# Concepts of Measure and Probability Theory

J.C. Keil | j@polyhelion.com

Palo Alto, CA, USA

September - November 2023

# Table of contents

# 1. Introduction

These slides are the enhanced version of a couple of talks I gave September to November 2023 in the course of a workshop on stochastic processes and machine learning. They are intended to give a structured overview of the measure and probability theory fundamentals to get a rigorous mathematical understanding of the topic. All errors are my own.

Have yet to find a convincing way to integrate the handwritten examples/calculations I did during the talks into the slides.

# 2. Measure Spaces

# Measure spaces

- Motivation: Volume does not work as a measure for general subsets of $\mathbb{R}^d$ (see Vitali sets). What sets should be measurable ? Properties ?
- Consider a set $\Omega$ and its power set $\mathcal{P}(\Omega)$. A subset $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is called a $\sigma$-algebra if
  - $\Omega \in \mathcal{A}$
  - $A \in \mathcal{A} \implies A^{\complement} \in \mathcal{A}$
  - $A_i \in \mathcal{A}$, $i \in I$ with $I$ countable $\implies \bigcup_I A_i \in \mathcal{A}$
- Immediate: $\varnothing \in \mathcal{A}$ and $\bigcap_I A_i \in \mathcal{A}$
- Elements of $\mathcal{A}$ are called measurable sets.
- $\{\varnothing, \Omega\}$ is the smallest $\sigma$-algebra, the largest one is $\mathcal{P}(\Omega)$.
- $\{\varnothing, A, A^{\complement}, \Omega\} = \sigma_\Omega(A)$ is the $\sigma$-algebra, generated by a $A \in \mathcal{A}$.
- A $\sigma$-algebra is either finite or uncountable (Axiom of choice needed to prove that).

# Measure spaces

- The ordered pair $(\Omega, \mathcal{A})$ is called a measurable space.
- Call a (countable) partition of an element $A \in \mathcal{A}$ a countable collection $\{A_i\}_{i \in I}$ of pairwise disjoint $A_i \in \mathcal{A}$ s.t. $A = \bigcup_{i \in I} A_i$. Notation: $A = \bigsqcup_{i \in I} A_i$.
- A function on a $\sigma$-algebra $\nu : \mathcal{A} \to \overline{\mathbb{R}} \equiv [-\infty, +\infty]$ is called a (countably additive) signed measure if
  - $\nu(\varnothing) = 0$
  - $\nu$ assumes at most one of the values $+\infty$ and $-\infty$ $(+\infty + (-\infty)$ is undefined)
  - If $A \in \mathcal{A}$, it holds for every (countable) partition $A = \bigsqcup_{i \in I} A_i$ that
$$\nu(\bigsqcup_{i \in I} A_i) = \sum_{i \in I} \nu(A_i)$$
(If the left-hand side has finite value, the series on the right converges absolutely.)

# Measure spaces

- Denote by $\mathcal{M}(\Omega, \mathcal{A})$ the set of signed measures and by $\mathcal{M}_f(\Omega, \mathcal{A}) \subset \mathcal{M}(\Omega, \mathcal{A})$ its subset of finite signed measures, i.e. measures $\nu \in \mathcal{M}(\Omega, \mathcal{A})$ s.t. $|\nu(\omega)| < +\infty \ \ \forall \omega \in \Omega$. While the former set is not an $\mathbb{R}$-vector space (because it is not closed under addition), the latter is.
- Be $\mathcal{M}^+(\Omega, \mathcal{A}) \subseteq \mathcal{M}(\Omega, \mathcal{A})$ the subset of nonnegative measures i.e. measures with the property $\nu(A) \in \overline{\mathbb{R}}_+ \equiv [0, +\infty]$ for all $A \in \mathcal{A}$. It is a convex pointed cone.
- Notation: A '$+$' sub- or superscript will always mean 'nonnegative' i.e. $\geq 0$.
- If $\nu \in \mathcal{M}^+(\Omega, \mathcal{A})$, it has some additional useful properties:
  Monotonicity: $A, A' \in \mathcal{A}$ with $A \subseteq A'$ implies $\nu(A) \leq \nu(A')$
  $\sigma$ - Subadditivity: For $A_i \in \mathcal{A}$, $i \in I$ with $I$ countable, there is

$$\nu(\bigcup_I A_i) \leq \sum_I \nu(A_i)$$

# Measure spaces

- The support of a measure $\nu$ is defined as

$$\text{supp}(\nu) \equiv \{\omega \in \Omega \mid U_\omega \text{ is an open neighbourhood of } \omega \implies \nu(U_\omega) > 0\}$$

- A measure $\nu$ is called degenerate if $\dim \text{supp}(\nu) < \dim \Omega$, nondegenerate otherwise.
- A measure $\nu$ is called a finite measure if $\nu(\Omega) < +\infty$ ($|\nu(\Omega)| < +\infty$ if it is signed). $\nu$ is called a probability measure if $\nu(\Omega) = 1$ and it has $[0, 1]$ as the range of values.
- A measure $\nu$ on a measurable space $(\Omega, \mathcal{A})$ is called $\sigma$-finite if there exists a countable cover $\Omega = \bigcup_I A_i$ with $A_i \in \mathcal{A}$ and $\nu(A_i) < +\infty$.
- The counting measure $\# : A \mapsto |A|$ is finite if $\Omega$ is finite and $\sigma$-finite if $\Omega$ is countable.

# Measure spaces

- Denote by $\mathcal{M}^+_\sigma(\Omega, \mathcal{A})$, $\mathcal{M}^+_f(\Omega, \mathcal{A})$, $\mathcal{M}^+_1(\Omega, \mathcal{A})$ the sets of $\sigma$-finite, finite and probability measures respectively.

$$\mathcal{M}^+_1(\Omega, \mathcal{A}) \subseteq \mathcal{M}^+_f(\Omega, \mathcal{A}) \subseteq \mathcal{M}^+_\sigma(\Omega, \mathcal{A}) \subseteq \mathcal{M}^+(\Omega, \mathcal{A}) \subseteq \mathcal{M}(\Omega, \mathcal{A})$$

- $\mathcal{M}^+(\Omega, \mathcal{A})$, $\mathcal{M}^+_\sigma(\Omega, \mathcal{A})$ and $\mathcal{M}^+_f(\Omega, \mathcal{A})$ are pointed convex cones.
- Let $(\Omega, \mathcal{A})$ be a measurable space. Together with a measure $\nu \in \mathcal{M}^+(\Omega, \mathcal{A})$ it is called a measure space, denoted by $(\Omega, \mathcal{A}, \nu)$. If $\nu \in \mathcal{M}^+_1(\Omega, \mathcal{A})$, $(\Omega, \mathcal{A}, \nu)$ is called a probability space.
- $\Omega$ is called the sample space of the probability space $(\Omega, \mathcal{A}, \nu)$ and $\mathcal{A}$ its event space.

