Does data interpolation contradict statistical optimality?

Mikhail Belkin
The Ohio State University

Alexander Rakhlin
MIT

Alexandre B. Tsybakov
CREST, ENSAE

Abstract

We show that classical learning methods interpolating the training data can achieve optimal rates for the problems of nonparametric regression and prediction with square loss.

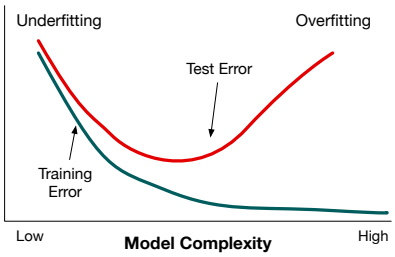
1 Introduction

In this paper, we exhibit estimators that interpolate the data, yet achieve optimal rates of convergence for the problems of nonparametric regression and prediction with square loss. This curious observation goes against the usual (or, folklore?) intuition that a good statistical procedure should forego the exact fit to data in favor of a more smooth representation. The family of estimators we consider do exhibit a bias-variance trade-off with a tuning parameter, yet this "regularization" co-exists in harmony with data interpolation.

Motivation for this work is the recent focus within the machine learning community on the out-of-sample performance of neural networks. These flexible models are typically trained to fit the data exactly (either in their sign or in the actual value), yet they predict well on unseen data. The conundrum has served both as a source of excitement about the "magical" properties of neural networks, as well as a call for the development of novel statistical techniques to resolve it.

So, should we be surprised to find a procedure that fits any amount of data yet generalizes well? An answer is immediate: No. We can take any procedure with good out-of-sample performance and modify it on the training points to fit the outcome variable. Such a modification on a zero-measure set (under appropriate assumptions) has no effect on the out-of-sample performance. One can argue, however, that this construction is not "natural." The aim of this paper is to show that a classical local estimation procedure satisfies the desiderata: for an appropriate choice of a kernel, the method interpolates the data, yet achieves optimal rates of convergence in the minimax sense. What is surprising, the optimal rate is achieved pointwise. Through this pedagogical example we emphasize that the degree to which a procedure fits the data can be completely decopled from the notion of overfitting.

Perhaps, some of the misconceptions regarding the generalization ability of learning methods that fit training data too well can be attributed to an (incorrect) interpretation of the familiar bias-variance cartoon we find in textbooks (see e.g. [6]):



In fact, low training error does not necessarily imply that the model is too complex and we are in the overfitting regime.

Let (X, Y) be a random pair on $\mathbb{R}^d \times \mathbb{R}$ with distribution P_{XY}, and let $f(x) = \mathbb{E}[Y|X = x]$ be the regression function. A goal of nonparametric estimation is to construct an estimate f_n of f, given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn independently from P_{XY}. A classical approach to this problem is kernel smoothing. In particular, the Nadaraya-Watson estimator [9, 13] is defined as

$$f_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}, \qquad (1)$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel function and $h > 0$ is a bandwidth and we assume that the denominator does not vanish. Appropriate choices of K and h lead to optimal rates of estimation, under various assumptions, and we refer the reader to [12] and references therein.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

We consider singular kernels that approach infinity when their argument tends to zero. It has been observed, at least since [11], that the resulting function in (1) interpolates the data. We will focus on the particular kernel

$$K(u) \triangleq \|u\|^{-a} \mathbf{I}\{\|u\| \leq 1\}, \tag{2}$$

for some $a > 0$. Here, $\|\cdot\|$ denotes the Euclidean norm. Our results can be extended to other related singular kernels, for example, to

$$K(u) \triangleq \|u\|^{-a} [1 - \|u\|]_+^2 \tag{3}$$

where $[c]_+ = \max\{c, 0\}$, and

$$K(u) \triangleq \|u\|^{-a} \cos^2(\pi \|u\| /2) \mathbf{I}\{\|u\| \leq 1\}, \tag{4}$$

considered in [8, 7]. Also, $\|\cdot\|$ can be any norm on $\mathbb{R}^d$, not necessarily the Euclidean norm.

Our main result, stated precisely in the next section and proved in Section 3, establishes that

$$\mathbb{E} \|f_n - f\|_{L_2(P_X)}^2 \triangleq \mathbb{E}(f_n(X) - f(X))^2 \leq C n^{-\frac{2\beta}{2\beta+d}}$$

whenever the regression function f belongs to a Hölder class with parameter $\beta \in (0, 1]$, and under additional assumptions stated below. Here C is a constant that does not depend on n and P_X is the marginal distribution of X. The rate $n^{-\frac{2\beta}{2\beta+d}}$ is the classical minimax optimal rate for Hölder classes [12].

