Projection estimators

(notes by Pavel Chigansky)

1. The heuristics

1.1. Regression. Consider the nonparametric regression model

$$Y_j = f(X_j) + \varepsilon_j, \quad j = 1, ..., n$$

with deterministic design points $X_j \in [0,1]$ and i.i.d. zero mean random variables ε_j's with known finite variance $\mathrm{Var}(\varepsilon_1) = \sigma^2$. Assuming that $f \in L^2([0,1])$, it can be expanded into norm convergent series of any orthonormal basis (ϕ_j) in $L^2([0,1])$ (see Appendix A),

$$f = \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \phi_j. \tag{1.1}$$

In this chapter we will consider the Fourier basis, which will be particularly convenient for the purpose of MSE analysis. The limitations of this basis will be explained in the next chapter, where the *wavelet* bases will be introduced.

Truncating the series (1.1) to the first N terms, yields an approximation of f, whose accuracy improves as N increases. Assuming the uniform design $X_j = j/n$,

$$\theta_j := \langle f, \phi_j \rangle = \int_0^1 f(x)\phi_j(x)dx \approx \frac{1}{n}\sum_{m=1}^{n} f(X_m)\phi_j(X_m),$$

where the Riemann integral is approximated by a finite sum. Since $f(X_j)$'s cannot be observed, they are replaced with Y_j's, thus obtaining the *projection* estimator

$$\widehat{f}_{n,N}(x) = \sum_{j=1}^{N} \widehat{\theta}_j \phi_j(x), \tag{1.2}$$

where $\widehat{\theta}_j = \frac{1}{n}\sum_{m=1}^{n} Y_m \phi_j(X_m)$. The integer N plays the role of the bandwidth, since the approximation error (the bias) decreases when N increases. The projection estimator is linear

$$\widehat{f}_{n,N}(x) = \sum_{m=1}^{n} Y_m W_{nm}(x) \tag{1.3}$$

with the weights $W_{nm}(x) = \frac{1}{n}\sum_{j=1}^{N} \phi_j(X_m)\phi_j(x)$.

---

lecture notes for “Advanced Statistical Models B” course.

1.2. Density estimation. For the density estimation problem with unknown density $p \in L^2(D)$ on some subset $D \subseteq \mathbb{R}$, the density estimator can be constructed similarly. Let (ϕ_j) be an orthonormal complete basis in $L^2(D)$. Motivated by the LLN, the scalar products $c_j := \langle p, \phi_j \rangle$ can be approximated by

$$\widehat{c}_j := \frac{1}{n} \sum_{m=1}^{n} \phi_j(X_m)$$

and the projection estimator in this case is

$$\widehat{p}_{n,N}(x) = \sum_{j=1}^{N} \widehat{c}_j \phi_j(x). \tag{1.4}$$

2. Sobolev ellipsoids

The function subspaces, convenient for analysis of the projection estimator, are the Sobolev spaces with integer smoothness degree β. Below $L^p(A)$ with $A \subseteq \mathbb{R}$ denotes the space of p-integrable functions

$$L^p(A) = \left\{ f : \int_A |f(x)|^p dx < \infty \right\},$$

and $C_\infty(A)$ is the space of smooth functions,

$$C_\infty(A) = \left\{ f : \exists f^{(j)}(x) \quad \forall x \in A,\ \forall j \in \mathbb{N} \right\}.$$

If A is a bounded interval, then all $C_\infty(A)$ function are bounded, being continuous.

2.1. Weak derivative.

DEFINITION 2.1. *A function $g \in L^1([a,b])$ is a weak derivative of a function $f \in L^1([a,b])$ if*

$$\int_a^b f(x) \frac{d}{dx} h(x) dx = -\int_a^b g(x) h(x) dx$$

for any $h \in C_\infty([a,b])$ with $h(a) = h(b) = 0$.

Note that the weak derivative can be modified on a Lebesgue measure null set, and hence is not defined uniquely as a function, but rather as an equivalence class. If f is differentiable on $[a,b]$ in the usual sense with derivative $f' \in L^1([a,b])$, it is also weakly differentiable and its weak derivative coincides with f' a.e. This is easily seen through the integration by parts formula

$$\int_a^b f(x) \frac{d}{dx} h(x) dx = f(x)h(x)\Big|_a^b - \int_a^b f'(x) h(x) dx = -\int_a^b f'(x) h(x) dx.$$

Below weak derivatives will be denoted by primes, e.g. f', whenever exist. The following example demonstrates that a function does not have to be differentiable at all points in $[a,b]$ to have a weak derivative.