# Measure spaces

- For a signed measure $\nu \in \mathcal{M}(\Omega, \mathcal{A})$ define its variation $|\nu|$:

$$|\nu|(A) \equiv \sup \left( \sum |\nu(A_i)| \right)$$

  where the supremum runs across all partitions $\{A_i\}$ of $A$. Obviously is $|\nu(A)| \leq |\nu|(A)$.
- $|\nu|$ is a nonnegative measure on $\mathcal{A}$, i.e. $|\nu| \in \mathcal{M}^+(\Omega, \mathcal{A})$. It is finite if $\nu$ is.
- $\mathcal{M}_f(\Omega, \mathcal{A})$ can be made a Banach space with the total variation norm $\|\nu\| \equiv |\nu|(\Omega)$.

# Measure spaces

- Be $E$ an $\mathbb{R}$-Banach space. An $E$-valued (countably additive) vector measure on a measurable space $(\Omega, \mathcal{A})$ is a generalization of a finite measure. It is a function $\nu : \mathcal{A} \to E$ s.t.
  - $\nu(\emptyset) = 0$
  - If $A \in \mathcal{A}$, it holds for every (countable) partition $A = \bigsqcup_{i \in I} A_i$ that
  $$\nu(\bigsqcup_{i \in I} A_i) = \sum_{i \in I} \nu(A_i)$$
  (The series on the right converges in norm of $E$.)
- The $E$-valued vector measures w.r.t. $(\Omega, \mathcal{A})$ form an $\mathbb{R}$-vector space, denoted by $\mathcal{M}(\Omega, \mathcal{A}; E)$.

# Measure spaces

- The variation of an $E$-valued vector measure $\nu \in \mathcal{M}(\Omega, \mathcal{A}; E)$:

$$|\nu|(A) \equiv \sup \left( \sum \|\nu(A_i)\|_E \right)$$

  the supremum runs across all partitions $\{A_i\}$ of $A$.
- $\|\nu(A)\|_E \leq |\nu|(A)$ and $|\nu| \in \mathcal{M}^+(\Omega, \mathcal{A})$.
- $|\nu|$ is not necessarily finite like in the case $E = \mathbb{R}$. If it is, $\nu$ is called of bounded variation.
- The subspace $\mathcal{M}_b(\Omega, \mathcal{A}; E) \subseteq \mathcal{M}(\Omega, \mathcal{A}; E)$ of measures of bounded variation can be made a Banach space with the total variation norm $\|\nu\| \equiv |\nu|(\Omega)$.
- $\mathcal{M}_b(\Omega, \mathcal{A}; \mathbb{R}) = \mathcal{M}(\Omega, \mathcal{A}; \mathbb{R}) = \mathcal{M}_f(\Omega, \mathcal{A})$

# Measure spaces

- If not explicitly stated otherwise, all measures are assumed nonnegative and real-valued.
- A measure $\nu$ on a measurable space $(\Omega, \mathcal{A})$ is called singular w.r.t a measure $\mu$, notation $\nu \perp \mu$, if there exists a set $A \in \mathcal{A}$ such that $\nu(A^{\complement}) = 0$ and $\mu(A) = 0$.
- $\nu$ is called (singular) discrete (w.r.t. $\mu$) if in addition $A$ is at most countable and all $\{a\}$ for every $a \in A$ are measurable. singular continuous (w.r.t. $\mu$) otherwise.
- Let $\delta_\omega$ be the Dirac measure, assigning a measure of 1 to any set containing $\omega$ and a measure of 0 to any other set. Then a measure $\nu$ is singular discrete (w.r.t. $\mu$) in the above sense if $\nu = \sum_i c_i \, \delta_{a_i}$ for some countable $A = \{a_1, a_2, \dots\} \in \mathcal{A}$ and $\mu(\{a_i\}) = 0$.
- The Cantor measure is an example of a singular continuous measure w.r.t. the Lebesgue measure (see below).

# Measure spaces

- A measurable set $A \in \mathcal{A}$ is called a $\nu$-atom if $\nu(A) > 0$ and if it holds for every $A' \in \mathcal{A}$ with $A' \subset A$ (and therefore $\nu(A') < \nu(A)$) that either $\nu(A') = 0$ or $\nu(A \setminus A') = 0$.
- A measure $\nu$ is called purely atomic or simply atomic if every $A \in \mathcal{A}$ with $\nu(A) > 0$ contains a $\nu$-atom. A measure which has no atoms is called nonatomic, atomless or diffuse.
- Every measure can be shown to be the sum of an atomic and a nonatomic measure. If the measure is $\sigma$-finite, this representation is unique.
- A $\sigma$-finite atomic measure $\nu$ has only countably many atoms. It is called a discrete measure if it is a countable, positive weighted sum of Dirac measures

$$\nu = \sum_{i \in I} m_i \, \delta_{\omega_i} \text{ with } \omega_i \in \Omega$$

The $\omega_i$ are common points of the atoms (mod $\nu$-zero). As intersections of atoms the $\{\omega_i\}$ are elements of $\mathcal{A}$.

# Measure spaces

- Assume all singletons $\{\omega\}, \omega \in \Omega$, are elements of $\mathcal{A}$ (This is the case for Borel $\sigma$-algebras $\mathcal{B}(\Omega)$ (see below) for example). A measure $\nu$ on $\mathcal{A}$ is called a continuous measure if $\nu(\{\omega\}) = 0$ for every $\omega$. A nonatomic measure is necessarily continuous but not every continuous measure is nonatomic.
- The counting measure $\#$ on a measurable space $(S, \mathcal{P}(S))$ ($S$ a countable set) is a discrete measure. Atoms and singletons are equivalent.
- If $\Omega$ is a separable metric space and $\nu$ a $\sigma$-finite measure on $\mathcal{B}(\Omega)$, every atom is a union of a singleton with measure greater zero and a null set.

# Measure spaces

- A continuous Radon measure (definition see below) on a Borel space $(\Omega, \mathcal{B}(\Omega))$ is nonatomic. In particular $\lambda^d$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.
- For a locally compact Hausdorff space $\Omega$, a general Radon measure $\nu$ on $\mathcal{B}(\Omega)$ has a unique decomposition $\nu = \nu_d + \nu_c$ where $\nu_d$ and $\nu_c$ are a discrete and continuous Radon measures respectively.

# Measure spaces

- For a measure space $(\Omega, \mathcal{A}, \nu)$ be $\mathcal{N}_\nu \subset \mathcal{P}(\Omega)$ the set of all sets contained in a $\nu$-null set. The completion of $\mathcal{A}$ w.r.t. $\nu$ is the $\sigma$-algebra $\widehat{\mathcal{A}}_\nu \equiv \sigma(\mathcal{A} \cup \mathcal{N}_\nu)$. $\mathcal{A} \subseteq \widehat{\mathcal{A}}_\nu$ with $A \mapsto A \cup \varnothing$ (We write $\widehat{\mathcal{A}}$ if the measure is obvious). $\nu$ is called complete if $\mathcal{A} = \widehat{\mathcal{A}}_\nu$ that is if $\mathcal{N}_\nu \subseteq \mathcal{A}$. The extension of $\nu$ to $\widehat{\mathcal{A}}_\nu$, denoted by $\widehat{\nu}$, is a measure defined by $\widehat{\nu}(A \cup N) \equiv \nu(A)$. $(\Omega, \widehat{\mathcal{A}}, \widehat{\nu})$ is called the completion of $(\Omega, \mathcal{A}, \nu)$.
- Having null sets with nonmeasurable subsets can lead to unwanted effects. See the the section on Borel- and Lebesgue-measurable sets below.

