CHAPTER 3

Misspecified Linear Models

Arguably, the strongest assumption that we made in Chapter 2 is that the regression function $f(x)$ is of the form $f(x) = x^\top \theta^*$. What if this assumption is violated? In reality, we do not really believe in the linear model and we hope that good statistical methods should be *robust* to deviations from this model. This is the problem of model misspecified linear models.

Throughout this chapter, we assume the following model:

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \dots, n\,, \tag{3.1}$$

where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$ is sub-Gaussian with variance proxy σ^2. Here $X_i \in \mathbb{R}^d$. When dealing with fixed design, it will be convenient to consider the vector $g \in \mathbb{R}^n$ defined for any function $g : \mathbb{R}^d \to \mathbb{R}$ by $g = (g(X_1), \dots, g(X_n))^\top$. In this case, we can write for any estimator $\hat{f} \in \mathbb{R}^n$ of f,

$$\mathsf{MSE}(\hat{f}) = \frac{1}{n}|\hat{f} - f|_2^2\,.$$

Even though the model may not be linear, we are interested in studying the statistical properties of various linear estimators introduced in the previous chapters: $\hat{\theta}^{\text{LS}}, \hat{\theta}^{\text{LS}}_K, \hat{\theta}^{\text{LS}}_X, \hat{\theta}^{\text{BIC}}, \hat{\theta}^{\mathcal{L}}$. Clearly, even with an infinite number of observations, we have no chance of finding a consistent estimator of f if we don't know the correct model. Nevertheless, as we will see in this chapter something can still be said about these estimators using *oracle inequalities*.

3.1 ORACLE INEQUALITIES

Oracle inequalities

As mentioned in the introduction, an oracle is a quantity that cannot be constructed without the knowledge of the quantity of interest, here: the regression function. Unlike the regression function itself, an oracle is constrained to take a specific form. For all matter of purposes, an oracle can be viewed as an estimator (in a given family) that can be constructed with an infinite amount of data. This is exactly what we should aim for in misspecified models.

When employing the least squares estimator $\hat{\theta}^{\text{LS}}$, we constrain ourselves to estimating functions that are of the form $x \mapsto x^\top \theta$, even though f itself may not be of this form. Therefore, the oracle $\hat{f}$ is the linear function that is the closest to f.

Rather than trying to approximate f by a linear function $f(x) \approx \theta^\top x$, we make the model a bit more general and consider a dictionary $\mathcal{H} = \{\varphi_1, \dots, \varphi_M\}$ of functions where $\varphi_j : \mathbb{R}^d \to \mathbb{R}$. In the case, we can actually remove the assumption that $X \in \mathbb{R}^d$. Indeed, the goal is now to estimate f using a linear combination of the functions in the dictionary:

$$f \approx \varphi_\theta := \sum_{j=1}^{M} \theta_j \varphi_j \,.$$

Remark 3.1. If $M = d$ and $\varphi_j(X) = X^{(j)}$ returns the jth coordinate of $X \in \mathbb{R}^d$ then the goal is to approximate $f(x)$ by $\theta^\top x$. Nevertheless, the use of a dictionary allows for a much more general framework.

Note that the use of a dictionary does not affect the methods that we have been using so far, namely penalized/constrained least squares. We use the same notation as before and define

1. The least squares estimator:

$$\hat{\theta}^{\text{LS}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^M} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \varphi_\theta(X_i)\big)^2 \tag{3.2}$$

2. The least squares estimator constrained to $K \subset \mathbb{R}^M$:

$$\hat{\theta}^{\text{LS}}_K \in \operatorname*{argmin}_{\theta \in K} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \varphi_\theta(X_i)\big)^2$$

3. The BIC estimator:

$$\hat{\theta}^{\text{BIC}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^M} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \varphi_\theta(X_i)\big)^2 + \tau^2 |\theta|_0 \Big\} \tag{3.3}$$

4. The Lasso estimator:

$$\hat{\theta}^{\mathcal{L}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^M} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \varphi_\theta(X_i)\big)^2 + 2\tau |\theta|_1 \Big\} \tag{3.4}$$

Definition 3.2. Let $R(\cdot)$ be a risk function and let $\mathcal{H} = \{\varphi_1, \dots, \varphi_M\}$ be a dictionary of functions from $\mathbb{R}^d$ to $\mathbb{R}$. Let K be a subset of $\mathbb{R}^M$. The *oracle* on K with respect to R is defined by $\varphi_{\bar{\theta}}$, where $\bar{\theta} \in K$ is such that

$$R(\varphi_{\bar{\theta}}) \leq R(\varphi_\theta)\,, \qquad \forall\, \theta \in K\,.$$

Moreover, $R_K = R(\varphi_{\bar{\theta}})$ is called *oracle risk* on K. An estimator $\hat{f}$ is said to satisfy an oracle inequality (over K) with remainder term ϕ in expectation (resp. with high probability) if there exists a constant $C \geq 1$ such that

$$\mathbb{E} R(\hat{f}) \leq C \inf_{\theta \in K} R(\varphi_\theta) + \phi_{n,M}(K)\,,$$

or

$$\mathbb{P}\big\{ R(\hat{f}) \leq C \inf_{\theta \in K} R(\varphi_\theta) + \phi_{n,M,\delta}(K) \big\} \geq 1 - \delta\,, \qquad \forall\ \delta > 0$$

respectively. If $C = 1$, the oracle inequality is sometimes called *exact*.

Our goal will be to mimic oracles. The finite sample performance of an estimator at this task is captured by an oracle inequality.

Oracle inequality for the least squares estimator

While our ultimate goal is to prove sparse oracle inequalities for the BIC and Lasso estimator in the case of misspecified model, the difficulty of the extension to this case for linear models, is essentially already captured for the least squares estimator. In this simple case, can even obtain an exact oracle inequality.

Theorem 3.3. *Assume the general regression model* (3.1) *with* $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. *Then, the least squares estimator* $\hat{\theta}^{\text{LS}}$ *satisfies for some numerical constant* $C > 0$,

$$\mathsf{MSE}(\varphi_{\hat{\theta}^{\text{LS}}}) \leq \inf_{\theta \in \mathbb{R}^M} \mathsf{MSE}(\varphi_\theta) + C \frac{\sigma^2 M}{n} \log(1/\delta)$$

with probability at least $1 - \delta$.

Proof. Note that by definition

$$|Y - \varphi_{\hat{\theta}^{\text{LS}}}|_2^2 \leq |Y - \varphi_{\bar{\theta}}|_2^2$$

where $\varphi_{\bar{\theta}}$ denotes the orthogonal projection of f onto the linear spam of $\varphi_1, \dots, \varphi_n$. Since $Y = f + \varepsilon$, we get

$$|f - \varphi_{\hat{\theta}^{\text{LS}}}|_2^2 \leq |f - \varphi_{\bar{\theta}}|_2^2 + 2\varepsilon^\top (\varphi_{\hat{\theta}^{\text{LS}}} - \varphi_{\bar{\theta}})$$

Moreover, by Pythagoras's theorem, we have

$$|f - \varphi_{\hat{\theta}^{\text{LS}}}|_2^2 - |f - \varphi_{\bar{\theta}}|_2^2 = |\varphi_{\hat{\theta}^{\text{LS}}} - \varphi_{\bar{\theta}}|_2^2 \,.$$

It yields

$$|\varphi_{\hat{\theta}^{\text{LS}}} - \varphi_{\bar{\theta}}|_2^2 \leq 2\varepsilon^\top (\varphi_{\hat{\theta}^{\text{LS}}} - \varphi_{\bar{\theta}}) \,.$$

Using the same steps as the ones following equation (2.5) for the well specified case, we get

$$|\varphi_{\hat{\theta}^{\text{LS}}} - \varphi_{\bar{\theta}}|_2^2 \lesssim \frac{\sigma^2 M}{n} \log(1/\delta)$$

with probability $1 - \delta$. The result of the lemma follows. □

Sparse oracle inequality for the BIC estimator

The techniques that we have developed for the linear model above also allows to derive oracle inequalities.

