MIT 9.520/6.860
Statistical Learning Theory and Applications

Class 0: Mathcamp

Lorenzo Rosasco

$\mathbb{R}^D$

We like $\mathbb{R}^D$ because we can

- add elements $v + w$
- multiply by numbers $3v$
- take scalar products $v^T w = \sum_{j=1}^{D} v^j w^j$
- …and norms $\|v\| = \sqrt{v^T v} = \sqrt{\sum_{j=1}^{D} (v^j)^2}$
- …and distances $d(v, w) = \|v - w\| = \sum_{j=1}^{D} (v^j - w^j)^2$.

We want to do the same thing with $D = \infty$…

Vector Space

- A **vector space** is a set V with binary operations

$$+ \colon V \times V \to V \quad \text{and} \quad \cdot : \mathbb{R} \times V \to V$$

such that for all $a, b \in \mathbb{R}$ and $v, w, x \in V$:

1. $v + w = w + v$
2. $(v + w) + x = v + (w + x)$
3. There exists $0 \in V$ such that $v + 0 = v$ for all $v \in V$
4. For every $v \in V$ there exists $-v \in V$ such that $v + (-v) = 0$
5. $a(bv) = (ab)v$
6. $1v = v$
7. $(a + b)v = av + bv$
8. $a(v + w) = av + aw$

- Example: $\mathbb{R}^n$, space of polynomials, space of functions.

Inner Product

- An **inner product** is a function $\langle \cdot, \cdot \rangle \colon V \times V \to \mathbb{R}$ such that for all $a, b \in \mathbb{R}$ and $v, w, x \in V$:
  1. $\langle v, w \rangle = \langle w, v \rangle$
  2. $\langle av + bw, x \rangle = a\langle v, x \rangle + b\langle w, x \rangle$
  3. $\langle v, v \rangle \geqslant 0$ and $\langle v, v \rangle = 0$ if and only if $v = 0$.
- $v, w \in V$ are orthogonal if $\langle v, w \rangle = 0$.
- Given $W \subseteq V$, we have $V = W \oplus W^{\perp}$, where $W^{\perp} = \{\, v \in V \mid \langle v, w \rangle = 0 \text{ for all } w \in W \,\}$.
- Cauchy-Schwarz inequality: $\langle v, w \rangle \leqslant \langle v, v \rangle^{1/2} \langle w, w \rangle^{1/2}$.

Norm

- A **norm** is a function $\| \cdot \|: V \to \mathbb{R}$ such that for all $a \in \mathbb{R}$ and $v, w \in V$:
    1. $\|v\| \geqslant 0$, and $\|v\| = 0$ if and only if $v = 0$
    2. $\|av\| = |a| \, \|v\|$
    3. $\|v + w\| \leqslant \|v\| + \|w\|$
- Can define norm from inner product: $\|v\| = \langle v, v \rangle^{1/2}$.

Metric

- A **metric** is a function $d \colon V \times V \to \mathbb{R}$ such that for all $v, w, x \in V$:
  1. $d(v, w) \geqslant 0$, and $d(v, w) = 0$ if and only if $v = w$
  2. $d(v, w) = d(w, v)$
  3. $d(v, w) \leqslant d(v, x) + d(x, w)$
- Can define metric from norm: $d(v, w) = \|v - w\|$.

Basis

- $B = \{v_1, \ldots, v_n\}$ is a **basis** of V if every $v \in V$ can be uniquely decomposed as

$$v = a_1 v_1 + \cdots + a_n v_n$$

for some $a_1, \ldots, a_n \in \mathbb{R}$.

- An orthonormal basis is a basis that is orthogonal ($\langle v_i, v_j \rangle = 0$ for $i \neq j$) and normalized ($\|v_i\| = 1$).

Hilbert Space, overview

- Goal: to understand Hilbert spaces (complete inner product spaces) and to make sense of the expression

$$f = \sum_{i=1}^{\infty} \langle f, \phi_i \rangle \phi_i, \quad f \in \mathcal{H}$$

- Need to talk about:
  1. Cauchy sequence
  2. Completeness
  3. Density
  4. Separability

Cauchy Sequence

- Recall: $\lim_{n\to\infty} x_n = x$ if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $\|x - x_n\| < \epsilon$ whenever $n \geqslant N$.
- $(x_n)_{n\in\mathbb{N}}$ is a **Cauchy sequence** if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $\|x_m - x_n\| < \epsilon$ whenever $m, n \geqslant N$.
- Every convergent sequence is a Cauchy sequence (why?)

Completeness

- A normed vector space V is **complete** if every Cauchy sequence converges.
- Examples:
  1. $\mathbb{Q}$ is not complete.
  2. $\mathbb{R}$ is complete (axiom).
  3. $\mathbb{R}^n$ is complete.
  4. Every finite dimensional normed vector space (over $\mathbb{R}$) is complete.

