Technische Universität München

Summary of the lecture MA4800

Foundations in Data Analysis

Instructors: Prof. Felix Krahmer and Dr. Anna Veselovska

Contents

1 Linear Algebra Review

1.1 The setup

- We work on $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$.
- $A^H = \overline{(A^T)}$.
- A Hermitian matrix A satisfies $A = A^H$.
- $A^{(i)}$ are rows and $A_{(j)}$ are the columns.
- $A^{(i)} = (a_{ij})_{j \in J}$ and $A_{(j)} = (a_{ij})_{i \in I} = \left(A^T\right)^{(j)}$
- The matrix-vector product between $A \in \mathbb{K}^{I \times J}$ and $x \in \mathbb{K}^I$ results in the vector in $Ax \in K^{\mathbb{P}ime}$ with entries

1.2 Matrices

$$(Ax)_i = \sum_{j \in J} a_{ij} x_j.$$

1.3 Matrix multiplication

The matrix-matrix product between $A \in \mathbb{K}^{I \times J}$ and $B \in \mathbb{K}^{J \times L}$ yields the matrix in $\mathbb{K}^{I \times L}$ with entries

$$(AB)_{i\ell} = \sum_{j \in J} A_{ij} B_{j\ell}.$$

2 The Singular Value Decomposition

2.1 The leading singular vector

2.2 Principal components

2.3 Further singular vectors

2.4 Best k-rank approximation

2.5 The power method

Lemma 2.1. *Let $x \in \mathbb{R}^d$ be a unit d-dimensional vector of components $x = (x_1, \ldots, x_d)$ with respect to the canonical basis and picked uniformly at random from the sphere $\{x : \|x\|_2 = 1\}$. The probability that $|x_1| \geq \alpha > 0$ is at least $1 - C\alpha\sqrt{d}$ for some absolute constant.*

Proof

We want the probability of y picked uniformly at random from

$$B^d(1) = \left\{y \in \mathbb{R}^d, ||y||_2 \leq 1\right\}$$

satisfies $|y_1| > \alpha$. In other words, we are looking for the fraction of $B^d(1)$ that satisfies $|y_1| > \alpha$. This corresponds to

$$V_\alpha := \mathrm{Vol}(B^d(1) \cap \{y : |y_1| \leq \alpha\})$$

$$= \int_{y \in B^d(1) \cap \{y : |y_1| \leq \alpha\}} 1 dy$$

$$= \int_{-\alpha}^{\alpha} \left(\int_{\mathbb{R}^{d-1}} 1_{y_2^2 + \ldots + y_d^2 \leq 1 - y_1^2} \, dy_2 \ldots dy_d \right) dy_1$$

$$= \int_{-\alpha}^{\alpha} \mathrm{Vol}\left(B^{d-1}\left(\sqrt{1 - y_1^2} \right) \right) dy_1$$

Replacing $\mathrm{Vol}\left(B^{d-1}\left(\sqrt{1 - y_1^2} \right) \right)$ with $(\sqrt{1-y_1^2})^{d-1} \mathrm{Vol}\left(B^{d-1}(1) \right)$ since the volume the unit ball with a factor proportional to radius in the power of $d-1$.

$$= \int_{-\alpha}^{\alpha} (\sqrt{1-y_1^2})^{d-1} \mathrm{Vol}\left(B^{d-1}(1) \right) dy_1$$

$$= \mathrm{Vol}\left(B^{d-1}(1) \right) \int_{-\alpha}^{\alpha} (1 - y_1^2)^{(d-1)/2} dy_1$$

In the integral part, $\int_{-\alpha}^{\alpha} (1 - y_1^2)^{(d-1)/2} dy_1$, notice that $(1 - y_1^2)^{(d-1)/2} < 1$ in the whole integration domain. Thus we can write

$$= \mathrm{Vol}\left(B^{d-1}(1) \right) \int_{-\alpha}^{\alpha} (1 - y_1^2)^{(d-1)/2} dy_1$$

$$\leq \mathrm{Vol}\left(B^{d-1}(1) \right) \int_{-\alpha}^{\alpha} 1 dy_1$$

$$= 2\alpha \mathrm{Vol}\left(B^{d-1}(1) \right)$$

Recall that volume of unit ball in d dimensions is asymptotically

$$V_1 = \frac{1}{\sqrt{d\pi}} \left(\frac{2\pi e}{d} \right)^{d/2}$$

Hence the probability $p = \mathbb{P}(\alpha \leq |y_1|)$ we are interested in satisfies asymptotically