# Measure spaces

- For two measures $\nu$ and $\mu$ on a measurable space $(\Omega, \mathcal{A})$ we say $\nu$ is absolutely continuous with respect to $\mu$, notation $\nu \ll \mu$, if $\mu(A) = 0$ for any $A \in \mathcal{A}$ implies $\nu(A) = 0$
- When $\nu \ll \mu$ and $\mu \ll \nu$, we call the measures equivalent and write $\nu \sim \mu$. Equivalent measures agree on which sets have measure zero.
- Every $\sigma$-finite measure $\nu$ is equivalent to a probability measure, so to a finite measure in particular.

# Borel spaces

- For a Hausdorff space $(\Omega, \mathcal{T})$ with $\mathcal{T}$ the topology (i.e. a system of 'open sets') the associated Borel $\sigma$-algebra $\mathcal{B}(\Omega, \mathcal{T})$ is the $\sigma$-algebra generated by $\mathcal{T}$, i.e. $\mathcal{B}(\Omega, \mathcal{T}) \equiv \sigma_\Omega(\mathcal{T})$. $\mathcal{B}(\Omega, \mathcal{T})$ is the smallest $\sigma$-algebra containing $\mathcal{T}$. The Hausdorff property ensures that all compact sets of $\Omega$ are closed and therefore part of $\mathcal{B}(\Omega, \mathcal{T})$.
- Usually the topology of $\Omega$ is understood and $\mathcal{T}$ is omitted. $\mathcal{B}(\Omega, \mathcal{T})$ is denoted by $\mathcal{B}(\Omega)$.
- The measurable space $(\Omega, \mathcal{B}(\Omega))$ is called a Borel space.
- Borel spaces ensure a degree of compatibility between measurable sets and the underlying topology. For the interlocking of measurability and topology see the introduction of related measure properties below.

# Borel spaces

- Intuition: $\mathcal{B}(\Omega, \mathcal{T})$ contains the subsets of $\Omega$ you reasonably want to be able to measure with regards to the topology $\mathcal{T}$.
- Is every $\sigma$-algebra on a space $\Omega$ identical to a $\mathcal{B}(\Omega, \mathcal{T})$ for some topology $\mathcal{T}$ on $\Omega$ ? No, counterexamples can be constructed.
- Notation for spaces of measures: $\mathcal{M}(\Omega) \equiv \mathcal{M}(\Omega, \mathcal{B}(\Omega))$, $\mathcal{M}^+(\Omega) \equiv \mathcal{M}^+(\Omega, \mathcal{B}(\Omega))$, etc.
- An element of $\mathcal{M}^+(\Omega, \mathcal{B}(\Omega))$ is called a Borel measure.

# Borel spaces

- Be $\nu \in \mathcal{M}^+(\Omega)$ a Borel measure.
- $\nu$ is called locally finite if for every point $\omega \in \Omega$ there is an open neighbourhood $U_\omega \in \mathcal{B}(\Omega)$ with $\nu(U_\omega) < +\infty$.
  By definition it implies $\nu(K) < +\infty$ for all compact sets $K \in \mathcal{B}(\Omega)$
- $\nu$ is called inner regular if for every $B \in \mathcal{B}(\Omega)$

$$\nu(B) = \sup\{\nu(K) \mid K \subset B, \ \text{K compact }\}$$

- $\nu$ is called outer regular if for every $B \in \mathcal{B}(\Omega)$

$$\nu(B) = \inf\{\nu(U) \mid B \subset U, \ \text{U open }\}$$

- $\nu$ is called regular if it is both inner and outer regular.

# Borel spaces

- If $\Omega$ is a locally compact Hausdorff space, then locally finite is equivalent to being finite on all compact sets from $\mathcal{B}(\Omega)$.
- A measure $\nu \in \mathcal{M}^+(\Omega)$ is called a Radon measure if it is locally finite and inner regular. The subset $\mathcal{M}_R^+(\Omega) \subset \mathcal{M}^+(\Omega)$ of Radon measures is a pointed convex subcone.
- The Radon measures on a locally compact Hausdorff space are precisely the inner regular Borel measures, finite on compact sets from $\mathcal{B}(\Omega)$.
- If $\Omega$ is a complete separable metric space, every Borel measure is Radon.
- The Dirac measure $\delta_\omega$ is a Radon measure on any Borel space.
- The Borel measure $\mu^d$ is a Radon measure on $\mathcal{B}(\mathbb{R}^d)$ but the Lebesgue measure $\lambda^d$ on $\mathcal{L}(\mathbb{R}^d)$ is not (see definitions and explanation on slides below)

# Borel spaces

- A topological space is called separable if it has an at most countable dense subset.
- A measure space $(\Omega, \mathcal{A}, \nu)$ is called separable if $\mathcal{A}$ is generated by an at most countable set of subsets of $\Omega$.
- A topological space is called a Polish space if it is homeomorphic to a separable complete metric space.
- All separable Banach spaces are Polish spaces.
- A locally compact Hausdorff space is Polish iff it is second countable.
- On Polish spaces all finite measures and therefore all probability measures are Radon measures.
- All Radon measures on Polish spaces are $\sigma$-finite.

# Borel spaces

- $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is the canonical Borel space on $\mathbb{R}^d$ aka $\mathbb{R}^d$ with the Borel $\sigma$-algebra $\mathcal{B}(\mathbb{R}^d)$, generated by the open intervals $\prod_{i=1}^d (a_i, b_i)$ with $a_i, b_i \in \mathbb{Q}$.
- Among the many Borel measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ there is a unique Borel measure $\mu^d$, 'the' Borel measure, assigning to every right half-open interval its volume. $\mu^d$ is not a probability measure but it is by definition locally finite and $\sigma$-finite.
- All Borel measures on $\mathbb{R}^d$ are regular (since $\mathbb{R}^d$ with the euclidian metric is a Polish space).

# Borel spaces

- With a function $p \geq 0$ on $\mathbb{R}^d$, $p\,\mu^d$ is a measure equivalent to $\mu^d$ if $\mu^d(p^{-1}(0)) = 0$.
- For the Borel space $(\overline{\mathbb{R}}, \mathcal{B}(\overline{\mathbb{R}}))$ there is $\mathcal{B}(\overline{\mathbb{R}}) \cap \mathbb{R} = \mathcal{B}(\mathbb{R})$.
- Every translation-invariant measure $\tilde{\mu}^d$ on $\mathcal{B}(\mathbb{R}^d)$ with a unit cube volume $\tilde{\mu}^d(U^d) < \infty$ is a multiple of $\mu^d$, $\tilde{\mu}^d = \mu^d(U^d)\,\mu^d$ to be precise. $\mu^d$ is the only translation-invariant measure on $\mathcal{B}(\mathbb{R}^d)$ with $\mu^d(U^d) = 1$.