EXAMPLE 2.2. Suppose f is a function on $[a,b]$ with derivative $f' \in L^1([a,b])$ at all points except for $x_0 \in (a,b)$ and f is continuous at x_0. Then for any $h \in C_\infty([a,b])$ with $h(a) = h(b) = 0$,

$$\int_a^b f(x)h'(x)dx = \int_a^{x_0} f(x)h'(x)dx + \int_{x_0}^b f(x)h'(x)dx =$$

$$f(x)h(x)\Big|_a^{x_0} - \int_a^{x_0} f'(x)h(x)dx + f(x)h(x)\Big|_{x_0}^b - \int_{x_0}^b f'(x)h(x)dx =$$

$$-\int_a^b f'(x)h(x)dx,$$

where the last equality holds by continuity of f at x_0. This implies that f has a weak derivative, which coincides with f' at all $x \in [a,b] \setminus \{x_0\}$ and can be defined arbitrarily at x_0. The same calculation shows that if f is differentiable everywhere except for a discrete set of points at which it is continuous, it's weak derivative coincides with the classical derivative almost everywhere with respect to the Lebesgue measure. The calculation above also suggests that if f has a jump discontinuity at x_0, then f does not have a weak derivative (check!). ■

LEMMA 2.3. *If f is absolutely continuous with respect to the Lebesgue measure, then it has a weak derivative which coincides with the R-N derivative a.e.*

PROOF. Let f' be the R-N derivative of f, then for any test function h as above

$$\int_a^b h(x)f'(x)dx = \int_a^b h(x)df(x) = -\int_a^b f(x)h'(x)dx,$$

where the second equality is obtained by integration by parts for Lebesgue–Stieltjes integrals. □

Absolute continuity in this lemma is crucial: a function may have a classical derivative a.e., but not a weak derivative. For example, Cantor's distribution function has zero derivative off the Cantor set, but does not have a weak derivative. Do not think, however, that highly irregular functions cannot have weak derivatives. For example, the Dirichlet function $f(x) = \mathbf{1}_{\{x \in \mathbb{Q}\}}$ on $[0,1]$ is nowhere continuous, and a fortiori nowhere differentiable. However, it's weak derivative is $f' \equiv 0$, since $\int_0^1 f(x)h'(x)dx = 0$ for any test function h as above.

The weak derivative satisfies many rules of classical calculus, e.g. derivative formulas of sums and products remain true.

2.2. Sobolev spaces.

DEFINITION 2.4. *A function $f : [0,1] \to \mathbb{R}$ belongs to the Sobolev class $W(\beta, L)$ with $L > 0$ and integer $\beta > 0$, if it has weak derivative of order β and $\|f^{(\beta)}\|_2 \le L$. If in addition $f^{(j)}(0) = f^{(j)}(1)$ for $j = 0, \dots, \beta - 1$, then the Sobolev class is called periodic, denoted $W^{\mathrm{per}}(\beta, L)$.*

The classical result from harmonic analysis asserts that any periodic function $f \in L^2([0,1])$ can be expanded into Fourier series (1.1) of the trigonometric basis functions (see Appendix A.5)

$$\begin{aligned}\phi_1(x) &= 1 \\ \phi_{2k}(x) &= \sqrt{2}\cos(2\pi kx) \\ \phi_{2k+1}(x) &= \sqrt{2}\sin(2\pi kx), \quad k = 1,2,\dots\end{aligned} \tag{2.1}$$

and $\theta_j = \langle f, \phi_j\rangle$ belongs to

$$\ell^2 = \Big\{\theta \in \mathbb{R}^\infty : \sum_{j=1}^\infty \theta_j^2 < \infty\Big\}.$$

Moreover, if $\theta \in \ell^1$, the series is pointwise convergent. The following result characterises functions from $W^{\text{per}}(\beta, L)$ in terms of their Fourier coefficients.

LEMMA 2.5. *A function $f(\cdot)$ belongs to $W^{\text{per}}(\beta, L)$ if and only if its Fourier coefficients belong to the ellipsoid*

$$\Theta(\beta, Q) = \Big\{\theta \in \ell^2 : \sum_{j=1}^\infty a_j^2\theta_j^2 \le Q\Big\}, \tag{2.2}$$

where $Q = L^2/\pi^{2\beta}$ and

$$a_j = \begin{cases} j^\beta, & j \text{ is odd}, \\ (j-1)^\beta, & j \text{ is even}. \end{cases}$$

PROOF. See Problem 6. □

If the classical derivative were used instead of the weak derivative in the definition of $W^{\text{per}}(\beta, L)$, the set of functions whose Fourier coefficients belong to the above ellipsoid would be strictly larger.