Theorem 3.4. *Assume the general regression model (3.1) with $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Then, the BIC estimator $\hat{\theta}^{\text{BIC}}$ with regularization parameter*

$$\tau^2 = \frac{16\sigma^2}{\alpha n} \log(6eM) \,, \alpha \in (0,1) \tag{3.5}$$

satisfies for some numerical constant $C > 0$,

$$\mathsf{MSE}(\varphi_{\hat{\theta}^{\text{BIC}}}) \leq \inf_{\theta \in \mathbb{R}^M} \Big\{ \frac{1+\alpha}{1-\alpha} \mathsf{MSE}(\varphi_\theta) + \frac{C\sigma^2}{\alpha(1-\alpha)n} |\theta|_0 \log(eM) \Big\} + \frac{C\sigma^2}{\alpha(1-\alpha)n} \log(1/\delta)$$

with probability at least $1 - \delta$.

Proof. Recall the the proof of Theorem 2.14 for the BIC estimator begins as follows:

$$\frac{1}{n}|Y - \varphi_{\hat{\theta}^{\text{BIC}}}|_2^2 + \tau^2 |\hat{\theta}^{\text{BIC}}|_0 \leq \frac{1}{n}|Y - \varphi_\theta|_2^2 + \tau^2 |\theta|_0 \,.$$

This is true for any $\theta \in \mathbb{R}^M$. It implies

$$|f - \varphi_{\hat{\theta}^{\text{BIC}}}|_2^2 + n\tau^2 |\hat{\theta}^{\text{BIC}}|_0 \leq |f - \varphi_\theta|_2^2 + 2\varepsilon^\top (\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta) + n\tau^2 |\theta|_0 \,.$$

Note that if $\hat{\theta}^{\text{BIC}} = \theta$, the result is trivial. Otherwise,

$$\begin{aligned} 2\varepsilon^\top (\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta) &= 2\varepsilon^\top \Big(\frac{\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta}{|\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta|_2} \Big) |\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta|_2 \\ &\leq \frac{2}{\alpha} \Big[\varepsilon^\top \Big(\frac{\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta}{|\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta|_2} \Big) \Big]^2 + \frac{\alpha}{2} |\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta|_2^2 \,, \end{aligned}$$

where we use Young's inequality $2ab \leq \frac{2}{\alpha}a^2 + \frac{\alpha}{2}b^2$ valif for $a, b \geq 0$, $\alpha > 0$. Next, since

$$\frac{\alpha}{2}|\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta|_2^2 \leq \alpha|\varphi_{\hat{\theta}^{\text{BIC}}} - f|_2^2 + \alpha|\varphi_\theta - f|_2^2\,,$$

we get for $\alpha < 1$,

$$\begin{aligned}(1-\alpha)|\varphi_{\hat{\theta}^{\text{BIC}}} - f|_2^2 &\leq (1+\alpha)|\varphi_\theta - f|_2^2 + n\tau^2|\theta|_0 \\ &\quad + \frac{2}{\alpha}\left[\varepsilon^\top \mathcal{U}(\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta)\right]^2 - n\tau^2|\hat{\theta}^{\text{BIC}}|_0 \\ &\leq (1+\alpha)|\varphi_\theta - f|_2^2 + 2n\tau^2|\theta|_0 \\ &\quad + \frac{2}{\alpha}\left[\varepsilon^\top \mathcal{U}(\varphi_{\hat{\theta}^{\text{BIC}}-\theta})\right]^2 - n\tau^2|\hat{\theta}^{\text{BIC}} - \theta|_0\end{aligned}$$

We conclude as in the proof of Theorem 2.14. □

A similar oracle can be obtained in expectation (exercise).

The interpretation of this theorem is enlightening. It implies that the BIC estimator will mimic the best tradeoff between the approximation error $\mathsf{MSE}(\varphi_\theta)$ and the complexity of θ as measured by its sparsity. In particular this result, sometimes called *sparse oracle inequality* implies the following oracle inequality. Define the oracle $\bar{\theta}$ to be such that

$$\mathsf{MSE}(\varphi_{\bar{\theta}}) = \min_{\theta \in \mathbb{R}^M} \mathsf{MSE}(\varphi_\theta)$$

then, with probability at least $1 - \delta$,

$$\mathsf{MSE}(\varphi_{\hat{\theta}^{\text{BIC}}}) \leq \frac{1+\alpha}{1-\alpha}\mathsf{MSE}(\varphi_{\bar{\theta}}) + \frac{C\sigma^2}{\alpha(1-\alpha)n}\left[|\bar{\theta}|_0 \log(eM) + \log(1/\delta)\right]\Big\}$$

If the linear model happens to be correct, then, simply, $\mathsf{MSE}(\varphi_{\bar{\theta}}) = 0$.

Sparse oracle inequality for the Lasso

To prove an oracle inequality for the Lasso, we need incoherence on the design. Here the design matrix is given by the $n \times M$ matrix Φ with elements $\Phi_{i,j} = \varphi_j(X_i)$.

Theorem 3.5. *Assume the general regression model (3.1) with $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Moreover, assume that there exists an integer k such that the matrix Φ satisfies assumption $\mathsf{INC}(k)$ holds. Then, the Lasso estimator $\hat{\theta}^{\mathcal{L}}$ with regularization parameter given by*

$$2\tau = 8\sigma\sqrt{\frac{2\log(2M)}{n}} + 8\sigma\sqrt{\frac{2\log(1/\delta)}{n}} \tag{3.6}$$

satisfies for some numerical constant C,

$$\mathsf{MSE}(\varphi_{\hat\theta^{\mathcal{L}}}) \le \inf_{\substack{\theta\in\mathbb{R}^M\\ |\theta|_0\le k}} \Big\{ \frac{1+\alpha}{1-\alpha}\mathsf{MSE}(\varphi_\theta) + \frac{C\sigma^2}{\alpha(1-\alpha)n}|\theta|_0\log(eM)\Big\} + \frac{C\sigma^2}{\alpha(1-\alpha)n}\log(1/\delta)$$

with probability at least $1-\delta$.

Proof. From the definition of $\hat\theta^{\mathcal{L}}$, it holds for any $\theta\in\mathbb{R}^M$,

$$\frac{1}{n}|Y-\varphi_{\hat\theta^{\mathcal{L}}}|_2^2 \le \frac{1}{n}|Y-\varphi_\theta|_2^2 + 2\tau|\theta|_1 - 2\tau|\hat\theta^{\mathcal{L}}|_1\,.$$

Adding $\tau|\hat\theta^{\mathcal{L}}-\theta|_1$ on each side and multiplying by n, we get

$$|\varphi_{\hat\theta^{\mathcal{L}}}-f|_2^2 - |\varphi_\theta - f|_2^2 + n\tau|\hat\theta^{\mathcal{L}}-\theta|_1 \le 2\varepsilon^\top(\varphi_{\hat\theta^{\mathcal{L}}}-\varphi_\theta) + n\tau|\hat\theta^{\mathcal{L}}-\theta|_1 + 2n\tau|\theta|_1 - 2n\tau|\hat\theta^{\mathcal{L}}|_1\,. \tag{3.7}$$

Next, note that $\mathsf{INC}(k)$ for any $k\ge 1$ implies that $|\varphi_j|_2\le 2\sqrt{n}$ for all $j=1,\dots,M$. Applying Hölder's inequality using the same steps as in the proof of Theorem 2.15, we get that with probability $1-\delta$, it holds

$$2\varepsilon^\top(\varphi_{\hat\theta^{\mathcal{L}}}-\varphi_\theta) \le \frac{n\tau}{2}|\hat\theta^{\mathcal{L}}-\theta|_1$$

Therefore, taking $S=\operatorname{supp}(\theta)$ to be the support of θ, we get that the right-hand side of (3.7) is bounded by

$$\begin{aligned} &\le 2n\tau|\hat\theta^{\mathcal{L}}-\theta|_1 + 2n\tau|\theta|_1 - 2n\tau|\hat\theta^{\mathcal{L}}|_1\\ &= 2n\tau|\hat\theta^{\mathcal{L}}_S-\theta|_1 + 2n\tau|\theta|_1 - 2n\tau|\hat\theta^{\mathcal{L}}_S|_1\\ &\le 4n\tau|\hat\theta^{\mathcal{L}}_S-\theta|_1 \end{aligned} \tag{3.8}$$

with probability $1-\delta$.

