ORFE 524: Statistical Theory and Methods
Final Homework

Zachary Hervieux-Moore

Thursday 12$^{\text{th}}$ January, 2017

Exercise 1: Consider the nonparametric regression model $Y = f(X) + \nu$, $X \in \mathcal{X} \subseteq \mathbb{R}^2$, $Y \in \mathbb{R}$, where the noise ν is independent of X and satisfies $\mathbb{E}[\nu] = 0$. We assume that $f : \mathcal{X} \to \mathbb{R}$ belongs to a function class $\mathcal{F} \doteq \operatorname{span}\{f_k : \mathcal{X} \to \mathbb{R}\}_{k=1}^{\infty}$, i.e., every $g \in \mathcal{F}$ is obtained as $g = \sum_{k=1}^{\infty} \alpha_k f_k$.

Let $\{(X_i, Y_i)\}_{i=1}^n$ be n i.i.d. observations of the nonparametrix regression model; our goal is to estimate the function f, in other words, the unknown coefficients α_k. The point of this problem is to compare two different estimators defied as follows: for a fixed integer $N \geq 1$, let

$$\widehat{f}_N = \textstyle\sum_{k=1}^N \widehat{\alpha}_k f_k, \text{ and } \tilde{f}_N = \textstyle\sum_{k=1}^N \tilde{\alpha}_k f_k,$$

where $\widehat{\alpha}_k = n^{-1} \sum_{i=1}^n Y_i f_k(X_i)$ and $\{\tilde{\alpha}_k\}_{k=1}^N$ is the least-squares estimator (LSE) defined by

$$\{\tilde{\alpha}_k\}_{k=1}^N \in \arg \min_{\alpha_1,\ldots,\alpha_N} \frac{1}{2n} \sum_{i=1}^n \left[Y_i - \sum_{k=1}^N \alpha_k f_k(X_i) \right]^2 \tag{1}$$

Let $\widehat{\alpha} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_N)^T$ and $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_N)^T$.

1) Give a sufficient condition so that the minimization in equation (1) has a unique solution $\tilde{\alpha}$.

2) Under the condition of question (1), express $\tilde{\alpha}$ in terms of $\widehat{\alpha}$.

3) Suppose $\{f_k\}_{k=1}^{\infty}$ is an orthonormal system with respect to the distribution of X, i.e.,

$$\langle f_j, f_k \rangle \doteq \mathbb{E}[f_j(X) f_k(X)] = 0 \text{ for } j \neq k,$$
$$\|f_j\|^2 \doteq \langle f_j, f_j \rangle = \mathbb{E}[f_j^2(X)] = 1$$

   Show that for fixed N and $x \in \mathbb{R}^d$, $|\widehat{f}_N(x) - \tilde{f}_N(x)| \xrightarrow{p} 0$ as $n \to \infty$.

4) If in addition $\mathcal{X}$ is compact, argue that for fixed N, $\sup_{x \in \mathcal{X}} |\widehat{f}_N(x) - \tilde{f}_N(x)| \xrightarrow{a.s.} 0$ as $n \to \infty$.

   Remark: *You would need to show, for the last two questions, that the condition (1) holds in probability, respectively a.s., as $n \to \infty$.*

5) Now suppose that $f \in \text{span}\{f_k\}_{k=1}^N$, i.e., f has a finite expansion up to N. Show that $\widehat{f}_N$, $\tilde{f}_N$ are both unbiased for f, i.e., for any $x \in \mathcal{X}$, $f(x) = \mathbb{E}[\widehat{f}_N(x)] = \mathbb{E}[\tilde{f}_N(x)]$.

   Remark: *For* $\widehat{f}_N$, you can directly use the results of a previous homework.

Answer:

1) First, let us write the least-squares estimator in matrix form

$$\tilde{\alpha} = \arg\min_{\alpha} \frac{1}{2n} \|Y - f(X)\alpha\|$$

   Where $\tilde{\alpha}$ is $N \times 1$, Y is $n \times 1$, $f(X)$ is $n \times N$ and $f(X)_{ij} = f_j(X_i)$, and α is $N \times 1$. Now, this is the standard LSE problem which has the result that

$$\tilde{\alpha} = (f(X)^T f(X))^{-1} f(X)^T Y$$

   This yields a unique solution when $f(X)^T f(X)$ is invertible, i.e., full rank.

2) Again, we write $\widehat{\alpha}$ in matrix form for simplicity

$$\widehat{\alpha} = \frac{1}{n} f(X)^T Y$$

   Now it is easy to see that

$$\tilde{\alpha} = n(f(X)^T f(X))^{-1} \widehat{\alpha}$$

3) We need to show that condition 1) holds in probability. The general goal is to pick enough samples to guarantee that $f(X)^T f(X)$ is invertible. First, let's assume that the f_k's are not equal almost everywhere. That is, there is some measurable set that is different between them. Then, there is some non zero probability, say p, that X_i is picked in this region. Thus, every sample taken is a Bernoulli trial whether or not the f_k's are equal. The question can then be phrased, do these Bernoulli trials hit success N times? The answer is of course yes and is due to the law of large number that the sample average will converge to p. Thus,

if we need N successes, we will need on the order of N/p of samples. For a large number of functions, one needs to pick the smallest overlap among all pairwise combinations but the same reasoning holds since N is finite. Therefore, the first condition holds almost surely and hence in probability.

