Learning Algorithm Interpolating the Data is Optimal for Nonparametric Regression and Prediction with Square Loss

Runcheng Huang
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Illinois, IL 61801
rhuang22@illinois.edu

Abstract

In this paper, I will show that a learning algorithm interpolating the data is optimal for nonparametric regression and prediction with square loss [1]. I will briefly go over the proof of an upper bound in the original paper [1]. Then, I will give some examples and also visualize these examples.

1 Introduction

The original paper [1] showed that the Nadaraya-Watson estimator [4] with some particular kernels achieve optimal rates of convergence for some regression and prediction problems. I will show the main results in the next section.

2 Main Results

Firstly, I will give some definitions that are used in the main result for the nonparametric regression. Then I will show the main result for the nonparametric regression.

2.1 Definitions

The definition of the Hölder class has a lot of versions. I will use the definition in the original paper [1].

Hölder class

$\mathcal{F}_{(\beta,L)}$ is a class of functions mapping from $\mathbb{R}^d$ to $\mathbb{R}$ which satisfy the following conditions.
For any $f \in \mathcal{F}_{(\beta,L)}$, $L > 0$ and $\beta \in (0, 2]$,
If $\beta \in (0, 1]$, $\forall x \in \mathbb{R}^d$ and $\forall y \in \mathbb{R}^d$ then

$$|f(x) - f(y)| \leq L\|x - y\|^\beta \tag{1}$$

If $\beta \in (1, 2]$, $\forall x \in \mathbb{R}^d$ and $\forall y \in \mathbb{R}^d$ then

$$|f(x) - f(y) - \langle \nabla f(y), x - y\rangle| \leq L\|x - y\|^\beta \tag{2}$$

$$\text{and } f \text{ is continuously differentiable} \tag{3}$$

Here are some observations of the Hölder class. When $\beta \in (0, 1]$, the definition of Hölder class generalizes the definition of Lipschitz continuity. $\mathcal{F}_{(1,L)}$ is a L-Lipschitz function class. It is easy to show that the only function class that satisfy $|f(x) - f(y)| \leq L\|x - y\|^\beta$ with $\beta > 1$ is the constant function class.

Nadaraya-Watson estimator

We use a modified definition of Nadaraya-Watson estimator in this paper [1]. We will use the framework of regression in the lecture notes [2]. Let $Z^n = (Z_1, ..., Z_n)$ is a training set, where $Z_i = (X_i, Y_i)$ and $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$. $Z_i = (X_i, Y_i)$ are distributed with unknown distribution P_{XY}. The kernel smoothing Algorithm $\mathcal{A}$ maps Z^n to a Nadaraya-Watson estimator f_n as follows:

$$f_n(x) = \begin{cases} Y_i & \text{if there exists } i \in [n] \text{ such that } x = X_i, \\ 0 & \text{if } \sum_{i=1}^n K(\frac{x - X_i}{h}) = 0, \\ \frac{\sum_{i=1}^n Y_i K(\frac{x-X_i}{h})}{\sum_{i=1}^n K(\frac{x-X_i}{h})} & \text{otherwise.} \end{cases} \tag{4}$$

where $h > 0$ is the bandwidth and $K : \mathbb{R}^d \to \mathbb{R}$ is the kernel function.

Kernels used in the Nadaraya-Watson estimator

The upper bound of the convergence rate of the Nadaraya-Watson estimator with the specific kernels is proved in the paper [1]. Here are the specific kernels [5].

Note that all the result in the following sections will use this particular kernel.

$$K(x) = \|x\|^{-a} I_{\|x\| \leq 1} \text{ , where } a > 0 \tag{5}$$

The author also stated that the upper bound of the convergence rate is also held for the following kernels.

$$K(x) = \|x\|^{-a}(1 - \|x\|)_+^2 \text{ , where } a > 0 \tag{6}$$

and this kernel provided in this paper [3]:

$$K(x) = \|x\|^{-a} I_{\|x\| \leq 1} \cos^2\left(\pi \frac{\|x\|}{2}\right) \text{ , where } a > 0 \tag{7}$$

Note that all these kernels are called singular kernels, which means that when x goes to zero, $K(x)$ goes to infinity.