Our result also yields a curious conclusion for the problem of prediction with square loss. Observe that excess loss—an object studied in Statistical Learning Theory—with respect to a Hölder class $\Sigma(\beta, L)$, formally defined below, can be written as

$$\begin{aligned}
&\mathbb{E}(f_n(X) - Y)^2 - \inf_{g \in \Sigma(\beta, L)} \mathbb{E}(g(X) - Y)^2 \\
&= \mathbb{E}(f_n(X) - f(X))^2 - \inf_{g \in \Sigma(\beta, L)} \mathbb{E}(g(X) - f(X))^2 \\
&= \mathbb{E}(f_n(X) - f(X))^2
\end{aligned}$$

under the assumption that the model is *well-specified* (that is, the regression function is in the class). We remark that the estimator f_n is *improper*, in the sense that it does not itself belong to the Hölder class (its smoothness depends on h and, hence, on n). In conclusion, despite the fact that f_n is improper and fits the data exactly, it attains optimal rates for excess loss. We refer the reader to [10] for further discussion of optimal rates in nonparametric estimation and statistical learning.

Prior work Within the context of pattern classification, the 1-Nearest-Neighbor classifier is an example of an interpolating rule. It is shown in [3] that the limit (as n tends to infinity) of the classification risk is no more than twice the Bayes risk. To make k-Nearest-Neighbor rules consistent, one is required to increase k with n [4, 2], in which case the rule is no longer interpolating.

The idea of interpolating the data using singular kernels appears already in [11] and was further developed in [8, 7], among others. These works were focusing on deterministic properties of the interpolants and no statistical guarantees have been established until [5] have shown consistency of the estimator (1) for the singular kernel $K(u) = \|u\|^{-d}$, however, without finite sample guarantees. The recent work of [1] proves the first (to the best of our knowledge) non-asymptotic rates for interpolating procedures, yet the guarantees are suboptimal. The present paper shows that statistical optimality of interpolating rules can indeed be achieved and it holds under rather standard nonparametric assumptions on the regression function.

2 Main Results

We start with a definition.

Definition 1. For $L > 0$ and $\beta \in (0, 2]$, the (β, L)-Hölder class, denoted by $\Sigma(\beta, L)$, is defined as follows:

- If $\beta \in (0, 1]$, the class $\Sigma(\beta, L)$ consists of functions $f : \mathbb{R}^d \to \mathbb{R}$ satisfying
$$\forall x, y \in \mathbb{R}^d, \quad |f(x) - f(y)| \leq L \|x - y\|^\beta . \tag{5}$$
- If $\beta \in (1, 2]$, the class $\Sigma(\beta, L)$ consists of continuously differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ satisfying for all $x, y \in \mathbb{R}^d$
$$|f(x) - f(y) - \langle \nabla f(y), x - y \rangle| \leq L \|x - y\|^\beta \tag{6}$$
where $\langle \cdot, \cdot \rangle$ denotes the inner product.

We assume the following.

(A1) For any $x \in \mathbb{R}^d$, the expectation $\mathbb{E}[Y|X = x] = f(x)$ exists and $\mathbb{E}[\xi^2|X = x] \leq \sigma_\xi^2 < \infty$, where $\xi = Y - \mathbb{E}[Y|X] = Y - f(X)$.

(A2) The marginal density $p(\cdot)$ of X exists and satisfies $0 < p_{\min} \leq p(x) \leq p_{\max}$ for all x on its support.

The Nadaraya-Watson estimator for a singular kernel K is defined as

$$f_n(x) = \begin{cases} Y_i & \text{if } x = X_i \text{ for some } i \in [n] \\ 0 & \text{if } \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) = 0, \\ \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)} & \text{otherwise.} \end{cases} \tag{7}$$

Note that setting here $f_n(X_i) = Y_i$ is not an artificial perturbation. Indeed, for kernel (2) or other kernel with singularity at 0, it is just a proper way of completing the definition of $f_n(\cdot)$ by continuity.

The two main results for this estimator are now stated.

Theorem 1. *Assume that $f \in \Sigma(\beta, L_f)$ for $\beta \in (0, 1]$, $L_f > 0$. Let Assumptions (A1) and (A2) be satisfied, and $0 < a < d/2$. Then for any fixed $x_0 \in \mathbb{R}^d$ in the support of p the estimator (7) with kernel (2) and bandwidth $h = n^{-\frac{1}{2\beta+d}}$ satisfies*

$$\mathbb{E}[(f_n(x_0) - f(x_0))^2] \le C n^{-\frac{2\beta}{2\beta+d}}$$

where $C > 0$ is a constant that does not depend on n.