$$p = \frac{V_\alpha}{V_1} \mathbb{P}opto \frac{2\alpha \frac{1}{\sqrt{(d-1)\pi}} \left(\frac{2\pi e}{d-1} \right)^{(d-1)/2}}{\frac{1}{\sqrt{d\pi}} \left(\frac{2\pi e}{d} \right)^{d/2}} = \frac{2\alpha \frac{1}{\sqrt{(d-1)\pi}} \left(\frac{2\pi e}{d-1} \right)^{(d-1)/2}}{\frac{1}{\sqrt{d\pi}} \left(\frac{2\pi e}{d} \right)^{(d-1)/2} \left(\frac{2\pi e}{d} \right)^{1/2}}$$

We simplify the last term

$$= 2\alpha * \left(\frac{d}{d-1} \right)^{1/2} * \left(\frac{d}{d-1} \right)^{(d-1)/2} * \left(\frac{d}{2\pi e} \right)^{1/2}$$

$$= 2\alpha * \left(\frac{d}{\sqrt{2\pi e(d-1)}} \right) * \left(\frac{d}{d-1} \right)^{(d-1)/2}$$

Since $\frac{d}{d-1} = 1 + \frac{1}{d-1}$

$$= 2\alpha * \left(\frac{d}{\sqrt{2\pi e(d-1)}} \right) * \left(1 + \frac{1}{d-1} \right)^{(d-1)/2}$$

We modify the power of the same term, to show it as

$$= 2\alpha * \left(\frac{d}{\sqrt{2\pi e(d-1)}}\right) * \left(\left(1 + \frac{1}{d-1}\right)^{(d-1)}\right)^{1/2}$$

Recall that

$$e = \lim_{n\to\infty} (1 + 1/n)^n$$

Thus this term is bounded with $\sqrt{e}$

$$\leq 2\alpha * \left(\frac{d}{\sqrt{2\pi e(d-1)}}\right) * \sqrt{e}$$

We reformulate as

$$= \alpha\sqrt{d}\sqrt{\frac{2d}{\pi(d-1)}}$$

Since $\sqrt{\frac{d}{d-1}} \leq 2$ for $d \geq 2$

$$\leq \frac{2\sqrt{2}}{\pi}\alpha\sqrt{d}$$

Given that all of this only holds asymptotically; we might need another multiplicative constant to make it hold in general. Hence the constant C in the theorem.

$$p \leq C\alpha\sqrt{d}$$

This bounds the probability $p = \mathbb{P}(\alpha \leq |y_1|) \leq C\alpha\sqrt{d}$. Considering the probability of the complement event the bounds $1 - \mathbb{P}(\alpha > |y_1|) \leq C\alpha\sqrt{d}$ can be stated as

$$1 - C\alpha\sqrt{d} \leq \mathbb{P}(\alpha > |y_1|).$$

Remark 2.2. *Notice that in the previous result essentially shows also that, independently of the dimension d, the* $x_1 = \langle x, u_1\rangle$ *component of a random unit vector* x *with respect to any orthonormal basis* $\{u_1, ..., u_d\}^1$ *is bounded away from zero with overwhelming probability.*

Remark 2.3. *Consider the isometric mapping* $(a, b) \to a + bi$ *from* $\mathbb{R}^2$ *to* $\mathbb{C}$*. The previous result extends to random unit vectors in* $\mathbb{C}^d$ *simplify by modifying the statement as follows: The probability that, for a randomly chosen unit vector* $z \in \mathbb{C}^d$, $|z_1| \geq \alpha > 0$ *holds is at least* $1 - C\alpha\sqrt{2d} = 1 - C'\alpha\sqrt{d}$.

It is important to note that remark 2.2 and remark 2.3 holds with any orthonormal basis by rotating it to coincide with the canonical basis.