# Borel spaces

- The completion of the measure space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu^d)$ is $(\mathbb{R}^d, \mathcal{L}(\mathbb{R}^d), \lambda^d)$ where $\mathcal{L}(\mathbb{R}^d) \equiv \widehat{\mathcal{B}(\mathbb{R}^d)}_{\mu^d}$ is the Lebesgue $\sigma$-algebra and $\lambda^d \equiv \widehat{\mu^d}$ the Lebesgue measure. As described above, $\mathcal{L}(\mathbb{R}^d)$ consists of sets $B \cup N$ with $B \in \mathcal{B}(\mathbb{R}^d)$ and $N \in \mathcal{N}_{\mu^d}$. $\mathcal{B}(\mathbb{R}^d) \subseteq \mathcal{L}(\mathbb{R}^d)$ with $B \mapsto B \cup \varnothing$. $\lambda^d$ is defined as $\lambda^d(B \cup N) \equiv \mu^d(B)$ with $\lambda^d_{|\mathcal{B}(\mathbb{R}^d)} = \mu^d$.
- $\mathcal{L}(\mathbb{R}^d)$ is important from a systematic point of view but impractical in many use-cases (transformation of random variables, issues with continuity, ...). $\mathcal{B}(\mathbb{R}^d)$ and $\lambda^d$ will be our weapons of choice for most of the time. We will write $\lambda^d$ even if we actually mean $\lambda^d_{|\mathcal{B}(\mathbb{R}^d)}$.

# Borel spaces

- A Borel measure with convenient properties like the Lebesgue measure does not exist for an infinite-dimensional $(\Omega, \mathcal{B}(\Omega))$.
- $\mathcal{L}(\mathbb{R}^d)$ is the Borel $\sigma$-algebra of a topology $\mathcal{T}$, i.e. $\mathcal{L}(\mathbb{R}^d) = \mathcal{B}(\mathbb{R}^d, \mathcal{T})$. The open sets of that topology are of the form $U - N$, where $U$ is an open set of the standard topology and $N$ is a $\mu^d$-measure-zero-set (see $F_\sigma$ and $G_\delta$ sets).

## Borel spaces

$$\begin{array}{ccccc}
\mathcal{B}(\mathbb{R}^d) & \subset & \mathcal{L}(\mathbb{R}^d) & \subset & \mathcal{P}(\mathbb{R}^d) \\
\downarrow & & \downarrow & & \downarrow \\
|\mathcal{B}(\mathbb{R}^d)| & & |\mathcal{L}(\mathbb{R}^d)| & & |\mathcal{P}(\mathbb{R}^d)| \\
\| & & \| & & \| \\
\beth_1 \equiv 2^{\beth_0} & < & \beth_2 \equiv 2^{\beth_1} & = & \beth_2 \equiv 2^{\beth_1}
\end{array}$$

- $$\begin{aligned} \beth_0 &= && = |\mathbb{N}| \\ \beth_1 &= 2^{\beth_0} && = |\mathbb{R}| \end{aligned}$$
- $\beth_i \geq \aleph_i$
- General Continuum Hypothesis: $\beth_i = \aleph_i$

# Borel spaces

- The inclusions in the first row of the above diagram are all strict.
- For non-Borel but Lebesgue-measurable sets, see Cantor set based examples.
- For non-Lebesgue-measurable subsets of $\mathbb{R}^d$ see Vitali sets (Axiom of Choice is needed here).
- Last row shows $\mathcal{L}(\mathbb{R}^d)$ is significantly larger than $\mathcal{B}(\mathbb{R}^d)$.
- There are no $\sigma$-algebras with cardinality $\beth_0$ (see remarks above). Cardinality is finite or $\geq \beth_1$.

# Measurable Functions

- Be $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$ measurable spaces. A function $f\colon (\Omega_1, \mathcal{A}_1) \to (\Omega_2, \mathcal{A}_2)$ is called $(\mathcal{A}_1, \mathcal{A}_2)$-measurable if $f^{-1}(A) \in \mathcal{A}_1$ for every $A \in \mathcal{A}_2$.
- Composition:
  For $(\mathcal{A}_1, \mathcal{A}_2)$-measurable $f\colon (\Omega_1, \mathcal{A}_1) \to (\Omega_2, \mathcal{A}_2)$ and $(\mathcal{A}_3, \mathcal{A}_4)$-measurable $g : (\Omega_3, \mathcal{A}_3) \to (\Omega_4, \mathcal{A}_4)$ the composition $g \circ f$ is $(\mathcal{A}_1, \mathcal{A}_4)$-measurable if $\mathcal{A}_3 \subseteq \mathcal{A}_2$. In particular if $\mathcal{A}_3 = \mathcal{A}_2$ of course.
- A $(\mathcal{T}_1, \mathcal{T}_2)$-continuous function $f\colon (\Omega_1, \mathcal{T}_1) \to (\Omega_2, \mathcal{T}_2)$ is a $(\mathcal{B}(\Omega_1, \mathcal{T}_1), \mathcal{B}(\Omega_2, \mathcal{T}_2))$-measurable function $f\colon (\Omega_1, \mathcal{B}(\Omega_1, \mathcal{T}_1)) \to (\Omega_2, \mathcal{B}(\Omega_2, \mathcal{T}_2))$. A $(\mathcal{B}(\Omega_1, \mathcal{T}_1), \mathcal{B}(\Omega_2, \mathcal{T}_2))$-measurable function $f\colon (\Omega_1, \mathcal{B}(\Omega_1, \mathcal{T}_1)) \to (\Omega_2, \mathcal{B}(\Omega_2, \mathcal{T}_2))$ is called a Borel measurable function. Not every $(\mathcal{B}(\Omega_1, \mathcal{T}_1), \mathcal{B}(\Omega_2, \mathcal{T}_2))$-measurable function is $(\mathcal{T}_1, \mathcal{T}_2)$-continuous.

# Measurable Functions

- The indicator function $1_A : \Omega \to \{0,1\} \subset \mathbb{R}$ w.r.t. a set $A \in \mathcal{P}(\Omega)$, defined by

$$1_A(\omega) \equiv \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

  is $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-measurable iff $A \in \mathcal{A}$.
- A function $f : (\Omega, \mathcal{A}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is $(\mathcal{A}, \mathcal{B}(\mathbb{R}^d))$-measurable iff all component functions $f_i : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ are $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-measurable (see section on product measurable spaces below).
- The sum and product of $(\mathcal{A}, \mathcal{B}(\mathbb{R}^d))$-measurable functions $f, g : (\Omega, \mathcal{A}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is $(\mathcal{A}, \mathcal{B}(\mathbb{R}^d))$-measurable. So are the pointwise maximum and minimum functions $\max(f, g)$ and $\min(f, g)$.