Theorem 2. *Assume that $f \in \Sigma(\beta, L_f)$ for $\beta \in (1, 2]$, $L_f > 0$. Let Assumptions (A1) and (A2) be satisfied, and $0 < a < d/2$. Assume in addition that, for all x, y in the support of p, we have $|p(x) - p(y)| \le L_p \|x - y\|^{\beta-1}$, $L_p > 0$. Then for any fixed $x_0 \in \mathbb{R}^d$ such that the Euclidean ball of radius h centered at x_0 is contained in the support of p, the estimator (7) with kernel (2) and bandwidth $h = n^{-\frac{1}{2\beta+d}}$ satisfies*

$$\mathbb{E}[(f_n(x_0) - f(x_0))^2] \le C n^{-\frac{2\beta}{2\beta+d}}$$

where $C > 0$ is a constant that does not depend on n.

In particular, the pointwise mean squared error (MSE) bound of Theorem 1 immediately implies that the integrated MSE with respect to the marginal distribution of X satisfies

$$\mathbb{E} \int_{\mathbb{R}^d} (f_n(x) - f(x))^2 p(x) dx \le C n^{-\frac{2\beta}{2\beta+d}},$$

assuming that f is bounded on the support of the marginal density p.

3 Proofs

Without loss of generality, consider the problem of estimating $f(x_0)$ at $x_0 = 0$, assuming it is in the support of p and $|f(x_0)| < \infty$.

Consider the event

$$\mathcal{E} = \left\{ \sum_{i=1}^n K_h(X_i) \neq 0 \right\} = \{\exists i = 1, \dots, n : \|X_i\| \le h\}$$

and observe that

$$P\left(\bar{\mathcal{E}}\right) \le \left(1 - C p_{\min} h^d\right)^n \le \exp\left\{-C p_{\min} n h^d\right\}$$

for a constant $C > 0$ that does not depend on n. On the event $\bar{\mathcal{E}}$, we have $f_n(0) = 0$ and thus the contribution to expected risk is at most $M_{\mathcal{E}} = f(0)^2 \exp\left\{-C p_{\min} n h^d\right\}$, a lower-order term compared to the remaining calculations.

On the event $\mathcal{E}$, the estimator $f_n(0)$ is equal to

$$\bar{f}_n(0) = \frac{\sum_{i=1}^n Y_i K_h(X_i)}{\sum_{i=1}^n K_h(X_i)}$$

(modulo an event of zero probability with respect to the joint distribution of $X_1, \dots, X_n$), where

$$K_h(x) \triangleq K(x/h).$$

Set $\xi_i = Y_i - f(X_i)$. Let $\mathbb{E}_Y$ denote the expectation with respect to $Y_1, \dots, Y_n$, conditional on $X_1, \dots, X_n$. We have the following "bias-variance" decomposition

$$\begin{aligned}
&\mathbb{E}[(f_n(0) - f(0))^2] \\
&\le \mathbb{E}[(\bar{f}_n(0) - \mathbb{E}_Y \bar{f}_n(0) + \mathbb{E}_Y \bar{f}_n(0) - f(0))^2 \mathbf{I}\{\mathcal{E}\}] + M_{\mathcal{E}} \\
&= \mathbb{E}[(\bar{f}_n(0) - \mathbb{E}_Y \bar{f}_n(0))^2 \mathbf{I}\{\mathcal{E}\}] \\
&+ \mathbb{E}[(\mathbb{E}_Y \bar{f}_n(0) - f(0))^2 \mathbf{I}\{\mathcal{E}\}] + M_{\mathcal{E}} \ .
\end{aligned}$$

It holds that, on the event $\mathcal{E}$,

$$\mathbb{E}_Y \bar{f}_n(0) = \frac{\sum_{i=1}^n f(X_i) K_h(X_i)}{\sum_{i=1}^n K_h(X_i)}$$

and, hence, the variance term is

$$\begin{aligned}
\sigma^2(0) &\triangleq \mathbb{E}[(\bar{f}_n(0) - \mathbb{E}_Y \bar{f}_n(0))^2 \mathbf{I}\{\mathcal{E}\}] \qquad (8) \\
&= \mathbb{E}\left[\left(\frac{\sum_{i=1}^n \xi_i K_h(X_i)}{\sum_{i=1}^n K_h(X_i)} \right)^2 \mathbf{I}\{\mathcal{E}\} \right] \le \sigma_\xi^2 \sigma_X^2,
\end{aligned}$$

where

$$\sigma_X^2 \triangleq n \mathbb{E}\left[\frac{K_h^2(X_1)}{\left(\sum_{i=1}^n K_h(X_i)\right)^2} \mathbf{I}\{\mathcal{E}\} \right].$$

On the other hand, the bias[1] is

$$\begin{aligned}
b^2(0) &\triangleq \mathbb{E}[(\mathbb{E}_Y \bar{f}_n(0) - f(0))^2 \mathbf{I}\{\mathcal{E}\}] \qquad (9) \\
&= \mathbb{E}\left[\left(\frac{\sum_{i=1}^n (f(X_i) - f(0)) K_h(X_i)}{\sum_{i=1}^n K_h(X_i)} \right)^2 \mathbf{I}\{\mathcal{E}\} \right].
\end{aligned}$$

The following lemmas control each of the above expressions under various assumptions on f and the marginal density p. We will denote by C positive constants that can vary from line to line.