It implies that either $\mathsf{MSE}(\varphi_{\hat\theta^{\mathcal{L}}})\le\mathsf{MSE}(\varphi_\theta)$ or that

$$|\hat\theta^{\mathcal{L}}_{S^c}-\theta_{S^c}|_1 \le 3|\hat\theta^{\mathcal{L}}_S-\theta_S|_1\,.$$

so that $\theta=\hat\theta^{\mathcal{L}}-\theta$ satisfies the cone condition (2.17). Using now the Cauchy-Schwarz inequality and Lemma 2.17 respectively, assume that $|\theta|_0\le k$, we get

$$4n\tau|\hat\theta^{\mathcal{L}}_S-\theta|_1 \le 4n\tau\sqrt{|S|}|\hat\theta^{\mathcal{L}}_S-\theta|_2 \le 4\tau\sqrt{2n|\theta|_0}|\varphi_{\hat\theta^{\mathcal{L}}}-\varphi_\theta|_2\,.$$

Using now the inequality $2ab\le\frac{2}{\alpha}a^2+\frac{\alpha}{2}b^2$, we get

$$\begin{aligned} 4n\tau|\hat\theta^{\mathcal{L}}_S-\theta|_1 &\le \frac{16\tau^2 n|\theta|_0}{\alpha} + \frac{\alpha}{2}|\varphi_{\hat\theta^{\mathcal{L}}}-\varphi_\theta|_2^2\\ &\le \frac{16\tau^2 n|\theta|_0}{\alpha} + \alpha|\varphi_{\hat\theta^{\mathcal{L}}}-f|_2^2 + \alpha|\varphi_\theta-f|_2^2 \end{aligned}$$

Combining this result with (3.7) and (3.8), we find

$$(1-\alpha)\mathsf{MSE}(\varphi_{\hat{\theta}^{\mathcal{L}}}) \le (1+\alpha)\mathsf{MSE}(\varphi_\theta) + \frac{16\tau^2|\theta|_0}{\alpha}\,.$$

To conclude the proof of the bound with high probability, it only remains to divide by $1-\alpha$ on both sides of the above inequality. The bound in expectation follows using the same argument as in the proof of Corollary 2.9. □

Maurey's argument

From the above section, it seems that the Lasso estimator is strictly better than the BIC estimator as long as incoherence holds. Indeed, if there is no sparse θ such that $\mathsf{MSE}(\varphi_\theta)$ is small, Theorem 3.4 is useless. In reality, no one really believes in the existence of sparse vectors but rater of approximately sparse vectors. Zipf's law would instead favor the existence of vectors θ with absolute coefficients that decay polynomially when ordered from largest to smallest in absolute value. This is the case for example if θ has a small ℓ_1 norm but is not sparse. For such θ, the Lasso estimator still enjoys slow rates as in Theorem 2.15, which can be easily extended to the misspecified case (see Problem 3.2). Fortunately, such vectors can be well approximated by sparse vectors in the following sense: for any vector $\theta \in \mathbb{R}^M$ such that $|\theta|_1 \le 1$, there exists a vector θ' that is sparse and for which $\mathsf{MSE}(\varphi_{\theta'})$ is not much larger than $\mathsf{MSE}(\varphi_\theta)$. The following theorem quantifies exactly the tradeoff between sparsity and MSE. It is often attributed to B. Maurey and was published by Pisier [Pis81]. This is why it is referred to as *Maurey's argument*.

Theorem 3.6. *Let $\{\varphi_1, \dots, \varphi_M\}$ be a dictionary normalized in such a way that*

$$\max_{1\le j\le M} |\varphi_j|_2 \le D\sqrt{n}\,.$$

Then for any integer k such that $1 \le k \le M$ and any positive R, we have

$$\min_{\substack{\theta\in\mathbb{R}^M \\ |\theta|_0\le 2k}} \mathsf{MSE}(\varphi_\theta) \le \min_{\substack{\theta\in\mathbb{R}^M \\ |\theta|_1\le R}} \mathsf{MSE}(\varphi_\theta) + \frac{D^2R^2}{k}\,.$$

Proof. Define

$$\bar{\theta} \in \operatorname*{argmin}_{\substack{\theta\in\mathbb{R}^M \\ |\theta|_1\le R}} |\varphi_\theta - f|_2^2$$

and assume without loss of generality that $|\bar{\theta}_1| \ge |\bar{\theta}_2| \ge \ldots \ge |\bar{\theta}_M|$.

Now decompose $\bar{\theta} = \theta^{(1)} + \theta^{(2)}$, where $\mathrm{supp}(\theta^{(1)}) \subset \{1 \ldots, k\}$ and $\mathrm{supp}(\theta^{(2)}) \subset \{k+1, \dots, M\}$. In particular it holds

$$\varphi_{\bar{\theta}} = \varphi_{\theta^{(1)}} + \varphi_{\theta^{(2)}}\,.$$

Moreover, observe that

$$|\theta^{(2)}|_1 = \sum_{j=k+1}^{M} |\bar{\theta}_j| \le R$$

Let now $U \in \mathbb{R}^n$ be a random vector with values in $\{0, \pm R\varphi_1, \dots, \pm R\varphi_M\}$ defined by

$$\mathbb{P}(U = R\mathrm{sign}(\theta_j^{(2)})\varphi_j) = \frac{|\theta_j^{(2)}|}{R}, \quad j = k+1, \dots, M$$
$$\mathbb{P}(U = 0) = 1 - \frac{|\theta^{(2)}|_1}{R}.$$

Note that $\mathbb{E}[U] = \varphi_{\theta^{(2)}}$ and $|U|_2 \le RD\sqrt{n}$. Let now $U_1, \dots, U_k$ be k independent copies of U define

$$\bar{U} = \frac{1}{k}\sum_{i=1}^{k} U_i .$$

Note that $\bar{U} = \varphi_{\tilde{\theta}}$ for some $\tilde{\theta} \in \mathbb{R}^M$ such that $|\tilde{\theta}|_0 \le k$. Therefore, $|\theta^{(1)} + \tilde{\theta}|_0 \le 2k$ and

$$\begin{aligned}
\mathbb{E}|f - \varphi_{\theta^{(1)}} - \bar{U}|_2^2 &= \mathbb{E}|f - \varphi_{\theta^{(1)}} - \varphi_{\theta^{(2)}} + \varphi_{\theta^{(2)}} - \bar{U}|_2^2 \\
&= \mathbb{E}|f - \varphi_{\theta^{(1)}} - \varphi_{\theta^{(2)}}|_2^2 + |\varphi_{\theta^{(2)}} - \bar{U}|_2^2 \\
&= |f - \varphi_{\tilde{\theta}}|_2^2 + \frac{\mathbb{E}|U - \mathbb{E}[U]|_2^2}{k} \\
&\le |f - \varphi_{\tilde{\theta}}|_2^2 + \frac{(RD\sqrt{n})^2}{k}
\end{aligned}$$

To conclude the proof, note that

$$\mathbb{E}|f - \varphi_{\theta^{(1)}} - \bar{U}|_2^2 = \mathbb{E}|f - \varphi_{\theta^{(1)}+\tilde{\theta}}|_2^2 \ge \min_{\substack{\theta \in \mathbb{R}^M \\ |\theta|_0 \le 2k}} |f - \varphi_\theta|_2^2$$

and to divide by n. □

Maurey's argument implies the following corollary.