3. Upper bound for MISE

The MISE risk of the projection estimator (1.2) can be written as

$$\begin{aligned}\mathbb{E}_f\big\|\widehat{f}_{n,N} - f\big\|_2^2 = &\mathbb{E}_f \int_0^1 \Big(\sum_{j=1}^N \widehat{\theta}_j\phi_j(x) - \sum_{j=1}^\infty \theta_j\phi_j(x)\Big)^2 dx = \\ &\mathbb{E}_f \int_0^1 \Big(\sum_{j=1}^N (\widehat{\theta}_j - \theta_j)\phi_j(x) - \sum_{j=N+1}^\infty \theta_j\phi_j(x)\Big)^2 dx \stackrel{\dagger}{=} \\ &\sum_{j=1}^N \mathbb{E}_f(\widehat{\theta}_j - \theta_j)^2 + \sum_{j=N+1}^\infty \theta_j^2 = \\ &\sum_{j=1}^N \mathrm{Var}_f(\widehat{\theta}_j) + \sum_{j=1}^N (\mathbb{E}_f\widehat{\theta}_j - \theta_j)^2 + \sum_{j=N+1}^\infty \theta_j^2,\end{aligned}$$

where † holds by orthonormaity of ϕ_j's. Here

$$\mathbb{E}_f\widehat{\theta}_j - \theta_j = \frac{1}{n}\sum_{m=1}^n f(X_m)\phi_j(X_m) - \theta_j := \delta_j. \tag{3.1}$$

and, with $\sigma^2 := \mathrm{Var}_f(\xi_1)$,

$$\mathrm{Var}_f(\widehat{\theta}_j) = \mathrm{Var}_f\left(\frac{1}{n}\sum_{m=1}^{n} \varepsilon_m \phi_j(X_m)\right) = \frac{\sigma^2}{n^2}\sum_{m=1}^{n}\phi_j(X_m)^2 \le 2n^{-1}\sigma^2,$$

where the inequality holds, since all ϕ_j's are bounded by $\sqrt{2}$. Consequently

$$\mathbb{E}_f\|\widehat{f}_{n,N} - f\|_2^2 \le 2\frac{N}{n}\sigma^2 + \sum_{j=1}^{N}\delta_j^2 + \sum_{j=N+1}^{\infty}\theta_j^2. \tag{3.2}$$

The first and the second terms increase with N and decrease with n. The last term decreases with N regardless of n.

Since we already know that the MSE should be expected to decrease at rate is $n^{2\beta/(2\beta+1)}$, uniformly on classes of functions with β derivatives, it makes sense to let N depend on n and choose $N = [cn^{1/(2\beta+1)}]$, with an arbitrary constant $c > 0$. In this case the first term in (3.2) is correctly bounded by $C_1 n^{-2\beta/(2\beta+1)}$ with a constant C_1. It is left to argue that the rest of the terms (3.2) decrease at this rate as well.

Since a_j's from Lemma 2.5 form an nondecreasing sequence, the last term in (3.2) satisfies

$$\sum_{j=N+1}^{\infty}\theta_j^2 \le \frac{1}{a_{N+1}^2}\sum_{j=N+1}^{\infty} a_j^2\theta_j^2 \le \frac{1}{a_{N+1}^2}Q \le Q(N+1)^{-2\beta} \le C_2 n^{-\frac{2\beta}{2\beta+1}},$$

and hence its contribution is of the same order.

To bound the second term in (3.2) we need to estimate the asymptotic growth of the approximation errors δ_j with n. To this end, hereafter we will assume the uniform design $X_j = j/n$, $j \in \{1, ..., n\}$. We will need the following simple property of the trigonometric basis.