3.1 Bounding the Variance

Lemma 1. Let Assumptions (A1) and (A2) hold. Then,

$$\sigma^2(0) \le \frac{C \sigma_\xi^2}{n h^d}. \qquad (10)$$

[1] To be precise, this term includes variance due to random X, as will be clear from Lemma 3.

Proof. Introduce the random variables

$$\eta_i = \mathbf{I}\{\|X_i\| \le h\}.$$

They are i.i.d. and follow the Bernoulli distribution with parameter

$$\bar{p} \triangleq P(\|X_1\| \le h) \ge c_0 p_{\min} h^d$$

where $c_0 > 0$ depends only on d. Then

$$\sigma_X^2 \le n\mathbb{E}\left[\frac{K_h^2(X_1)}{\left(\sum_{i=1}^n K_h(X_i)\right)^2}\mathbf{I}\left\{\sum_{i=1}^n \eta_i \le \frac{n\bar{p}}{2}\right\}\mathbf{I}\{\mathcal{E}\}\right] + n\mathbb{E}\left[\frac{4}{(n\bar{p})^2}K_h^2(X_1)\right] \tag{11}$$

where we have used the fact that

$$K_h(X_i) \ge \eta_i, \quad i = 1, \dots, n.$$

Change of variables yields

$$n\mathbb{E}[K_h^2(X_1)] \le nh^d p_{\max}\int_{\mathbb{R}^d} K^2(u)du. \tag{12}$$

Since the kernel K is radially symmetric and supported on the unit Euclidean ball, the last expression is bounded from above by

$$Cnh^d p_{\max}\int_0^1 r^{-2a}r^{d-1}dr \le C_2 nh^d$$

whenever $d-2a-1 > -1$ (equivalently, $a < d/2$). Here C, C_2 are positive constants depending only on d. It follows that

$$n\mathbb{E}\left[\frac{4}{(n\bar{p})^2}K_h^2(X_1)\right] \le \frac{4}{(c_0 p_{\min} nh^d)^2}C_2 nh^d \le \frac{C}{nh^d}.$$

To conclude the proof, we analyze the first term in (11):

$$\begin{aligned} &n\mathbb{E}\left[\frac{K_h^2(X_1)}{\left(\sum_{i=1}^n K_h(X_i)\right)^2}\mathbf{I}\left\{\sum_{i=1}^n \eta_i \le \frac{n\bar{p}}{2}\right\}\mathbf{I}\{\mathcal{E}\}\right] \\ &\le nP\left(\sum_{i=1}^n \eta_i \le \frac{n\bar{p}}{2}\right) \\ &= nP\left(\sum_{i=1}^n \eta_i - n\bar{p} \le \frac{n\bar{p}}{2}\right). \end{aligned}$$

By Bernstein's inequality, the last expression is at most

$$\begin{aligned} &n\exp\left\{-\frac{(n\bar{p}/2)^2}{2(n\bar{p}(1-\bar{p}) + n\bar{p}/3)}\right\} \\ &\le n\exp\left\{-\frac{3n\bar{p}}{32}\right\} \le n\exp\left\{-Cnh^d\right\}. \end{aligned}$$

□

3.2 Bounding the Bias

Lemma 2. Let $\beta \in (0, 1]$, $L_f > 0$, and assume that $f \in \Sigma(\beta, L_f)$. Then

$$b^2(0) \le L_f^2 h^{2\beta}.$$

Proof. Since $f \in \Sigma(\beta, L_f)$ we have, on the event $\mathcal{E}$,

$$\begin{aligned} &\left|\frac{\sum_{i=1}^n (f(X_i) - f(0))K_h(X_i)}{\sum_{i=1}^n K_h(X_i)}\right| \\ &\le \left|\frac{\sum_{i=1}^n L_f \|X_i\|^\beta K_h(X_i)}{\sum_{i=1}^n K_h(X_i)}\right| \\ &\le L_f h^\beta. \end{aligned}$$

The last step holds because the kernel K_h is zero outside of the Euclidean ball of radius h. □

Lemma 2 can be extended to smoothness $\beta \in (1, 2]$ under an additional assumption on the marginal density.