Corollary 3.7. *Assume that the assumptions of Theorem 3.4 hold and that the dictionary $\{\varphi_1, \dots, \varphi_M\}$ is normalized in such a way that*

$$\max_{1 \le j \le M} |\varphi_j|_2 \le \sqrt{n}\,.$$

Then there exists a constant $C > 0$ such that the BIC estimator satisfies

$$\mathsf{MSE}(\varphi_{\hat{\theta}^{\mathrm{BIC}}}) \le \inf_{\theta \in \mathbb{R}^M} \Big\{ 2\mathsf{MSE}(\varphi_\theta) + C\Big[\frac{\sigma^2 |\theta|_0 \log(eM)}{n} \wedge \sigma|\theta|_1 \sqrt{\frac{\log(eM)}{n}}\Big]\Big\} + C\frac{\sigma^2 \log(1/\delta)}{n}$$

with probability at least $1 - \delta$.

Proof. Choosing $\alpha = 1/3$ in Theorem 3.4 yields

$$\mathsf{MSE}(\varphi_{\hat{\theta}^{\text{BIC}}}) \le 2 \inf_{\theta \in \mathbb{R}^M} \left\{ \mathsf{MSE}(\varphi_\theta) + C \frac{\sigma^2 |\theta|_0 \log(eM)}{n} \right\} + C \frac{\sigma^2 \log(1/\delta)}{n}$$

For any $\theta' \in \mathbb{R}^M$, it follows from Maurey's argument that there exist $\theta \in \mathbb{R}^M$ such that $|\theta|_0 \le 2|\theta'|_0$ and

$$\mathsf{MSE}(\varphi_\theta) \le \mathsf{MSE}(\varphi_{\theta'}) + \frac{2|\theta'|_1^2}{|\theta|_0}$$

It implies that

$$\mathsf{MSE}(\varphi_\theta) + C \frac{\sigma^2 |\theta|_0 \log(eM)}{n} \le \mathsf{MSE}(\varphi_{\theta'}) + \frac{2|\theta'|_1^2}{|\theta|_0} + C \frac{\sigma^2 |\theta|_0 \log(eM)}{n}$$

Taking infimum on both sides, we get

$$\begin{aligned} &\inf_{\theta \in \mathbb{R}^M} \left\{ \mathsf{MSE}(\varphi_\theta) + C \frac{\sigma^2 |\theta|_0 \log(eM)}{n} \right\} \\ &\qquad \le \inf_{\theta' \in \mathbb{R}^M} \left\{ \mathsf{MSE}(\varphi_{\theta'}) + C \min_k \left(\frac{|\theta'|_1^2}{k} + C \frac{\sigma^2 k \log(eM)}{n} \right) \right\}. \end{aligned}$$

To control the minimum over k, we need to consider three cases for the quantity

$$\bar{k} = \frac{|\theta'|_1}{\sigma} \sqrt{\frac{\log M}{n}}$$

1. If $1 \le \bar{k} \le M$, then we get

$$\min_k \left(\frac{|\theta'|_1^2}{k} + C \frac{\sigma^2 k \log(eM)}{n} \right) \le C\sigma |\theta'|_1 \sqrt{\frac{\log(eM)}{n}}$$

2. If $\bar{k} \le 1$, then

$$|\theta'|_1^2 \le C \frac{\sigma^2 \log(eM)}{n},$$

which yields

$$\min_k \left(\frac{|\theta'|_1^2}{k} + C \frac{\sigma^2 k \log(eM)}{n} \right) \le C \frac{\sigma^2 \log(eM)}{n}$$

3. If $\bar{k} \ge M$, then

$$\frac{\sigma^2 M \log(eM)}{n} \le C \frac{|\theta'|_1^2}{M}.$$

Therefore, on the one hand, if $M \ge \frac{|\theta|_1}{\sigma \sqrt{\log(eM)/n}}$, we get

$$\min_k \left(\frac{|\theta'|_1^2}{k} + C \frac{\sigma^2 k \log(eM)}{n} \right) \le C \frac{|\theta'|_1^2}{M} \le C\sigma |\theta'|_1 \sqrt{\frac{\log(eM)}{n}}.$$

On the other hand, if $M \leq \frac{|\theta|_1}{\sigma\sqrt{\log(eM)/n}}$, then for any $\Theta \in \mathbb{R}^M$, we have

$$\frac{\sigma^2|\theta|_0 \log(eM)}{n} \leq \frac{\sigma^2 M \log(eM)}{n} \leq C\sigma|\theta'|_1\sqrt{\frac{\log(eM)}{n}}\,.$$

□

Note that this last result holds for any estimator that satisfies an oracle inequality with respect to the ℓ_0 norm such as the result of Theorem 3.4. In particular, this estimator need not be the BIC estimator. An example is the Exponential Screening estimator of [RT11].

Maurey's argument allows us to enjoy the best of both the ℓ_0 and the ℓ_1 world. The rate adapts to the sparsity of the problem and can be even generalized to ℓ_q-sparsity (see Problem 3.3). However, it is clear from the proof that this argument is limited to squared ℓ_2 norms such as the one appearing in MSE and extension to other risk measures is non trivial. Some work has been done for non Hilbert spaces [Pis81, DDGS97] using more sophisticated arguments.

3.2 NONPARAMETRIC REGRESSION

So far, the oracle inequalities that we have derived do not deal with the approximation error $\mathsf{MSE}(\varphi_\theta)$. We kept it arbitrary and simply hoped that it was small. Note also that in the case of linear models, we simply assumed that the approximation error was zero. As we will see in this section, this error can be quantified under natural smoothness conditions if the dictionary of functions $\mathcal{H} = \{\varphi_1, \ldots, \varphi_M\}$ is chosen appropriately. In what follows, we assume for simplicity that $d = 1$ so that $f : \mathbb{R} \to \mathbb{R}$ and $\varphi_j : \mathbb{R} \to \mathbb{R}$.

Fourier decomposition

Historically, nonparametric estimation was developed before high-dimensional statistics and most results hold for the case where the dictionary $\mathcal{H} = \{\varphi_1, \ldots, \varphi_M\}$ forms an orthonormal system of $L_2([0,1])$:

$$\int_0^1 \varphi_j^2(x)\mathrm{d}x = 1\,, \quad \int_0^1 \varphi_j(x)\varphi_k(x)\mathrm{d}x = 0,\ \forall\, j \neq k\,.$$

We will also deal with the case where $M = \infty$.

When $\mathcal{H}$ is an orthonormal system, the coefficients $\theta_j^* \in \mathbb{R}$ defined by

$$\theta_j^* = \int_0^1 f(x)\varphi_j(x)\mathrm{d}x\,,$$

are called *Fourier coefficients* of f.

Assume now that the regression function f admits the following decomposition

$$f = \sum_{j=1}^{\infty} \theta_j^* \varphi_j \,.$$

There exists many choices for the orthonormal system and we give only two as examples.

Example 3.8. *Trigonometric basis.* This is an orthonormal basis of $L_2([0,1])$. It is defined by

$$\begin{aligned} \varphi_1 &\equiv 1 \\ \varphi_{2k}(x) &= \sqrt{2}\cos(2\pi k x)\,, \\ \varphi_{2k+1}(x) &= \sqrt{2}\sin(2\pi k x)\,, \end{aligned}$$

for $k = 1, 2, \ldots$ and $x \in [0,1]$. The fact that it is indeed an orthonormal system can be easily check using trigonometric identities.

The next example has received a lot of attention in the signal (sound, image, ...) processing community.

Example 3.9. *Wavelets.* Let $\psi : \mathbb{R} \to \mathbb{R}$ be a sufficiently smooth and compactly supported function, called "*mother wavelet*". Define the system of functions

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)\,, \quad j,k \in \mathbb{Z}\,.$$

It can be shown that for a suitable ψ, the dictionary $\{\psi_{j,k}, j,k \in \mathbb{Z}\}$ forms an orthonormal system of $L_2([0,1])$ and sometimes a basis. In the latter case, for any function $g \in L_2([0,1])$, it holds

$$g = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \theta_{jk}\psi_{jk}\,, \quad \theta_{jk} = \int_0^1 g(x)\psi_{jk}(x)\mathrm{d}x\,.$$

The coefficients θ_{jk} are called *wavelet coefficients* of g.