# Measurable Functions

- A constant function $f\colon \Omega_1 \to \Omega_2$ with $f(\omega_1) = \omega \in \Omega_2$ is always measurable. For any two $\sigma$-algebras $\mathcal{A}_1, \mathcal{A}_2$ on $\Omega_1$ and $\Omega_2$ respectively and $A \in \mathcal{A}_2$ there is $f^{-1}(A) = \Omega_1$ if $\omega \in A$ and $f^{-1}(A) = \varnothing$ if $\omega \notin A$.
- For a not necessarily countable set $I$ and a family of measurable spaces $\{(\Omega_i, \mathcal{A}_i)\}_{i\in I}$ consider a family of functions $\Gamma = \{\gamma_i\}_{i\in I}$ with $\gamma_i : \Omega \to (\Omega_i, \mathcal{A}_i)$. The $\sigma$-algebra

$$\sigma_\Omega(\Gamma) \equiv \sigma_\Omega\left(\bigcup_{i\in I} \gamma_i^{-1}(\mathcal{A}_i)\right) \subseteq \sigma_\Omega\left(\prod_{i\in I} \mathcal{A}_i\right) \subseteq \mathcal{P}(\Omega)$$

  is called the $\sigma$-algebra, generated by $\Gamma$. It is the smallest $\sigma$-algebra $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ for which the $\gamma_i$ are $(\mathcal{A}, \mathcal{A}_i)$-measurable.
- If $I$ is countable and the $\gamma_i : (\Omega, \mathcal{A}) \to (\Omega_i, \mathcal{A}_i)$ are $(\mathcal{A}, \mathcal{A}_i)$-measurable functions for some $\sigma$-algebra $\mathcal{A}$, then $\sigma_\Omega(\Gamma) = \bigcup_{i\in I} \gamma_i^{-1}(\mathcal{A}_i) \subseteq \mathcal{A}$ is already a $\sigma$-algebra.

# Measurable Functions

- If $i : U \hookrightarrow (\Omega, \mathcal{A})$, then $\sigma_U(i) = \mathcal{A}_{|U}$.
- If $\Omega = \prod_{i \in I} \Omega_i$ and $\Gamma$ consists of the projections $\pi_i$ then $\sigma_\Omega(\Gamma)$ is called the cylinder $\sigma$-algebra w.r.t. $(\Omega, \Gamma)$, denoted by $\mathcal{C}yl(\Omega, \Gamma)$ (see section on product measurable spaces for more details).
- If $\Omega$ is a Hausdorff space $(\Omega, \mathcal{T}_\Omega)$, be $\mathcal{C}(\Omega, \mathcal{T}_\Omega)$ the set space real-valued functions on $\Omega$, continuous w.r.t. the topology $\mathcal{T}_\Omega$. $\mathcal{C}_c(\Omega, \mathcal{T}_\Omega)$ the subspace of continuous functions $f \in \mathcal{C}(\Omega, \mathcal{T}_\Omega)$ with compact support $\mathrm{supp}(f) \equiv \overline{\{\omega \in \Omega \mid f(\omega) \neq 0\}}$. Denote $\mathcal{C}_0(\Omega, \mathcal{T}_\Omega)$ the space of continuous functions on $\Omega$, vanishing at infinity. These are functions $f \in \mathcal{C}(\Omega, \mathcal{T}_\Omega)$, for which there exists a compact set $K \subset \Omega$ for every $\epsilon > 0$ s.t. $|f(\omega)| < \epsilon$ for all $\omega \in \Omega - K$. Be $\mathcal{C}_b(\Omega, \mathcal{T}_\Omega)$ the space of bounded functions from $\mathcal{C}(\Omega, \mathcal{T}_\Omega)$.

# Measurable Functions

- Obviously

$$\mathcal{C}_c(\Omega, \mathcal{T}_\Omega) \subseteq \mathcal{C}_0(\Omega, \mathcal{T}_\Omega) \subseteq \mathcal{C}_b(\Omega, \mathcal{T}_\Omega) \subseteq \mathcal{C}(\Omega, \mathcal{T}_\Omega)$$

  If $\Omega$ is compact, these spaces coincide.

- It holds that

$$\sigma_\Omega(\mathcal{C}_c(\Omega, \mathcal{T}_\Omega)) \subseteq \sigma_\Omega(\mathcal{C}_0(\Omega, \mathcal{T}_\Omega)) \subseteq \sigma_\Omega(\mathcal{C}_b(\Omega, \mathcal{T}_\Omega)) \subseteq \sigma_\Omega(\mathcal{C}(\Omega, \mathcal{T}_\Omega))$$

  If $\Omega$ is locally compact

$$\sigma_\Omega(\mathcal{C}_c(\Omega, \mathcal{T}_\Omega)) = \sigma_\Omega(\mathcal{C}_0(\Omega, \mathcal{T}_\Omega)) \subseteq \sigma_\Omega(\mathcal{C}_b(\Omega, \mathcal{T}_\Omega)) = \sigma_\Omega(\mathcal{C}(\Omega, \mathcal{T}_\Omega))$$

  First equality follows from the density theorem (see below).
  Second equality see definition of $\sigma_\Omega$.
  The inclusion is usually strict (see cardinality argument for discrete spaces).

# Measurable Functions

- $\mathcal{B}a(\Omega, \mathcal{T}_\Omega) \equiv \sigma_\Omega(\mathcal{C}(\Omega, \mathcal{T}_\Omega))$ is called the Baire $\sigma$-algebra.
- Obviously is $\mathcal{B}a(\Omega, \mathcal{T}_\Omega) \subseteq \mathcal{B}(\Omega, \mathcal{T}_\Omega)$. In general this inclusion is strict. If $(\Omega, d)$ is a metric space, it holds that $\mathcal{B}a(\Omega, \mathcal{T}_d) = \mathcal{B}(\Omega, \mathcal{T}_d)$ for the metric induced topology $\mathcal{T}_d$.
- An element of $\mathcal{M}^+(\Omega, \mathcal{B}a(\Omega, \mathcal{T}_\Omega))$ is called a Baire measure.

# Integration

- On a measurable space $(\Omega, \mathcal{A})$ consider functions

$$w(\omega) = \sum_{i=1}^{m} a_i \, 1_{A_i}(\omega)$$

with $m < +\infty$, $a_i \in \mathbb{R}$ and disjoint $A_i \equiv w^{-1}(a_i) \in \mathcal{A}$. These functions are called simple functions and they are obviously $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-measurable.

- A function $f\colon (\Omega, \mathcal{A}) \to \mathbb{R}$ is $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-measurable iff there exists a sequence of simple functions $(w_n)_{n\in\mathbb{N}}$ s.t.

$$\lim_{n\to+\infty} |f(\omega) - w_n(\omega)| = 0 \qquad \forall \omega \in \Omega$$

For a nonnegative $f$ the sequence can be taken to be nonnegative increasing i.e. $0 \leq w_1 \leq w_2 \leq \cdots \leq f$.

# Integration

- $f : (\Omega, \mathcal{A}, \nu) \to \mathbb{R}$ is $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-measurable if there exists a sequence of simple functions $(w_n)_{n \in \mathbb{N}}$ converging $\nu$-a.e. against $f$. $\nu$ needs to be a complete measure for this to hold.
- Denote $S(\Omega, \mathcal{A})$ the $\mathbb{R}$-vector space of simple functions and $S^+(\Omega, \mathcal{A}) \subset S(\Omega, \mathcal{A})$ the convex cone of nonnegative simple functions (i.e. $a_i \geq 0$).
- The integral of a nonnegative simple function $w \in S^+(\Omega, \mathcal{A})$ over a subset $V \subseteq \Omega$ w.r.t. a measure $\nu$ on $\mathcal{A}$ is defined as :

$$\int_V \nu(d\omega)\, w(\omega) \equiv \sum_{i=1}^{m} a_i\, \nu(A_i \cap V)$$

Note that $a_i = 0$ or $\nu(A_i \cap V) = +\infty$ is allowed here.
Convention: $0 \cdot +\infty = 0$.