Lemma 3. Let $\beta \in (1, 2]$, $L_f > 0$, and $f \in \Sigma(\beta, L_f)$. Assume that the density p of the marginal distribution of X satisfies $p \in \Sigma(\beta - 1, L_p)$, and $p(x) \ge p_{\min} > 0$ for all x in the support of p. Then

$$b^2(0) \le (L_f + \|\nabla f(0)\| L_p p_{\min}^{-1})h^{2\beta} + \sigma_X^2.$$

Proof. We write (9) as $b^2(0) = \mathbb{E}\left[\sum_{i,j=1}^n G_i G_j \mathbf{I}\{\mathcal{E}\}\right]$ where

$$G_i = \frac{(f(X_i) - f(0))K_h(X_i)}{\sum_{i=1}^n K_h(X_i)}.$$

For $i \ne j$ we can write

$$\begin{aligned} &\mathbb{E}[G_i G_j \mathbf{I}\{\mathcal{E}\}] \\ &= \mathbb{E}\left[(f(X_i) - f(0))(f(X_j) - f(0))A(X_i, X_j)\right] \end{aligned}$$

where

$$A(X_i, X_j) = \frac{K_h(X_i)K_h(X_j)}{\left(\sum_{i=1}^n K_h(X_i)\right)^2}\mathbf{I}\{\mathcal{E}\} \ge 0.$$

We omit for brevity the dependence of $A(X_i, X_j)$ on $(X_k, k \ne i, k \ne j)$. Thus,

$$\begin{aligned} &\mathbb{E}'[G_i G_j \mathbf{I}\{\mathcal{E}\}] \\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}(f(x_i) - f(0))(f(x_j) - f(0)) \\ &\qquad \times A(x_i, x_j)p(x_i)p(x_j)dx_i dx_j \end{aligned}$$

where $\mathbb{E}'$ denotes the conditional expectation over (X_i, X_j) for fixed $(X_k, k \ne i, k \ne j)$. Let us define

$$R(x_i) = f(x_i) - f(0) - \langle \nabla f(0), x_i \rangle$$

and $R(x_j) = f(x_j) - f(0) - \langle \nabla f(0), x_j \rangle$.

Then

$$\begin{aligned}
&\mathbb{E}'[G_i G_j \mathbf{I}\{\mathcal{E}\}] \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \langle \nabla f(0), x_i\rangle \langle \nabla f(0), x_j\rangle \\
&\qquad\qquad \times A(x_i,x_j)p(x_i)p(x_j)dx_i dx_j \\
&+ 2\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \langle \nabla f(0), x_i\rangle R(x_j) \\
&\qquad\qquad \times A(x_i,x_j)p(x_i)p(x_j)dx_i dx_j \\
&+ \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} R(x_i)R(x_j)A(x_i,x_j)p(x_i)p(x_j)dx_i dx_j
\end{aligned}$$

where the factor 2 arises from symmetry considerations. Now observe that

$$\int_{\mathbb{R}^d} \langle \nabla f(0), x_i\rangle A(x_i,x_j)p(0)dx_i = 0$$

for any x_j since the function under the integral is odd for any fixed $(X_k, k \neq i, k \neq j)$. Applying this observation for both x_i and x_j in the first term of the above decomposition, as well as for the second term, we obtain

$$\begin{aligned}
&\mathbb{E}'[G_i G_j \mathbf{I}\{\mathcal{E}\}] \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \langle \nabla f(0), x_i\rangle \langle \nabla f(0), x_j\rangle \\
&\qquad\qquad \times A(x_i,x_j)(p(x_i)-p(0))(p(x_j)-p(0))dx_i dx_j \\
&+ 2\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \langle \nabla f(0), x_i\rangle R(x_j) \\
&\qquad\qquad \times A(x_i,x_j)(p(x_i)-p(0))p(x_j)dx_i dx_j \\
&+ \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} R(x_i)R(x_j)A(x_i,x_j)p(x_i)p(x_j)dx_i dx_j.
\end{aligned}$$

Condition (6) implies that $|R(x_i)| \le L_f \|x_i\|^\beta$. Next, recall that A is zero whenever either $\|x_i\| > h$ or $\|x_j\| > h$. Using Cauchy-Schwarz inequality for the inner products and the Hölder assumption on p, we conclude that

$$\begin{aligned}
&\mathbb{E}'[G_i G_j \mathbf{I}\{\mathcal{E}\}] \\
&\le B^2 L_p^2 h^{2\beta}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} A(x_i,x_j)dx_i dx_j \\
&+ 2BL_f L_p h^{2\beta}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} A(x_i,x_j)p(x_j)dx_i dx_j \\
&+ L_f^2 h^{2\beta}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} A(x_i,x_j)p(x_i)p(x_j)dx_i dx_j
\end{aligned}$$

