs the result obtained in Lemma 14, and is a direct application of the "η-net" idea (Van der Vaart and Wellner, 2000). Because the result in Lemma 14 can be applied to both the regular squared norm method and the data sparsity method, the result in Lemma 17 holds for both methods as well. Before discussing Lemma 17 and its proof, we first introduce the Hoeffding's Inequality.

Proposition 16 *(Hoeffding's Inequality). Let X be a random variable with mean 0 and range in $[a,b]$. Then for any fixed $z>0$, $E(\exp(zX)) \le \exp(z^2(b-a)^2/8)$.*

Lemma 17 *The empirical Rademacher complexity $\hat{R}_n(\mathcal{F}_n(s))$ for $\tilde{f}_n$ satisfies that*

$$\hat{R}_n(\mathcal{F}_n(s)) \le 2sn^{-1/4}+\max(\tau,1-\tau)(s+t)\sqrt{2^{11}n^{-1/2}+(\log(5)/n)+(\log(n)/4n)}.$$

Proof of Lemma 17: Let $R=\frac{\hat{R}_n\mathcal{F}_n(s)}{2s}$. We consider the following function $h_f(\cdot)=(2s)^{-1}\rho_\tau(f-\cdot)$, and the corresponding class $\mathcal{H}_R=\{h_f : f\in\mathcal{F}_n(s)\}$. From the proof of Lemma 14, one can verify that the entropy number of $\mathcal{H}_R$ is bounded by that of $\mathcal{G}$, because the check function is upper bounded by $L(f,\cdot)$ defined after the proof of Lemma 13. Next, we construct the smallest η-net of $\mathcal{H}_R$, $\mathcal{Y}$, such that for all $f\in\mathcal{F}_n(s)$, there exists an element $g_\mathcal{Y}\in\mathcal{Y}$ with the $L_2(T_X)$ distance between f and $g_\mathcal{Y}$ smaller than η, for any arbitrary empirical measure T_X. Therefore, one can verify that

$$\begin{aligned}
R \le& \frac{1}{2s}E_\sigma[\sup_{g_\mathcal{Y}}\frac{1}{n}\sum_{i=1}^n\sigma_i\rho_\tau(g_\mathcal{Y}(x_i)-y_i)]\\
&+\frac{1}{2s}E_\sigma[\sup_{f,g_\mathcal{Y}\text{ close}}\frac{1}{n}\sum_{i=1}^n\sigma_i\big(\rho_\tau(f(x_i)-y_i)-\rho_\tau(g_\mathcal{Y}(x_i)-y_i)\big)]\\
:=&R_1+R_2.
\end{aligned}$$

Here "$f, g_{\mathcal{Y}}$ close" means that the $L_2(T_X)$ distance between f and $g_{\mathcal{Y}}$ is smaller than η. Consequently, we have that R_2 is bounded by η, by the Hölder's Inequality.

Now we bound R_1. To this end, let z be a positive number to be determined later. From the Jensen's Inequality, we have

$$\begin{aligned}\exp(2sznR_1) \leq & E_\sigma \exp[z \sup_{g_{\mathcal{Y}}} \sum_{i=1}^n \sigma_i \rho_\tau(g_{\mathcal{Y}}(x_i) - y_i)] \\ \leq & \sum_{g_{\mathcal{Y}}} E_\sigma[\exp(z \sum_{i=1}^n \sigma_i \rho_\tau(g_{\mathcal{Y}}(x_i) - y_i))] \\ \leq & \sum_{g_{\mathcal{Y}}} \prod_{i=1}^n E_{\sigma_i}[\exp(z\sigma_i \rho_\tau(g_{\mathcal{Y}}(x_i) - y_i)].\end{aligned}$$

Observe that $E_{\sigma_i}[\sigma_i \rho_\tau(g_{\mathcal{Y}}(x_i) - y_i)] = 0$, and

$$-\max(\tau, 1-\tau)(s+t) \leq \rho_\tau(g_{\mathcal{Y}}(x_i) - y_i) \leq \max(\tau, 1-\tau)(s+t).$$