Theorem 2.4. *Let* $A \in \mathbb{K}^{I\times J}$ *and* $x \in \mathbb{K}^I$*. Let* V *be the space spanned by the left singular vectors of* A *corresponding to singular values greater than* $(1-\epsilon)\sigma_1$*. Let* $m \in \Omega\left(\frac{\ln(d/\epsilon)}{\epsilon}\right)$*. Let* w^* *be the unit vector after* m *iterations of the power method, namely,*

$$w^* = \frac{\left(AA^H\right)^m x}{||\left(AA^H\right)^m x||_2} \tag{1}$$

The probability that w^* *has a component of at most* l*, where* $l \in O\left(\frac{\epsilon}{\alpha d}\right)$*, orthogonal to* V *is at least* $1 - C\alpha\sqrt{d}$ *i.e.* $1 - C\alpha\sqrt{d} < \mathbb{P}\left(\|Proj_{V^\perp}(w^*)\|_2 < l\right)$.

Proof

Let the SVD of A be given by

$$A = \sum_{k=1}^{r} \sigma_k u_k v_k^H$$

If the rank of A is less than $n = |I|$ we complete the orthonormal set of vectors $\{u_1, ..., u_r\}$ into a full orthogonal basis $\{u_1, ..., u_n\}$ of the n-dimensional space. We can expand x in the terms of this basis as

$$x = \sum_{k=1}^{n} \langle x, u_k \rangle u_k$$

We set $\sigma_k = 0$ for $k > r$ so that we can write A as

$$A = \sum_{k=1}^{n} \sigma_k u_k v_k^H$$

It follows that

$$(AA^H)^m x = \sum_{k=1}^{n} \sigma_k^{2m} u_k u_k^H x = \sum_{k=1}^{n} \sigma_k^{2m} u_k \langle x, u_k \rangle$$

By lemma 2.1, remark 2.2 and remark 2.3 one

has $|\langle x_1, u_1 \rangle| \geq \alpha > 0$ with probability at least $1 - C\alpha\sqrt{d}$. We choose r_ϵ such that $\sigma_1, ..., \sigma_{r_\epsilon}$ are the singular values of A that are greater or equal to $(1-\epsilon)\sigma_1$ and $\sigma_{r_\epsilon+1}, ..., \sigma_n$ are those that are less than $(1-\epsilon)\sigma_1$. Notice that $V = \text{span}\{\sigma_1, ..., \sigma_{r_\epsilon}\}$ and $V^\perp = \text{span}\{\sigma_{r_{\epsilon+1}}, ..., \sigma_n\}$. The component of w^* orthogonal to $V^\perp$ is $\text{Proj}_{V^\perp}(w^*)$ which can be written as

$$\text{Proj}_{V^\perp}(w^*) = \frac{\text{Proj}_{V^\perp}\left(\left(AA^H\right)^m x\right)}{\|\left(AA^H\right)^m x\|_2} \tag{2}$$

We find denominator of eq. (2) by Pythagoras-Fourier theorem

$$||(AA^H)^m x||_2^2 = \sum_{k=1}^{n} \sigma_k^{4m} |\langle x, u_k \rangle|^2 \tag{3}$$

$$\sum_{k=1}^{n} \sigma_k^{4m} |\langle x, u_k \rangle|^2 \geq \sigma_1^{4m} |\langle x, u_1 \rangle|^2 \geq \sigma_1^{4m} \alpha^2 \tag{4}$$

with probability at least $1 - C\alpha\sqrt{d}$. To find the nominator of eq. (2),

we check component of $(AA^H)^m x$ that is orthogonal to $V = \mathit{span}\{u_1, \ldots, u_{r_\epsilon}\}$, namely,

$$\text{Proj}_{V^\perp}\left(\left(AA^H\right)^m x\right) = \sum_{k=1}^{n} \sigma_k^{2m} |\langle x, u_k \rangle|_2 = \sum_{k=r_\epsilon+1}^{n} \sigma_k^{2m} |\langle x, u_k \rangle|_2 \tag{5}$$

$$\text{Proj}_{V^\perp}\left(\left(AA^H\right)^m x\right) \leq (1-\epsilon)^{2m} \sigma_1^{2m} \sum_{k=r_\epsilon+1}^{n} |\langle x, u_k \rangle|_2 \leq (1-\epsilon)^{2m} \sigma_1^{2m} \tag{6}$$

since $\sum_{k=r_\epsilon+1}^{n} \leq ||x||_2^2 = 1$ and $(1-\epsilon)\sigma_1 > \sigma_k$ for $r_\epsilon < k$.