2.2 Assumption

In order to derive the upper bound of the error of Nadaraya-Watson estimator, we need to make some assumptions on P_{XY}, which is the distribution of $Z_i = (X_i, Y_i)$.

Here is the first assumption used in paper [1] to derive the main results.

Assumption 1

For any $x \in \mathbb{R}^d$, the ground truth regression function $E[Y|X = x] = f(x)$ exist and $E[\xi^2|X = x] \leq \sigma_\xi^2$, where $\xi = Y - E[Y|X]$ is the error term and the variance σ_ξ^2 is finite.

Here is the second assumption used in paper [1] to derive the main results.

Assumption 2

The marginal probability density function of X exist; we use $p(x)$ to donate the marginal probability density function of X. $p(x)$ is bounded on its support. In other words, $p(x)$ satisfy the following inequality:

$$0 < p_{min} \leq p(x) \leq p_{max} \tag{8}$$

2.3 Results

Now, I will present the results in the paper [1].

Theorem 1

We assume that the Assumption 1 and Assumption 2 are true. For any $f \in \mathcal{F}_{(\beta,L)}$, $\beta \in (0,1]$, $0 < a < \frac{d}{2}$, $h = n^{-\frac{1}{2\beta+d}}$ and $\forall x_0 \in \mathbb{R}^d$ fixed, we have the following inequality:

$$E[(f_n(x_0) - f(x_0))^2] \leq Cn^{-\frac{2\beta}{2\beta+d}} \text{ , where } C > 0 \text{ is a constant that is independent on } n \tag{9}$$

Theorem 2

We assume that the Assumption 1 and Assumption 2 are true. For any $f \in \mathcal{F}_{(\beta,L_1)}$, $\beta \in (1,2]$, $0 < a < \frac{d}{2}$, $h = n^{-\frac{1}{2\beta+d}}$, the probability density function $p \in \mathcal{F}_{(\beta-1,L_2)}$ and $\forall x_0 \in \mathbb{R}^d$ fixed, we have the following inequality:

$$E[(f_n(x_0) - f(x_0))^2] \leq Cn^{-\frac{2\beta}{2\beta+d}} \text{ , where } C > 0 \text{ is a constant that is independent on } n \tag{10}$$

2.4 Results for nonparametric regression

With Theorem 1 and Theorem 2, it is clear that:

$$E\|f_n(x) - f(x)\|^2_{L^2(P)} \leq Cn^{-\frac{2\beta}{2\beta+d}} \text{ , where } C > 0 \text{ is a constant that is independent on } n \tag{11}$$

Then we can conclude that the Learning Algorithm interpolating the data using the Nadaraya-Watson estimator with some particular singular kernels is optimal for nonparametric regression.

2.5 Results for prediction with square loss

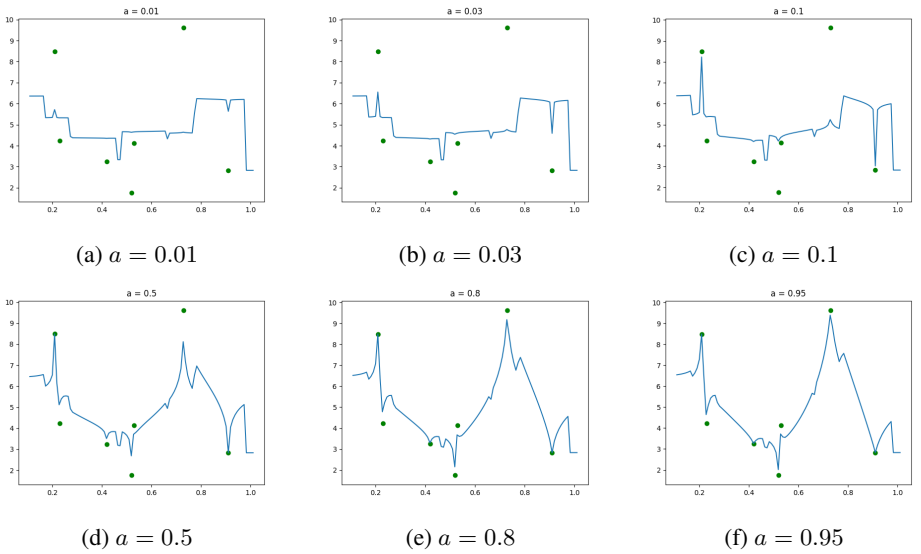
Based on the result for nonparametric regression [1], we get:

$$\begin{aligned}
&E[(f_n(X) - Y)]^2 - \inf_{g \in \mathcal{F}_{(\beta,L)}} E[(g(X) - Y)^2] \\
&= E[(f_n(X) - f(X))]^2 - \inf_{g \in \mathcal{F}_{(\beta,L)}} E[(g(X) - f(X))^2] \qquad \text{substitute } Y = f(X) \\
&= E[(f_n(X) - Y)]^2
\end{aligned}$$

Then we can conclude that the Learning Algorithm interpolating the data using the Nadaraya-Watson estimator with some particular singular kernels is optimal for prediction with square loss.

3 Examples

In this section, I will visualize the convergence rate of the learning algorithm mentioned in the above section. In order to see it clearly, I will set $d = 1$ and then $x, y \in \mathbb{R}$.



(a) $a = 0.01$ (b) $a = 0.03$ (c) $a = 0.1$

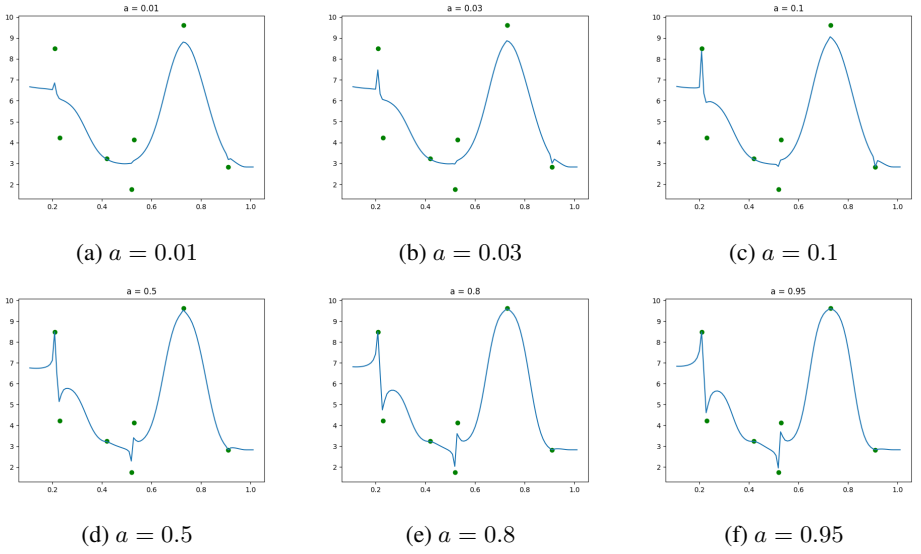
(d) $a = 0.5$ (e) $a = 0.8$ (f) $a = 0.95$

Figure 1: $\boldsymbol{K(x) = \|x\|^{-a} I_{\|x\| \leq 1}}$, with different $\boldsymbol{a}$ values

3.1 Comparison on a

I will show how a in the kernels affect the Learning algorithm. In the following graph, I will set the bandwidth $h = 0.25$. As shown in Figure [1], the learning algorithm fits the training data better when a increases.

As we can see in Figure [2], the learning algorithm fits the training data better when a increases. However, the difference between kernel $K(x) = \|x\|^{-a} I_{\|x\| \leq 1}$ and kernel $K(x) = \|x\|^{-a}(1-\|x\|)_+^2$ is that kernel $K(x) = \|x\|^{-a}(1 - \|x\|)_+^2$ is not sensitive to the value of a. It is more robust on a.



