Math Camp 2: Functional analysis

Sayan Mukherjee, Alessandro Verri, Alex Rakhlin

Dense

Let A and B be subspaces of a metric space $\mathbb{R}$. A is said to be **dense** in B if $\bar{A} \subset B$. $\bar{A}$ is the closure of the subset A. In particular A is said to be *everywhere dense* in $\mathbb{R}$ if $\bar{A} = R$.

A point $x \in \mathbb{R}$ is called a *contact point* of a set $A \in \mathbb{R}$ if every neighborhood of x contains at least on point of A. The set of all contact points of a set A denoted by $\bar{A}$ is called the *closure* of A.

Examples

1. The set of all rational points is dense in the real line.

2. The set of all polynomials with rational coefficients is dense in $C[a, b]$.

3. Let K be a positive definite Radial Basis Function then the functions

$$f(x) = \sum_{i=1}^{n} c_i K(x - x_i)$$

is dense in L_2.

Note: A hypothesis space that is dense in L_2 is a desired property of any approximation scheme.

Separable

A metric space is said to be **separable** if it has a countable everywhere dense subset.

Examples:

1. The spaces $\mathbb{R}^1$, $\mathbb{R}^n$, $L_2[a,b]$, and $C[a,b]$ are all separable.

2. The set of real numbers is separable since the set of rational numbers is a countable subset of the reals and the set of rationals is is everywhere dense.

Completeness

A sequence of functions f_n is *fundamental* if $\forall \epsilon > 0$ $\exists N_\epsilon$ such that

$$\forall n \text{ and } m > N_\epsilon, \quad \rho(f_n, f_m) < \epsilon.$$

A metric space is **complete** if all fundamental sequences converge to a point in the space.

C, L^1, and L^2 are complete. That C_2 is not complete, instead, can be seen through a counterexample.

Incompleteness of C_2

Consider the sequence of functions $(n = 1, 2, ...)$

$$\phi_n(t) = \begin{cases} -1 & \text{if } -1 \leq t < -1/n \\ nt & \text{if } -1/n \leq t < 1/n \\ 1 & \text{if } 1/n \leq t \leq 1 \end{cases}$$

and assume that ϕ_n converges to a continuous function ϕ in the metric of C_2. Let

$$f(t) = \begin{cases} -1 & \text{if } -1 \leq t < 0 \\ 1 & \text{if } 0 \leq t \leq 1 \end{cases}$$

Incompleteness of C_2 (cont.)

Clearly,

$$\left(\int (f(t)-\phi(t))^2 dt\right)^{1/2} \leq \left(\int (f(t)-\phi_n(t))^2 dt\right)^{1/2} + \left(\int (\phi_n(t)-\phi(t))^2 dt\right)^{1/2}.$$

Now the l.h.s. term is strictly positive, because $f(t)$ is not continuous, while for $n \to \infty$ we have

$$\int (f(t)-\phi_n(t))^2 dt \to 0.$$

Therefore, contrary to what assumed, ϕ_n cannot converge to ϕ in the metric of C_2.

Completion of a metric space

Given a metric space $\mathbb{R}$ with closure $\bar{\mathbb{R}}$, a complete metric space $\mathbb{R}^*$ is called a **completion** of $\mathbb{R}$ if $\mathbb{R} \subset \mathbb{R}^*$ and $\bar{\mathbb{R}} = \mathbb{R}^*$.

Examples

1. The space of real numbers is the completion of the space of rational numbers.

2. Let K be a positive definite Radial Basis Function then L_2 is the completion the space of functions
$$f(x) = \sum_{i=1}^{n} c_i K(x - x_i).$$

Compact spaces

A metric space is **compact** *iff* it is *totally bounded* and *complete*.

Let $\mathbb{R}$ be a metric space and ϵ any positive number. Then a set $A \subset \mathbb{R}$ is said to be an ϵ*-net* for a set $M \subset \mathbb{R}$ if for every $x \in M$, there is at least one point $a \in A$ such that $\rho(x, a) < \epsilon$.

Given a metric space $\mathbb{R}$ and a subset $M \subset \mathbb{R}$ suppose M has a finite ϵ-net for every $\epsilon > 0$. Then M is said to be *totally bounded*.

