03 Local Polynomials

We saw previously that, under some conditions (including assuming the regression function is Lipschitz), simple kNN and kernel estimators can achieve rates of the order $n^{-1/(2+d)}$ on the unsquared scale, and that this rate is optimal for Lipschitz regression functions. Here we explore one way to add more structure beyond Lipschitz continuity, how this can allow for better error guarantees, and some estimators based on local polynomials that achieve optimal rates.

1 Holder smoothness

We have discussed regression under the assumption that the conditional expectation function is Lipschitz, i.e., that

$$|\mu(x) - \mu(x')| \leq L\|x - x'\|$$

for all x, x' in the support of X, for some constant L. This ensures that the regression function at nearby points cannot be too different; note that a differentiable function is Lipschitz iff its derivative is bounded, while a Lipschitz function is differentiable almost everywhere.

However the class of Lipschitz functions is quite large, and one can imagine different functions with roughly the same first-order derivative behavior but quite different smoothness or higher-order derivative behavior. For example both functions $\mu(x) = \sin(Lx)$ and $\mu(x) = Lx$ have first derivatives bounded by L, whereas higher-order derivatives scale with L^k for the former and are simply zero for the latter. One might expect that ruling out more wiggly functions could help allow for better estimation error.

A classic way to characterize more structure beyond Lipschitz continuity is through a Holder condition that imposes bounded higher-order derivatives. Consider the one-dimensional X case first. There, the Holder class $\mathcal{H}(s, L)$ (for s a positive integer, $L > 0$) denotes the set of functions with all derivatives up to order $s - 1$ bounded, and the $s - 1$ order derivatives Lipschitz, i.e., all f such that

$$\left|\partial_x^\alpha f(x)\right| \leq L \text{ for all } \alpha \leq s-1, \text{ and } \left|\partial_x^{s-1} f(x) - \partial_x^{s-1} f(x')\right| \leq L\|x - x'\|.$$

Note that the $(s-1)$-order derivative is Lipschitz continuous and so is itself differentiable almost everywhere, with any such derivative bounded by L; therefore, s-order bounded derivatives exist almost everywhere for s-smooth functions.

We have made two simplifications in the above definition: that $f : \mathbb{R} \to \mathbb{R}$ (i.e., dimension $d = 1$) and that $s \in \mathbb{N}$ is an integer. Both can be relaxed as follows, with a bit more notation. Let $\lfloor s \rfloor$ denote the largest integer strictly smaller than s, and for $k \in \mathbb{N}^d$ let $D^k = \partial_{x_1}^{k_1} ... \partial_{x_d}^{k_d}$ denote the partial derivative operator. Then, more generally, we say $f \in \mathcal{H}(s, L)$ for $s, L > 0$ if

$$\left|D^\alpha f(x)\right| \leq L \text{ , and } \left|D^\beta f(x) - D^\beta f(x')\right| \leq L\|x - x'\|^{s-\lfloor s \rfloor}$$

for all $\sum_j \alpha_j \leq \lfloor s \rfloor$ and all $\sum_j \beta_j = \lfloor s \rfloor$, and for all x, x'. Note when s is an integer then $s - \lfloor s \rfloor = 1$. Sometimes we will omit the constant L and simply say $f \in \mathcal{H}(s, L)$ is s-smooth.

2 Local polynomial construction

Kernel estimators estimate $\mu(x)$ with a weighted average of nearby points, with weights defined by a kernel function $K : \mathbb{R}^d \to \mathbb{R}$ and what is nearby quantified by a bandwidth parameter h. We saw that these estimators can be optimal for Lipschitz functions, but the analysis does not make clear how faster rates could be attainable for s-smooth functions, and in fact Problem 5.1 of Györfi et al. [2002] (page 82) shows that boxcar kernel estimators cannot be optimal for $\mathcal{H}(2, L)$ (i.e., $s \geq 2$) at least in terms of L_2 norm error.

Thus kernel estimators cannot track higher-order derivatives, essentially because they estimate the regression function with a constant in each window. This issue can be ameliorated by instead fitting a curve in each window, e.g., a polynomial whose degree depends on the desired smoothness s. This approach is called local polynomial estimation.