LEMMA 3.1. *The Fourier basis* (2.1) *satisfies*

$$\frac{1}{n}\sum_{m=1}^{n}\phi_i(m/n)\phi_j(m/n) = \begin{cases} 1 & i = j \\ 0 & i \ne j \end{cases}, \quad i, j \in \{1, ..., n-1\}.$$

PROOF. The claimed formula is checked by direct calculation, using the representations $\cos(x) = (e^{\mathbf{i}x} + e^{-\mathbf{i}x})/2$ and $\sin(x) = (e^{\mathbf{i}x} - e^{-\mathbf{i}x})/(2i)$. □

Using this lemma we can write for all $j \le N$,

$$\begin{aligned}\delta_j =& \frac{1}{n}\sum_{m=1}^{n} f(m/n)\phi_j(m/n) - \theta_j \stackrel{\dagger}{=} \\ & \sum_{i=1}^{\infty}\theta_i \frac{1}{n}\sum_{m=1}^{n}\phi_i(m/n)\phi_j(m/n) - \theta_j \stackrel{\ddagger}{=} \\ & \sum_{i=n}^{\infty}\theta_i \frac{1}{n}\sum_{m=1}^{n}\phi_i(m/n)\phi_j(m/n),\end{aligned}$$

where we plugged $f(x)=\sum_{i=1}^\infty \theta_i\phi_i(x)$ in † and the equality ‡ holds by Lemma 3.1, since for the choice of N as above, $N<n$ for all n large enough and hence also $j<n$. Hence for all $j\le N$,

$$|\delta_j| \le \sum_{i=n}^\infty |\theta_i| \frac{1}{n}\Big|\sum_{m=1}^n \phi_i(m/n)\phi_j(m/n)\Big| \le 2\sum_{i=n}^\infty |\theta_i| = 2\sum_{i=1}^\infty a_i|\theta_i|\frac{1}{a_i}\mathbf{1}_{\{i\ge n\}} \stackrel{\dagger}{\le}$$
$$2\Big(\sum_{i=1}^\infty a_i^2\theta_i^2\Big)^{1/2}\Big(\sum_{i=n}^\infty \frac{1}{a_i^2}\Big)^{1/2} \le 2\sqrt{Q}\Big(\sum_{i=n}^\infty \frac{1}{(i-1)^{2\beta}}\Big)^{1/2} \stackrel{\ddagger}{\le} Cn^{-\beta+1/2},$$

where † holds by the Cauchy-Schwarz inequality and ‡ is obtained by the standard estimation of series by integrals. Consequently, the third term in (3.2) satisfies

$$\sum_{j=1}^N \delta_j^2 \le C^2 N n^{-2\beta+1} \le C_3 n^{\frac{1}{2\beta+1}} n^{-2\beta+1} = C_3 n^{-\frac{2\beta}{2\beta+1}+2(1-\beta)}.$$

Hence its contribution is exactly of the right order for $\beta=1$ and it is asymptotically negligible for $\beta>1$. Thus we obtained the following bound for the MISE risk of the projection estimator

$$\sup_{f\in W^{\mathrm{per}}(\beta,L)} \mathbb{E}_f \Big\| n^{\frac{\beta}{2\beta+1}}\big(\widehat{f}_n - f\big)\Big\|^2 < C(\beta,L)<\infty,$$

for any $\beta\ge 1$ and any $L>0$.

Computer Experiment

Implement the density estimator (1.4) using the Fourier basis (2.1) and the bandwidth selector from Problem 4.

(1) Generate a sample from a smooth density on $[0,1]$ of size $n=100$. Apply first your estimator with some small N and plot the obtained estimator versus the true density. By trial and error, tune N to obtain the best estimate as it visually appears to you. Repeat for $n=1000$. Explain the results.

(2) Define the discrete MISE

$$\mathrm{dMISE}(N) = \mathbb{E}_p \frac{1}{n}\sum_{j=1}^n \big(\widehat{p}_{n,N}(j/n) - p(j/n)\big)^2.$$

For each $N\in\{1,\dots,30\}$ approximate $\mathrm{dMISE}(N)$ by MC averaging over $M=10{,}000$ trials for the true density you chose. Plot the obtained sequence and find the optimal value of N, call it N^*. Compare it to the one you obtained in (1) after manual tuning. Apply your estimator using N^* to the same data set you generated in (1) and add the plot to the figure. Comment.

(3) Apply your estimator using the bandwidth selector routine to the dataset in (1). Compare the suggested value of N to those obtained above. Add the plot to the figure. Summarize the results.

Exercises

PROBLEM 1 (Fourier series). Translating and stretching the functions in (2.1) gives the Fourier basis in $L^2([-1,1])$,

$$\phi_1(x) = 1/\sqrt{2}$$
$$\phi_{2k}(x) = \cos(\pi kx)$$
$$\phi_{2k+1}(x) = \sin(\pi kx), \quad k = 1,2,...$$

(1) Check that these functions are orthonormal.