A compact space has a finite ϵ-net for all $\epsilon > 0$.

Examples

1. In Euclidean n-space, $\mathbb{R}^n$, total boundedness is equivalent to boundedness. If $M \subset \mathbb{R}$ is bounded then M is contained in some hypercube Q. We can partition this hypercube into smaller hypercubes with sides of length ϵ. The vertices of the little cubes from a finite $\sqrt{n}\epsilon/2$-net of Q.

2. This is not true for infinite-dimensional spaces. The unit sphere Σ in l_2 with constraint

$$\sum_{n=1}^{\infty} x_n^2 = 1,$$

is bounded but not totally bounded. Consider the points

$$e_1 = (1, 0, 0, ...),\ e_2 = (0, 1, 0, 0, ...), ...,$$

where the n-th coordinate of e_n is one and all others are zero. These points lie on Σ but the distance between any two is $\sqrt{2}$. So Σ cannot have a finite ϵ-net with $\epsilon < \sqrt{2}/2$.

3. Infinite-dimensional spaces maybe totally bounded. Let Π be the set of points $x = (x_1, ..., x_n, ..)$ in l_2 satisfying the inequalities

$$|x_1| < 1,\ |x_2| < \frac{1}{2}, ...,\ |x_n| < \frac{1}{2^{n-1}}, ...$$

The set Π called the *Hilbert cube* is an example of an infinite-dimensional totally bounded set. Given any $\epsilon > 0$, choose n such that

$$\frac{1}{2^{n+1}} < \frac{\epsilon}{2},$$

and with each point

$$x = (x_1, ..., x_n, ..)$$

is Π associate the point

$$x^* = (x_1, ..., x_n, 0, 0, ...). \qquad (1)$$

Then

$$\rho(x, x^*) = \sqrt{\sum_{k=n+1}^{\infty} x_k^2} < \sqrt{\sum_{k=n}^{\infty} \frac{1}{4^k}} < \frac{1}{2^{n-1}} < \frac{\epsilon}{2}.$$

The set Π^* of all points in Π that satisfy (1) is totally bounded since it is a bounded set in n-space.

4. The RKHS induced by a kernel K with an infinite number of positive eigenvalues that decay exponentially is compact. In this case, our vector $x = (x_1, ..., x_n, ..)$ can

be written in terms of its basis functions, the eigenvectors of K. Now for the RKHS norm to be bounded

$$|x_1| < \mu_1,\ |x_2| < \mu_2, ...,\ |x_n| < \mu_n, ...$$

and we know that $\mu_n = O(n^{-\alpha})$. So we have the case analogous to the Hilbert cube and we can introduce a point

$$x^* = (x_1, ..., x_n, 0, 0, ...) \tag{2}$$

in a bounded n-space which can be made arbitrarily close to x.

Compactness and continuity

A family Φ of functions ϕ defined on a closed interval $[a, b]$ is said to be *uniformly bounded* if for $K > 0$

$$|\phi(x)| < K$$

for all $x \in [a, b]$ and all $\phi \in \Phi$.

A family Φ of functions ϕ is *equicontinuous* of for any given $\epsilon > 0$ there exists $\delta > 0$ such that $|x - y| < \delta$ implies

$$|\phi(x) - \phi(y)| < \epsilon$$

for all $x, y \in [a, b]$ and all $\phi \in \Phi$.

Arzela's theorem: A necessary and sufficient condition for a family Φ of continuous functions defined on a closed interval $[a, b]$ to be (relatively) compact in $C[a, b]$ is that Φ is uniformly bounded and equicontinuous.