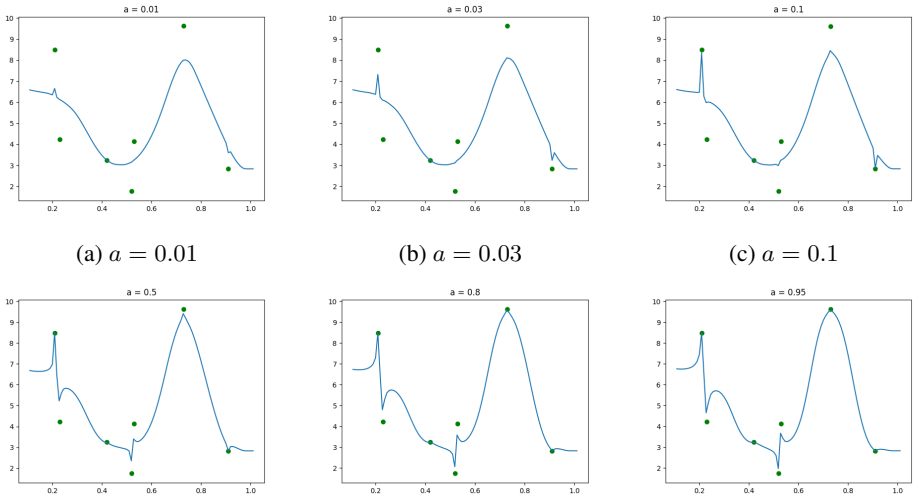
(a) $a = 0.01$ (b) $a = 0.03$ (c) $a = 0.1$

(d) $a = 0.5$ (e) $a = 0.8$ (f) $a = 0.95$

Figure 2: $K(x) = \|x\|^{-a}(1 - \|x\|)_+^2$, with different a values

In Figure [3], the learning algorithm fits the training data better when a increases. Note that the fitting curve of this kernel $K(x) = \|x\|^{-a} I_{\|x\| \leq 1} \cos^2\left(\pi \frac{\|x\|}{2}\right)$ is almost exactly the same as the fitting

curve of previous kernel $K(x) = \|x\|^{-a}(1 - \|x\|)_+^2$. The reason for this is that $1 - \|x\|$ is a good approximation of $\cos\left(\pi \frac{\|x\|}{2}\right)$ near the origin.



(a) $a = 0.01$ (b) $a = 0.03$ (c) $a = 0.1$

(d) $a = 0.5$ (e) $a = 0.8$ (f) $a = 0.95$

Figure 3: $\boldsymbol{K(x) = \|x\|^{-a} I_{\|x\|\leq 1} \cos^2\left(\pi \frac{\|x\|}{2}\right)}$, with different $\boldsymbol{a}$ values

3.2 Comparison on kernel

I will show how the different kernels affect Learning algorithm. In the following graph, I will set the bandwidth $h = 0.15$ and $a = 0.55$. As shown in Figure [4], the testing errors for all three kernels are pretty low. The fitting curve for kernel(b) and kernel(c) is smoother than that of kernel(a). As I mentioned in the previous part, the fitting curve of kernel(b) and kernel(c) are almost the same.

3.3 Noise in training data set

Professor Hajek suggested that I should explore how the noise in the training data would affect the result of the learning algorithm. I will visualize the results as following. I will add random Gaussian noise to all data points in the training data set and check how this will affect the learned predictor.

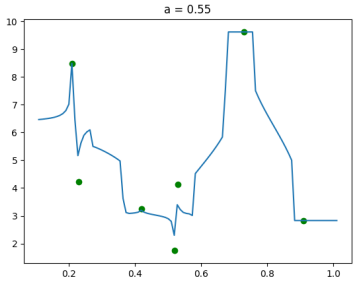
In this section, I use green dots to represent the correct training data set. I use red dots to represent the training data set with noise. I use the solid line to represent the predictor learned from the correct training data set. I use the dashed line to represent the predictor learned from the training data set with noise.