(2) Find the Fourier coefficients of the function $f(x) = \operatorname{sign}(x)$.

(3) Find the Fourier coefficients of the function $f(x) = 1 - |x|$.

(4) Plot the approximations for the above functions, obtained by truncating the Fourier series to 10 and 50 first terms. Compare and explain.

PROBLEM 2. Fill in the details in the proof of Lemma 3.1.

PROBLEM 3 (Reproducing property). A trigonometric polynomial of degree q on the interval $[0,1]$ is a linear combination of all integer powers of $\sin(2\pi x)$ and $\cos(2\pi x)$ up to the power q. Since any power is itself a linear combination of sines and cosines of all frequencies $2\pi j$, $j \in \{0,...,q\}$, it follows that any trigonometric polynomial has the form

$$P(x) = \sum_{k=1}^{N} b_k \phi_k(x),$$

where ϕ_j's are the Fourier basis functions (2.1) and $N = 2q$. Prove that if $N \le n-1$, the projection estimator based on the Fourier basis has the following reproducing property

$$\sum_{m=1}^{n} P(X_m) W_{nm}(x) = P(x), \quad x \in [0,1],$$

where $W_{nm}(x)$ are defined in (1.3) with the uniform design $X_m = m/n, m \in \{1,...,n\}$.

PROBLEM 4. ([1, Exercise 1.9]) Let $X_1,...,X_n$ be an i.i.d. sample from density $p \in L_2([0,1])$. Cosnider the projection estimator (1.4).

(1) Show that $\widehat{c}_j$ are unbiased estimators of the coefficients $c_j = \langle \phi_j, p\rangle$ and find their variances.

(2) Express the MISE of the estimator (1.4) as a function of p and ϕ_j's. Denote it by $\mathrm{MISE}(N)$.

(3) Derive an unbiased risk estimation method. Show that

$$\mathbb{E}_p(\widehat{J}(N)) = \mathrm{MISE}(N) - \int_0^1 p^2(x)dx,$$

where

$$\widehat{J}(N)=\frac{1}{n-1}\sum_{j=1}^{N}\Big(\frac{2}{n}\sum_{i=1}^{n}\phi_j^2(X_i)-(n+1)\widehat{c}_j^2\Big).$$

Propose a data-driven selector for N.

Hint: you may need Parseval's identity, $\|p\|_2^2=\sum_{j=1}^\infty c_j^2$.

(4) Suppose now that $\{\phi_j\}_{j\in\mathbb{N}}$ is the Fourier trigonometric basis. Show that the MISE is bounded by

$$\frac{N+1}{n}+\rho_N,$$

where $\rho_N=\sum_{j=N+1}^\infty c_j^2$. Use this bound to prove that uniformly over the class of all densities p belonging to $W^{\text{per}}(\beta,L)$ with $\beta>0$ and $L>0$, the MISE is of order $O(n^{-2\beta/(2\beta+1)})$ for an appropriate choice of $N=N_n$.

PROBLEM 5. ([1, Exercise 1.10]) Consider the nonparametric regression model under the assumptions as in the text and suppose f belongs to the class $W^{\text{per}}(\beta,L)$ with $\beta\ge 2$. The aim of this problem is to study the *weighted* projection estimator

$$\widehat{f}_{n,\lambda}(x)=\sum_{j=1}^{n}\lambda_j\widehat{\theta}_j\varphi_j(x),$$

where $\lambda=(\lambda_j)$ is a sequence of weights.

(1) Prove that the MISE of $\widehat{f}_{n,\lambda}$ is minimized with respect to λ_j's at

$$\lambda_j^*=\frac{\theta_j(\theta_j+\delta_j)}{\varepsilon^2+(\theta_j+\delta_j)^2},\quad j=1,\dots,n,$$

where $\varepsilon^2=\sigma^2/n$ and δ_j's are defined in (3.1). Note that these weights correspond to the *oracle*, rather than an estimator, since they involve the unknowns θ_j.

(2) Check that the corresponding value of the risk is

$$\text{MISE}(\lambda^*)=\sum_{j=1}^{n}\frac{\varepsilon^2\theta_j^2}{\varepsilon^2+(\theta_j+\delta_j)^2}+\rho_n,$$

where $\rho_n=\sum_{j=n+1}^\infty\theta_j^2$.