Linear space

A set L of elements $x, y, z, ...$ is a **linear space** if the following three axioms are satisfied:

1. Any two elements $x, y \in L$ uniquely determine a third element in $x + y \in L$ called the sum of x and y such that
   (a) $x + y = y + x$ (commutativity)
   (b) $(x + y) + z = x + (y + z)$ (associativity)
   (c) An element $0 \in L$ exists for which $x + 0 = x$ for all $x \in L$
   (d) For every $x \in L$ there exists an element $-x \in L$ with the property $x + (-x) = 0$

2. Any number α and any element $x \in L$ uniquely determine an element $\alpha x \in L$ called the product such that
   (a) $\alpha(\beta x) = \beta(\alpha x)$
   (b) $1x = x$

3. Addition and multiplication follow two distributive laws
   (a)$(\alpha + \beta)x = \alpha x + \beta x$
   (b)$\alpha(x + y) = \alpha x + \alpha y$

Linear functional

A functional, $\mathcal{F}$, is a function that maps another function to a real-value

$$\mathcal{F} : f \to \mathbb{R}.$$

A linear functional defined on a linear space L, satisfies the following two properties

1. Additive: $\mathcal{F}(f+g) = \mathcal{F}(f) + \mathcal{F}(g)$ for all $f, g \in L$

2. Homogeneous: $\mathcal{F}(\alpha f) = \alpha \mathcal{F}(f)$

Examples

1. Let $\mathbb{R}^n$ be a real n-space with elements $x = (x_1, ..., x_n)$, and $a = (a_1, ..., a_n)$ be a fixed element in $\mathbb{R}^n$. Then

   $$\mathcal{F}(x) = \sum_{i=1}^{n} a_i x_i$$

   is a linear functional

2. The integral

   $$\mathcal{F}[f(x)] = \int_a^b f(x)p(x)dx$$

   is a linear functional

3. Evaluation functional: another linear functional is the

Dirac delta function

$$\delta_t[f(\cdot)] = f(t).$$

Which can be written

$$\delta_t[f(\cdot)] = \int_a^b f(x)\delta(x-t)dx.$$

4. Evaluation functional: a positive definite kernel in a RKHS

$$\mathcal{F}_t[f(\cdot)] = (K_t, f) = f(t).$$

This is simply the reproducing property of the RKHS.

Fourier Transform

The *Fourier Transform* of a real valued function $f \in L_1$ is the complex valued function $\tilde{f}(\omega)$ defined as

$$\mathcal{F}[f(x)] = \tilde{f}(\omega) = \int_{-\infty}^{+\infty} f(x)\ e^{-j\omega x} dx.$$

The FT $\tilde{f}$ can be thought of as a representation of the information content of $f(x)$. The original function f can be obtained through the *inverse Fourier Transform* as

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \tilde{f}(\omega)\ e^{j\omega x} d\omega.$$