Therefore, by the Hoeffding's Inequality, we have

$$\begin{aligned}\exp(2sznR_1) \leq & \sum_{g_{\mathcal{Y}}} \prod_{i=1}^n \exp(\frac{z^2(2\max(\tau, 1-\tau)(s+t))^2}{8}) \\ \leq & |\mathcal{Y}| \exp(\frac{nz^2(\max(\tau, 1-\tau)(s+t))^2}{2}).\end{aligned}$$

Equivalently, we have $2snR_1 \leq \frac{\log|\mathcal{Y}|}{z} + \frac{nz(\max(\tau,1-\tau)(s+t))^2}{2}$. Choose $z = \frac{\sqrt{2\log|\mathcal{Y}|}}{\max(\tau,1-\tau)(s+t)\sqrt{n}}$, and we have

$$2snR_1 \leq \max(\tau, 1-\tau)(s+t)\sqrt{2n\log|\mathcal{Y}|},$$

or equivalently,

$$R_1 \leq \frac{\max(\tau, 1-\tau)(s+t)\sqrt{2\log|\mathcal{Y}|}}{2s\sqrt{n}} \leq \frac{\max(\tau, 1-\tau)(s+t)}{2s\sqrt{n}} \sqrt{2(C_0\eta^{-2} + \log(5/\eta))}.$$

After we combine the bounds of R_1 and R_2, choose $\eta = n^{-1/4}$, the results then follows. □

Next, we focus on the data sparsity method. In Lemma 18 we would prove that the empirical Rademacher complexity of the functional space in (5) can be smaller than that of (4). As we will see, the technique we use only works for the data sparsity constraint. This bound then leads to another finite sample bound on the prediction error, namely, $\mu = s\sqrt{\frac{2\log(2n+2)}{n}}$. As discussed in the main text, when n is small or moderate, this bound can be much better than the one derived from Lemma 14.

Lemma 18 *The empirical Rademacher complexity $\hat{R}_n(\mathcal{F}_n(s))$ for $\hat{f}_n$ in (5) satisfies that*

$$\hat{R}_n(\mathcal{F}_n(s)) \leq s\max(\tau, 1-\tau)\sqrt{\frac{2\log(2n+2)}{n}}.$$

Proof of Lemma 18: We first consider the empirical Rademacher complexity of the functional space $\mathcal{F}_n(s)$ without taking the check loss function into consideration. To this end, let us define $\hat{R}_f(\mathcal{F}_n(s)) = E_\sigma[\sup_{f\in\mathcal{F}_n(s)} \frac{1}{n}\sum_{i=1}^n \sigma_i f(x_i)]$. Recall the definition of $\bar{\alpha}$ from Section 4, and define the augmented vector $\tilde{K}_i = (1, K_i)$ for $i = 1,\ldots,n$, where K_i is the i^{th} row of the gram matrix K. We can now rewrite $\hat{R}_f(\mathcal{F}_n(s))$ as

$$\begin{aligned}\hat{R}_f(\mathcal{F}_n(s)) &= E_\sigma[\sup_{\|\bar{\alpha}\|_1\le s} \frac{1}{n}\sum_{i=1}^n \sigma_i \bar{\alpha}^T \tilde{K}_i] \\ &= \frac{s}{n} E_\sigma \|\sum_{i=1}^n \sigma_i \tilde{K}_i\|_\infty \\ &= \frac{s}{n} E_\sigma[\max_{\sigma', j=1,\ldots,n+1} \sigma' \sum_{i=1}^n \sigma_i \tilde{K}_i(j)],\end{aligned}$$

where $\|\cdot\|_\infty$ is the L_∞ norm, $\tilde{K}_i(j)$ is the j^{th} element of $\tilde{K}_i$, and σ' is an independent Rademacher variable. Notice that the new functional space defined by

$$\{\sigma'(\tilde{K}_1(j), \tilde{K}_2(j), \ldots, \tilde{K}_n(j))^T;\ j = 1,\ldots,n+1, \sigma' \in \{\pm 1\}\}$$

consists of $2n+2$ elements.