The simplest example is given by the *Haar system* obtained by taking ψ to be the following piecewise constant function (see Figure 3.1). We will not give more details about wavelets here but refer simply point the interested reader to [Mal09].

$$\psi(x) = \begin{cases} 1 & 0 \le x < 1/2 \\ -1 & 1/2 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Sobolev classes and ellipsoids

We begin by describing a class of smooth functions where smoothness is understood in terms of its number of derivatives. Recall that $f^{(k)}$ denotes the k-th derivative of f.

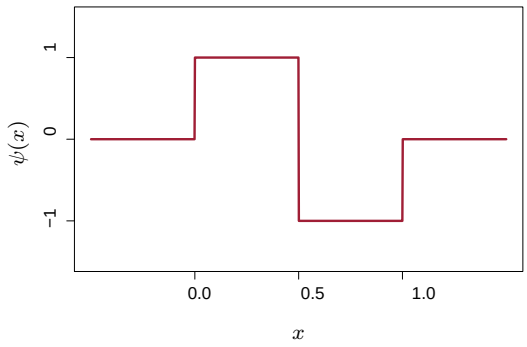

Figure 3.1. The Haar mother wavelet

Definition 3.10. Fix parameters $\beta \in \{1, 2, \dots\}$ and $L > 0$. The Sobolev class of functions $W(\beta, L)$ is defined by

$$W(\beta, L) = \Big\{ f : [0,1] \to \mathbb{R} \ : \ f \in L_2([0,1]), f^{(\beta-1)} \text{ is absolutely continuous and } \int_0^1 [f^{(\beta)}]^2 \le L^2, f^{(j)}(0) = f^{(j)}(1), j = 0, \dots, \beta - 1 \Big\}$$

Any function $f \in W(\beta, L)$ can represented[1] as its Fourier expansion along the trigonometric basis:

$$f(x) = \theta_1^* \varphi_1(x) + \sum_{k=1}^{\infty} \left(\theta_{2k}^* \varphi_{2k}(x) + \theta_{2k+1}^* \varphi_{2k+1}(x) \right), \quad \forall\, x \in [0,1],$$

where $\theta^* = \{\theta_j^*\}_{j \ge 1}$ is in the space of squared summable sequence $\ell_2(\mathbb{N})$ defined by

$$\ell_2(\mathbb{N}) = \Big\{ \theta \ : \ \sum_{j=1}^{\infty} \theta_j^2 < \infty \Big\}.$$

For any $\beta > 0$, define the coefficients

$$a_j = \begin{cases} j^\beta & \text{for } j \text{ even} \\ (j-1)^\beta & \text{for } j \text{ odd} \end{cases} \tag{3.9}$$

Thanks to these coefficients, we can define the Sobolev class of functions in terms of Fourier coefficients.

[1] In the sense that

$$\lim_{k \to \infty} \int_0^1 |f(t) - \sum_{j=1}^{k} \theta_j \varphi_j(t)|^2 \mathrm{d}t = 0$$

Theorem 3.11. *Fix $\beta \geq 1$ and $L > 0$ and let $\{\varphi_j\}_{j\geq 1}$ denote the trigonometric basis of $L_2([0,1])$. Moreover, let $\{a_j\}_{j\geq 1}$ be defined as in (3.9). A function $f \in W(\beta, L)$ can be represented as*

$$f = \sum_{j=1}^{\infty} \theta_j^* \varphi_j \,,$$

where the sequence $\{\theta_j^*\}_{j\geq 1}$ belongs to Sobolev ellipsoid of $\ell_2(\mathbb{N})$ defined by

$$\Theta(\beta, Q) = \Big\{ \theta \in \ell_2(\mathbb{N}) \,:\, \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq Q \Big\}$$

for $Q = L^2/\pi^{2\beta}$.

Proof. Let us first recall the definition of the Fourier coefficients $\{s_k(j)\}_{k\geq 1}$ of the jth derivative $f^{(j)}$ of f for $j = 1, \ldots, \beta$:

$$\begin{aligned}
s_1(j) &= \int_0^1 f^{(j)}(t)\mathrm{d}t = f^{(j-1)}(1) - f^{(j-1)}(0) = 0 \,, \\
s_{2k}(j) &= \sqrt{2} \int_0^1 f^{(j)}(t) \cos(2\pi k t)\mathrm{d}t \,, \\
s_{2k+1}(j) &= \sqrt{2} \int_0^1 f^{(j)}(t) \sin(2\pi k t)\mathrm{d}t \,,
\end{aligned}$$

The Fourier coefficients of f are given by $\theta_k = s_k(0)$.

Using integration by parts, we find that

$$\begin{aligned}
s_{2k}(\beta) &= \sqrt{2} f^{(\beta-1)}(t) \cos(2\pi k t)\Big|_0^1 + (2\pi k)\sqrt{2} \int_0^1 f^{(\beta-1)}(t) \sin(2\pi k t)\mathrm{d}t \\
&= \sqrt{2}[f^{(\beta-1)}(1) - f^{(\beta-1)}(0)] + (2\pi k)\sqrt{2} \int_0^1 f^{(\beta-1)}(t) \sin(2\pi k t)\mathrm{d}t \\
&= (2\pi k) s_{2k+1}(\beta - 1) \,.
\end{aligned}$$

Moreover,

$$\begin{aligned}
s_{2k+1}(\beta) &= \sqrt{2} f^{(\beta-1)}(t) \sin(2\pi k t)\Big|_0^1 - (2\pi k)\sqrt{2} \int_0^1 f^{(\beta-1)}(t) \cos(2\pi k t)\mathrm{d}t \\
&= -(2\pi k) s_{2k}(\beta - 1) \,.
\end{aligned}$$

In particular, it yields

$$s_{2k}(\beta)^2 + s_{2k+1}(\beta)^2 = (2\pi k)^2 \big[s_{2k}(\beta - 1)^2 + s_{2k+1}(\beta - 1)^2 \big]$$

By induction, we find that for any $k \geq 1$,

$$s_{2k}(\beta)^2 + s_{2k+1}(\beta)^2 = (2\pi k)^{2\beta} \big(\theta_{2k}^2 + \theta_{2k+1}^2 \big)$$

Next, it follows for the definition (3.9) of a_j that

$$\sum_{k=1}^{\infty}(2\pi k)^{2\beta}\big(\theta_{2k}^2+\theta_{2k+1}^2\big) = \pi^{2\beta}\sum_{k=1}^{\infty}a_{2k}^2\theta_{2k}^2 + \pi^{2\beta}\sum_{k=1}^{\infty}a_{2k+1}^2\theta_{2k+1}^2$$
$$= \pi^{2\beta}\sum_{j=1}^{\infty}a_j^2\theta_j^2\,.$$

Together with the Parseval identity, it yields

$$\int_0^1\big(f^{(\beta)}(t)\big)^2\mathrm{d}t = \sum_{k=1}^{\infty}s_{2k}(\beta)^2+s_{2k+1}(\beta)^2 = \pi^{2\beta}\sum_{j=1}^{\infty}a_j^2\theta_j^2\,.$$

To conclude, observe that since $f\in W(\beta,L)$, we have

$$\int_0^1\big(f^{(\beta)}(t)\big)^2\mathrm{d}t \le L^2\,,$$

so that $\theta\in\Theta(\beta,L^2/\pi^{2\beta})$. □

It can actually be shown that the reciprocal is true, that is any function with Fourier coefficients in $\Theta(\beta,Q)$ belongs to if $W(\beta,L)$ but we will not be needing this.

In what follows, we will define smooth functions as functions with Fourier coefficients (with respect to the trigonometric basis) in a Sobolev ellipsoid. By extension, we write $f\in\Theta(\beta,Q)$ in this case and consider any real value for β.