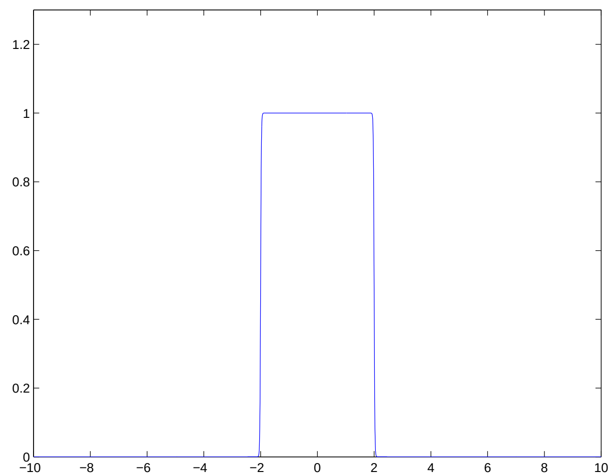
Properties

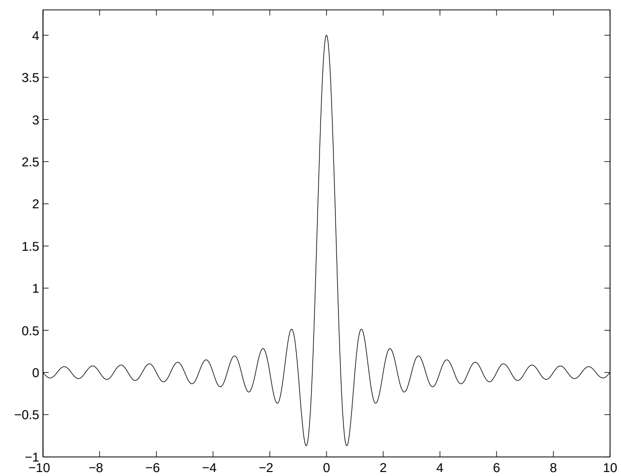
$$
\begin{aligned}
f(at) &\Leftrightarrow \frac{1}{|a|} F\left(\frac{\omega}{a}\right) \\
f^*(t) &\Leftrightarrow F^*(\omega) \\
F(t) &\Leftrightarrow 2\pi f(-\omega) \\
f(t-t_0) &\Leftrightarrow F(\omega) e^{-jt_0\omega} \\
f(t)e^{j\omega_0 t} &\Leftrightarrow F(\omega-\omega_0) \\
\frac{d^n f(t)}{dt^n} &\Leftrightarrow (j\omega)^n F(\omega) \\
(-jt)^n f(t) &\Leftrightarrow \frac{d^n F(\omega)}{d\omega^n} \\
\int_{-\infty}^{\infty} f_1(\tau) f_2(t-\tau) d\tau &\Leftrightarrow F_1(\omega) F_2(\omega) \\
\int_{-\infty}^{\infty} f^*(\tau) f(t+\tau) d\tau &\Leftrightarrow |F(\omega)|^2
\end{aligned}
$$

Properties

The box and the sinc

$$f(t) = 1 \text{ if } -a \leq t \leq a \text{ and } 0 \text{ otherwise}$$
$$F(\omega) = \frac{2\sin(a\omega)}{\omega}.$$