Next, by applying the same technique as we used in the proof of Lemma 17 to bound R_1, we can show that

$$\hat{R}_f(\mathcal{F}_n(s)) \le s\sqrt{\frac{2\log(2n+2)}{n}}.$$

The rest of the proof is to apply the Talagrand's lemma (Lemma 4.2 on page 78 in Mohri et al., 2012). In particular, as the check loss function is $\max(\tau, 1-\tau)$-Lipschitz, we have

$$\hat{R}_n(\mathcal{F}_n(s)) \le \max(\tau,1-\tau)\hat{R}_f(\mathcal{F}_n(s)) \le s\max(\tau,1-\tau)\sqrt{\frac{2\log(2n+2)}{n}}.$$

This completes the proof. ■

Appendix F. Proof of Corollary 10

With the t defined in Corollary 10, one can verify that with probability at least $1-\delta/2$, all the errors $\varepsilon_i;\ i = 1,\ldots,n$ are in $[-t,t]$. Conditioning on this, the claim follows from Theorem 9. ■

Appendix G. Proof of Proposition 12

The proof follows directly from that of Theorem 1 in Li et al. (2007) and is omitted. ■

Section 2	**Methodology**
x, X	Predictor variable
y, Y	Response
n	Number of observation
p	Dimensionality of x
τ	Quantile level
$\varepsilon(\cdot)$	Noise, may depend on x
$f_0(\cdot)$	Defined as $Y = f_0 + \varepsilon$
D	Domain of f_0
$\rho_\tau(\cdot)$	The check function
f_{true}	The population minimizer of the check function
$J(\cdot)$	Penalty on the regression function
λ, s	Tuning parameters
$\mathcal{F}$	Functional class
$\mathbb{R}$	The real line
$\mathcal{H}$	A RKHS
$\|\cdot\|_{\mathcal{H}}$	The norm in the RKHS $\mathcal{H}$
b	Intercept
f'	Regression function in $\mathcal{H}$ without an explicit intercept
f	Regression function in $\mathcal{H} \oplus \mathbb{R}$
$K(\cdot,\cdot)$	The kernel function
K	The gram matrix
$\alpha = (\alpha_1, \ldots, \alpha_n)$	The kernel function coefficients
$\hat{f}_n$	The estimated regression function using the proposed data sparsity constraint
$\tilde{f}_n$	The estimated regression function using the squared norm penalty

Table 11: Important notation introduced in Section 2.

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Annals of Statistics*, 36(2):489–531, 2008.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM. ISBN 0-89791-497-X. doi: http://doi.acm.org/10.1145/130385.130401. URL `http://doi.acm.org/10.1145/130385.130401`.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Section 3	**Statistical Theory**
$\mathcal{F}_n(s)$	$\{f = f' + b : f'(x) = \sum_{i=1}^n \alpha_i K(x, x_i); \lvert b\rvert + \sum_{i=1}^n \lvert\alpha_i\rvert \le s\}$
$\mathcal{F}(\infty)$	$\lim_{s\to\infty} \lim_{n\to\infty} \mathcal{F}_n(s)$
$f_n^{(s)}$	$\operatorname{arginf}_{f\in\mathcal{F}_n(s)} E\rho_\tau(Y - f(X))$
$f^{(\infty)}$	$\operatorname{arginf}_{f\in\mathcal{F}(\infty)} E\rho_\tau(Y - f(X))$
$e(f_1, f_2)$	$E\rho_\tau(Y - f_1(X)) - E\rho_\tau(Y - f_2(X))$
$d_{n,s}$	$e(f_n^{(s)}, f^{(\infty)})$, the approximation error between $\mathcal{F}_n(s)$ and $\mathcal{F}(\infty)$
$\hat{f}_n^*$	The estimated function using the data sparsity constraint, without penalty on $\lvert b\rvert$
$\lVert f\rVert_{L_2(Q)}$	$(\int f^2 dQ)^{1/2}$, the $L_2(Q)$ norm of f
$A(\infty)$	$E\big(\rho_\tau(Y - f^{(\infty)})\big) - E\big(\rho_\tau(Y - f_{\text{true}})\big)$
a_i	Constants
D_i	A partition of D, and $f_{\text{true}} = a_i$ on D_i
a	Upper bound on $\lvert f_{\text{true}}\rvert$
$\text{dis}(x, D_j)$	The distance between the point x and the set D_j
ψ_x	$\min_{j\neq i} \text{dis}(x, D_j)$
$B(x, \psi_x)$	The ball centered at x with radius ψ_x
σ	Kernel parameter for Gaussian/Laplacian kernels
P_X	The marginal distribution of X
t	The upper bound of $\lvert\varepsilon\rvert$ in Assumption A
δ	A small probability
μ	$\min\left(2\sqrt{n^{-1/2}\big(\log(n)+1\big)}, \sqrt{n^{-1}\big(2^{11}n^{1/4} + 2\log(5) + 0.5\log(n)\big)}\right)$
Φ_ε	The common cumulative distribution function of ε