We now write out the problem

$$|\widehat{f}_N(x) - \tilde{f}_N(x)| = |f(x)^T\widehat{\alpha} - f(x)^T\tilde{\alpha}|$$

However, we just showed that $\tilde{\alpha} \xrightarrow{p} n(f(X)^T f(X))^{-1}\widehat{\alpha}$. So

$$|f(x)^T\widehat{\alpha} - f(x)^T\tilde{\alpha}| \xrightarrow{p} |f(x)^T\widehat{\alpha} - f(x)^T n(f(X)^T f(X))^{-1}\widehat{\alpha}|$$

By the law of large numbers $n(f(X)^T f(X))^{-1} \xrightarrow{p} \mathbb{E}[f(X)^T f(X))]^{-1}$. By the orthogonality assumption, $\mathbb{E}[f(X)^T f(X))]^{-1} = I$. This yields

$$|f(x)^T\widehat{\alpha} - f(x)^T n(f(X)^T f(X))^{-1}\widehat{\alpha}| \xrightarrow{p} |f(x)^T\widehat{\alpha} - f(x)^T\widehat{\alpha}| = 0$$

4) By invoking the law of large numbers in the previous question, almost sure convergence is also achieved for condition 1). The same proof then follows but the supremum may be taken since it is on a compact set and thus a supremum will be achieved.

5) We showed that $\widehat{f}_N$ is unbiased in Homework 6. Then, we know that $\tilde{f}_N$ is the solution to the LSE which we know is unbiased since there is only a finite number of terms.

Exercise 2: Consider a function $f : \mathbb{R} \to \mathbb{R}$, twice continuously differentiable on $[-1, 1]$. Let's elucidate properties of f on the compact set $[-1, 1]$.

1) Show that f is $(\lambda, 2)$-Holder, for some λ, i.e.,

$$\exists \lambda > 0 \text{ such that } \forall x, x' \in [-1, 1],\ |f'(x) - f'(x')| \leq \lambda |x - x'|$$
$$\text{(Condition (1))}$$

2) For any $x_0 \in [-1, 1]$, let $f_{x_0}(x) = f(x_0) + f'(x_0)(x - x_0)$ be the Taylor expansion of f at x_0. Show that f satisfies

$$\exists \lambda > 0 \text{ such that } \forall x, x' \in [-1, 1],\ |f_x(x') - f(x')| \leq \lambda |x - x'|^2$$
$$\text{(Condition (2))}$$

3) Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be the classes of functions $\mathbb{R} \to \mathbb{R}$ satisfying conditions (1) and (2) respectively. Consider a regression setting on variables (X, Y), where $X \sim$ Uniform $[-1, 1]$, $f(x) = \mathbb{E}[Y|X = x]$ is in $\mathcal{F}_1$, or alternatively is in $\mathcal{F}_2$. How do the minimax rates for these two classes $(\mathcal{F}_1, \mathcal{F}_2)$ compare and why? (One should be of equal or larger order).

Answer:

1) The result follows from the fact that f is twice continuously differentiable. Because the second derivative is continuous on a compact set, it is bounded. That is, $f''(x) \leq \lambda$. By the mean value theorem, for all $x, x' \in [-1, 1]$ there exists some c such that

$$|f'(x) - f'(x')| \leq |f''(c)||x - x'| \leq \lambda |x - x'|$$

2) We first note that $f_x(x') - f(x')$ is simply the remainder term of the Taylor expansion. So

$$|f_x(x') - f(x')| = |R_1(x')|$$

This is upper bounded by

$$|R_1(x')| \leq \frac{|f''(c)|}{2!}|x' - x|^2 = \lambda |x' - x|^2$$

Which is exactly the result we desired.

3) We note that the second class has the following peculiarity.

$$\begin{aligned} |f_x(x') - f(x')| &\leq \lambda |x' - x|^2 \\ \Leftrightarrow \frac{|f_x(x') - f(x')|}{|x' - x|} &\leq \lambda |x' - x| \end{aligned}$$

Taking the limit as $x' \to x$ produces the definition of the derivative on the left hand side. Thus

$$f'(x) = \lim_{x' \to x} \lambda |x' - x| = 0$$

So, $\mathcal{F}_2$ is the family of constants. Obviously, constants satisfy condition 1). Thus, $\mathcal{F}_1$ is a much "richer" family than $\mathcal{F}_2$. This also means that the minimax rate for $\mathcal{F}_1$ is much slower than $\mathcal{F}_2$. That is, the rate is of equal or larger order than $\mathcal{F}_2$ since it takes longer to get to 0.