By using eq. (3) and eq. (5) we find squared norm of the component of w^* orthogonal to V, that is $\|\text{Proj}_{V^\perp}(w^*)\|_2^2$, as

$$\|\text{Proj}_{V^\perp}(w^*)\|_2^2 = \frac{\sum_{k=r_\epsilon+1}^{n} \sigma_k^{4m} |\langle x, u_k \rangle|^2}{\sum_{k=1}^{n} \sigma_k^{4m} |\langle x, u_k \rangle|^2}$$

We bound this term by using the relations eq. (4) and eq. (6)

$$\|\mathrm{Proj}_{V^\perp}(w^*)\|_2^2 = \frac{\sum_{k=r_\epsilon+1}^{n} \sigma_k^{4m} |\langle x, u_k\rangle|^2}{\sum_{k=1}^{n} \sigma_k^{4m} |\langle x, u_k\rangle|^2} \le \frac{(1-\epsilon)^{4m}\sigma_1^{4m}}{\alpha^2 \sigma_1^{4m}} = \frac{(1-\epsilon)^{4m}}{\alpha^2}$$

Thus, by taking the square root we have

$$\|\mathrm{Proj}_{V^\perp}(w^*)\|_2 \le \frac{(1-\epsilon)^{2m}}{\alpha}$$

In terms of *Big O* notation we have

$$\|\mathrm{Proj}_{V^\perp}(w^*)\|_2 \in \mathcal{O}\left(\frac{(1-\epsilon)^{2m}}{\alpha}\right)$$

Notice that $1-\epsilon$ is a linear approximation of $e^{-\epsilon}$. Similarly, $(1-\epsilon)^{2m}$ approximates e^{-2m} for small ϵ. Using this approximation,

$$\|\mathrm{Proj}_{V^\perp}(w^*)\|_2 \in \mathcal{O}\left(\frac{e^{-2\epsilon m}}{\alpha}\right)$$

Recall that $m \in \Omega\left(\frac{\ln(d/\epsilon)}{\epsilon}\right)$. This is another way of saying there exists m_0 and some constant $c > 0$ such that $m \ge c\frac{\ln(d/\epsilon)}{d}$ for all $m > m_0$. Similarly, this also means there exists m_0 and some constant $c > 0$ such that $-\frac{m}{c} \le -\frac{\ln(d/\epsilon)}{d}$ for all $m > m_0$. Since exponentiation is a non-decreasing function $e^{-\frac{m}{c}} \le e^{-\frac{\ln(d/\epsilon)}{\epsilon}} = e^{\frac{\ln(\epsilon/d)}{\epsilon}} = (\epsilon/d)^{1/\epsilon}$. We have

$$e^{-\frac{m}{c}} \le (\epsilon/d)^{1/\epsilon}$$

for some constant $c > 0$ and $m > m_0$. We take the power of ϵ of both sides

$$e^{-\frac{m\epsilon}{c}} \le \frac{\epsilon}{d}$$

Let $c_1 = c/2$

$$e^{-\frac{2m\epsilon}{c_1}} \le \frac{\epsilon}{d}$$

For some constant $e^{-1/c_1} > 0$ and all $m > m_0$. We divide both sides with α

$$\frac{e^{-\frac{2m\epsilon}{c_1}}}{\alpha} \le \frac{\epsilon}{\alpha d}$$

$$\alpha^{-1} e^{-\frac{2m\epsilon}{c_1}} \le \frac{\epsilon}{\alpha d}$$

Which means

$$\frac{e^{-2m\epsilon}}{\alpha} \in \mathcal{O}\left(\frac{\epsilon}{\alpha d}\right)$$

Consequently

$$\|\mathrm{Proj}_{V^\perp}(w^*)\|_2 \in \mathcal{O}\left(\frac{\epsilon}{\alpha d}\right)$$

2.6 Stability of the Singular Value Decomposition

Common scenario: data matrix of interest A, but one has only a perturbed version of it $\tilde{A} = A + E$ where E is going to be a relatively small quantifiable perturbation.

We start by considering Hermitian matrices $A \in \mathbb{K}^{I\times I}$, i.e. $A = A^H$. Recall that a nonzero vector $v \in K^I$ is an eigenvector of A if $Av = \lambda v$ for some scalar $\lambda \in \mathbb{K}$ called the corresponding eigenvalue. The following theorem establishes that Hermitian matrices have real eigenvalues and orthogonal eigenvectors.