(3) Prove that

$$\sum_{j=1}^{n}\frac{\varepsilon^2\theta_j^2}{\varepsilon^2+(\theta_j+\delta_j)^2}=(1+o(1))\sum_{j=1}^{n}\frac{\varepsilon^2\theta_j^2}{\varepsilon^2+\theta_j^2},\quad \text{as } \varepsilon\to 0.$$

(4) Prove that

$$\rho_n=(1+o(1))\sum_{j=n+1}^{\infty}\frac{\varepsilon^2\theta_j^2}{\varepsilon^2+\theta_j^2},\quad \text{as } \varepsilon\to 0.$$

(5) Deduce from the above results that

$$\mathrm{MISE}(\lambda^*) = A_n^*(1+o(1)), \quad n \to \infty,$$

where

$$A_n^* = \sum_{j=1}^{\infty} \frac{\varepsilon^2 \theta_j^2}{\varepsilon^2 + \theta_j^2}.$$

(6) Check that

$$A_n^* \leq \min_{N \geq 1} A_{n,N},$$

where $A_{n,N} = \frac{\sigma^2 N}{n} + \rho_N$.

PROBLEM 6. Prove Lemma 2.5, following the steps below.

(1) Show that $\Theta(\beta, Q) \subset \ell^1$ for $\beta \geq 1$ and hence the Fourier series

$$f(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x)$$

is absolutely convergent for $\theta \in \Theta(\beta, Q)$ and thus define a bounded function $[0,1] \mapsto \mathbb{R}$.

(2) Show that the function in (1) has $\beta - 1$ classical derivatives on $(0,1)$.

(3) Argue that for $\theta \in \Theta(\beta, Q)$ the series

$$g(x) = \sum_{j=1}^{\infty} \theta_j \phi_j^{(\beta)}(x)$$

is convergent in $L^2([0,1])$ and defines the weak derivative of $f^{(\beta-1)}$, i.e, the weak derivative $f^{(\beta)}$ of f.

Hint: show that partial sums form a Cauchy sequence in $L^2([0,1])$ and use completeness of this space.

(4) Show that the weak derivative found above satisfies $\|f^{(\beta)}\| \leq L$, thus concluding $\Theta(\beta, Q) \subseteq W^{\mathrm{per}}(\beta, L)$.

Hint: use Parseval's identity.

(5) Show that $W^{\mathrm{per}}(\beta, L) \subseteq \Theta(\beta, Q)$.

Appendix A. A brief on the Hilbert spaces

A.1. A refresh on Linear Algebra. A set V is called *linear space* if it is closed under linear combinations (with respect to scalars from $\mathbb{R}$). A *basis* of V is a linearly independent set B which spans V. A linear space will typically have many bases and all of them will have the same number of elements (vectors). If this number is finite, say equal $k \in \mathbb{N}$, V is said to have *dimension* k. If there is no finite basis, V is said to be *infinite dimensional*. By the definition of basis, any vector in V has unique representation as a linear combination of the basis vectors.

A function $p : V \times V \mapsto \mathbb{R}$ is called a *scalar (or inner) product* if it satisfies the following axioms.

(1) Symmetry: $p(u,v) = p(v,u)$ for any $u, v \in V$.

(2) Linearity: $p(au_1 + bu_2, v) = ap(u_1, v) + bp(u_2, v)$ for all $u_1, u_2, v \in V$.

(3) Positive definiteness: $p(u,u) > 0$ for all $u \neq 0$.

Note that by symmetry, $p(u,v)$ is linear with respect to both of its arguments, hence it is a *bilinear* form. A common notation for scalar products is $\langle u, v\rangle := p(u,v)$. Typically various scalar products can be defined on the same space.

Recall that a function $r : V \mapsto \mathbb{R}_+$ is a *norm* if it satisfies the following axioms.

(1) Homogeneity: $r(av) = |a| r(v)$ for any $v \in V$ and $a \in \mathbb{R}$.

(2) Triangular inequality: $r(u+v) \le r(u) + r(v)$, $u, v \in V$.

(3) Positive definiteness: $r(u,u) \ge 0$ for all $u \in \mathbb{R}$ and $r(u,u) = 0$ if and only if $u = 0$.

A common notation for norms is $\|u\| := r(u)$. Any scalar product defines a norm $\|v\| = \sqrt{\langle v, v\rangle}$, called the *Euclidean norm* (check!).