Table 12: Important notation introduced in Section 3.

B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators*, volume 98. Cambridge University Press, 1990.

R. Chappell. Fitting bent lines to data, with applications to allometry. *Journal of Theoretical Biology*, 138:235–256, 1989.

D.-R. Chen, Q. Wu, Y. Ying, and D.-X. Zhou. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, 5:1143–1175, 2004.

C. Cortes and V. N. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

F. Cucker and S. Smale. On the mathematical foundations of learning. *American Mathematical Society*, 39(1):1–49, 2002.

C. De Boor. *A Practical Guide to Splines*. Springer-Verlag, 1978.

H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. *Advances in Neural Information Processing Systems*, pages 155–161, 1997.

B. Efron, T. J. Hastie, I. Johnstone, and R. J. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–451, 2004.

P. H. C. Eilers and B. D. Marx. Flexible smoothing using B-splines and penalized likelihood. *Statistical Science*, 11(2):89–121, 1996.

Appendix	
$L(a,b)$	$\lvert a-b\rvert$
$g_f(\cdot)$	$(2s)^{-1}(L(\cdot,f)-L(\cdot,f_n^{(s)}))$, a scaled empirical process
$\mathcal{G}$	$\{g_f : f \in \mathcal{F}_n(s)\}$
T_X	The empirical measure of a training set
$\lVert f\rVert_{L_2(T_X)}$	$\left(\frac{1}{n}\sum_{i=1}^n \lvert f(x_i,y_i)\rvert^2\right)^{1/2}$
$\mathcal{M}$	A η-net with respect to the $\lVert\cdot\rVert_{L_2(T_X)}$ distance
$N(\eta,\mathcal{F},L_2(T_X))$	The η-covering number of $\mathcal{F}$ with respect to the $\lVert\cdot\rVert_{L_2(T_X)}$ distance
M	$\sqrt{2}n^{-1/2}\log(n)$
$\eta_{n,0}$	A number depending on n, chosen to satisfy (9)
C_0	2^{10}, a large constant
$P_n g_f$	Defined as $P_n g_f = \frac{1}{n}\sum_{i=1}^n g_f(y_i)$
Pg_f	$\int g_f$
P^*	The outer probability
N	$N = 2n$
$(W_1,\ldots,W_N)$	A permutation of $1,\ldots,N$ whose distribution is uniform
$\delta_{(X)_i}$	Dirac measure at the observation X_i
$\tilde{P}_{n,N}$	$\frac{1}{n}\sum_{i=1}^n \delta_{(X)_{W_i}}$
P_N	$\frac{1}{N}\sum_{i=1}^N \delta_{(X)_i}$
$P_{\lvert N}$	The conditional probability given N observations
T	A positive integer
$\eta_1,\ldots,\eta_T$	A sequence of T positive numbers
$\mathcal{G}_q$	η_q-net of $\mathcal{G}$
$\pi_q g$	The projection of g on $\mathcal{G}_q$
P_1, P_2, P_3	Three probabilities to be bounded
χ	A number such that $T\chi \le M/16$
σ_N^2	Defined as $\sigma_N^2 = P_N(\pi_q g_f - \pi_{q-1} g_f)^2$
$\mathcal{G}_{\mathcal{H},b}$	$\{(2s)^{-1}L(\cdot,f) : f \in \mathcal{F}_n(s)\}$
$\mathcal{G}_{\mathcal{H}}$	$\{(2s)^{-1}L(\cdot,f') : f' = \sum_{i=1}^n \alpha_i K(x_i,\cdot), \sum_{i=1}^n \lvert\alpha_i\rvert \le s\}$
$\mathcal{G}'_{\mathcal{H}}$	$\{(2s)^{-1}f' : f' = \sum_{i=1}^n \alpha_i K(x_i,\cdot), \sum_{i=1}^n \lvert\alpha_i\rvert \le s\}$
T'_X	The empirical measure of the set $(x_1,\ldots,x_n)$
$B_{\mathcal{H}}$	The unit ball in $\mathcal{H}$
$\mathcal{G}'_{\mathcal{H},b}$	$\{(2s)^{-1}f : f \in \mathcal{F}_n(s)\}$
$\mathcal{G}''$	A minimal $\eta/2$-net of $\mathcal{G}'_{\mathcal{H}}$