Theorem 2.5 (Spectral theorem for Hermitian matrices)**.** *If $A \in K^{I\times I}$ and $A = A^H$, then there exists an orthonormal basis $\{v_1, ..., v_n\}$ consisting of eigenvectors of A with real corresponding eigenvalues $\{\lambda_1, ..., \lambda_n\}$ such that*

$$A = \sum_{k=1} \lambda_k v_k {v_k}^H$$

This representation is called the spectral decomposition of A.

As a consequence of the spectral theorem, for Hermitian matrices, singular value decomposition and eigenvalue decomposition are closely related. Indeed, by denoting $u_k = v_k \operatorname{sign} \lambda_k$ and $\sigma_k = |\lambda_k|$, then

$$A = \sum_{k=1}^{n} \lambda_k v_k v_k^H = \sum_{k=1}^{n} \operatorname{sign} \lambda_k |\lambda_k| v_k v_k^H = \sum_{k=1}^{n} \sigma_k u_k v_k^H,$$

which is actually the SVD of A. Thus the SVD agrees with the spectral decomposition up to signs. A first step towards analyzing stability is hence to study stability of the spectral decomposition.

Weyl's Bounds

Assume $\tilde{A} = \tilde{A}^H$ is Hermitian and hence also E. As the eigenvalues of A are real, we can order in a non-increasing order $\lambda_1 \geq \lambda_2 \geq \cdot\cdot \geq \lambda_n$. Recall that we cannot assume that the eigenvalues are positive. Let's show

$$\lambda_1(\tilde{A}) = \max_{|v|_2^2=1} v^H(\tilde{A})v$$

Let $v = \tilde{v}_1$ from the spectral decomposition of $\tilde{A}$.

$$\tilde{v_1}^H \tilde{A}\tilde{v}_1 \leq \max_{|v|_2^2=1} v^H(\tilde{A})v$$
$$\tilde{\lambda}_1 \leq \max_{|v|_2^2=1} v^H(\tilde{A})v$$

Now consider the maximizer v^* decomposed into orthonormal basis vectors i.e. $v^* = \sum_j \alpha_j \tilde{v}_j$ where $\sum_j \alpha_j^2 = 1$. Let's plug it in

$$\max_{|v|_2^2=1} \left(\sum_j \overline{\alpha}_j \tilde{v}_j^H\right) \tilde{A} \left(\sum_j \alpha_j \tilde{v}_j\right)$$

Let $\tilde{A} = \sum_j \tilde{\lambda}_j \tilde{v}_j \tilde{v}_j^H$ be the spectral decomposition. Let's plug it in the equation.

$$\max_{|v|_2^2=1} \left(\sum_j \overline{\alpha}_j \tilde{v}_j^H\right) \sum_j \tilde{\lambda}_j \tilde{v}_j \tilde{v}_j^H \left(\sum_j \alpha_j \tilde{v}_j\right)$$

When the orthogonal vectors cancel out we will have the following

$$\max_{|v|_2^2=1} \sum_j \alpha_j^2 \tilde{\lambda}_j$$

Notice that

$$\sum_j \alpha_j^2 \tilde{\lambda}_j \leq \sum_j \alpha_j^2 \tilde{\lambda}_1 = \tilde{\lambda}_1$$

since $\sum_j \alpha_j^2 = 1$. Hence we have

$$\lambda_1(\tilde{A}) = \max_{|v|_2^2=1} v^H(\tilde{A})v$$

$$\lambda_1(\tilde{A}) = \max_{|v|_2^2=1} v^H(A+E)v$$

$$\max_{|v|_2^2=1} v^H(A+E)v \leq \max_{|v|_2^2=1} v^H(A)v + \max_{|v|_2^2=1} v^H(A)v$$

Thus we have

$$\lambda_1(\tilde{A}) \leq \lambda_1(A) + \lambda_1(E)$$

Theorem 2.6 (Weyl)**.** *If $A, E \in \mathbb{K}^{I\times I}$ are two Hermitian matrices, then for all $k = 1, ..., n$*

$$\lambda_k(A) + \lambda_n(E) \leq \lambda_k(A+E) \leq \lambda_k(A) + \lambda_1(E).$$