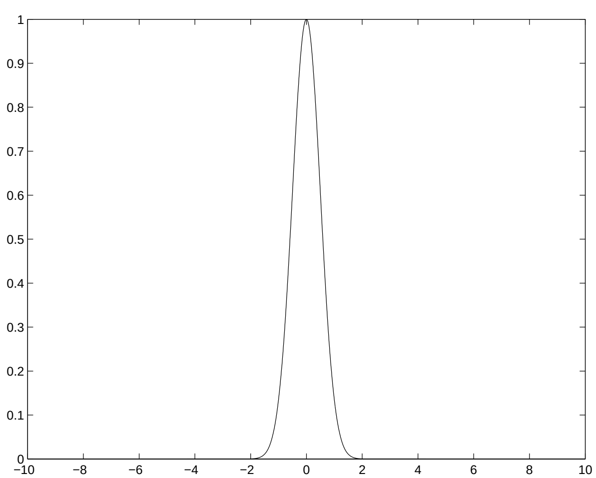


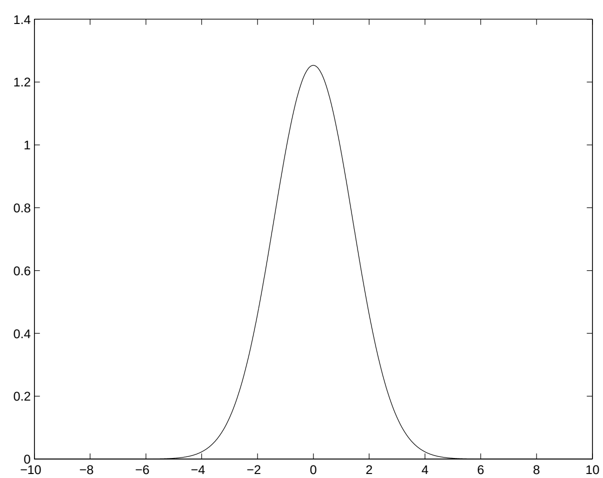
Properties

The Gaussian

$$f(t) = e^{-at^2}$$

$$F(\omega) = \sqrt{\frac{\pi}{a}} e^{-\omega^2/4a}.$$

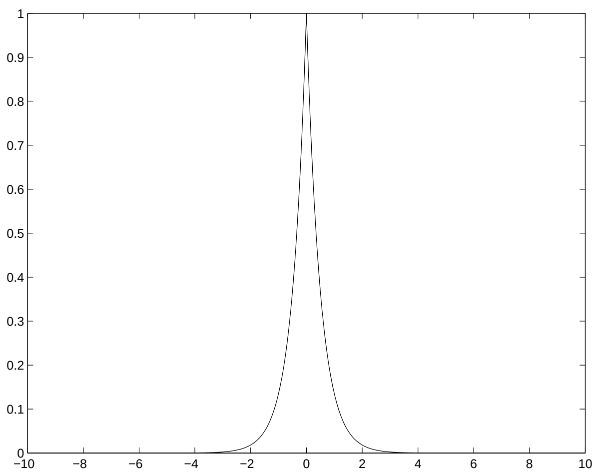


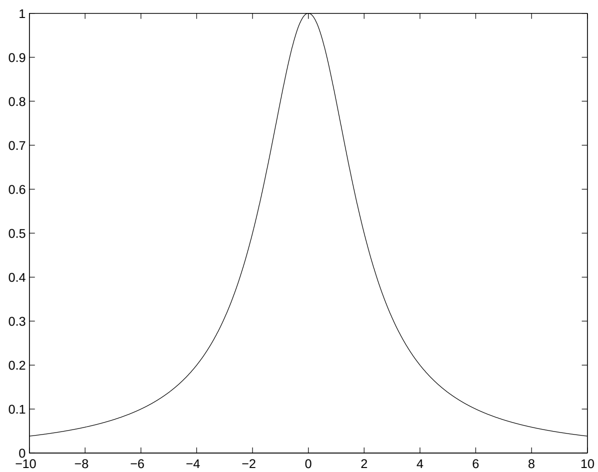


Properties

The Laplacian and Cauchy distributions

$$
\begin{aligned}
f(t) &= e^{-a|t|} \\
F(\omega) &= \frac{2a}{a^2 + \omega^2}.
\end{aligned}
$$





Fourier Transform in the distribution sense

With due care, the Fourier Transform can be defined in the distribution sense. For example, we have

- $\delta(x) \iff 1$
- $\cos(\omega_0 x) \iff \pi(\delta(\omega - \omega_0) + \delta(\omega + \omega_0))$
- $\sin(\omega_0 x) \iff j\pi(\delta(\omega + \omega_0) - \delta(\omega - \omega_0))$
- $U(x) \iff \pi\delta(\omega) - j/\omega$
- $|x| \iff -2/\omega^2$

Parseval's formula

If f is also square integrable, the *Fourier Transform* leaves the norm of f unchanged. Parseval's formula states that

$$\int_{-\infty}^{+\infty} |f(x)|^2 dx = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\tilde{f}(\omega)|^2 d\omega.$$

Fourier Transforms of functions and distributions

The following are Fourier transforms of some functions and distributions

- $f(x) = \delta(x) \iff \tilde{f}(\omega) = 1$
- $f(x) = \cos(\omega_0 x) \iff \tilde{f}(\omega) = \pi(\delta(\omega - \omega_0) + \delta(\omega + \omega_0))$
- $f(x) = \sin(\omega_0 x) \iff \tilde{f}(\omega) = i\pi(\delta(\omega + \omega_0) - \delta(\omega - \omega_0))$
- $f(x) = U(x) \iff \tilde{f}(\omega) = \pi\delta(\omega) - i/\omega$
- $f(x) = |x| \iff \tilde{f}(\omega) = -2/\omega^2$.

Functional differentiation

In analogy with standard calculus, the minimum of a functional can be obtained by setting equal to zero the *derivative* of the functional. If the functional depends on the derivatives of the unknown function, a further step is required (as the unknown function has to be found as the solution of a differential equation).

Functional differentiation

The derivative of a functional $\Phi[f]$ is defined

$$\frac{D\Phi[f]}{Df(s)} = \lim_{h \to 0} \frac{\Phi[f(t) + h\delta(t-s)] - \Phi[f(t)]}{h}.$$

Note that the derivative depends on the location s. For example, if $\Phi[f] = \int_{-\infty}^{+\infty} f(t)g(t)dt$

$$\frac{D\Phi[f]}{Df(s)} = \int_{-\infty}^{+\infty} g(t)\delta(t-s)dt = g(s).$$

Intuition

Let $f : [a, b] \to \mathbb{R}$, $a = x_1$ and $b = x_N$. The intuition behind this definition is that the functional $\Phi[f]$ can be thought of as the limit for $N \to \infty$ of the function of N variables

$$\Phi_N = \Phi_N(f_1, f_2, ..., f_N)$$

with $f_1 = f(x_1)$, $f_2 = f(x_2)$, ... $f_N = f(x_N)$.