Proposition 3.12. The Sobolev ellipsoids enjoy the following properties

(i) For any $Q>0$,

$$0<\beta'<\beta \;\Rightarrow\; \Theta(\beta,Q)\subset\Theta(\beta',Q)$$

(ii) For any $Q>0$,

$$\beta>\frac{1}{2} \;\Rightarrow\; f \text{ is continuous}$$

The proof is left as an exercise (Problem 3.5)

It turns out that the first functions in the trigonometric basis are orthonormal with respect to the inner product of L_2 but also to the inner predictor associated to fixed design $\langle f,g\rangle := \frac{1}{n}f(X_i)g(X_i)$ when the design is chosen to be *regular*, i.e., $X_i=(i-1)/n$, $i=1,\dots,n$.

Lemma 3.13. *Assume that $\{X_1,\dots,X_n\}$ is the regular design, i.e., $X_i=(i-1)/n$. Then, for any $M\le n-1$, the design matrix $\Phi=\{\varphi_j(X_i)\}_{\substack{1\le i\le n\\ 1\le j\le M}}$ satisfies the* ORT *condition.*

Proof. Note first that for any $j, j' \in \{1, \dots, n-1\}$, $j \neq j'$ the inner product $\varphi_j^\top \varphi_{j'}$ is of the form

$$\varphi_j^\top \varphi_{j'} = 2 \sum_{s=0}^{n-1} u_j(2\pi k_j s/n) v_{j'}(2\pi k_{j'} s/n)$$

where $k_j = \lfloor j/2 \rfloor$ is the integer part of $j/2$ for any $x \in \mathbb{R}$, $u_j(x), v_{j'}(x) \in \{\mathfrak{Re}(e^{\mathrm{i}x}), \mathfrak{Im}(e^{\mathrm{i}x})\}$.

Next, observe that if $k_j \neq k_{j'}$, we have

$$\sum_{s=0}^{n-1} e^{\frac{\mathrm{i}2\pi k_j s}{n}} e^{-\frac{\mathrm{i}2\pi k_{j'} s}{n}} = \sum_{s=0}^{n-1} e^{\frac{\mathrm{i}2\pi (k_j - k_{j'}) s}{n}} = 0\,.$$

Moreover, if we define the vectors $a, b, a', b' \in \mathbb{R}^n$ with coordinates such that $e^{\frac{\mathrm{i}2\pi k_j s}{n}} = a_s + \mathrm{i}b_s$ and $e^{\frac{\mathrm{i}2\pi k_{j'} s}{n}} = a'_s + \mathrm{i}b'_s$, we get

$$\sum_{s=0}^{n-1} e^{\frac{\mathrm{i}2\pi k_j s}{n}} e^{-\frac{\mathrm{i}2\pi k_{j'} s}{n}} = (a + \mathrm{i}b)^\top (a' - \mathrm{i}b') = a^\top a' + b^\top b' + \mathrm{i}\big[b^\top a' - a^\top b'\big]$$

and consequently that

$$\frac{1}{2}\varphi_j^\top \varphi_{j'} = a^\top a' + b^\top b' + \mathrm{i}\big[b^\top a' - a^\top b'\big]$$

with $|a|_2 |b|_2 = |a'|_2 |b'|_2 = 0$, i.e., either $a = 0$ or $b = 0$ and either $a' = 0$ or $b' = 0$. Therefore, in the case where $k_j \neq k_{j'}$, we have

$$a^\top a' = -b^\top b' = 0, \qquad b^\top a' = a^\top b' = 0$$

which implies $\varphi_j^\top \varphi_{j'} = 0$. To conclude the proof, it remains to deal with the case where $k_j = k_{j'}$. This can happen in two cases: $|j' - j| = 1$ or $j' = j$. In the first case, we have that $\{u_j(x), v_{j'}(x)\} = \{\mathfrak{Re}(e^{\mathrm{i}x}), \mathfrak{Im}(e^{\mathrm{i}x})\}$, i.e., one is a $\sin(\cdot)$ and the other is a $\cos(\cdot)$. Therefore,

$$\frac{1}{2}\varphi_j^\top \varphi_{j'} = a^\top a' + b^\top b' + \mathrm{i}\big[b^\top a' - a^\top b'\big] = 0$$

The final case is $j = j'$ for which, on the one hand,

$$\sum_{s=0}^{n-1} e^{\frac{\mathrm{i}2\pi k_j s}{n}} e^{\frac{\mathrm{i}2\pi k_j s}{n}} = \sum_{s=0}^{n-1} e^{\frac{\mathrm{i}4\pi k_j s}{n}} = 0$$

and on the other hand

$$\sum_{s=0}^{n-1} e^{\frac{\mathrm{i}2\pi k_j s}{n}} e^{\frac{\mathrm{i}2\pi k_j s}{n}} = |a + \mathrm{i}b|_2^2 = |a|_2^2 - |b|_2^2$$

so that $|a|^2 = |b|^2$. Moreover, by definition,

$$|\varphi_j|_2^2 = \begin{cases} 2|a|_2^2 & \text{if } j \text{ is even} \\ 2|b|_2^2 & \text{if } j \text{ is odd} \end{cases}$$

so that

$$|\varphi_j|_2^2 = 2\frac{|a|_2^2 + |b|_2^2}{2} = \sum_{s=0}^{n-1} \left|e^{\frac{\mathrm{i}2\pi k_j s}{n}}\right|^2 = n$$

Therefore, the design matrix Φ is such that

$$\Phi^\top \Phi = nI_M \,.$$

□

Integrated squared error

As mentioned in the introduction of this chapter, the smoothness assumption allows us to control the approximation error. Before going into the details, let us gain some insight. Note first that if $\theta \in \Theta(\beta, Q)$, then $a_j^2\theta_j^2 \to 0$ as $j \to \infty$ so that $|\theta_j| = o(j^{-\beta})$. Therefore, the θ_js decay polynomially to zero and it makes sense to approximate f by its truncated Fourier series

$$\sum_{j=1}^{M} \theta_j^* \varphi_j =: \varphi_{\theta^*}^M$$

for any fixed M. This truncation leads to a systematic error that vanishes as $M \to \infty$. We are interested in understanding the rate at which this happens.

The Sobolev assumption to control precisely this error as a function of the tunable parameter M and the smoothness β.

Lemma 3.14. *For any integer $M \geq 1$, and $f \in \Theta(\beta, Q)$, $\beta > 1/2$, it holds*

$$\|\varphi_{\theta^*}^M - f\|_{L_2}^2 = \sum_{j>M} |\theta_j^*|^2 \leq QM^{-2\beta} \,. \tag{3.10}$$

and for $M = n - 1$, we have

$$|\varphi_{\theta^*}^{n-1} - f|_2^2 \leq 2n\Big(\sum_{j\geq n} |\theta_j^*|\Big)^2 \lesssim Qn^{2-2\beta} \,. \tag{3.11}$$

Proof. Note that for any $\theta \in \Theta(\beta, Q)$, if $\beta > 1/2$, then

$$\begin{aligned} \sum_{j=2}^{\infty} |\theta_j| &= \sum_{j=2}^{\infty} a_j|\theta_j|\frac{1}{a_j} \\ &\leq \sqrt{\sum_{j=2}^{\infty} a_j^2\theta_j^2 \sum_{j=2}^{\infty} \frac{1}{a_j^2}} \quad \text{by Cauchy-Schwarz} \\ &\leq \sqrt{Q\sum_{j=1}^{\infty} \frac{1}{j^{2\beta}}} < \infty \end{aligned}$$

Since $\{\varphi_j\}_j$ forms an orthonormal system in $L_2([0,1])$, we have

$$\min_{\theta\in\mathbb{R}^M} \|\varphi_\theta - f\|_{L_2}^2 = \|\varphi_{\theta^*} - f\|_{L_2}^2 = \sum_{j>M} |\theta_j^*|^2 \,.$$