3 Basics on Probability

3.1 Motivation and single random variables

Motivating example

Power method for computing singular vectors requires us to choose a point at random on the sphere. But there are infinitely many points on the sphere Yet the sphere is too "small" to proceed via a density on the entire space. We model a random point on S^1 using the following random process. Recall that $S^n = \left\{x \in \mathbb{R}^{n+1} : ||x|| = 1\right\}$.
Consider an imaginary player X of darts, with no experience, throwing darts at random at the two dimensional disk $\Omega = B_r$ radius r and center O. At every point $\omega \in \Omega$ within the target hit by a dart we assign a point on the boundary of the target $X(\omega) = \frac{\omega}{||\omega||_2} \in S^1$. S^1 can be parameterized by the angle $\theta \in [0, 2\pi)$. Hence we consider the player X as a map from $\omega \in \Omega$ to θ

$$X : \Omega \to \mathbb{R}$$

Calculating probabilities

What is the probability of the event

$$B = \{w \in \Omega : \theta_1 \leq X(\omega) \leq \theta_2\}?$$

Expectations

For N attempts of the player the empirical average will be

$$\frac{1}{N}\sum_{i=1}^{N} X(\omega_i)$$

As $N \to \infty$ this yields the expectation

$$\mathbb{E}X := \frac{1}{2\pi}\int_0^{2\pi} \theta d\theta = \pi$$

Abstract probability theory

A probability space is given by a triplet $(\Omega, \Sigma, \mathbb{P})$. Where the sample space Ω is the set on which the probability is defined. Σ is the σ-algebra (a family of subsets of Ω) and defines the admissible events. $\mathbb{P}$ is the probability measure on (Ω, Σ), that is, it assigns to each event $B \in \Sigma$ to a value $\in [0, 1]$, the probability of the event through

$$\mathbb{P}(B) = \int_B d\mathbb{P}(\omega) = \int_\Omega 1_B(\omega) d\mathbb{P}(\omega)$$

The union bound

Consequence of the properties of a measure: For two disjoint events $B_1, B_2, B_1 \cap B_2 = \emptyset$

$$\mathbb{P}(B_1 \cup B_2) = \mathbb{P}(B_1) + \mathbb{P}(B_2)$$

Theorem 3.1. *The union bound (or Bonferroni's inequality, or Boole's inequality) states that for a collection of events $B_l \in \Sigma$, $l = 1, ..., n$, we have*

$$\mathbb{P}\left(\bigcup_{l=1}^{n} B_l\right) \leq \sum_{l=1}^{n} \mathbb{P}\left(B_l\right). \tag{7}$$

Proof

For two sets B_1 and B_2: Notice that these two sets are equal

$$B_1 \cup B_2 = (B_1 \setminus B_2) \cup B_2$$

If we write the probabilities of these events

$$\mathbb{P}\left(B_1 \cup B_2\right) = \mathbb{P}\left((B_1 \setminus B_2) \cup B_2\right)$$

Notice that $(B_1 \setminus B_2) \cap B_2 = \emptyset$

$$\mathbb{P}\left((B_1 \setminus B_2) \cup B_2\right) = \mathbb{P}\left(B_1 \setminus B_2\right) + \mathbb{P}\left(B_2\right)$$

Since $B_1 \setminus B_2 \subseteq B_1$ we can write the proof in one line as

$$\mathbb{P}\left(B_1 \cup B_2\right) = \mathbb{P}\left((B_1 \setminus B_2) \cup B_2\right) = \mathbb{P}\left(B_1 \setminus B_2\right) + \mathbb{P}\left(B_2\right) \leq \mathbb{P}(B_1) + \mathbb{P}(B_2)$$

The bound is obtained by the fact $\mathbb{P}\left(B_1 \setminus B_2\right) \leq \mathbb{P}\left(B_1\right)$. Assume for $n-1$

$$\mathbb{P}\left(\bigcup_{l=1}^{n-1} B_l\right) \leq \sum_{l=1}^{n-1} \mathbb{P}\left(B_l\right). \tag{8}$$

Let $\bigcup_{l=1}^{n-1} B_l = A$.

$$\mathbb{P}\left(A\right) \leq \sum_{l=1}^{n-1} \mathbb{P}\left(B_l\right). \tag{9}$$

Consider the union $A \cup B_n$

$$\mathbb{P}\left((A \setminus B_n) \cup B_n\right) = \mathbb{P}\left(A \setminus B_n\right) + \mathbb{P}(B_n)$$

Notice that $\mathbb{P}\left(A \setminus B_n\right) \leq \mathbb{P}\left(A\right)$

$$\mathbb{P}\left((A \setminus B_n) \cup B_n\right) \leq \mathbb{P}\left(A\right) + \mathbb{P}(B_n)$$