For $N \to \infty$, Φ depends on the entire function f. The dependence on the location brought in by the δ function corresponds to the partial derivative with respect to the variable f_k.

Functional differentiation (cont.)

If $\Phi[f] = f(t)$, the derivative is simply

$$\frac{D\Phi[f]}{Df(s)} = \frac{Df(t)}{Df(s)} = \delta(t-s).$$

Similarly to ordinary calculus, the minimum of a functional $\Phi[f]$ is obtained as the function solution to the equation

$$\frac{D\Phi[f]}{Df(s)} = 0.$$

Random variables

We are given a random variable $\xi \sim F$. To define a random variable you need three things:
1) a set to draw the values from, we'll call this Ω
2) a σ-algebra of subsets of Ω, we'll call this $\mathcal{B}$
3) a probability measure F on $\mathcal{B}$ with $F(\Omega) = 1$

So $(\Omega, \mathcal{B}, F)$ is a probability space and a random variable is a measurable function $X : \Omega \rightarrow \mathbb{R}$.

Expectations

Given a random variable $\xi \sim F$ the expectation is

$$\mathbb{E}\xi \equiv \int \xi dF.$$

Similarly the variance of the random variable $\sigma^2(\xi)$ is

$$\text{var}(\xi) \equiv \mathbb{E}(\xi - \mathbb{E}\xi)^2.$$

Law of large numbers

The law of large numbers tells us:

$$\lim_{\ell\to\infty} \frac{1}{\ell} \sum_{i=1}^{\ell} I_{[f(x_i)\neq y_i]} \to \mathbb{E}_{x,y} I_{[f(x)\neq y]}.$$

If $\ell\sigma \to \infty$ the Central Limit Theorem states:

$$\frac{\sqrt{\ell}(\frac{1}{\ell}\sum I - \mathbb{E}I)}{\sqrt{\mathrm{var} I}} \to N(0,1),$$

which implies

$$\left|\frac{1}{\ell}\sum I - \mathbb{E}I\right| \sim \frac{k}{\sqrt{\ell}}.$$

If $\ell\sigma \to c$ the Central Limit Theorem implies

$$\left|\frac{1}{\ell}\sum I - \mathbb{E}I\right| \sim \frac{k}{\ell}.$$

Useful Probability Inequalities

Jensen's inequality: if ϕ is a convex function, then

$$\phi(\mathbb{E}(X)) \leq \mathbb{E}(\phi(X)).$$

For $X \geq 0$,

$$\mathbb{E}(X) = \int_0^\infty \Pr(X \geq t)dt.$$

Markov's inequality: if $X \geq 0$, then

$$\Pr(X \geq t) \leq \frac{\mathbb{E}(X)}{t},$$

where $t \geq 0$.

Useful Probability Inequalities

Chebyshev's inequality (second moment): if X is arbitrary random variable and $t > 0$,

$$\text{Pr}(|X - \mathbb{E}(X)| \geq t) \leq \frac{var(X)}{t^2}.$$

Cauchy-Schwarz inequality: if $\mathbb{E}(X^2)$ and $\mathbb{E}(Y^2)$ are finite, then

$$|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

Useful Probability Inequalities

If X is a sum of independent variables, then X is better approximated by $\mathbb{E}(X)$ than predicted by Chebyshev's inequality. In fact, it's exponentially close!

Hoeffding's inequality:

Let $X_1, ..., X_n$ be independent bounded random variables, $a_i \leq X_i \leq b_i$ for any $i \in 1...n$. Let $S_n = \sum_{i=1}^{n} X_i$, then for any $t > 0$,

$$\Pr(|S_n - \mathbb{E}(S_n)| \geq t) \leq 2exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Playing with Expectations

Fix a function f, loss V, and dataset $S = \{z_1, ..., z_n\}$. The empirical loss of f on this data is $I_S[f] = \frac{1}{n}\sum_{i=1}^{n} V(f, z_i)$. The expected error of f is $I[f] = \mathbb{E}_z V(f, z)$. What is the expected empirical error with respect to a draw of a set S of size n?

$$\mathbb{E}_S I_S[f] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_S V(f, z_i) = \mathbb{E}_S V(f, z_1)$$