where $B = \|\nabla f(0)\|^2$. Using the lower bound $p_{\min}$ on the density, completing the square and taking the expectation with respect to $(X_k, k \neq i, k \neq j)$, we establish that $\mathbb{E}[G_i G_j \mathbf{I}\{\mathcal{E}\}]$ is bounded above by

$$h^{2\beta}(BL_p p_{\min}^{-1} + L_f)^2 \mathbb{E}\left[\frac{K_h(X_i)K_h(X_j)}{\left(\sum_{i=1}^n K_h(X_i)\right)^2}\mathbf{I}\{\mathcal{E}\}\right].$$

On the other hand, the sum of diagonal elements is

$$\sum_{i=1}^n \mathbb{E}[G_i^2 \mathbf{I}\{\mathcal{E}\}] = n\mathbb{E}\left[\frac{K^2(X_1)}{\left(\sum_{i=1}^n K_h(X_i)\right)^2}\mathbf{I}\{\mathcal{E}\}\right],$$

which is precisely the variance term σ_X^2. Finally,

$$\begin{aligned}
&\sum_{i\neq j} \mathbb{E}[G_i G_j \mathbf{I}\{\mathcal{E}\}] \\
&= h^{2\beta}(BL_p p_{\min}^{-1} + L_f)^2 \mathbb{E}\left[\frac{\sum_{i\neq j} K_h(X_i)K_h(X_j)}{\left(\sum_{i=1}^n K_h(X_i)\right)^2}\mathbf{I}\{\mathcal{E}\}\right] \\
&\le h^{2\beta}(BL_p p_{\min}^{-1} + L_f)^2 \mathbb{E}\left[\frac{\sum_{i,j=1}^n K_h(X_i)K_h(X_j)}{\left(\sum_{i=1}^n K_h(X_i)\right)^2}\mathbf{I}\{\mathcal{E}\}\right] \\
&\le h^{2\beta}(BL_p p_{\min}^{-1} + L_f)^2.
\end{aligned}$$

□

3.3 Proofs of Theorem 1 and 2

The two theorems follow immediately from Lemmas 1, 2, and 3 by balancing $n\exp\{-Cnh^d\} + \frac{C}{nh^d} + Ch^{2\beta}$ with $h = n^{-\frac{1}{2\beta+d}}$.

4 Discussion

We presented a proof of concept: an interpolating rule can achieve optimal rates for the problems of nonparametric estimation and prediction with square loss. Our proof technique extends to other kernels where the indicator over the unit Euclidean ball in (2) is replaced with a function that dominates an appropriately scaled indicator. The analysis also works for non-singular kernels under the assumption of square integrability (required only in Eq. (12)).

We observe that by thresholding the singular kernel at a value κ, one can control the degree of fitting the data in a manner that is decoupled from the bias-variance trade-off achieved through h.

We also remark that local polymnomial estimators [12] with an interpolating kernel as in (2) can be shown to achieve optimal rate $n^{-2\beta/(2\beta+1)}$ for all $\beta > 0$ and $d = 1$. The proof will be included in a full version of this paper.

While each pair (X_i, Y_i) is fit exactly by the proposed estimator, the influence of the datapoint is local. In aggregate, however, the function f_n is being "pulled" towards the true regression function f. Whether a similar phenomenon occurs in other interpolating rules—such as overparametrized neural networks—requires further investigation.

5 Visualization

The figures below show interpolations with kernels (2) and (3). While both achieve optimal rates of convergence in this simple one-dimensional problem, the latter kernel appears to be less irregular. Indeed, unlike (2), kernels (3) and (4) produce continuous functions.

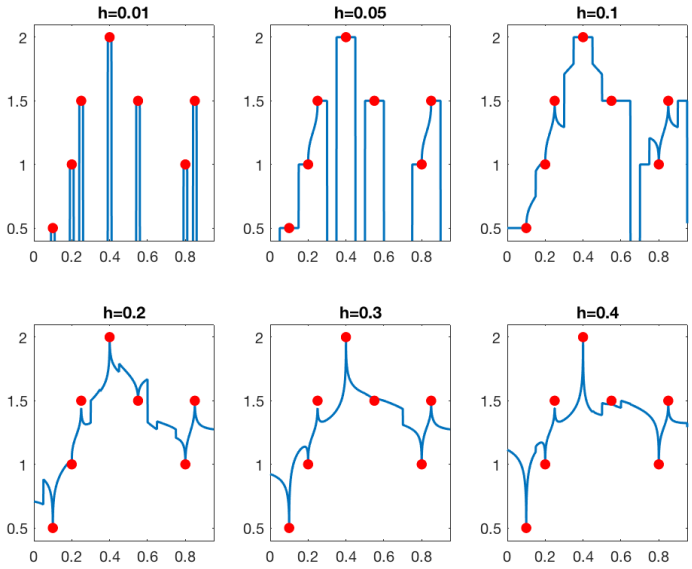


Figure 1: Interpolation with $K(u) = \|u\|^{-a} \mathbf{I}\{\|u\| \leq 1\}$, $a = 0.49$, and various values of h.