Using the bound from eq. (9) we have

$$\mathbb{P}\left((A \setminus B_n) \cup B_n\right) \leq \mathbb{P}\left(A\right) + \mathbb{P}(B_n) \leq \sum_{l=1}^{n-1} \mathbb{P}\left(B_l\right) + \mathbb{P}(B_n)$$

Hence we have

$$\mathbb{P}\left((A \setminus B_n) \cup B_n\right) = \mathbb{P}\left(\bigcup_{l=1}^{n} B_l\right) \leq \sum_{l=1}^{n} \mathbb{P}\left(B_l\right).$$

Random variables

A random variable X is a real-valued measurable function on (Ω, Σ). **Recall**: X is called measurable if the preimage

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$$

is contained in Σ for all Borel measurable subsets $A \subset \mathbb{R}$. This means reasonable events (values X can take) are contained in the σ-algebra.

Densities

The distribution function $F = F_X$ of X is defined as

$$F(t) = \mathbb{P}(X \leq t), \quad t \in \mathbb{R}. \tag{10}$$

A random variable X possesses a *probability density* function $\phi : \mathbb{R} \to \mathbb{R}_+$ if

$$\mathbb{P}(a < X \leq b) = \int_a^b \phi(t)\,\mathrm{d}t \quad \text{for all } a < b \in \mathbb{R} \tag{11}$$

Then the density function $\phi = \frac{dF(t)}{t}$. Note that not every random variable has a density.For example X with $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 0.5$. $F_X(t)$ is a partial function and isn't continuous.

Expectations revisited

The probability measure associated to a density ϕ is then given by $d\mathbb{P} = \varphi(\theta)d\theta$. Consequently, we can compute for a function g

$$\mathbb{E}g(X) := \int_\Omega g(X(\omega))\mathrm{d}\mathbb{P}(\omega) = \int_\mathbb{R} g(\theta)\varphi(\theta)d\theta.$$

The probability of an event $E = \{X \in A\}$ satisfies $\mathbb{P}(E) = \mathbb{E}1_E$ and hence

$$\mathbb{P}(E) = \int_A \varphi(\theta)d\theta$$

Moments

$\mathbb{E}X^p$ for $p > 0$ are called moments of X, while $\mathbb{E}|X|^p$ are called absolute moments. The quantity $\mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$ is called variance. For $1 \leq p \leq \infty$, $(\mathbb{E}|X|^p)^{1/p}$ defines a norm on the $L^p(\Omega, \mathbb{P})$-space of all p-integrable random variables, in particular, the triangle

$$(\mathbb{E}|X + Y|^p)^{1/p} \leq (\mathbb{E}|X|^p)^{1/p} + (\mathbb{E}|Y|^p)^{1/p} \tag{12}$$

holds for all p-integrable random variables X,Y on $(\Omega, \Sigma, \mathbb{P})$. Here, p-integrable random variables are random variables that have bounded p-th absolute moments.

Important results about random variables

Hoelder's inequality states that, for random variables X, Y on a common probability space and $p, q \geq 1$ with

$$\frac{1}{p} + \frac{1}{q} = 1$$

we have

$$|\mathbb{E}XY| \leq (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|Y|^q)^{1/q}$$

Let $X_n, n \in \mathbb{N}$, be a sequence of random variables such that X_n converges to X as $n \to \infty$ in the sense that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ almost surely (a.s.).

Theorem 3.2 (Lebesgue's dominated convergence theorem)**.** *If there exists a random variable Y with $\mathbb{E}|Y| < \infty$ such that $|X_n| < |Y|$ then almost surely $\lim_{n\to\infty} \mathbb{E}X_n = \mathbb{E}X$.*

Moment computations & Cavalieri's formula

Proposition 3.3. *The absolute moments of a random variable X can be expressed as*

$$\mathbb{E}|X|^p = p\int_0^\infty \mathbb{P}(|X| \geq t)t^{p-1}dt \quad p > 0. \tag{13}$$

Important tool for the proof: Fubini's theorem

Let $f : A \times B \to C$ be measurable, where (A, ν) and (B, μ) are measurable spaces. If $\int_{A\times B} |f(x, y)|d(\nu \otimes \mu)(x, y) < \infty$ then

$$\int_A \left(\int_B f(x,y)\, d\mu(y)\right) d\nu(x) = \int_B \left(\int_A f(x,y)\, d\nu(x)\right) d\mu(y)\,.$$