As shown in Figure [5], this kernel $K(x) = \|x\|^{-a} I_{\|x\|\leq 1}$ is very robust to noise. The robustness of this predictor is not dependent on the value of a.

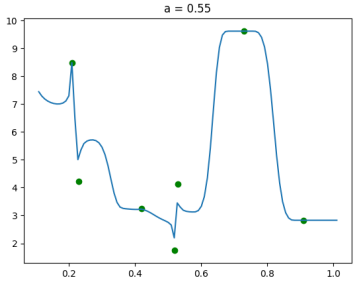
As shown in Figure [6], this kernel $K(x) = \|x\|^{-a}(1 - \|x\|)_+^2$ is very robust to noise. The robustness of this predictor is not dependent on the value of a.

In this example, I increase the variance of the random Gaussian noise that is added to each training data point.

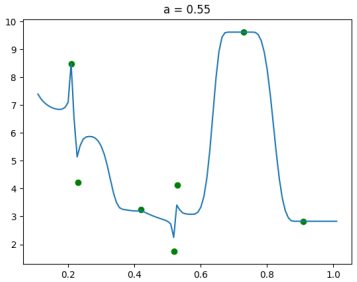
As shown in Figure [7], this kernel $K(x) = \|x\|^{-a} I_{\|x\|\leq 1} \cos^2\left(\pi \frac{\|x\|}{2}\right)$ is some kind robust to noise. The robustness of this predictor is dependent on the value of a. In this example, the magnitude of the noise is large. The predictor is robust to noise for small a values. The predictor is not robust to noise for large a values.



(a) $K(x) = \|x\|^{-a} I_{\|x\| \leq 1}$

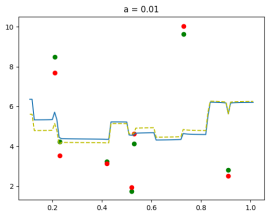


(b) $K(x) = \|x\|^{-a}(1 - \|x\|)_+^2$

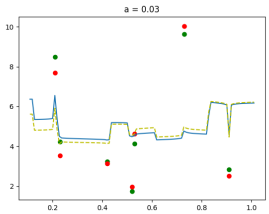


(c) $K(x) = \|x\|^{-a} I_{\|x\| \leq 1} \cos^2\left(\pi \frac{\|x\|}{2}\right)$

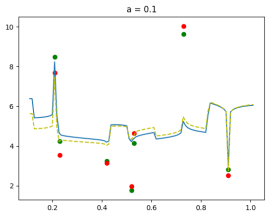
Figure 4: Three different kernels with $h = 0.15$ and $a = 0.55$