# Integration

- The integral of a $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-measurable function $f \geq 0$ over a subset $V \subseteq \Omega$ w.r.t. a measure $\nu$ on $\mathcal{A}$ is defined as :

$$\int_V \nu(d\omega)\, f(\omega) \equiv \sup_w \int_V \nu(d\omega)\, w(\omega)$$

  The supremum is being taken over all $w \in S^+(\Omega, \mathcal{A})$ with $w \leq f$. (Such $w$ exist, see above).
- For an arbitrary $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-measurable function $f$ with $f^+ \equiv \max(f, 0)$ and $f^- \equiv \max(-f, 0)$, the integral w.r.t. a subset $V \subseteq \Omega$ is defined by

$$\int_V d\nu\, f \equiv \int_V d\nu\, f^+ - \int_V d\nu\, f^-$$

# Integration

- $f = f^+ - f^-$ and $|f| = f^+ + f^-$
- The expected properties of an integral like linearity, monotonicity, etc. are met.
- The above integral w.r.t. a measure is sometimes called after its inventor the 'Lebesgue integral'. We will use this name only for integrals in the context $(\Omega, \mathcal{A}, \nu) = (\mathbb{R}^d, \mathcal{L}(\mathbb{R}^d), \lambda^d)$.

# Integration

- $f$ is called $\nu$-integrable if

$$\int_V d\nu\, f^+ < +\infty \quad \text{and} \quad \int_V d\nu\, f^- < +\infty$$

- If one of these integrals has an infinite value, the integral of $f$ is defined to be $+\infty$ or $-\infty$ repectively and $f$ is called $\nu$-quasi-integrable. If both integrals have an infinite value, the integral of $f$ is not defined.

# Integration

- $$\nu(A) = \int_\Omega d\nu \; 1_A$$
- $f : \Omega \to \mathbb{R}$ a $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-measurable function:
$$\int_\Omega d\delta_{\omega_0} \; f = f^+(\omega_0) - f^-(\omega_0) = f(\omega_0)$$
- With a general discrete measure $\nu = \sum_{i \in I} m_i \, \delta_{\omega_i}$, it holds that
$$\int_\Omega d\nu \; f = \sum_{i \in I} m_i \int_\Omega d\delta_{\omega_i} \; f = \sum_{i \in I} m_i \, f(\omega_i)$$

# 3. $L^p$ spaces

# $L^p$ spaces

- With a $p \in (0, +\infty]$, a function $f\colon \Omega \to \mathbb{R}$ on a measure space $(\Omega, \mathcal{A}, \nu)$ is called $p$-th power $\nu$-integrable, if it is $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-measurable and the following condition holds:

$$\|f\|_p \equiv \left\{ \int_\Omega \nu(d\omega)\, |f(\omega)|^p \right\}^{\frac{1}{p}} < +\infty$$

  respectively

$$\|f\|_\infty \equiv \operatorname*{ess\,sup}_{\omega \in \Omega} |f(\omega)| < +\infty$$

  (The essential supremum is the the actual supremum but 'ignoring' behaviour on a measure zero set of points.)
- The set of $\mathbb{R}$-valued $p$-th power $\nu$-integrable functions is denoted by $\mathcal{L}^p(\Omega, \mathcal{A}, \nu; \mathbb{R})$. Usually the functions are understood to be $\mathbb{R}$-valued and we simply write $\mathcal{L}^p(\Omega, \mathcal{A}, \nu)$, $\mathcal{L}^p_+(\Omega, \mathcal{A}, \nu)$ for the subset of nonnegative functions.

# $L^p$ spaces

- Be $f^+ \equiv \max(f, 0)$ and $f^- \equiv \max(-f, 0)$. Obviously is $f = f^+ - f^-$. Since $|f| = f^+ + f^-$, $f$ is $\nu$-integrable iff $|f|$ is $\nu$-integrable. I.e. $f$ is $\nu$-integrable iff $X \in \mathcal{L}^1(\Omega, \mathcal{A}, \nu; \mathbb{R})$.
- It holds that

$$c \in \mathbb{R} \text{ and } f \in \mathcal{L}^p(\Omega, \mathcal{A}, \nu) \Rightarrow c\, f \in \mathcal{L}^p(\Omega, \mathcal{A}, \nu)$$

and that

$$\begin{aligned} f, g \in \mathcal{L}^p(\Omega, \mathcal{A}, \nu) &\Rightarrow |f + g|^p \leq 2^{p-1}\left(|f|^p + |g|^p\right) \\ &\Rightarrow f + g \in \mathcal{L}^p(\Omega, \mathcal{A}, \nu) \end{aligned}$$

making $\mathcal{L}^p(\Omega, \mathcal{A}, \nu)$ an $\mathbb{R}$-vector space.

## $L^p$ spaces

- Hölder's inequality: For $p, q \in [1, +\infty]$ with $p^{-1} + q^{-1} = 1$ and functions $f \in \mathcal{L}^p(\Omega, \mathcal{A}, \nu)$, $g \in \mathcal{L}^q(\Omega, \mathcal{A}, \nu)$ there is $(fg) \in \mathcal{L}^1(\Omega, \mathcal{A}, \nu)$ and

$$\|fg\|_1 \leq \|f\|_p \, \|g\|_q$$

- Minkowski's inequality: For $p \in [1, +\infty]$ and $f, g \in \mathcal{L}^p(\Omega, \mathcal{A}, \nu)$ there is

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p$$

# $L^p$ spaces

- $(\mathcal{L}^p(\Omega, \mathcal{A}, \nu), \|\cdot\|_p)$ is a semi-normed space for $p \in [1, +\infty]$. For $p \in (0, 1)$ it is a semi-quasi-normed space.
- For $\mathcal{N} \subseteq \mathcal{L}^p(\Omega, \mathcal{A}, \nu)$, the linear subspace of functions which are zero $\nu$-a. e., be $L^p(\Omega, \mathcal{A}, \nu) \equiv \mathcal{L}^p(\Omega, \mathcal{A}, \nu)/\mathcal{N}$, the factor space, identifying functions equal $\nu$- a. e. It is the Kolmogorov quotient of the semi-normed space $\mathcal{L}^p(\Omega, \mathcal{A}, \nu)$ and is referred to as $L^p$ space. The Hölder and Minkowski inequalities from above still apply (with the same conditions regarding $p$ and $q$).