Exercise 3: Consider a nonparametric classification problem over jointly distributed (X, Y), where X is in some space $\mathcal{X}$, and $Y \in \{-1, 1\}$. We consider the following plug-in classifier:

Define the regression function $f(x) = \mathbb{E}[Y|X = x]$, and let $\widehat{f} : \mathcal{X} \to \mathbb{R}$ be an estimator of f based on i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^n$ of (X, Y). The classifier at $x_0 \in \mathcal{X}$ is then obtained as $\widehat{g}(x_0) = \text{sign}(\widehat{f}(x_0))$. We define the ℓ_2-regression error and the 0-1 classification error at a fixed x_0 as:

$$\ell_2[\widehat{f}(x_0)] := \mathbb{E}_{Y|X=x_0}[(\widehat{f}(x_0) - Y)^2]$$
$$\ell_{0,1}[\widehat{g}(x_0)] := \mathbb{E}_{Y|X=x_0}[1\{\widehat{g}(x_0) \neq Y\}]$$

1) Show that the best possible classifier at x_0 is $g(x_0) = \text{sign}(\widehat{f}(x_0))$, i.e., $g(x_0)$ minimizes $\ell_{0,1}(y)$ over any value $y \in \{-1, 1\}$.

2) Show that $|\widehat{f}(x_0) - f(x_0)| = \sqrt{\ell_2[\widehat{f}(x_0)] - \ell_2[f(x_0)]}$.

3) Show that the "excess" classification error can be bounded as:

$$\ell_{0,1}[\widehat{g}(x_0)] - \ell_{0,1}[g(x_0)] \leq |\widehat{f}(x_0) - f(x_0)|$$

4) Now we consider the general case where $Y \in \{a, b\}$ with $a < b$. Similarly, let $f(x) = \mathbb{E}[Y|X = x]$, and let $\widehat{f} : \mathcal{X} \to \mathbb{R}$ be an estimator. The best possible classifier at $x_0 \in \mathcal{X}$ is given as

$$g(x_0) = \begin{cases} a & \text{if } f(x_0) \leq (a+b)/2 \\ b & \text{if } f(x_0) > (a+b)/2 \end{cases},$$

and an estimate is $\widehat{g}(x_0) = \begin{cases} a & \text{if } \widehat{f}(x_0) \leq (a+b)/2 \\ b & \text{if } \widehat{f}(x_0) > (a+b)/2 \end{cases}$

Show that the excess classification error is now bounded as:

$$\ell_{0,1}[\widehat{g}(x_0)] - \ell_{0,1}[g(x_0)] \leq \frac{2}{b-a} \cdot |\widehat{f}(x_0) - f(x_0)|$$

Answer:

1) We write the $\ell_{0,1}$ loss in integral form

$$\begin{aligned}\ell_{0,1}[\widehat{g}(x_0)] &= \mathbb{E}_{Y|X=x_0}[1\{\widehat{g}(x_0) \neq Y\}]\\ &= \int_{\{-1,1\}} 1\{\widehat{g}(x_0) \neq Y|x_0\} dP(y|x_0)\\ &= 1_{\{\widehat{g}(x_0)=1\}} p(y=-1|x_0) + 1_{\{\widehat{g}(x_0)=-1\}} p(y=1|x_0)\end{aligned}$$

We wish to minimize this expression. Thus, we set $\widehat{g}(x_0)$ to be the indicator of the smaller conditional probability. Or

$$\arg\min_{\widehat{g}(x_0)} \ell_{0,1}[\widehat{g}(x_0)] = g(x_0) = \begin{cases} 1 & \text{if } p(y=-1|x_0) \leq p(y=1|x_0)\\ -1 & \text{if } p(y=-1|x_0) > p(y=1|x_0)\end{cases}$$

Which is precisly $\text{sign}(f(x_0))$. Since

$$\begin{aligned}f(x_0) = \mathbb{E}[Y|X=x_0] &= \int_{\{-1,1\}} y dP(y|x_0)\\ &= p(y=1|x_0) - p(y=-1|x_0)\end{aligned}$$

Thus, $\text{sign}(f(x_0))$ is precisely $g(x_0)$.

2) We start with the right hand side

$$\begin{aligned}&\sqrt{\ell_2[\widehat{f}(x_0)] - \ell_2[f(x_0)]}\\ &= \sqrt{\mathbb{E}_{Y|X=x_0}[(\widehat{f}(x_0) - Y)^2] - \mathbb{E}_{Y|X=x_0}[(f(x_0)-Y)^2]}\\ &= \sqrt{\mathbb{E}_{Y|X=x_0}[\widehat{f}(x_0)^2 - 2\widehat{f}(x_0)Y + Y^2] - \mathbb{E}_{Y|X=x_0}[f(x_0)^2 - 2f(x_0)Y + Y^2]}\\ &= \sqrt{\widehat{f}(x_0)^2 - 2\widehat{f}(x_0)f(x_0) + 1 - f(x_0)^2 + 2f(x_0)^2 - 1}\end{aligned}$$

Where in the last step we used linearity of conditional expectation and that $f(x_0)$ and $\widehat{f}(x_0)$ is measurable w.r.t. x_0. Also, we used the facts that $f(x_0) = \mathbb{E}[Y|X=x_0]$ and finally that $\mathbb{E}[Y^2|X=x_0] = 1$. Now simplifying

$$\begin{aligned}&= \sqrt{\widehat{f}(x_0)^2 - 2\widehat{f}(x_0)f(x_0) + f(x_0)^2}\\ &= \sqrt{(\widehat{f}(x_0) - f(x_0))^2}\\ &= |\widehat{f}(x_0) - f(x_0)|\end{aligned}$$