A linear space with a scalar product is called *scalar product* space. A pair of vectors u, v are *orthogonal* with respect to a particular scalar product if $\langle u, v\rangle = 0$. A basis of a linear space V is orthogonal if all of its elements are pairwise orthogonal. For k-dimensional linear space V an *orthogonal basis* $B = \{u_1, ..., u_k\}$ can be found by applying the Gram-Schmidt algorithm to an arbitrary basis. Expanding elements of V into orthogonal basis vectors is particularly simple: for any $v \in V$,

$$v = \sum_{j=1}^{k} \langle v, u_j\rangle u_j.$$

EXERCISE A.1.

(1) Show that $\mathbb{R}^n$ is a linear space.

(2) Show that the set of the following n vectors

$$B = \left\{ \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \cdots, \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \right\}$$

is a basis and hence $\dim(\mathbb{R}^n) = n$.

(3) Prove that the function $\langle u, v\rangle = \sum_{j=1}^{n} u_j v_j$ is a scalar product.

(4) Argue that B is not an orthogonal basis.

(5) Let e_j denote the vector in $\mathbb{R}^n$ with 1 in the j-th entry and zeros at all the others. Show that $E = \{e_1, ..., e_n\}$ is an orthogonal basis in $\mathbb{R}^n$.

(6) Show that the space $\mathbb{R}^{\mathbb{N}}$ of sequences $x = (x_1, x_2, ...)$ with $x_j \in \mathbb{R}$ is an infinite dimensional space. Find a basis in this space.

EXERCISE A.2. Let P^n be the space of polynomials of degree less or equal n on $\mathbb{R}$.

(1) Show that P^n is a linear space.

(2) Show that $B = \{1, t, ..., t^n\}$ is a basis in P^n and hence $\deg(P^n) = n+1$.

(3) Show that $\langle p, q\rangle = \int_0^1 p(s)q(s)ds$ is a scalar product on P^n.

(4) Argue that B is not an orthogonal basis.

(5) Find an orthogonal basis in P^2.

(6) Let P be the set of all polynomials. Show that it is a linear infinite-dimensional space.

Finally let us recall the following basic notions from analysis.

DEFINITION A.1. *A sequence $v_n \in V$ is Cauchy if*

$$\lim_n \sup_{m \geq n} \|v_n - v_m\| = 0.$$

Obviously any convergent sequence $v_n \to v \in V$ is Cauchy, since

$$\|v_n - v_m\| \leq \|v_n - v\| + \|v_m - v\| \xrightarrow{m,n\to\infty} 0.$$

The converse does not have to be true. For example, for $V = \mathbb{Q}$, the rational numbers, a Cauchy sequence may converge to an irrational number.

DEFINITION A.2. *A space V is complete if any Cauchy sequence in it converges to a limit in V.*

PROBLEM 7. Prove that $\mathbb{R}$ is complete.

Hint: for a Cauchy sequence x_n define $x = \overline{\lim}_n x_n$ and show that x is finite and $x = \lim_n x_n$.

It can be shown that $\mathbb{R}^k$ and P^k are complete for all $k \in \mathbb{N}$. However the space of all polynomials P is incomplete: the truncated Taylor's expansion for $\sin(x)$ converges in norm and hence is a Cauchy sequence in P, but the limit is not a polynomial.

A.2. The Hilbert space. Let us start with the definition.

DEFINITION A.3. *Hilbert space is a complete scalar product space.*

In many (but not all!) aspects, Hilbert spaces resemble the usual finite dimensional spaces. In particular, it is possible to find an infinitely countable orthogonal basis in some Hilbert spaces (which makes them separable). This is a very useful feature which allows to expand elements in the Hilbert space into series of the basis functions.

A.3. ℓ^2-space. Consider the linear space $\mathbb{R}^{\mathbb{N}}$ of all sequences $x = (x_1, x_2, \dots)$ with entries $x_j \in \mathbb{R}$. Consider the subset

$$\ell^2 = \left\{ x \in \mathbb{R}^{\mathbb{N}} : \sum_{j=1}^{\infty} x_j^2 < \infty \right\}.$$

It is easy to see that this is linear subspace (check!) and that the bilinear form

$$\langle x, y \rangle = \sum_{j=1}^{\infty} x_j y_j$$

is well defined and finite and is a scalar product. It can be checked that ℓ^2 is complete and thus is the Hilbert space. If we let e^j be the sequence whose j-the entry is 1 and all other entries are zeros, then for any $x \in \ell^2$, obviously,

$$x = \sum_{j=1}^{\infty} x_j e^j = \sum_{j=1}^{\infty} \langle x, e^j \rangle e^j. \tag{A.1}$$