When $\theta^* \in \Theta(\beta, Q)$, we have

$$\sum_{j>M} |\theta_j^*|^2 = \sum_{j>M} a_j^2 |\theta_j^*|^2 \frac{1}{a_j^2} \le \frac{1}{a_{M+1}^2} Q \le \frac{Q}{M^{2\beta}} \,.$$

To prove the second part of the lemma, observe that

$$|\varphi_{\theta^*}^{n-1} - f|_2 = \big|\sum_{j\ge n} \theta_j^* \varphi_j\big|_2 \le 2\sqrt{2n} \sum_{j\ge n} |\theta_j^*| \,,$$

where in the last inequality, we used the fact that for the trigonometric basis $|\varphi_j|_2 \le \sqrt{2n}, j \ge 1$ regardless of the choice of the design $X_1, \dots, X_n$. When $\theta^* \in \Theta(\beta, Q)$, we have

$$\sum_{j\ge n} |\theta_j^*| = \sum_{j\ge n} a_j |\theta_j^*| \frac{1}{a_j} \le \sqrt{\sum_{j\ge n} a_j^2 |\theta_j^*|^2} \sqrt{\sum_{j\ge n} \frac{1}{a_j^2}} \lesssim Q n^{\frac{1}{2}-\beta} \,.$$

□

Note the truncated Fourier series φ_{θ^*} is an oracle: this is what we see when we view f through the lens of functions with only low frequency harmonics.

To estimate φ_{θ^*}, consider the estimator $\varphi_{\hat\theta^{\text{LS}}}$ where

$$\hat\theta^{\text{LS}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^M} \sum_{i=1}^n \big(Y_i - \varphi_\theta(X_i)\big)^2 \,.$$

Which should be such that $\varphi_{\hat\theta^{\text{LS}}}$ is close to φ_{θ^*}. For this estimator, we have proved (Theorem 3.3) an oracle inequality for the MSE that is of the form

$$|\varphi_{\hat\theta^{\text{LS}}}^M - f|_2^2 \le \inf_{\theta\in\mathbb{R}^M} |\varphi_\theta^M - f|_2 + C\sigma^2 M \log(1/\delta) \,, \qquad C > 0 \,.$$

It yields

$$\begin{aligned} |\varphi_{\hat\theta^{\text{LS}}}^M - \varphi_{\theta^*}^M|_2^2 &\le 2(\varphi_{\hat\theta^{\text{LS}}}^M - \varphi_{\theta^*}^M)^\top (f - \varphi_{\theta^*}^M) + C\sigma^2 M \log(1/\delta) \\ &= 2(\varphi_{\hat\theta^{\text{LS}}}^M - \varphi_{\theta^*}^M)^\top (\sum_{j>M} \theta_j^* \varphi_j) + C\sigma^2 M \log(1/\delta) \\ &= 2(\varphi_{\hat\theta^{\text{LS}}}^M - \varphi_{\theta^*}^M)^\top (\sum_{j\ge n} \theta_j^* \varphi_j) + C\sigma^2 M \log(1/\delta) \,, \end{aligned}$$

where we used Lemma 3.13 in the last equality. Together with (3.11) and Young's inequality $2ab \le \alpha a^2 + b^2/\alpha, a, b \ge 0$ for any $\alpha > 0$, we get

$$2(\varphi_{\hat\theta^{\text{LS}}}^M - \varphi_{\theta^*}^M)^\top (\sum_{j\ge n} \theta_j^* \varphi_j) \le \alpha |\varphi_{\hat\theta^{\text{LS}}}^M - \varphi_{\theta^*}^M|_2^2 + \frac{C}{\alpha} Q n^{2-2\beta} \,,$$

for some positive constant C when $\theta^* \in \Theta(\beta, Q)$. As a result,

$$|\varphi^M_{\hat\theta^{\text{LS}}} - \varphi^M_{\theta^*}|_2^2 \lesssim \frac{1}{\alpha(1-\alpha)} Q n^{2-2\beta} + \frac{\sigma^2 M}{1-\alpha} \log(1/\delta) \tag{3.12}$$

for any $t \in (0,1)$. Since, Lemma 3.13 implies, $|\varphi^M_{\hat\theta^{\text{LS}}} - \varphi^M_{\theta^*}|_2^2 = n\|\varphi^M_{\hat\theta^{\text{LS}}} - \varphi^M_{\theta^*}\|^2_{L_2([0,1])}$, we have proved the following theorem.

Theorem 3.15. *Fix $\beta \geq (1+\sqrt{5})/4 \simeq 0.81, Q > 0, \delta > 0$ and assume the general regression model (3.1) with $f \in \Theta(\beta, Q)$ and $\varepsilon \sim \mathsf{subG}_n(\sigma^2), \sigma^2 \leq 1$. Moreover, let $M = \lceil n^{\frac{1}{2\beta+1}} \rceil$ and n be large enough so that $M \leq n-1$. Then the least squares estimator $\hat\theta^{\text{LS}}$ defined in (3.2) with $\{\varphi_j\}_{j=1}^M$ being the trigonometric basis, satisfies with probability $1-\delta$, for n large enough,*

$$\|\varphi_{\hat\theta^{\text{LS}}} - f\|^2_{L_2([0,1])} \lesssim n^{-\frac{2\beta}{2\beta+1}} + \sigma^2 \frac{\log(1/\delta)}{n}\,.$$

where the constant factors may depend on β, Q and σ. Moreover

$$\mathbb{E}\|\varphi_{\hat\theta^{\text{LS}}} - f\|^2_{L_2([0,1])} \lesssim n^{-\frac{2\beta}{2\beta+1}}\,.$$

Proof. Choosing $\alpha = 1/2$ for example and absorbing Q in the constants, we get from (3.12) and Lemma 3.13 that for $M \leq n-1$,

$$\|\varphi_{\hat\theta^{\text{LS}}} - \varphi_{\theta^*}\|^2_{L_2([0,1])} \lesssim n^{1-2\beta} + \sigma^2 \frac{M + \log(1/\delta)}{n}\,.$$

Using now Lemma 3.14 and $\sigma^2 \leq 1$, we get

$$\|\varphi_{\hat\theta^{\text{LS}}} - f\|^2_{L_2([0,1])} \lesssim M^{-2\beta} + n^{1-2\beta} + \frac{M + \sigma^2 \log(1/\delta)}{n}\,.$$

Taking $M = \lceil n^{\frac{1}{2\beta+1}} \rceil \leq n-1$ for n large enough yields

$$\|\varphi_{\hat\theta^{\text{LS}}} - f\|^2_{L_2([0,1])} \lesssim n^{-\frac{2\beta}{2\beta+1}} + n^{1-2\beta} + \sigma^2 \frac{\log(1/\delta)}{n}\,.$$

To conclude the proof, simply note that for the prescribed β, we have $n^{1-2\beta} \leq n^{-\frac{2\beta}{2\beta+1}}$. The bound in expectation can be obtained by integrating the tail bound. □

Adaptive estimation

The rate attained by the projection estimator $\varphi_{\hat\theta^{\text{LS}}}$ with $M = \lceil n^{\frac{1}{2\beta+1}} \rceil$ is actually optimal so, in this sense, it is a good estimator. Unfortunately, its implementation requires the knowledge of the smoothness parameter β which is typically unknown, to determine the level M of truncation. The purpose of *adaptive estimation* is precisely to adapt to the unknown β, that is to build an estimator

that does not depend on β and yet, attains a rate of the order of $Cn^{-\frac{2\beta}{2\beta+1}}$ (up to a logarithmic lowdown). To that end, we will use the oracle inequalities for the BIC and Lasso estimator defined in (3.3) and (3.4) respectively. In view of Lemma 3.13, the design matrix Φ actually satisfies the assumption ORT when we work with the trigonometric basis. This has two useful implications:

1. Both estimators are actually thresholding estimators and can therefore be implemented efficiently
2. The condition $\mathsf{INC}(k)$ is automatically satisfied for any $k \geq 1$.

These observations lead to the following corollary.