Table 13: Important notation introduced in the Appendix (Part 1).

P. H. C. Eilers and B. D. Marx. Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(6):637–653, 2010.

J. Fan and R. Li. Statistical challenges with high dimensionality: feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians*, 3:595–622, 2006.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: a library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

J. H. Friedman and B. W. Silverman. Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31:3–21, 1989.

Appendix	
ζ	Upper bound on f_0. Assumed in the proof of Corollary 5.
t^*	$t^* = \Phi_\varepsilon^{-1}\big(0.5+0.5(1-\delta/2)^{1/n}\big)$
s^*	$s^* = s+\zeta+t^*$
M^*	$M^* = \sqrt{2}n^{-1/2}\log(n)/t^*$
$V(x)$	$C\int_D \exp(\frac{\lvert x-x'\rvert^2}{2\sigma^2}) f_{\text{true}}(x')dx'$
C	A constant such that $V(x)$ can be used to estimate f_{true}
P'	The measure of a spherical Gaussian in D
$\bar{f}_{\text{true}}$	A piecewise constant function used to approximate f_{true}
$\sigma = (\sigma_i;\ i = 1,\ldots,n)$	A set of n Rademacher random variables, where $P(\sigma_i = 1) = 1/2$ and $P(\sigma_i = -1) = 1/2$
S	A sample of $(x_1,y_1),\ldots,(x_n,y_n)$ $i.i.d.$ from the joint distribution of X and Y
$\hat{R}_n(\mathcal{F}_n(s))$	$E_\sigma[\sup_{f\in\mathcal{F}_n(s)} \frac{1}{n}\sum_{i=1}^n \sigma_i\rho_\tau\big(Y-f(X)\big)]$
$R_n(\mathcal{F}_n(s))$	$E_S\hat{R}_n\big(\mathcal{F}_n(s)\big)$
$T_n(\delta)$	$\max(\tau,1-\tau)\Big(n^{-1}(2s^2+2t^2)\log(1/\delta)\Big)^{1/2}$
$\phi(S)$	$\sup_{f\in\mathcal{F}_n(s)}[E\rho_\tau(Y-f(X)) - \frac{1}{n}\sum_{i=1}^n \rho_\tau(y_i-f(x_i))]$
$S^{(i,x)}$	$\{(x_1,y_1),\ldots,(x_i',y_i),\ldots,(x_n,y_n)\}$, the difference between $S^{(i,x)}$ and S is only on the x value of their i^{th} pair
$S^{(i,y)}$	$\{(x_1,y_1),\ldots,(x_i,y_i'),\ldots,(x_n,y_n)\}$
R	$\frac{\hat{R}_n\mathcal{F}_n(s)}{2s}$
$h_f(\cdot)$	$(2s)^{-1}\rho_\tau(f-\cdot)$
$\mathcal{H}_R$	$\{h_f : f\in\mathcal{F}_n(s)\}$
$\mathcal{Y}$	The smallest η-net on $\mathcal{H}_R$
$g_{\mathcal{Y}}$	The projection of f on $\mathcal{Y}$
R_1	$\frac{1}{2s}E_\sigma[\sup_{g_{\mathcal{Y}}} \frac{1}{n}\sum_{i=1}^n \sigma_i\rho_\tau(g_{\mathcal{Y}}(x_i)-y_i)]$
R_2	$\frac{1}{2s}E_\sigma[\sup_{f,g_{\mathcal{Y}}\text{ close}} \frac{1}{n}\sum_{i=1}^n \sigma_i\big(\rho_\tau(f(x_i)-y_i)-\rho_\tau(g_{\mathcal{Y}}(x_i)-y_i)\big)]$
Λ	$(\Lambda_1,\ldots,\Lambda_{2n})$, convex combination parameters of any element in $\mathcal{G}'_{\mathcal{H}}$
$F_1,\ldots,F_k$	$i.i.d.$ random elements with $\text{pr}(F_1 = e_i) = \Lambda_i;\ i = 1,\ldots,2n$
$\bar{F}$	The average of $F_1,\ldots,F_k$