# $L^p$ spaces

- For $p \in [1, +\infty]$ $(L^p(\Omega, \mathcal{A}, \nu), \|\cdot\|_p)$ is a normed space with $\mathcal{N} = \ker \|\cdot\|_p$ . $(L^p(\Omega, \mathcal{A}, \nu), d_p)$ with distance $d_p(f, g) \equiv \|f - g\|_p$ is a complete metric space and $L^p(\Omega, \mathcal{A}, \nu)$ therefore a Banach space. In particular it is a complete Hausdorff locally convex topological vector space (tvs) with the metric induced topology.
- For $p \in (0, 1)$ $(L^p(\Omega, \mathcal{A}, \nu), \|\cdot\|_p)$ is only a quasi-normed space (with $\mathcal{N} = \ker \|\cdot\|_p$) because $\|\cdot\|_p$ does not obey to the Minkowski triangle inequality. But since there is $\|f + g\|_p^p \le \|f\|_p^p + \|g\|_p^p$, the metric space $(L^p(\Omega, \mathcal{A}, \nu), d_p^p)$ is complete and $L^p(\Omega, \mathcal{A}, \nu)$ a quasi-Banach (or p-Banach) space. In particular it is a complete Hausdorff tvs with the metric induced topology. It is not necessarily locally convex (see $L^p([0, 1], \mathcal{B}([0, 1], \lambda))$.

# $L^p$ spaces

- If $\nu$ is a finite measure, then $L^p(\Omega, \mathcal{A}, \nu) \subseteq L^q(\Omega, \mathcal{A}, \nu)$ and $\nu(\omega)^{-1/q} \|f\|_q \leq \nu(\omega)^{-1/p} \|f\|_p$ whenever $0 < q \leq p \leq +\infty$. This is not true in general.
- Consider $L^p(\Omega, \mathcal{A}, \nu)^\vee \equiv \mathcal{L}(L^p(\Omega, \mathcal{A}, \nu), \mathbb{R})$, the space of continuous linear forms on $L^p(\Omega, \mathcal{A}, \nu)$ aka the topological dual space of $L^p(\Omega, \mathcal{A}, \nu)$. It is a normed space and even a Banach space (because $\mathbb{R}$ is a complete metric space) with the operator norm

$$\|T\| \equiv \sup_{\|f\|_p = 1} |T(f)| \quad , T \in L^p(\Omega, \mathcal{A}, \nu)^\vee$$

- If $L^p(\Omega, \mathcal{A}, \nu)$ is a Banach space, $L^p(\Omega, \mathcal{A}, \nu)^\vee$ consists precisely of the bounded linear forms.

## $L^p$ spaces

- $L^p(\Omega, \mathcal{A}, \nu)^\vee \subseteq L^p(\Omega, \mathcal{A}, \nu)^* \equiv \mathrm{Hom}(L^p(\Omega, \mathcal{A}, \nu), \mathbb{R})$. $L^p(\Omega, \mathcal{A}, \nu)^*$ denotes the algebraic dual space of $L^p(\Omega, \mathcal{A}, \nu)$. In general the inclusion is strict. If $\dim L^p(\Omega, \mathcal{A}, \nu) < +\infty$ the two spaces are identical.
- If $L^p(\Omega, \mathcal{A}, \nu)$ is locally convex, the Hahn-Banach theorem ensures that $L^p(\Omega, \mathcal{A}, \nu)^\vee$ is large enough. For $p \in (0, 1)$ $L^p(\Omega, \mathcal{A}, \nu)$ is not locally convex in general (check $L^p([0,1], \mathcal{B}([0,1]), \lambda)^\vee = \varnothing$).

# $L^p$ spaces

- For $p \in (1, +\infty)$ and $p^{-1} + q^{-1} = 1$ the canonical mapping $\phi : L^q(\Omega, \mathcal{A}, \nu) \longrightarrow L^p(\Omega, \mathcal{A}, \nu)^\vee$

  $$h \mapsto \phi_h(f) \equiv \int_\Omega d\nu\, f h$$

  is an isometry (check via Hölder's inequality) and it is surjective (see Radon Nikodym Theorem on slides below) which makes it an isometric isomorphism of Banach spaces, i.e.

  $$L^q(\Omega, \mathcal{A}, \nu) = L^p(\Omega, \mathcal{A}, \nu)^\vee$$

- Reflexivity: $L^q(\Omega, \mathcal{A}, \nu) = L^q(\Omega, \mathcal{A}, \nu)^{\vee\vee}$

## $L^p$ spaces

- $L^1(\Omega, \mathcal{A}, \nu)^\vee = L^\infty(\Omega, \mathcal{A}, \nu)$ if $\nu$ is $\sigma$-finite. The dual space of $L^\infty(\Omega, \mathcal{A}, \nu)$ is usually much larger than $L^1(\Omega, \mathcal{A}, \nu)$ (axiom of choice assumed).
- Both $L^1(\Omega, \mathcal{A}, \nu)$ and $L^\infty(\Omega, \mathcal{A}, \nu)$ are not reflexive in general, but are so when finite dimensional (Every finite dimensional normed space is reflexive.)
- $L^2(\Omega, \mathcal{A}, \nu)$ in particular can be made a Hilbert space via the inner product

$$\langle f, g \rangle \equiv \int_\Omega d\nu \; f g \;, \; \|f\|_p = \sqrt{\langle f, f \rangle}$$

- Self-duality: $L^2(\Omega, \mathcal{A}, \nu) = L^2(\Omega, \mathcal{A}, \nu)^\vee$ (All Hilbert spaces are self-dual and reflexive.)

# $L^p$ spaces

- An F-space is a vector space with an F-norm whose induced metric is complete. In particular it is a tvs with the metric induced topology. An F-space is called a Frechet space if it is a locally convex tvs.
- With an abuse of notation define $\mathcal{L}^0(\Omega, \mathcal{A})$ to be the $\mathbb{R}$-vector space of $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-measurable functions and $L^0(\Omega, \mathcal{A}, \nu)$ the corresponding factor space, identifying functions equal $\nu$-almost everywhere.

# $L^p$ spaces

- If $\nu$ is $\sigma$-finite w.r.t. a covering $\Omega = \bigcup_{i \in I} \Omega_i$ with $\Omega_i \subseteq \Omega_{i+1}$, the F-norm

$$\|f\|_0 \equiv \sum_{i \in I} 2^{-i} \int_{\Omega_i} \nu(d\omega) \, \frac{|f(\omega)|}{1 + |f(\omega)|}$$

with the corresponding metric $d_0(f, g) \equiv \|f - g\|_0$ induces a topology on $L^0(\Omega, \mathcal{A}, \nu)$, the topology of (local) convergence in measure. $L^0(\Omega, \mathcal{A}, \nu)$ is an F-space but not a Frechet space. Convergence in the $d_0$ metric is equivalent to the local convergence in measure (see below).