Corollary 3.16. *Fix $\beta \geq (1+\sqrt{5})/4 \simeq 0.81, Q > 0, \delta > 0$ and n large enough to ensure $n-1 \geq \lceil n^{\frac{1}{2\beta+1}} \rceil$ assume the general regression model (3.1) with $f \in \Theta(\beta, Q)$ and $\varepsilon \sim \mathsf{subG}_n(\sigma^2), \sigma^2 \leq 1$. Let $\{\varphi_j\}_{j=1}^{n-1}$ be the trigonometric basis. Denote by $\varphi_{\hat{\theta}^{\mathrm{BIC}}}^{n-1}$ (resp. $\varphi_{\hat{\theta}^{\mathcal{L}}}^{n-1}$) the BIC (resp. Lasso) estimator defined in (3.3) (resp. (3.4)) over $\mathbb{R}^{n-1}$ with regularization parameter given by (3.5) (resp. (3.6)). Then $\varphi_{\hat{\theta}}^{n-1}$, where $\hat{\theta} \in \{\hat{\theta}^{\mathrm{BIC}}, \hat{\theta}^{\mathcal{L}}\}$ satisfies with probability $1-\delta$,*

$$\|\varphi_{\hat{\theta}}^{n-1} - f\|_{L_2([0,1])}^2 \lesssim n^{-\frac{2\beta}{2\beta+1}} + \sigma^2 \frac{\log(1/\delta)}{n}\,.$$

Moreover,

$$\mathbb{E}\|\varphi_{\hat{\theta}}^{n-1} - f\|_{L_2([0,1])}^2 \lesssim \sigma^2 \Big(\frac{\log n}{n}\Big)^{\frac{2\beta}{2\beta+1}}\,.$$

where constant factors may depend on β and Q.

Proof. For $\hat{\theta} \in \{\hat{\theta}^{\mathrm{BIC}}, \hat{\theta}^{\mathcal{L}}\}$, adapting the proofs of Theorem 3.4 for the BIC estimator and Theorem 3.5 for the Lasso estimator, for any $\theta \in \mathbb{R}^{n-1}$, with probability $1-\delta$

$$|\varphi_{\hat{\theta}}^{n-1} - f|_2^2 \leq \frac{1+\alpha}{1-\alpha}|\varphi_{\theta}^{n-1} - f|_2^2 + R(|\theta|_0)\,.$$

where

$$R(|\theta|_0) := \frac{C\sigma^2}{\alpha(1-\alpha)}|\theta_0|\log(en) + \frac{C\sigma^2}{\alpha(1-\alpha)}\log(1/\delta)$$

It yields

$$\begin{aligned}|\varphi_{\hat{\theta}}^{n-1} - \varphi_{\theta}^{n-1}|_2^2 &\leq \frac{2\alpha}{1-\alpha}|\varphi_{\theta}^{n-1} - f|_2^2 + 2(\varphi_{\hat{\theta}}^{n-1} - \varphi_{\theta}^{n-1})^\top(\varphi_{\theta}^{n-1} - f) + R(|\theta|_0) \\ &\leq \Big(\frac{2\alpha}{1-\alpha} + \frac{1}{\alpha}\Big)|\varphi_{\theta}^{n-1} - f|_2^2 + \alpha|\varphi_{\hat{\theta}}^{n-1} - \varphi_{\theta}^{n-1}|_2^2 + R(|\theta|_0)\,,\end{aligned}$$

where we used Young's inequality once again. Choose now $\alpha = 1/2$ and $\theta = \theta_M^*$, where θ_M^* is equal to θ^* on its first M coordinates and 0 otherwise so that $\varphi_{\theta_M^*}^{n-1} = \varphi_{\theta^*}^M$. It yields

$$|\varphi_{\hat{\theta}}^{n-1} - \varphi_{\theta_M^*}^{n-1}|_2^2 \lesssim |\varphi_{\theta_M^*}^{n-1} - f|_2^2 + R(M) \lesssim |\varphi_{\theta_M^*}^{n-1} - \varphi_{\theta^*}^{n-1}|_2^2 + |\varphi_{\theta^*}^{n-1} - f|_2^2 + R(M)$$

Next, it follows from (3.11) that $|\varphi_{\theta^*}^{n-1} - f|_2^2 \lesssim Qn^{2-2\beta}$. Together with Lemma 3.13, it yields

$$\|\varphi_{\hat{\theta}}^{n-1} - \varphi_{\theta_M^*}^{n-1}\|_{L_2([0,1])}^2 \lesssim \|\varphi_{\theta^*}^{n-1} - \varphi_{\theta_M^*}^{n-1}\|_{L_2([0,1])}^2 + Qn^{1-2\beta} + \frac{R(M)}{n}\,.$$

Moreover, using (3.10), we find that

$$\|\varphi_{\hat{\theta}}^{n-1} - f\|_{L_2([0,1])}^2 \lesssim M^{-2\beta} + Qn^{1-2\beta} + \frac{M}{n}\log(en) + \frac{\sigma^2}{n}\log(1/\delta)\,.$$

To conclude the proof, choose $M = \lceil (n/\log n)^{\frac{1}{2\beta+1}} \rceil$ and observe that the choice of β ensures that $n^{1-2\beta} \lesssim M^{-2\beta}$. This yields the high probability bound. The bound in expectation is obtained by integrating the tail. □

While there is sometimes a (logarithmic) price to pay for adaptation, it turns out that the extra logarithmic factor can be removed by a clever use of blocks (see [Tsy09, Chapter 3]). The reason why we get this extra logarithmic factor here is because we use a hammer that's too big. Indeed, BIC and Lasso allow for "holes" in the Fourier decomposition and we use a much weaker version of their potential.

3.3 PROBLEM SET

Problem 3.1. Show that the least-squares estimator $\hat{\theta}^{\mathrm{LS}}$ defined in (3.2) satisfies the following *exact* oracle inequality:

$$\mathbb{E}\mathsf{MSE}(\varphi_{\hat{\theta}^{\mathrm{LS}}}) \le \inf_{\theta \in \mathbb{R}^M} \mathsf{MSE}(\varphi_\theta) + C\sigma^2 \frac{M}{n}$$

for some constant M to be specified.

Problem 3.2. Assume that $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$ and the vectors φ_j are normalized in such a way that $\max_j |\varphi_j|_2 \le \sqrt{n}$. Show that there exists a choice of τ such that the Lasso estimator $\hat{\theta}^{\mathcal{L}}$ with regularization parameter 2τ satisfies the following *exact* oracle inequality:

$$\mathsf{MSE}(\varphi_{\hat{\theta}^{\mathcal{L}}}) \le \inf_{\theta \in \mathbb{R}^M} \left\{ \mathsf{MSE}(\varphi_\theta) + C\sigma|\theta|_1 \sqrt{\frac{\log M}{n}} \right\}$$

with probability at least $1 - M^{-c}$ for some positive constants C, c.

Problem 3.3. Let $\{\varphi_1, \dots, \varphi_M\}$ be a dictionary normalized in such a way that $\max_j |\varphi_j|_2 \le \sqrt{n}$. Show that for any integer k such that $1 \le k \le M$, we have

$$\min_{\substack{\theta \in \mathbb{R}^M \\ |\theta|_0 \le 2k}} \mathsf{MSE}(\varphi_\theta) \le \min_{\substack{\theta \in \mathbb{R}^M \\ |\theta|_{w\ell_q} \le 1}} \mathsf{MSE}(\varphi_\theta) + C_q D^2 \frac{\left(k^{\frac{1}{q}} - M^{\frac{1}{q}}\right)^2}{k},$$

where $|\theta|_{w\ell_q}$ denotes the weak ℓ_q norm and $\bar{q}$ is such that $\frac{1}{q} + \frac{1}{\bar{q}} = 1$.

Problem 3.4. Show that the trigonometric basis and the Haar system indeed form an orthonormal system of $L_2([0,1])$.

Problem 3.5. If $f \in \Theta(\beta, Q)$ for $\beta > 1/2$ and $Q > 0$, then f is continuous.

18.S997 High-dimensional Statistics
Spring 2015