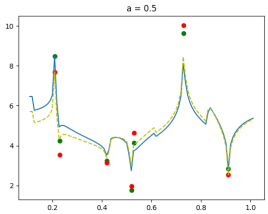
(a) $a = 0.01$



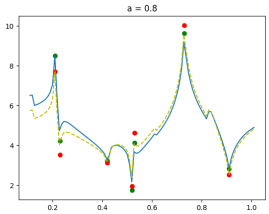
(b) $a = 0.03$



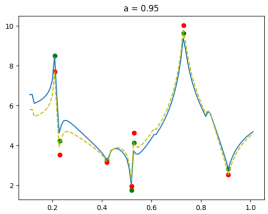
(c) $a = 0.1$



(d) $a = 0.5$



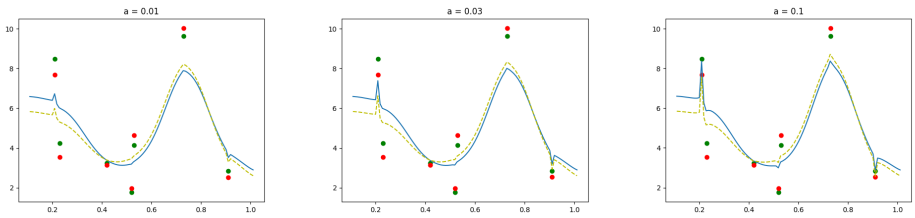
(e) $a = 0.8$



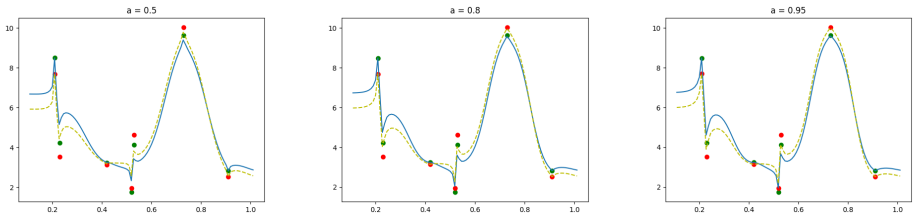
(f) $a = 0.95$

Figure 5: $\boldsymbol{K(x) = \|x\|^{-a} I_{\|x\| \leq 1}}$, with different $\boldsymbol{a}$ values

I will show how the different kernels affect the robustness of Learning algorithms. In the following graph, I will set the bandwidth $h = 0.20$ and $a = 0.32$. As shown in Figure [8], kernel(a) is least robust to noise and kernel(b) is most robust to noise.

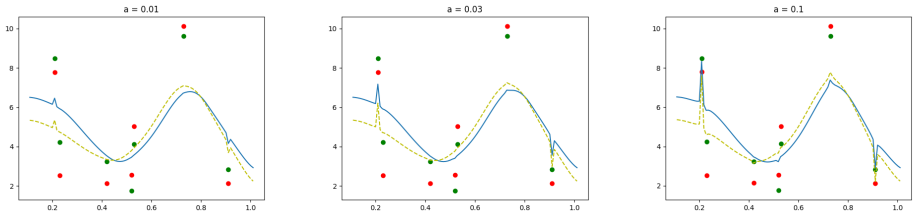


(a) $a = 0.01$ (b) $a = 0.03$ (c) $a = 0.1$

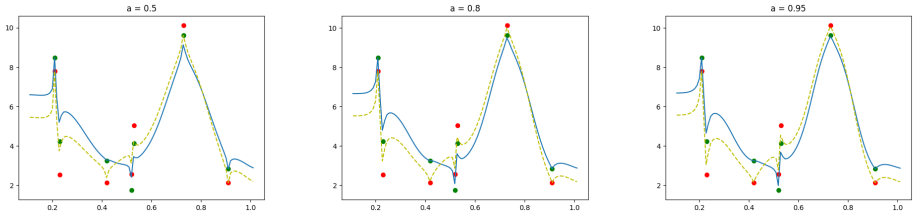


(d) $a = 0.5$ (e) $a = 0.8$ (f) $a = 0.95$

Figure 6: $K(x) = \|x\|^{-a}(1 - \|x\|)_+^2$, with different a values



(a) $a = 0.01$ (b) $a = 0.03$ (c) $a = 0.1$



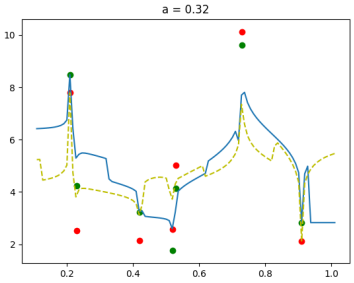
(d) $a = 0.5$ (e) $a = 0.8$ (f) $a = 0.95$

Figure 7: $\boldsymbol{K(x) = \|x\|^{-a} I_{\|x\|\leq 1} \cos^2\left(\pi \frac{\|x\|}{2}\right)}$, with different $\boldsymbol{a}$ values

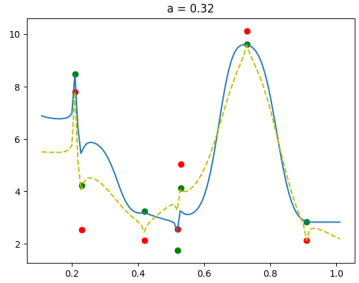
4 Proofs

In this section, I will briefly go over the proofs for Theorem 1 and Theorem 2 in the paper [1]. Even though the result is pretty straightforward. The proof of Theorem 1 and Theorem 2 is complicated. I will show the keystone of the proof.