Note that the standard kernel estimator $\widehat{\mu}(x) = \sum_i K\left(\frac{X_i - x}{h}\right) Y_i / \sum_j K\left(\frac{X_j - x}{h}\right)$ can also be written as

$$\widehat{\mu}(x) = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \Big(Y_i - \beta\Big)^2$$

i.e., as a locally weighted least squares regression of Y on a constant. This follows by taking a derivative with respect to β and setting to zero, yielding

$$0 = -\frac{2}{n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \Big(Y_i - \beta\Big);$$

solving for β gives the original estimator. Thus to fit a curve in each window one can use the local polynomial estimator $\widehat{\mu}(x) = \widehat{\beta}^T b(x)$, where

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \Big\{Y_i - \beta^{\mathrm{T}} b(X_i)\Big\}^2$$

and $b : \mathbb{R}^d \to \mathbb{R}^p$ is some basis vector, e.g., $b(x) = (1, x, x^2, ..., x^{p-1})$ when $d = 1$. This estimator can also be written in closed-form as

$$\widehat{\mu}(x) = b(x)^{\mathrm{T}} \mathbb{P}_n \left\{ b(X) \frac{1}{h^d} K\left(\frac{X - x}{h}\right) b(X)^{\mathrm{T}} \right\}^{-1} \mathbb{P}_n \left\{ b(X) \frac{1}{h^d} K\left(\frac{X - x}{h}\right) Y \right\}$$

where we use the notation $\mathbb{P}_n\{f(Z)\} = \sum_{i=1}^n f(Z_i)$ for sample averages. The extra scaling by $1/h^d$ will be useful in a later analysis. Right now we are being intentionally vague about the choice of basis b, but this will be clarified shortly (it will involve tensor products when $d > 1$).

Notice that this estimator takes the form $\sum_i w_i(x; X^n) Y_i$ with weights

$$w_i(x; X^n) = b(x)^{\mathrm{T}} \mathbb{P}_n \left\{ b(X) \frac{1}{h^d} K\left(\frac{X - x}{h}\right) b(X)^{\mathrm{T}} \right\}^{-1} b(X_i) \frac{1}{nh^d} K\left(\frac{X - x}{h}\right)$$

and so it is also a linear smoother, just like kNN and the simple kernel estimator studied before.

3 Basis & weights

Here we begin to analyze the local polynomial estimator, with the aim of showing it can achieve faster and indeed minimax optimal rates of convergence for s-smooth functions.

To start, we need to be a bit more careful about the choice of basis b. When $d = 1$ we can use the simple choice $b(x) = (1, x, x^2, ..., x^{\lfloor s \rfloor})^{\mathrm{T}}$ of polynomials up to degree $\lfloor s \rfloor$, or to reduce collinearity problems, the Legendre polynomial basis

$$b(x) = \{1, \sqrt{3}x, \sqrt{5/4}(3x^2 - 1), ..., \}^{\mathrm{T}}$$

which is orthonormal on $[0, 1]$, again up to degree $\lfloor s \rfloor$.

Remark 1. It is often helpful for analysis to use a localized basis $b(x) = b^*(1/2 + \frac{x - x_0}{h})$, where h is the bandwidth to be introduced shortly, x_0 is a target point, and b^* is a standard basis, e.g., orthonormal on $[0, 1]$. This maps the small interval $[x_0 \pm h/2]$ to the whole space $[0, 1]$, e.g., $b(x_0 - h/2) = b^*(0)$, $b(x_0 + h/2) = b^*(1)$, etc.

When $d > 1$, we need to take tensor products of the one-dimensional basis functions. Specifically we apply the one-dimensional basis $b(x)$ to each feature X_j, $j = 1, ..., d$, giving us $d \times s$ total features $\{b(x_1), ..., b(x_d)\}$, and then we take all possible interactions of these features, which gives s^d total terms. Strictly speaking, for the subsequent analysis, we do not need all possible interactions, only interactions yielding polynomials up to order $\lfloor s \rfloor$ (any larger order polynomial terms we can throw out). This leads to $\binom{d + \lfloor s \rfloor}{\lfloor s \rfloor}$ terms in total, which can be large when d and/or s are not too small; for example, when $d = 10$ and $s = 2$, we get 66 terms.

Now we have essentially everything we need to begin analyzing the estimator. To start out we give some properties of the weights, which are adapted from Proposition 1.12 and Lemma 1.3 of Tsybakov [2009].

The first result shows that the local polynomial estimator reproduces polynomials, i.e., if the outcomes $Y = f(X) = \beta^{\mathrm{T}} b(X)$ are noiseless polynomials, then the estimator simply returns $\widehat{\mu}(x) = f(x)$.