Hence (e^j) is an orthonormal basis in ℓ^2 and thus it is separable. Of course, there are other (orthogonal) bases as well. Note that the series in (A.1) converges both pointwise

$$x_i^N = \sum_{j=1}^{N} \langle x, e^j \rangle e^j \xrightarrow[N\to\infty]{} x_i, \quad \forall i \in \mathbb{N}$$

and in the norm

$$\|x^N - x\| = \left\| \sum_{j=N+1}^{\infty} \langle x, e^j \rangle e^j \right\| = \sum_{j=N+1}^{\infty} \langle x, e^j \rangle^2 = \sum_{j=N+1}^{\infty} x_j^2 \xrightarrow[N\to\infty]{} 0,$$

where we used orthogonality of e^j's and finiteness of the norm $\|x\| < \infty$.

A.4. L^2-space. For a domain $D \subseteq \mathbb{R}^d$ consider measurable functions $f : D \mapsto \mathbb{R}$ which are square integrable, i.e.,

$$\int_D f(x)^2 dx < \infty.$$

The subset of all such functions is a linear subspace. If we consider any two functions f and g in this subspace to be equivalent if they coincide a.e. on D, the map

$$\|f\|_2 = \left(\int_D f(x)^2 dx \right)^{1/2}$$

defines a norm and

$$\langle g, f \rangle = \int_D f(x) g(x) dx$$

is a scalar product (well defined due to Cauchy-Schwartz inequality). It can be shown that this space, denoted by $L^2(D)$, is complete and thus is a Hilbert space. Moreover, it is separable and it is possible to find various orthogonal bases in this space. For example, the already familiar Legendre polynomials can be shown to form a basis in $L^2([-1,1])$.

If ϕ_j is any orthonormal basis in $L^2(D)$ then the $L^2(D)$-norm of f equals the ℓ^2-norm of the coefficients in its expansion into the basis functions,

$$\int_D f(x)^2 dx = \sum_{j=1}^{\infty} \langle f, \phi_j \rangle^2.$$

This is the Parseval identity.

A.5. The Fourier basis in $L^2([0,1])$. A particularly useful basis for in $L^2([0,1])$ is the Fourier basis of complex exponentials

$$\phi_k(x) = \exp(\mathbf{i}2\pi kx), \quad k \in \mathbb{Z}, \quad x \in [0,1].$$

where $\mathbf{i} = \sqrt{-1}$ is the imaginary unit. The relevant $L^2([0,1])$ space here is of complex valued functions $f : [0,1] \mapsto \mathbb{C}$ is considered with the scalar product

$$\langle g, f \rangle = \int_0^1 g(x)\overline{f(x)}dx,$$

where for $\overline{a}$ stands for the complex conjugate of $a \in \mathbb{C}$. The corresponding norm is

$$\|f\| = \left(\int_0^1 |f(x)|^2 dx \right)^{1/2}.$$

It is straightforward to see that ϕ_j's are orthonormal and it can be shown that

$$f = \sum_{k=-\infty}^{\infty} \langle f, \phi_k \rangle \phi_k$$

where the series converges in norm. Other types of convergence is a more subtle matter, but often this series converges pointwise or even uniformly. This series representation is called the Fourier expansion of f and the constants $\langle f, \phi_k \rangle$ are its Fourier coefficients. Note that if f is real valued, by the Euler formula,

$$\overline{\langle f, \phi_k \rangle} = \langle f, \phi_{-k} \rangle = \int_0^1 f(x)\big(\cos(2\pi kx) - \mathbf{i}\sin(2\pi kx)\big) =: c_k(f) - \mathbf{i}s_k(f),$$

and a simple calculation shows that

$$\sum_{k=-\infty}^{\infty} \langle f, \phi_k \rangle \phi_k = \langle f, 1 \rangle + 2\sum_{k=1}^{\infty} \Big(c_k(f)\cos(2\pi kx) + s_k(f)\sin(2\pi kx) \Big).$$

Thus for real valued functions the Fourier expansion is a trigonometric series.

References

[1] AB Tsybakov, Introduction to Nonparametric Estimation, Springer Science & Business Media, 22 Oct 2008

[2] M. Pinsky, Introduction to Fourier Analysis and Wavelets, AMS 2009

Department of Statistics, The Hebrew University, Mount Scopus, Jerusalem 91905, Israel

E-mail address: Pavel.Chigansky@mail.huji.ac.il