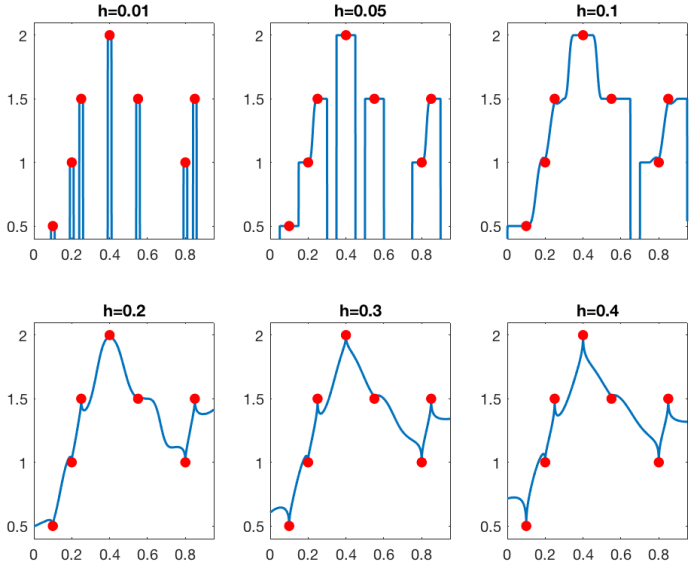


Figure 2: Interpolation with $K(u) = \|u\|^{-a} [1 - \|u\|]_+^2$, $a = 0.49$, and various values of h.

We now compare Figures 1 and 2 to those with a non-singular kernel. We remark that choices of bandwidth h differ depending on the kernel, and direct comparisons for the same value across kernels might not be meaningful.

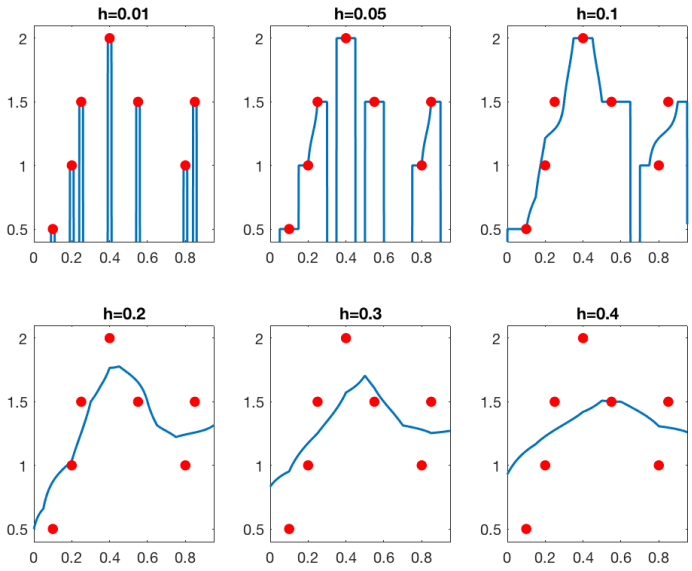


Figure 3: Comparison: non-singular Epanechnikov kernel $K(u) = (3/4)(1 - \|u\|^2)\mathbf{I}\{\|u\| \leq 1\}$.

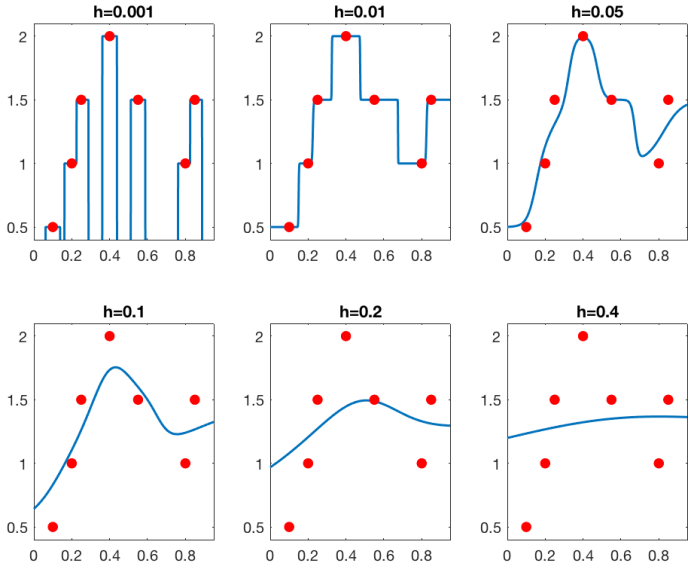


Figure 4: Comparison: non-singular Gaussian kernel $K(u) = (1/\sqrt{2\pi})\exp\left\{-\|u\|^2\right\}$. Note the altered choices of h.