Hilbert Space

- A **Hilbert space** is a complete inner product space.
- Examples:

1. $\mathbb{R}^n$
2. Every finite dimensional inner product space.
3. $\ell_2 = \{(a_n)_{n=1}^{\infty} \mid a_n \in \mathbb{R},\ \sum_{n=1}^{\infty} a_n^2 < \infty\}$
4. $L_2([0,1]) = \{f\colon [0,1] \to \mathbb{R} \mid \int_0^1 f(x)^2\, dx < \infty\}$

Density

- Y is **dense** in X if $\overline{Y} = X$.
- Examples:
  1. $\mathbb{Q}$ is dense in $\mathbb{R}$.
  2. $\mathbb{Q}^n$ is dense in $\mathbb{R}^n$.
  3. Weierstrass approximation theorem: polynomials are dense in continuous functions (with the supremum norm, on compact domains).

Separability

- X is **separable** if it has a countable dense subset.
- Examples:
  1. $\mathbb{R}$ is separable.
  2. $\mathbb{R}^n$ is separable.
  3. ℓ_2, $L_2([0,1])$ are separable.

Orthonormal Basis

- A Hilbert space has a countable orthonormal basis if and only if it is separable.
- Can write:

$$f = \sum_{i=1}^{\infty} \langle f, \phi_i \rangle \phi_i \quad \text{for all } f \in \mathcal{H}.$$

- Examples:
  1. Basis of ℓ_2 is $(1, 0, \ldots,), (0, 1, 0, \ldots), (0, 0, 1, 0, \ldots), \ldots$
  2. Basis of $L_2([0, 1])$ is $1, 2\sin 2\pi nx, 2\cos 2\pi nx$ for $n \in \mathbb{N}$

Maps

Next we are going to review basic properties of maps on a Hilbert space.

- functionals: $\Psi : \mathcal{H} \to \mathbb{R}$
- linear operators $A : \mathcal{H} \to \mathcal{H}$, such that $A(af + bg) = aAf + bAg$, with $a, b \in \mathbb{R}$ and $f, g \in \mathcal{H}$.

Representation of Continuous Functionals

Let $\mathcal{H}$ be a Hilbert space and $g \in \mathcal{H}$, then

$$\Psi_g(f) = \langle f, g \rangle, \qquad f \in \mathcal{H}$$

is a continuous linear functional.

Riesz representation theorem

The theorem states that every continuous linear functional Ψ can be written uniquely in the form,

$$\Psi(f) = \langle f, g \rangle$$

for some appropriate element $g \in \mathcal{H}$.

Matrix

- Every linear operator $L\colon \mathbb{R}^m \to \mathbb{R}^n$ can be represented by an $m \times n$ matrix A.
- If $A \in \mathbb{R}^{m\times n}$, the transpose of A is $A^\top \in \mathbb{R}^{n\times m}$ satisfying
$$\langle Ax, y\rangle_{\mathbb{R}^m} = (Ax)^\top y = x^\top A^\top y = \langle x, A^\top y\rangle_{\mathbb{R}^n}$$
for every $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$.
- A is symmetric if $A^\top = A$.

Eigenvalues and Eigenvectors

- Let $A \in \mathbb{R}^{n\times n}$. A nonzero vector $v \in \mathbb{R}^n$ is an eigenvector of A with corresponding eigenvalue $\lambda \in \mathbb{R}$ if $Av = \lambda v$.
- Symmetric matrices have real eigenvalues.
- **Spectral Theorem:** Let A be a symmetric $n \times n$ matrix. Then there is an orthonormal basis of $\mathbb{R}^n$ consisting of the eigenvectors of A.
- Eigendecomposition: $A = V\Lambda V^\top$, or equivalently,

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^\top.$$

Singular Value Decomposition

- Every $A \in \mathbb{R}^{m \times n}$ can be written as

$$A = U\Sigma V^\top,$$

where $U \in \mathbb{R}^{m \times m}$ is orthogonal, $\Sigma \in \mathbb{R}^{m \times n}$ is diagonal, and $V \in \mathbb{R}^{n \times n}$ is orthogonal.

- Singular system:

$$\begin{aligned} Av_i &= \sigma_i u_i & \qquad AA^\top u_i &= \sigma_i^2 u_i \\ A^\top u_i &= \sigma_i v_i & \qquad A^\top A v_i &= \sigma_i^2 v_i \end{aligned}$$

Matrix Norm

- The spectral norm of $A \in \mathbb{R}^{m \times n}$ is

$$\|A\|_{\mathrm{spec}} = \sigma_{\max}(A) = \sqrt{\lambda_{\max}(AA^\top)} = \sqrt{\lambda_{\max}(A^\top A)}.$$

- The Frobenius norm of $A \in \mathbb{R}^{m \times n}$ is

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2}.$$

Positive Definite Matrix

A real symmetric matrix $A \in \mathbb{R}^{m \times m}$ is positive definite if

$$x^T A x > 0, \quad \forall x \in \mathbb{R}^m.$$

A positive definite matrix has positive eigenvalues.

Note: for positive semi-definite matrices $>$ is replaced by $\geqslant$.