3) Start by writing the definition of $\ell_{0,1}$ as done in part 1)

$$\begin{aligned}
&\ell_{0,1}[\widehat{g}(x_0)] - \ell_{0,1}[g(x_0)] \\
&= 1_{\{\widehat{g}(x_0)=1\}} p(y=-1|x_0) + 1_{\{\widehat{g}(x_0)=-1\}} p(y=1|x_0) \\
&\quad - 1_{\{g(x_0)=1\}} p(y=-1|x_0) - 1_{\{g(x_0)=-1\}} p(y=1|x_0) \\
&= (1_{\{\widehat{g}(x_0)=1\}} - 1_{\{g(x_0)=1\}}) p(y=-1|x_0) \\
&\quad + (1_{\{\widehat{g}(x_0)=-1\}} - 1_{\{g(x_0)=-1\}}) p(y=1|x_0) \qquad (\star)
\end{aligned}$$

Now consider the following two cases. If $\widehat{g}(x_0) = g(x_0)$, then the above is 0 and the inequality is trivially satisfied. Now consider the case when $\widehat{g}(x_0) \neq g(x_0)$. Then we have

$$\begin{aligned}
\widehat{g}(x_0) = 1 &\implies (\star) = p(y=-1|x_0) - p(y=1|x_0) = -f(x_0) \\
\widehat{g}(x_0) = -1 &\implies (\star) = -p(y=-1|x_0) + p(y=1|x_0) = f(x_0)
\end{aligned}$$

In both case, we have the following inequalities since $\widehat{f}(x_0) \geq 0$ when $\widehat{g}(x_0) = 1$ and $-\widehat{f}(x_0) \geq 0$ when $\widehat{g}(x_0) = -1$.

$$\begin{aligned}
\widehat{g}(x_0) = 1 &\implies (\star) = p(y=-1|x_0) - p(y=1|x_0) \leq \widehat{f}(x_0) - f(x_0) \\
\widehat{g}(x_0) = -1 &\implies (\star) = -p(y=-1|x_0) + p(y=1|x_0) \leq f(x_0) - \widehat{f}(x_0)
\end{aligned}$$

Combining these inequalities yield the result $\ell_{0,1}[\widehat{g}(x_0)] - \ell_{0,1}[g(x_0)] \leq |\widehat{f}(x_0) - f(x_0)|$.

4) The proof is similar to part 3). First we note that

$$f(x_0) = ap(y=a|x_0) + bp(y=b|x_0).$$

Following the same steps as part 3) we get

$$\begin{aligned}
&\ell_{0,1}[\widehat{g}(x_0)] - \ell_{0,1}[g(x_0)] \\
&= (1_{\{\widehat{g}(x_0)=b\}} - 1_{\{g(x_0)=b\}}) p(y=a|x_0) \\
&\quad + (1_{\{\widehat{g}(x_0)=a\}} - 1_{\{g(x_0)=a\}}) p(y=b|x_0) \qquad (\star)
\end{aligned}$$

Again, if $\widehat{g}(x_0) = g(x_0)$ then $(\star)$ is trivially satisfied. Now consider the case when $\widehat{g}(x_0) \neq g(x_0)$. For the first inequality, assume $\widehat{g}(x_0) = a$. Then

$$(\star) = -p(y=a|x_0) + p(y=b|x_0)$$

Note that

$$\begin{aligned} a \cdot (\star) &= -f(x_0) + (a+b)p(y = b|x_0) \\ \text{and } -b \cdot (\star) &= -f(x_0) + (a+b)p(y = a|x_0) \end{aligned}$$

Adding the above equations yield

$$\begin{aligned} (a-b) \cdot (\star) &= -2f(x_0) + (a+b) \\ (\star) &= \frac{2}{b-a} f(x_0) - \frac{a+b}{b-a} \end{aligned}$$

Note that since $\widehat{g}(x_0) = a$ by assumption, then $\widehat{f}(x_0) \leq (a+b)/2$. This is equivalent to $2\widehat{f}(x_0)/(b-a) \leq (a+b)/(b-a)$. We conclude that

$$(\star) \leq \frac{2}{b-a}(f(x_0) - \widehat{f}(x_0))$$

Now we assume $\widehat{g}(x_0) = b$. Following the same logic. We have

$$(\star) = p(y = a|x_0) - p(y = b|x_0)$$

Then

$$\begin{aligned} (b-a)(\star) &= -2f(x_0) + (a+b) \\ (\star) &= \frac{a+b}{b-a} - \frac{2}{b-a} f(x_0) \end{aligned}$$

Now, since $\widehat{g}(x_0) = b$, this implies $\widehat{f}(x_0) > (a+b)/2$ which is equivalent to $2\widehat{f}(x_0)/(b-a) > (a+b)/(b-a)$. We conclude that

$$(\star) \leq \frac{2}{b-a}(\widehat{f}(x_0) - f(x_0))$$

Putting both inequalities together yields the result

$$\ell_{0,1}[\widehat{g}(x_0)] - \ell_{0,1}[g(x_0)] \leq \frac{2}{b-a} \cdot |\widehat{f}(x_0) - f(x_0)|$$