Lemma 1. *Let*

$$w_i(x; X^n) = b(x)^{\mathrm{T}} \mathbb{P}_n \left\{ b(X) \frac{1}{h^d} K\left(\frac{X - x}{h}\right) b(X)^{\mathrm{T}} \right\}^{-1} b(X_i) \frac{1}{nh^d} K\left(\frac{X_i - x}{h}\right)$$

denote the weights of a local polynomial estimator built from basis $b(x)$. If $b(x)$ is a polynomial basis up to order $\lfloor s \rfloor$ then this estimator reproduces polynomials up to order $\lfloor s \rfloor$ in the sense that

$$\sum_{i=1}^{n} w_i(x; X^n) f(X_i) = f(x)$$

for any polynomial of the form $f(x) = \beta^{\mathrm{T}} b(x)$.

Proof. Note that $\sum_{i=1}^n w_i(x;X^n)b(X_i)^{\mathrm{T}}\beta$ equals

$$\begin{aligned} &b(x)^{\mathrm{T}}\mathbb{P}_n\left\{b(X)\frac{1}{h^d}K\left(\frac{X-x}{h}\right)b(X)^{\mathrm{T}}\right\}^{-1}\mathbb{P}_n\left\{b(X_i)\frac{1}{h^d}K\left(\frac{X-x}{h}\right)b(X)^{\mathrm{T}}\right\}\beta \\ &\quad = b(x)^{\mathrm{T}}\beta \end{aligned}$$

as long as the Gram matrix is invertible. □

The next result shows that the weights have certain important localizing properties. In what follows we use the operator norm $\left\|\widehat{Q}_{hx}^{-1}\right\|$, which is just the largest eigenvalue of the real symmetric matrix $\widehat{Q}_{hx}^{-1}$, or equivalently, the reciprocal of the smallest eigenvalue of $\widehat{Q}_{hx}$.

Lemma 2. *Define*

$$\begin{aligned} \widehat{Q}_{hx} &\equiv \mathbb{P}_n\left\{b(X)\frac{1}{h^d}K\left(\frac{X-x}{h}\right)b(X)^{\mathrm{T}}\right\} \\ \lambda_n &\equiv \left\|\widehat{Q}_{hx}^{-1}\right\| \\ \xi_n &\equiv \mathbb{P}_n\left(\|X-x\|\leq h\right)/h^d, \end{aligned}$$

and assume (i) $|K(u)|\leq K_{max}\mathbb{1}(\|u\|\leq 1)$ *and (ii)* $\sup_x\|b(x)\|\leq B$*. Then the local polynomial weights* $w_i(x;X^n)\equiv b(x)^{\mathrm{T}}\widehat{Q}_{hx}^{-1}b(X_i)\frac{1}{nh^d}K\left(\frac{X-x}{h}\right)$ *satisfy, for* $C=K_{\max}B^2$*:*

1. $\max_i|w_i(x;X^n)|\leq C\lambda_n/nh^d$.
2. $\sum_i|w_i(x;X^n)|\leq C\lambda_n\xi_n$.
3. $w_i(x;X^n)=0$ *when* $\|X_i-x\|>h$.

Proof. We have

$$\begin{aligned} |w_i(x;X^n)| &= \left|b(x)^{\mathrm{T}}\widehat{Q}_{hx}^{-1}b(X_i)\frac{1}{nh^d}K\left(\frac{X-x}{h}\right)\right| \\ &\leq \frac{K_{\max}}{nh^d}\|b(x)\|^2\left\|\widehat{Q}_{hx}^{-1}\right\|\mathbb{1}\left(\|X_i-x\|\leq h\right) \\ &\leq \frac{K_{\max}}{nh^d}B^2\lambda_n\mathbb{1}\left(\|X_i-x\|\leq h\right) \end{aligned}$$

using the submultiplicative property of the operator norm, along with Assumptions (i)-(ii). □

In particular Property 1 of Lemma 2 shows that no weight is too large, scaling like $1/nh^d$ disregarding other factors, Property 2 shows the sum of the weights is not too large, only depending on the product $\lambda_n\xi_n$, and Property 3 shows they are localized just like those of a simple kernel estimator. Note these properties are reminiscent of the conditions of Stone's theorem.

4 Conditional bias & variance

Now we analyze the pointwise conditional bias of the local polynomial estimator, exploiting the properties derived above.