Linear Operator

- An operator $L\colon \mathcal{H}_1 \to \mathcal{H}_2$ is linear if it preserves the linear structure.
- A linear operator $L\colon \mathcal{H}_1 \to \mathcal{H}_2$ is bounded if there exists $C > 0$ such that

$$\|Lf\|_{\mathcal{H}_2} \leqslant C\|f\|_{\mathcal{H}_1} \quad \text{for all } f \in \mathcal{H}_1.$$

- A linear operator is continuous if and only if it is bounded.

Adjoint and Compactness

- The adjoint of a bounded linear operator $L\colon \mathcal{H}_1 \to \mathcal{H}_2$ is a bounded linear operator $L^*\colon \mathcal{H}_2 \to \mathcal{H}_1$ satisfying

$$\langle Lf, g\rangle_{\mathcal{H}_2} = \langle f, L^* g\rangle_{\mathcal{H}_1} \text{ for all } f \in \mathcal{H}_1, g \in \mathcal{H}_2.$$

- L is self-adjoint if $L^* = L$. Self-adjoint operators have real eigenvalues.
- A bounded linear operator $L\colon \mathcal{H}_1 \to \mathcal{H}_2$ is compact if the image of the unit ball in $\mathcal{H}_1$ has compact closure in $\mathcal{H}_2$.

Spectral Theorem for Compact Self-Adjoint Operator

- Let $L\colon \mathcal{H} \to \mathcal{H}$ be a compact self-adjoint operator. Then there exists an orthonormal basis of $\mathcal{H}$ consisting of the eigenfunctions of L,
$$L\phi_i = \lambda_i \phi_i$$
and the only possible limit point of λ_i as $i \to \infty$ is 0.

- Eigendecomposition:
$$L = \sum_{i=1}^{\infty} \lambda_i \langle \phi_i, \cdot \rangle \phi_i.$$

Probability Space

A triple $(\Omega, \mathcal{A}, P)$, where Ω is a set,

$\mathcal{A}$ a Sigma Algebra, i.e. a family of subsets of Ω s.t.

- $\mathcal{X}, \emptyset \in \mathcal{A}$,
- $A \in \mathcal{A} \Rightarrow \Omega \backslash A \in \mathcal{A}$,
- $A_i \in \mathcal{A}, i = 1, 2 \cdots \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{A}$.

P a probability measure, i.e a function $P : \mathcal{A} \to [0, 1]$

- $P(\mathcal{X}) = 1$ (hence and $P(\emptyset) = 0$),
- Sigma additivity: If $A_i \in \mathcal{A}, i = 1, 2 \ldots$ are disjoint, then

$$P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Real Random Variables (RV)

A measurable function $X : \Omega \to \mathbb{R}$, i.e. mapping elements of the sigma algebra in open subsets of $\mathbb{R}$.

- Law of a random variable: probability measure on $\mathbb{R}$ defined as

  $$\rho(I) = P(X^{-1}(I))$$

  for all open subsets $I \subset \mathbb{R}$.

- Probability density function of a probability measure ρ on X: a function $p : \mathbb{R} \to \mathbb{R}$ such that

  $$\int_I d\rho(x) = \int_I p(x)dx$$

  for open subsets $I \subset \mathbb{R}$.

Convergence of Random Variables

X_i, $i = 1, 2, \ldots$, a sequence of random variables.

- Convergence in probability:

$$\forall \epsilon \in (0, \infty), \quad \lim_{i \to \infty} \mathbb{P}\left(|X_i - X| > \epsilon\right) = 0.$$

- Almost Sure Convergence:

$$\mathbb{P}\left(\lim_{i \to \infty} X_i = X\right) = 1.$$

Law of Large Numbers

X_i, $i = 1, 2, \ldots$, sequence of independent copies of a random variable X

Weak Law of Large Numbers:

$$\forall \epsilon \in (0, \infty), \quad \lim_{n \to \infty} \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mathbb{E}[X] \right| > \epsilon \right) = 0.$$

Strong Law of Large Numbers:

$$\mathbb{P}\left(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mathbb{E}[X] \right) = 1.$$

Concentration Inequalities

X, be a random variable $\forall \epsilon \in (0, \infty)$

- Markov's inequality: if $X > 0$

$$\mathbb{P}\left(X \geqslant \epsilon\right) \leqslant \frac{\mathbb{E}[X]}{\epsilon}$$

- Chebysev's inequality: If $\mathbb{V}ar[X] < \infty$

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geqslant \epsilon\right) \leqslant \frac{\mathbb{V}ar[X]}{\epsilon^2}$$

Concentration Inequalities for Sums

$X_1, \dots, X_n$ identical independent random variables with expectation $\mathbb{E}[X]$.

Chebysev's inequality can be applied to $\frac{1}{n}\sum_{i=1}^{n} X_i$ to get

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| \geqslant \epsilon\right) \leqslant \frac{\mathbb{V}ar[X]}{\epsilon^2 n}$$

A stronger results holds if $|X_i| < c$.

- Höeffding's inequality:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| \geqslant \epsilon\right) \leqslant 2e^{-\frac{\epsilon^2 n}{2c^2}}$$