Exercise 4: The problem builds on the last one. We are now interested in bounding the classification error at a fixed point x_0. Assume throughout that the marginal $X \sim \text{Uniform}[-1,1]$, and $x_0 = 0$. Again, $Y \in \{-1,1\}$ is to be predicted at $x_0 = 0$. We now assume that the regression function $f : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable on $[-1,1]$ (as in the previous problem above). We consider a *local linear* classifier (over i.i.d. observations $\{(X_i, Y_i\}_{i=1}^n$ of (X,Y)) obtained as follows:

First, for a bandwidth $h > 0$, we define a new data representation by mapping any $x \in [-1,1]$ to a vector $\tilde{x} = (1, x/h)^T$. Thus we map $x_0 = 0$ to $\tilde{x}_0 = (1,0)^T$ and X_i to $\tilde{X}_i$, $i \in [n]$. Then we define a local linear regression estimate $f_h(0) = \widehat{\theta}^T \tilde{x}_0$ for $f(0)$, where $\widehat{\theta}$ is obtained by weighted-least-squares as

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^2} \sum_{i=1}^n [Y_i - \theta^T \tilde{X}_i]^2 \cdot \omega_{h,i}$$

Here the weight $\omega_{h,i}$ is determined as follows. Let $n_h := \sum_{i=1}^n 1\{|X_i| \le h\}$, i.e., the number of sample points in the interval $[-h,h]$ containing $x_0 = 0$, then

$$\omega_{h,i} = \frac{1}{n_h} 1\{|X_i| \le h\} \text{ if } n_h > 0, \text{ otherwise } \omega_{h,i} = \frac{1}{n} \text{ if } n_h = 0$$

The final classifier is obtained as $g_h(0) = \text{sign}(f_h(0))$, estimating $g(0) = \text{sign}(f(0))$. In what follows, we evaluate the performance of this estimate by considering the ℓ_2-regression error and 0-1 classification error defined in problem 3).

To simplify the notation, let $\mathbf{X} = (X_1, \ldots, X_n)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ denote the random samples viewed as vectors, and let the $n \times 2$ matrix $\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_n)^T$ denote the corresponding transformation of $\mathbf{X}$. Also, define the 2×2 matrix $B_h = \tilde{\mathbf{X}}^T W_h \tilde{\mathbf{X}}$.

1) Assume $B_h \succ 0$. Show that $\widehat{\theta} = B_h^{-1}(\tilde{\mathbf{X}}^T W_h)\mathbf{Y}$

2) Assume $B_h \succ 0$. Show that $f_h(0)$ is linear in $\mathbf{Y}$, i.e.,

$$f_h(0) = \sum_{i=1}^n \alpha_{h,i} \cdot Y_i, \text{ where } \alpha_{h,i} = \omega_{h,i} \cdot \tilde{x}_0^T B_h^{-1} \tilde{X}_i$$

3) Reproducibility of linear functions: Assume $B_h \succ 0$. Suppose, only for this question, that $f(x)$ were affine, and that we were to regress on $\{X_i, f(X_i)\}_{i=1}^n$; then since $\alpha_{h,i}$ depends only on X_i's, we would obtain $f_h(0) = \sum_{i=1}^n \alpha_{h,i} \cdot f(X_i)$.

   a) Show that, in this case, we have $f_h(0) = f(0)$. (Hint: $f(X_i) = f(0) + X_i \cdot f'(0)$ since f is affine, i.e., $f(X_i) = \theta^T \tilde{X}_i$ for some θ. Show that $\widehat{\theta} = \theta$).

   b) Since f were assumed to be an arbitrary affine function, deduce from above that $\sum_{i=1}^n \alpha_{h,i} = 1$.

4) Regression variance: Suppose $n_h > 0$ and $\lambda_{min}(B_h) \geq \lambda_0$ for some λ_0 independent of X_i's.

   a) Show that $\sum_{i=1}^n |\alpha_{h,i}| \leq C(\lambda_0)$ and that $\max_i |\alpha_{h,i}| \leq C(\lambda_0)/n_h$ for some constant $C(\lambda_0)$.

   b) Fix $\mathbf{X} = \{X_i\}_{i=1}^n$. Let $\tilde{f}_h(0) = \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[f_h(0)] = \sum_{i=1}^n \alpha_{h,i} \cdot f(X_i)$. Deduce from the above that

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[(f_h(0) - \tilde{f}_h(0))^2] \leq \frac{C(\lambda_0)}{n_h} \text{ for some constant } C(\lambda_0)$$

5) Regression bias: Let $\tilde{f}_h(0)$ be defined as in the previous question. Assume that $n_h > 0$ and that $\lambda_{\min}(B_h) \geq \lambda_0$ for some λ_0 independent of the X_i's. Show that

$$|\tilde{f}_h(0) - f(0)| \leq C(\lambda_0) \cdot \|f''\|_\infty \cdot h^2, \text{ where } \|f''\|_\infty = \sup_{x \in [-1,1]} |f''(x)|$$

   Hint: consider parts 3), 4), and Problem 2).