Table 14: Important notation introduced in the Appendix (Part 2)

D. Gervini. Free-knot spline smoothing for functional data. *Journal of the Royal Statistical Society: Series B*, 68(4):671–687, 2006.

C. Gu. *Smoothing Spline ANOVA Models*. Springer-Verlag, 2002.

M. H. Hansen and C. Kooperberg. Spline adaptation in extended linear models. *Statistical Science*, 17(1):2–51, 2002.

T. J. Hastie, R. J. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

X. He, P. Ng, and S. Portnoy. Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society: Series B*, 60(3):537–550, 1998.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

S. S. Keerthi and C.-J. Lin. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation*, 15(7):1667–1689, 2003.

G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971.

R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978.

R. Koenker and O. Geling. Reappraising medfly longevity: a quantile regression survival analysis. *Journal of the American Statistical Association*, 96(454):458–468, 2001.

R. Koenker and K. Hallock. Quantile regression: an introduction. *Journal of Economic Perspectives*, 15(4):43–56, 2001.

R. Koenker, P. Ng, and S. Portnoy. Quantile smoothing splines. *Biometrika*, 81:673–680, 1994.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, 34(6):2593–2656, 2006.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1):1–50, 2002.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer, 1991.

Y. Li and J. Zhu. L1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17:163–185, 2008.

Y. Li, Y. Liu, and J. Zhu. Quantile regression in reproducing kernel Hilbert spaces. *Journal of the American Statistical Association*, 102(477):255–268, 2007.

W. Mao and L. H. Zhao. Free-knot polynomial splines with confidence intervals. *Journal of the Royal Statistical Society: Series B*, 65(4):901–919, 2003.

C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, 1989.

S. Mendelson. A few notes on statistical learning theory. In *Advanced Lectures on Machine Learning*, pages 1–40. Springer, 2003.

H. Q. Minh. Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338, 2010.

S. Miyata and X. Shen. Free-knot splines and adaptive knot selection. *Journal of the Japan Statistical Society*, 35(2):303–324, 2005.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT press, 2012.

D. Nychka, G. Gray, P. Haaland, D. Martin, and M. OConnell. A nonparametric regression approach to syringe grading for quality improvement. *Journal of the American Statistical Association*, 90: 1171–1178, 1995.

V. I. Paulsen. An introduction to the theory of reproducing kernel Hilbert spaces. 2009. Technical report.

S. Rosset. Bi-level path following for cross validated solution of kernel quantile regression. *Journal of Machine Learning Research*, 10:2473–2505, 2009.

D. Ruppert. Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11:735–757, 2002.

D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge University Press, 2003.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. MIT Press, 2002.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

S. Smale and D.-X. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1(1):1–25, 2003.

A. J. Smola and B. Schölkopf. On a kernel-based method for pattern recognition, regression, approximation, and operator inversion. *Algorithmica*, 22(1-2):211–231, 1998.

A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2002.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Annals of Statistics*, 35(2):575–607, 2007.

I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Trans. Inform. Theory*, 52:4635–4643, 2006.

M. Stitson, A. Gammerman, V. Vapnik, V. Vovk, C. Watkins, and J. Weston. Support vector regression with ANOVA decomposition kernels. *Advances in Kernel Methods–Support Vector Learning*, pages 285–292, 1999.

I. Takeuchi, Q. V. Le, T. D. Sears, and A. J. Smola. Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264, 2006.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

A. W. Van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes with Application to Statistics*. Springer, 2000.

V. Vapnik, S. E. Golowich, and A. J. Smola. Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*, pages 281–287, 1997.

G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In *Advances in Kernel Methods Support Vector Learning*, pages 69–88. MIT Press, 1999.