For simplicity first consider the $d = 1$ case. There we can use the Taylor expansion

$$\mu(X_i) = \mu(x) + \sum_{j=1}^{\lfloor s \rfloor - 1} \frac{D^j \mu(x)}{j!}(X_i - x)^j + \frac{D^{\lfloor s \rfloor}\mu(x + \tau(X_i - x))}{\lfloor s \rfloor!}(X_i - x)^{\lfloor s \rfloor}$$

to obtain

$$\begin{aligned}
\mathbb{E}\{\widehat{\mu}(x) \mid X^n\} - \mu(x) &= \sum_{i=1}^n w_i(x; X^n)\mu(X_i) - \mu(x) = \sum_{i=1}^n w_i(x; X^n)\Big\{\mu(X_i) - \mu(x)\Big\} \\
&= 0 + \sum_{i=1}^n w_i(x; X^n) \frac{D^{\lfloor s \rfloor}\mu(x + \tau(X_i - x)) - D^{\lfloor s \rfloor}\mu(x)}{\lfloor s \rfloor!}(X_i - x)^{\lfloor s \rfloor} \\
&\leq \sum_{i=1}^n |w_i(x; X^n)| \frac{L\|X_i - x\|^s}{\lfloor s \rfloor!} \mathbb{1}(\|X_i - x\| \leq h) \\
&\leq \left(\frac{LC}{\lfloor s \rfloor!}\right) \lambda_n \xi_n h^s
\end{aligned}$$

where the second line used the polynomial reproducing property of Lemma 1, the third the Holder assumption on μ and Property 3 of Lemma 2, and the last line Properties 1-2 of Lemma 2. When $d > 1$ one can make use of the multivariate Taylor expansion

$$\begin{aligned}
\mu(X_i) = \mu(x) &+ \sum_{0 < |j| \leq \lfloor s \rfloor - 1} \frac{D^j \mu(x)}{j!}(X_i - x)^j \\
&+ \sum_{|k| = \lfloor s \rfloor} \frac{\lfloor s \rfloor}{k!} \int_0^1 (1 - t)^{\lfloor s \rfloor - 1} D^k \mu(x + t(X_i - x))\, dt (X_i - x)^k
\end{aligned}$$

and similarly obtain

$$\begin{aligned}
\mathbb{E}\{\widehat{\mu}(x) \mid X^n\} - \mu(x) &= \sum_{i=1}^n w_i(x; X^n)\mu(X_i) - \mu(x) = \sum_{i=1}^n w_i(x; X^n)\Big\{\mu(X_i) - \mu(x)\Big\} \\
&= 0 + \sum_{i=1}^n w_i(x; X^n) \sum_{|k| = \lfloor s \rfloor} \frac{\lfloor s \rfloor}{k!} \int_0^1 (1 - t)^{\lfloor s \rfloor - 1} \Big\{D^k \mu(x + t(X_i - x)) - D^k \mu(x)\Big\}\, dt (X_i - x)^k \\
&\lesssim \sum_{i=1}^n |w_i(x; X^n)| \|X_i - x\|^s \mathbb{1}(\|X_i - x\| \leq h) \;\lesssim\; h^s \lambda_n \xi_n
\end{aligned}$$

using the polynomial reproducing property of Lemma 1, and Properties 2-3 of Lemma 2.

For the pointwise conditional variance, assuming $\mathrm{var}(Y \mid X = x) \leq \sigma^2$, we have

$$\begin{aligned}
\mathrm{var}\{\widehat{\mu}(x) \mid X^n\} &= \sum_{i=1}^n w_i(x; X^n)^2 \mathrm{var}(Y \mid X_i) \\
&\leq \sigma^2 \max_i |w_i(x; X^n)| \sum_{i=1}^n |w_i(x; X^n)| \\
&\leq \sigma^2 \left(\frac{C\lambda_n}{nh^d}\right) C\lambda_n \xi_n
\end{aligned}$$

using the assumed variance bound together with Properties 1-2 of Lemma 2.

Therefore we have the following bound on the conditional (pointwise) error

$$\begin{aligned}
\mathbb{E}\{|\widehat{\mu}(x) - \mu(x)| \mid X^n\}^2 &\leq \mathbb{E}\Big[\{\widehat{\mu}(x) - \mu(x)\}^2 \mid X^n\Big] \\
&= \mathbb{E}\{\widehat{\mu}(x) - \mu(x) \mid X^n\}^2 + \mathrm{var}\{\widehat{\mu}(x) \mid X^n\} \;\lesssim\; h^{2s}(\lambda_n \xi_n)^2 + \frac{\lambda_n^2 \xi_n}{nh^d}
\end{aligned}$$

In a fixed design setting we might treat λ_n, ξ_n as bounded as an assumption about the design.