6) We are now ready to conclude. First, we need to check that the above conditions, hold for sufficiently large n, namely that $n_h > 0$, and $\lambda_{\min}(B_h) \geq \lambda_0$.

   a) Let $h \in (0,1)$ be fixed. Notice that $n_h/n \xrightarrow{p} h$. Why? Deduce from this that

$$P\left(\frac{1}{n_h} \leq \frac{1}{n \cdot h}\right) \xrightarrow{n\to\infty} 1, \text{ and therefore that } P(n_h > 0) \xrightarrow{n\to\infty} 1$$

b) Show that there exists $\lambda_0 > 0$ such that $P(\lambda_{\min}(B_h) \geq \lambda_0) \xrightarrow{n\to\infty} 1$. Hint: $B_h = B_{h,1} 1\{n_h > 0\} + B_{h,2} 1\{n_h = 0\}$, where $B_{h,1}$ is a sample counterpart to the conditional "covariance-type" matrix

$$\mathbb{E}[\tilde{X}\tilde{X}^T | X \in [-h,h]] \doteq (1/h) \cdot \mathbb{E}[\tilde{X}\tilde{X}^T 1\{X \in [-h,h]\}] = \begin{bmatrix} 1 & 0 \\ 0 & 1/3 \end{bmatrix}$$

c) Conclude that, for fixed $h \in (0,1)$, $\exists N_0$, such that, $\forall n > N_0$

$$\mathbb{E}_{\mathbf{X},\mathbf{Y}} \left[\ell_{0,1}[g_h(0)] - \ell_{0,1}[g(0)]\right] \leq C \left(\frac{1}{n \cdot h} + h^4\right)^{1/2}$$

for a constant C independent of n and h.

(In other words, we can make the excess classification error arbitrarily small by picking h sufficiently small, and n sufficiently large w.r.t. h. In fact, much stronger results hold ... if you're curious).

Answer:

1) We first note that

$$\widehat{\theta} \in \arg\min_{\theta\in\mathbb{R}^2} \sum_{i=1}^{n} [Y_i - \theta^T \tilde{X}_i]^2 \cdot \omega_{h,i} = \arg\min_{\theta\in\mathbb{R}^2} \sum_{i=1}^{n} [Y_i' - \theta^T \tilde{X}_i']^2$$

where $Y_i' = Y_i\sqrt{w_{i,h}}$ and $\tilde{X}_i' = \tilde{X}_i\sqrt{w_{i,h}}$. Then the solution is simply the solution to normal least squares regression, $(\tilde{\mathbf{X}}\mathbf{'}^T\tilde{\mathbf{X}}\mathbf{'})^{-1}\tilde{\mathbf{X}}\mathbf{'}^T Y'$ where $\mathbf{Y'} = \sqrt{W_h}\mathbf{Y}$ and $\tilde{\mathbf{X}}\mathbf{'} = \sqrt{W_h}\tilde{\mathbf{X}}$ where $\sqrt{W_h} = \text{diag}(\sqrt{w_{1,h}}, \ldots, \sqrt{w_{n,h}})$. Pluggin this in results in

$$\begin{aligned}
\widehat{\theta} &= ((\sqrt{W_h}\tilde{\mathbf{X}})^T \sqrt{W_h}\tilde{\mathbf{X}})^{-1} (\sqrt{W_h}\tilde{\mathbf{X}})^T \sqrt{W_h}\mathbf{Y} \\
&= (\tilde{\mathbf{X}}^T \sqrt{W_h}^T \sqrt{W_h}\tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \sqrt{W_h}\sqrt{W_h}\mathbf{Y} \\
&= (\tilde{\mathbf{X}}^T W_h \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T W_h \mathbf{Y} \\
&= B_h^{-1} (\tilde{\mathbf{X}}^T W_h) \mathbf{Y}
\end{aligned}$$

2) Using the previous result, we have

$$\begin{aligned}
f_h(0) &= \widehat{\theta}^T \tilde{x}_0 \\
&= (B_h^{-1} (\tilde{\mathbf{X}}^T W_h) Y)^T \tilde{x}_0
\end{aligned}$$

This is a scalar so we can take the transpose without changing the answer.

$$= \tilde{x}_0^T B_h^{-1} (\tilde{\mathbf{X}}^T W_h) Y$$

Since W_h is diagonal, we have the nice simplification that $\tilde{\mathbf{X}}^T W_h = (\tilde{X}_1 w_{1,h}, \dots, \tilde{X}_n w_{n,h})$. Writing as sum notation,

$$f_h(0) = \sum_{i=1}^{n} \alpha_{h,i} \cdot Y_i, \text{ where } \alpha_{h,i} = \omega_{h,i} \cdot \tilde{x}_0^T B_h^{-1} \tilde{X}_i$$

3)

a) Writing down the optimization problem again

$$\widehat{\theta} \in \arg\min_{\theta' \in \mathbb{R}^2} \sum_{i=1}^{n} [Y_i - \theta'^T \tilde{X}_i]^2 \cdot \omega_{h,i}$$

However, since we are regressing on $f(X_i)$ we have

$$\widehat{\theta} \in \arg\min_{\theta' \in \mathbb{R}^2} \sum_{i=1}^{n} [f(X_i) - \theta'^T \tilde{X}_i]^2 \cdot \omega_{h,i}$$

By the hint and the fact that $f(x)$ is affine, we have $f(X_i) = \theta^T \tilde{X}_i$.