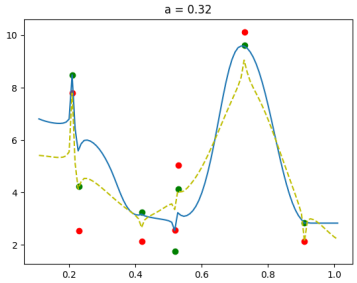
The proofs involve the techniques we learned in the ECE 543 class. We can first bound the Bias of the error term. Then we bound the variance.



(a) $K(x) = \|x\|^{-a} I_{\|x\| \leq 1}$



(b) $K(x) = \|x\|^{-a}(1 - \|x\|)_+^2$



(c) $K(x) = \|x\|^{-a} I_{\|x\| \leq 1} \cos^2\left(\pi \frac{\|x\|}{2}\right)$

Figure 8: Three different kernels with $h = 0.15$ and $a = 0.55$

To bound the Bias, we need to proof the following [1]:

$$\text{Let } b^2(0) = E[(E_Y \bar{f}_n(0) - f(0))^2 I_\xi]$$

$$\text{where } \bar{f}_n(0) = \frac{\sum_{i=1}^n Y_i K(\frac{-X_i}{h})}{\sum_{i=1}^n K(\frac{-X_i}{h})}, \xi = Y - E[Y|X] \text{ as mentioned in Assumption 1}$$

$$\text{When } \beta \in (0, 1] \text{ and } \forall f \in \mathcal{F}_{(\beta,L)}, \text{we can prove: } b^2(0) \leq L^2 h^{2\beta}$$

$$\text{When } \beta \in (1, 2], \forall f \in \mathcal{F}_{(\beta,L_1)}, \text{ the probability density function of } X\text{: } p(x) \in \mathcal{F}_{(\beta-1,L_2)}, \text{ and } p(x) \geq p_{min},$$
$$\text{we can prove: } b^2(0) \leq (L_1 + \|\nabla f(0)\| L_2 p_{min}^{-1}) h^{2\beta} + \sigma_X^2$$

I will not go over these two proof, because it is very long and you can read it in the original paper [1]. We bound the Bias by completing these proofs.

To bound the Variance, we need to proof the following [1]:

$$\text{Let } \sigma^2(0) = E[(E_Y \bar{f}_n(0) - \bar{f}_n(0))^2 I_\xi]$$

$$\text{where } \bar{f}_n(0) = \frac{\sum_{i=1}^n Y_i K(\frac{-X_i}{h})}{\sum_{i=1}^n K(\frac{-X_i}{h})}, \xi = Y - E[Y|X] \text{ as mentioned in Assumption 1}$$

$$\text{we can prove: } \sigma^2(0) \leq \frac{C\sigma_\xi^2}{nh^d}$$

I will not prove this result, since the proof is very long. The proof is in the original paper [1].

Once we bounded the bias and variance, we can get Theorem 1 and Theorem 2 immediately.

5 Discussion

In this paper, I summarized the results in the original paper [1]. I briefly go over the proofs of Theorem 1 and Theorem 2. I worked on a lot of examples and visualized these examples. There are some works we can do in the future. The upper bound proved in the original paper [1] only works on a very narrow class of learning algorithms. Specifically, the bound only works on the Nadaraya-Watson estimator with some particular singular kernels. We may use the prove techniques in the paper and extend the upper bound to a larger set of estimators.

Acknowledgments

I would like to thank Professor Bruce Hajek for the great materials presented in ECE 543 [2] class. Also, the lecture notes for the prerequisite course ECE 534 Random Process is also great.

References

[1] Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? 2018.

[2] Bruce Hajek and Maxim Raginsky. Ece 543: Statistical learning theory. 2019.

[3] Peter Lancaster and Kes Salkauskas. Surfaces generated by moving least squares methods. *Mathematics of computation*, 37(155):141–158, 1981.

[4] Elizbar A Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9(1):141–142, 1964.

[5] Donald Shepard. A two dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pages 517–524. ACM, 1968.