Proof

Using Fubini's theorem we derive

$$\begin{aligned}\mathbb{E}|X|^p &= \int_\Omega |X(\omega)|^p d\mathbb{P}(\omega)\\ &= \int_\Omega \int_0^{|X(\omega)|^p} 1dx\, d\mathbb{P}(\omega)\\ &= \int_\Omega \int_0^\infty \mathbb{1}_{|X(\omega)|^p > x} dx\, d\mathbb{P}(\omega)\\ &= \int_0^\infty \int_\Omega \mathbb{1}_{|X(\omega)|^p > x} d\mathbb{P}(\omega)\ dx\\ &= \int_0^\infty \mathbb{P}(|X|^p \geq x)\ dx\end{aligned}$$

Let $t^p = x$, we use the change of variables trick $p\, t^{p-1}\, dt = dx$

$$\begin{aligned}&\int_0^\infty \mathbb{P}(|X|^p \geq x)\ dx\\ &= p\int_0^\infty \mathbb{P}(|X|^p \geq t^p)\, t^{p-1}\, dt\\ &= p\int_0^\infty \mathbb{P}(|X| \geq t)\, t^{p-1}\, dt\end{aligned}$$

Hence we have proved proposition 3.3.

Corollary 3.4. *For a random variable X the expectation satisfies*

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X \geq t)\, dt - \int_0^\infty \mathbb{P}(X \leq -t)\, dt\,. \tag{14}$$

Proof

Write $X = X_+ + X_-$ where $X_+ = X1_{X\geq 0}$ and $X_- = X1_{X<0}$. Consequently $\mathbb{E}X = \mathbb{E}X_+ + \mathbb{E}X_- = \mathbb{E}|X_+| - \mathbb{E}|X_-|$. Applying proposition 3.3 to both yields (for $p = 1$) the corollary 3.4.

Tail bounds and Markov inequality

The function $t \to \mathbb{P}(|X| \geq t)$ is called the **tail** of X. The tail can be estimated by expectations and moments via the Markov inequality.

Theorem 3.5. *Let X be a random variable. Then*

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}|X|}{t} \text{ for all } t > 0. \tag{15}$$

Proof

Consider the term

$$t\mathbb{P}\left(|X| \geq t\right) \tag{16}$$

since $\mathbb{P}\left(|X| \geq t\right) = \mathbb{E}\mathbb{1}_{|X|\geq t}$ this term eq. (16)
can be written as

$$t\mathbb{P}\left(|X| \geq t\right) = t\mathbb{E}\left[\mathbb{1}_{|X|\geq t}\right] = \mathbb{E}\left[t\mathbb{1}_{|X|\geq t}\right] \tag{17}$$

Notice that

$$t\mathbb{1}_{|X|\geq t} \leq |X|$$

always holds. This also means

$$\mathbb{E}\left[t\mathbb{1}_{|X|\geq t}\right] \leq \mathbb{E}|X|$$

Notice that left hand side is same as eq. (16). Hence we have

$$\begin{aligned} t\mathbb{P}\left(|X| \geq t\right) &\leq \mathbb{E}|X| \\ \mathbb{P}\left(|X| \geq t\right) &\leq \frac{\mathbb{E}|X|}{t}. \end{aligned}$$

Remark 3.6. *As an important consequence we note that for $p > 0$*

$$\mathbb{P}(|X| \geq t) = \mathbb{P}(|X|^p \geq t^p) \leq t^{-p}\mathbb{E}|X|^p \quad \textit{for all} \quad t > 0$$

The special case $p = 2$ is referred to as the ***Chebyshev*** *inequality.*

Remark 3.7. *For $\theta > 0$ we obtain that for all $t \in \mathbb{R}$*

$$\mathbb{P}(X \geq t) = \mathbb{P}(\exp(\theta X) \geq \exp(\theta t)) \leq \exp(-\theta t)\mathbb{E}\exp(\theta X)\,.$$

The function $\theta \to \mathbb{E}\exp(\theta X)$ is usually called the ***Laplace transform*** *or the* ***moment generating function*** *of X.*

Gaussian Random Variables

A normally distributed random variable or Gaussian random variable X has probability density function

$$\psi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right).$$

Its distribution is often denoted by $\mathcal{N}(\mu, \sigma)$. It has mean $\mathbb{E}X = \mu$ and variance $\mathbb{E}(X-\mu)^2 = \sigma^2$.