$$\widehat{\theta} \in \arg\min_{\theta' \in \mathbb{R}^2} \sum_{i=1}^{n} \theta^T \tilde{X}_i - \theta'^T \tilde{X}_i]^2 \cdot \omega_{h,i}$$

It is obvious from this statement that $\theta' = \theta$ minimizes the problem. Thus, $\widehat{\theta} = \theta$ as desired. We conclude that $f_h(0) = \widehat{\theta}^T \tilde{x}_0 = \theta^T \tilde{x}_0 = f(0)$.

b) Since $f(x)$ is affine, we have that $f_h(0) = \sum_{i=1}^n \alpha_{h,i} \cdot f(X_i)$. However, $f(X_i) = \theta^T \tilde{x}_0$ and $f_h(0) = \widehat{\theta}^T \tilde{x}_0$ by definition. Since $\widehat{\theta} = \theta$ from the previous part, we have

$$f_h(0) = \sum_{i=1}^{n} \alpha_{h,i} \theta^T \tilde{x}_0 = \theta^T \tilde{x}_0$$

$$\implies \sum_{i=1}^{n} \alpha_{h,i} = 1$$

4)

a) We wish to upper bound $\sum_{i=1}^{n}|\alpha_{h,i}|$. Begin with the definition of

$$\sum_{i=1}^{n}|\alpha_{h,i}| = \sum_{i=1}^{n}|w_{h,i}\tilde{x}_0^T B_h^{-1}\tilde{X}_i|$$

Since $n_h > 0$, then $w_{h,i} = \frac{1}{n_h}1\{|X_i| \leq h\}$. We want to upper bound this expression, in the worse case $X_i = h$ for all i that satisfy the indicator. So we have $w_{h,i} = \frac{1}{n_h}$ for all i.

$$\leq \sum_{i=1}^{n_h}|\frac{1}{n_h}\tilde{x}_0^T B_h^{-1}\tilde{X}_i|$$

Now, we know that $\tilde{x}_0^T = (1,0)$ and $\tilde{X}_i = (1, X_i/h)^T = (1,1)$ since we upper bounded X_i by h. Multiplying the matrices yields

$$\begin{aligned} &\leq \frac{1}{n_h}\sum_{i=1}^{n_h}|B_{h_{11}}^{-1} + B_{h_{12}}^{-1}| \\ &\leq |B_{h_{11}}^{-1}| + |B_{h_{12}}^{-1}| \qquad (\star)(\star) \end{aligned}$$

If one calculates $B_h = \tilde{\mathbf{X}}^T W_h \tilde{\mathbf{X}}$, the result is

$$\begin{bmatrix} \sum_i w_{i,h} & \sum_i w_{i,h}x_i/h \\ \sum_i w_{i,h}x_i/h & \sum_i w_{i,h}x_i^2/h^2 \end{bmatrix}$$

Thus, after inversion, the top row is the smallest row sum. It is a well known result that the minimal row sum is a lower bound of $\lambda_{\max}$. Thus,

$$\leq \lambda_{\max}(B_h^{-1})$$

Since $B \succ 0$ then all the eigenvalues are positive and $\lambda_{\max}(B_h^{-1}) = 1/\lambda_{\min}(B_h)$. Which is lesser than λ_0 by assumption. We finish it off with

$$\leq \lambda_{\max}(B_h^{-1}) = 1/\lambda_{\min}(B_h) \leq C(\lambda_0)$$

For $\max_i|\alpha_{h,i}|$ one can use all the steps but simply use max instead of summing across all i. Without the sum, $\frac{1}{n_h}$ does not disappear at $(\star)(\star)$ and so we are left with the result $\max_i|\alpha_{h,i}| \leq C(\lambda_0)/n_h$.

b) We expand the variance

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[(f_h(0) - \tilde{f}_h(0))^2]$$
$$= \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[f_h(0)^2] - \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[f_h(0)]^2$$

To upper bound, we ignore the second term since it is always negative. So

$$\leq \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[f_h(0)^2] = \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[(\sum_{i=1}^{n} \alpha_{h,i} Y_i)^2]$$

We now note that the Y_i's are independent and so the covariances are 0.

$$= \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\sum_{i=1}^{n} \alpha_{h,i}^2 Y_i^2]$$

Now, $Y_i^2 = 1$ for all i. We also now upper bound each $\alpha_{h,i}$ by $\max_i \alpha_{h,i}$. We change the sum from n to n_h to indicate that not all $\alpha_{h,i} \neq 0$.

$$\leq \mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\sum_{i=1}^{n_h} \max_i \alpha_{h,i}^2]$$

Using the bound in the previous part

$$\leq n_h \frac{C(\lambda_0)^2}{n_h^2} = \frac{C(\lambda_0)^2}{n_h}$$

5) We rewrite the bias as

$$|\tilde{f}_h(0) - f(0)| = |\sum_{i=1}^{n} \alpha_{h,i} f(X_i) - f(0)|$$

Approximating $f(X_i)$ using it's Taylor series at 0 results in

$$\leq |\sum_{i=1}^{n} \alpha_{h,i}(f(0) + X_i f'(0) + R_1(X_i)) - f(0)|$$

Where $R_1(X_i) = f''(z)x^2/2$ is the Lagrange remainder term. Applying triangle inequality

$$\leq |\sum_{i=1}^{n} \alpha_{h,i} f(0)| + |\sum_{i=1}^{n} \alpha_{h,i} X_i f'(0)| + |\sum_{i=1}^{n} \alpha_{h,i} R_1(X_i))| + |f(0)|$$