5 Random design

We have shown that

$$\mathbb{E}\{|\widehat{\mu}(x) - \mu(x)| \mid X^n\} \;\lesssim\; h^s \lambda_n \xi_n + \frac{\lambda_n \sqrt{\xi_n}}{\sqrt{nh^d}}$$

What are the implications in a random design where (λ_n, ξ_n) are random variables? If $\xi_n = O_{\mathbb{P}}(1)$ and $\lambda_n = O_{\mathbb{P}}(1)$ then unconditionally we have

$$|\widehat{\mu}(x) - \mu(x)| = O_{\mathbb{P}}\left(h^s + \frac{1}{\sqrt{nh^d}}\right).$$

This follows from the fact that for generic random variable Z_n if $\mathbb{E}(|Z_n| \mid X^n)| \lesssim r_n = r_n(X^n)$ then $Z_n = O_{\mathbb{P}}(r_n)$ by iterated expectation and Markov's inequality; in other words, unconditional convergence in probability follows from conditional bounds.

Now we show that indeed $\xi_n = O_{\mathbb{P}}(1)$ and $\lambda_n = O_{\mathbb{P}}(1)$ under some conditions. For the first we have

$$\mathbb{E}(\xi_n) = \frac{n}{nh^d} \mathbb{P}(\|X_i - x\| \leq h) \leq \frac{B}{h^d} \int \mathbb{1}(\|t - x\| \leq h)\, dt \leq C$$

where the last two inequalities follow from an assumption that the density of X is upper bounded by B, and since the volume of a ball in $\mathbb{R}^d$ with radius h is proportional to h^d. Thus Markov's inequality yields

$$\mathbb{P}(\xi_n \geq M_\epsilon) \leq \mathbb{E}(\xi_n)/M_\epsilon \leq \epsilon$$

if we take $M_\epsilon = C/\epsilon$.

For the second, we assume the eigenvalues of $Q = \mathbb{E}(\widehat{Q})$ are bounded above and below away from zero (this holds if the density of X is bounded above/below near x) and use the matrix law of large numbers (which tells us $\|\widehat{Q} - Q\| \xrightarrow{p} 0$ as long as $nh^d \to \infty$, see Tropp 2012) to conclude the result as in the following lemma.

Lemma 3. *For any matrix $\widehat{Q}$, if $\|\widehat{Q} - Q\| \xrightarrow{p} 0$ and $\|Q^{-1}\| \leq C < \infty$ then $\|\widehat{Q}^{-1}\| = O_{\mathbb{P}}(1)$.*

Proof. First note that

$$\begin{aligned}
\|\widehat{Q}^{-1}\| &\leq \|\widehat{Q}^{-1} - Q^{-1}\| + \|Q^{-1}\| \\
&= \|\widehat{Q}^{-1}(Q - \widehat{Q})Q^{-1}\| + \|Q^{-1}\| \\
&\leq \|\widehat{Q}^{-1}\|\|\widehat{Q} - Q\|\|Q^{-1}\| + \|Q^{-1}\|
\end{aligned}$$

by the triangle inequality and the submultiplicative property of the operator norm, which implies $\|\widehat{Q}^{-1}\| \leq \frac{\|Q^{-1}\|}{1 - \|\widehat{Q} - Q\|\|Q^{-1}\|}$. Therefore we have

$$\begin{aligned}
\mathbb{P}\left(\|\widehat{Q}^{-1}\| \geq 2\|Q^{-1}\|\right) &\leq \mathbb{P}\left(\frac{\|Q^{-1}\|}{1 - \|\widehat{Q} - Q\|\|Q^{-1}\|} \geq 2\|Q^{-1}\|\right) \\
&= \mathbb{P}\left(\|\widehat{Q} - Q\| \geq \frac{1}{2\|Q^{-1}\|}\right)
\end{aligned}$$

and when $\|Q^{-1}\| < \infty$ so that $\delta \equiv 1/2\|Q^{-1}\| > 0$, the above probability is no more than ϵ for all $n > N_\epsilon$, since $\|\widehat{Q} - Q\| \xrightarrow{p} 0$. □

References

L. Györfi, M. Kohler, A. Krzykaz, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, 2002.

A. B. Tsybakov. *Introduction to Nonparametric Estimation.* New York: Springer, 2009.