Figure 5 below shows a comparison between the interpolating kernel 3 and the Gaussian kernel for binary-valued data. We observe the more global effect that each point has on the behavior of the solution with the Gaussian kernel, in comparison to the singular kernel. Understanding properties of the plug-in classifier $\text{sign}(f_n)$ under various margin conditions appears to be an interesting direction of further research.

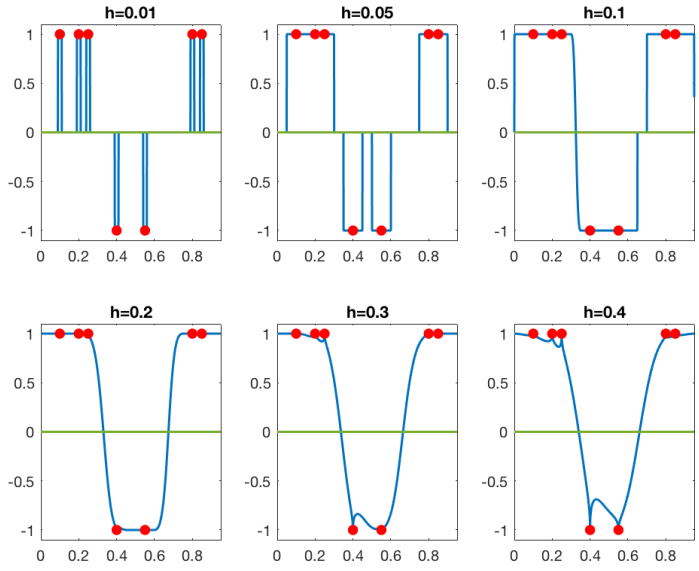


Figure 5: Interpolation with $K(u) = \|u\|^{-a}[1-\|u\|]_+^2$, $a = 0.49$, for binary-valued Y.

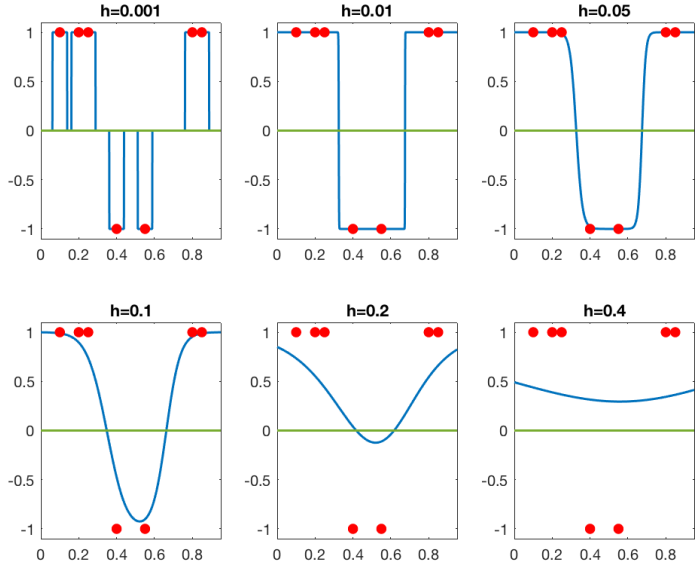


Figure 6: Comparison: non-singular Gaussian kernel $K(u) = (1/\sqrt{2\pi})\exp\left\{-\|u\|^2\right\}$ for binary-valued Y. Note the altered choices of h.

References

[1] Mikhail Belkin, Daniel Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *arXiv preprint arXiv:1806.05161*, 2018. 2

[2] Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 3437–3445, 2014. 2

[3] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967. 2

[4] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Springer, 1996. 2

[5] Luc Devroye, Laszlo Györfi, and Adam Krzyżak. The Hilbert kernel regression estimate. *Journal of Multivariate Analysis*, 65(2):209–227, 1998. 2

[6] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2016. 1

[7] V Katkovnik. *Nonparametric identification and smoothing of data (Local approximation methods)*. Nauka, Moscow, 1985. 2

[8] Peter Lancaster and Kes Salkauskas. Surfaces generated by moving least squares methods. *Mathematics of computation*, 37(155):141–158, 1981. 2

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

[10] Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2): 789–824, 2017. 2

[11] 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. 2

[12] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. 1, 2, 5

[13] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964. 1