Now we bound each term applying the previous results

$$\begin{aligned}
&|\sum_{i=1}^{n} \alpha_{h,i} f(0)| \leq C(\lambda_0) \cdot \|f\|_\infty \\
&|\sum_{i=1}^{n} \alpha_{h,i} X_i f'(0)| \leq C(\lambda_0) \cdot \|f'\|_\infty \cdot h \\
&|\sum_{i=1}^{n} \alpha_{h,i} R_1(X_i))| \leq C(\lambda_0) \cdot \|f''\|_\infty \cdot h^2/2 \\
&|f(0)| \leq \|f\|_\infty
\end{aligned}$$

Thus, we can encapsulate all the other terms in some constant $C'(\lambda_0)$ and get the result

$$|\tilde{f}_h(0) - f(0)| \leq C'(\lambda_0) \cdot \|f''\|_\infty \cdot h^2$$

6)

a) Let $h \in (0, 1)$ be fixed. Then, since $X_i \sim \text{Uniform}[-1, 1]$, each X_i has probability $2h/2 = h$ to fall in $[-h, h]$. By the law of large numbers, we get that $n_h/n \xrightarrow{p} h$. Then, since $\lim_{n\to\infty} n_h \neq 0$, we have that

$$P\left(\frac{1}{n_h} \leq \frac{1}{n \cdot h}\right) \xrightarrow{n\to\infty} 1$$

Thus taking the limit,

$$\begin{aligned}
\lim_{n\to\infty} P\left(\frac{1}{n_h} \leq \frac{1}{n \cdot h}\right) &= \lim_{n\to\infty} P\left(\frac{1}{n_h} \leq 0\right) \\
&= \lim_{n\to\infty} P\left(\frac{1}{n_h} \leq 0\right) = 1 \\
&\implies \lim_{n\to\infty} P\left(n_h > 0\right) = 1
\end{aligned}$$

b) From the previous part, we showed that

$$1_{\{n_h>0\}} \xrightarrow{n\to\infty} 1 \text{ and}$$
$$1_{\{n_h=0\}} \xrightarrow{n\to\infty} 0$$

Since $P(n_h > 0) = \mathbb{E}[1_{\{n_h>0\}}]$. Thus, the only way for the left hand side to approach 1 is if the indicator approaches 1. Now, using the hint, this means that

$$\lim_{n\to\infty} B_h = \lim_{n\to\infty} B_{h,1}1\{n_h > 0\} + B_{h,2}1\{n_h = 0\} = B_{h,1}$$

It is given that

$$B_{h,1} = \begin{bmatrix} 1 & 0 \\ 0 & 1/3 \end{bmatrix}$$

From this, it is obvious that the smallest eigenvalue is 1/3. We put everything together to conclude

$$P(\lambda_{\min}(B_h) \geq 1/3) \xrightarrow{n\to\infty} 1$$

Where $\lambda_0 = 1/3$.

c) We begin by using the inequalities shown in Problem 3).

$$\mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[\ell_{0,1}[g_h(0)] - \ell_{0,1}[g(0)]\right]$$

Applying Problem 3) part 3)

$$\leq \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[|f_h(0) - f(0)|\right]$$

Adding in additional terms

$$\leq \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[|f_h(0) - \tilde{f}_h(0) + \tilde{f}_h(0) - f(0)|\right]$$

Applying Problem 3) part 2)

$$\leq \mathbb{E}_{\mathbf{X},\mathbf{Y}}\sqrt{\ell_2[f_h(0) - \tilde{f}_h(0)] - \ell_2[\tilde{f}_h(0) - f(0)]}$$

Change the subtraction to an addition

$$\leq \mathbb{E}_{\mathbf{X},\mathbf{Y}}\sqrt{\ell_2[f_h(0)-\tilde{f}_h(0)]+\ell_2[\tilde{f}_h(0)-f(0)]}$$

$$= \mathbb{E}_{\mathbf{X},\mathbf{Y}}\sqrt{\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[f_h(0)-\tilde{f}_h(0)-Y]^2+\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\tilde{f}_h(0)-f(0)-Y]^2}$$

Since $|Y|$ is dominated by 1, then there exists a constant C such that

$$\leq C\mathbb{E}_{\mathbf{X},\mathbf{Y}}\sqrt{\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[f_h(0)-\tilde{f}_h(0)]^2+\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\tilde{f}_h(0)-f(0)]^2}$$

We bound the first term by the results of part 4) and the second term by the results of part 5)

$$\leq C\mathbb{E}_{\mathbf{X},\mathbf{Y}}\sqrt{\frac{C'(\lambda_0)}{n_h}+C''(\lambda_0)^2\cdot\|f''\|_\infty^2\cdot h^4}$$

We can also apply the results of part 6) to bound $1/n_h$ by $1/(n\cdot h)$. Collecting all the constants and dropping the expectation as there are no more random variables yields

$$\leq C^*\sqrt{\frac{1}{n\cdot h}+h^4}$$

Which is the result desired.