H. Wang and X. He. Detecting differential expressions in genechip microarray studies: a quantile approach. *Journal of the American Statistical Association*, 102(477):104–112, 2007.

L. Wang and X. Shen. On l1-norm multi-class support vector machines: methodology and theory. *Journal of the American Statistical Association*, 102:595–602, 2007.

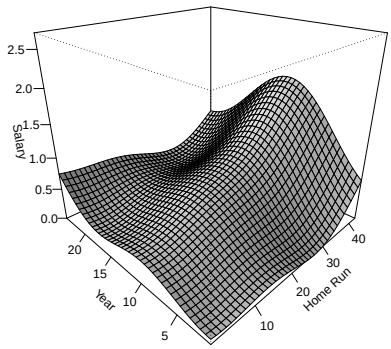
Y. Wei and X. He. Conditional growth charts. *Annals of Statistics*, 34(5):2069–2097, 2006.

M. Yuan. GACV for quantile smoothing splines. *Computational Statistics and Data Analysis*, 5: 813–829, 2006.

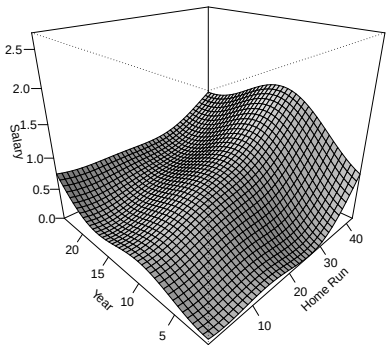
D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

S. Zhou, X. Shen, and D. A. Wolfe. Local asymptotics for regression splines and confidence regions. *Annals of Statistics*, 26(5):1760–1782, 1998.

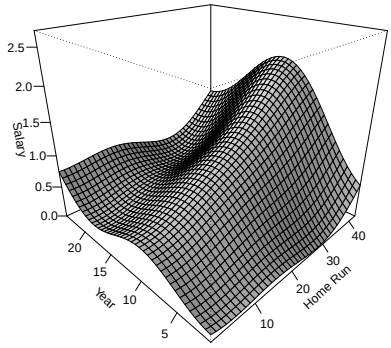
H. Zou and T. J. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.



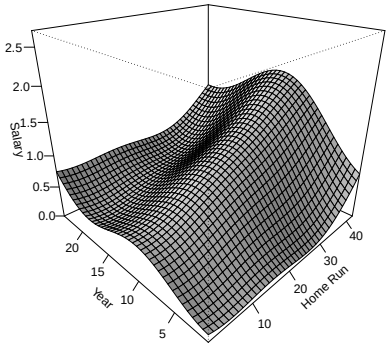
(a) Squared norm penalty with $\tau = 0.25$.



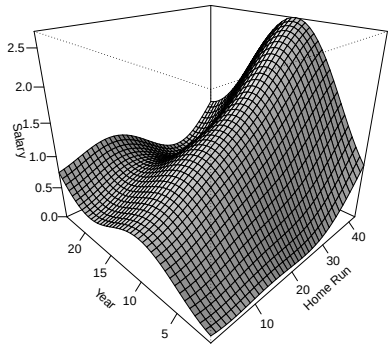
(b) Data sparsity constraint with $\tau = 0.25$.



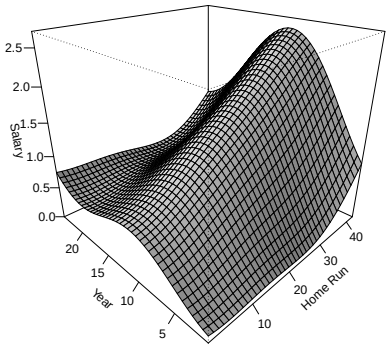
(c) Squared norm penalty with $\tau = 0.5$.



(d) Data sparsity constraint with $\tau = 0.5$.



(e) Squared norm penalty with $\tau = 0.75$.



(f) Data sparsity constraint with $\tau = 0.75$.

Figure 7: Estimated salary for the Baseball data using the number of home runs and the number of years played as the predictors.