## $L^p$ spaces

- For $p \geq 0$ the spaces $L^p(\Omega, \mathcal{A}, \nu)$ are F-spaces. For $p \geq 1$ they are in addition locally convex and therefore Frechet spaces (even Banach spaces).
- $L^0(\Omega, \mathcal{A}, \nu) \supset \cup_{p>0} L^p(\Omega, \mathcal{A}, \nu)$.(Inclusion is strict but union is dense inside $L^0(\Omega, \mathcal{A}, \nu)$)

# $L^p$ spaces

- Consider a function $f \in L^0(\Omega, \mathcal{A}, \nu)$ with $\|f\|_\infty > 0$. If there exists an $r \in (0, +\infty)$ s.t. $f \in L^r(\Omega, \mathcal{A}, \nu) \cap L^\infty(\Omega, \mathcal{A}, \nu)$, then $f \in L^s(\Omega, \mathcal{A}, \nu)$ for all $s \geq r$ and
$$\lim_{p \to +\infty} \|f\|_p = \|f\|_\infty$$
- Be $f \in L^0(\Omega, \mathcal{A}, \nu)$ with $\nu(\Omega) = 1$ and $\|f\|_\infty > 0$. Since $\nu$ is finite $L^s(\Omega, \mathcal{A}, \nu) \subseteq L^r(\Omega, \mathcal{A}, \nu)$ for $0 \leq r \leq s \leq +\infty$. If there exists an $s \in (0, +\infty)$ s.t. $f \in L^s(\Omega, \mathcal{A}, \nu)$, then $f \in L^r(\Omega, \mathcal{A}, \nu)$ for all $r \leq s$ and with $\exp(-\infty)$ defined to be 0
$$\lim_{p \to 0^+} \|f\|_p = \exp\left(\int_\Omega d\nu \, \log(|f|)\right)$$
The limit is the geometric mean of $f$.

## $L^p$ spaces

- For $p \in [1, +\infty)$, the spaces $L^p(\Omega, \mathcal{A}, \nu)$ are separable if the measurable space $(\Omega, \mathcal{A}, \nu)$ is separable. $L^\infty(\Omega, \mathcal{A}, \nu)$ is only separable if it is finite-dimensional i.e. $\Omega$ consists of a finite number of $\nu$-atoms.
- If the space $L^p(\Omega, \mathcal{A}, \nu)$ with $p \in [1, +\infty)$ is separable, there is a countable set $\Gamma$ s.t. $L^p(\Omega, \mathcal{A}, \nu)$ is either isometrically isomorphic to $L^p([0,1], \mathcal{B}([0,1]), \lambda) \oplus_p \ell^p(\Gamma)$ or to $\ell^p(\Gamma)$. If in addition $(\Omega, \mathcal{A}, \nu)$ is atom free, $\Gamma$ can be chosen to be empty and therefore $L^p(\Omega, \mathcal{A}, \nu)$ is isometrically isomorphic to $L^p([0,1], \mathcal{B}([0,1]), \lambda)$.
- If the measure space is a Borel space, i.e. $(\Omega, \mathcal{A}, \nu) = (\Omega, \mathcal{B}(\Omega), \nu)$, an abbreviated notation is used: $L^p(\Omega, \nu) \equiv L^p(\Omega, \mathcal{B}(\Omega), \nu)$.

# Sequence spaces

- With $I$ an arbitrary countable set and $\#$ the counting measure, $(I, \mathcal{P}(I), \#)$ becomes a measurable space. The spaces $L^p(I, \mathcal{P}(I), \#)$ are called sequence spaces, denoted by $\ell^p(I)$. $\#$ is $\sigma$-finite precisely because $I$ is countable.
- $L^p(I, \mathcal{P}(I), \#) = \mathcal{L}^p(I, \mathcal{P}(I), \#) \quad$ (because $\mathcal{N} = \{0\}$).
- $\ell^0(I) = \mathcal{L}^0(I, \mathcal{P}(I), \#) \overset{!}{=} \mathrm{Fun}(I, \mathbb{R})$ (with $\mathcal{P}(I)$ as $\sigma$-algebra, every function $f\colon I \to \mathbb{R}$ is $(\mathcal{P}(I), \mathcal{B}(\mathbb{R}))$-measurable).
- $$\|f\|_0 = \sum_{i \in I} 2^{-i} \int_{\{i\}} \#(dj)\, \frac{|f(j)|}{1 + |f(j)|} = \sum_{i \in I} 2^{-i} \frac{|f(i)|}{1 + |f(i)|}$$
- For $p \in (0, +\infty)$:

$$\|f\|_p = \left( \int_I \#(di)\, |f(i)|^p \right)^{\frac{1}{p}} = \left( \sum_{i \in I} |f(i)|^p \right)^{\frac{1}{p}}$$

# Sequence spaces

- $$\|f\|_\infty = \sup_{i \in I} |f(i)|$$
- Since $\mathrm{Fun}(I, \mathbb{R}) = \mathbb{R}^I$, we identify functions $f : I \to \mathbb{R}$ with sequences $\{x_i\}_{i \in I} \in \mathbb{R}^I$ ($x_i = f(i)$).
- $\|\{x_i\}_{i \in I}\|_0 = \sum_{i \in I} 2^{-i} |x_i| / (1 + |x_i|)$ .
- $\|\{x_i\}_{i \in I}\|_p = \left(\sum_{i \in I} |x_i|^p\right)^{\frac{1}{p}}$ for $p \in (0, +\infty)$.
- $\|\{x_i\}_{i \in I}\|_\infty = \sup_{i \in I} |x_i|$.
- $\ell^p(I)$ is the space of sequences $\{x_i\}_{i \in I} \in \mathbb{R}^I$, obeying $\sum_{i \in I} |x_i|^p < +\infty$ for $p \in (0, +\infty)$ and $\sup_{i \in I} |x_i| < +\infty$ for $p = +\infty$.
- Componentwise multiplication gives the $\ell^p(I)$ a Banach algebra structure.

# Sequence spaces

- $\ell^p(n) \equiv \ell^p(\{1, \ldots, n\})$, $\ell^p(n) = (\mathbb{R}^n, \|\cdot\|_p)$
- $\ell^p \equiv \ell^p(\mathbb{N})$
- Day Theorem: For $p \in (0,1]$ there is $(\ell^p)^\vee = \ell^\infty$.
- For $p \in (1, +\infty)$ it holds that $(\ell^p)^\vee = \ell^q$ with $p^{-1} + q^{-1} = 1$.
- For all $q \in (0, +\infty]$ there is $\ell^q \subset \ell^0$, and $\ell^q \subset \ell^p$, whenever $0 < q < p \leq +\infty$, showing that the finiteness condition on the measure is necessary and $\sigma$-finiteness does not suffice.
- $\ell^p$ is separable for $p \in (0, +\infty)$, $\ell^\infty$ is not.

# Sequence spaces

- Every separable Banach space is isometrically isomorphic to a quotient space $\ell^1/M$, $M$ a closed subspace of $\ell^1$.
- Every separable Banach space can be isometrically embedded into $\ell^\infty$.

# Bochner spaces

- For functions with values in an arbitrary $\mathbb{K}$-Banach space $(E, \|\cdot\|_E)$, spaces $L^p(\Omega, \mathcal{A}, \nu; E)$ with the expected properties can be defined in a similar way. They are called Bochner-Lebesgue spaces.
- A function $f\colon \Omega \to E$ is called strongly measurable or Bochner-Lebesgue measurable if there exists a sequence of simple functions

$$s_n(\omega) = \sum_{i=1}^{m} c_i \, 1_{A_i}(\omega) \text{ with } c_i \in E,\ A_i \in \mathcal{A}$$

such that

$$\lim_{n \to +\infty} \|f(\omega) - s_n(\omega)\|_E = 0 \quad \forall \omega \in \Omega$$

# Bochner spaces

- Denote by $\mathcal{L}^0(\Omega, \mathcal{A}, \nu; E)$ the space of strongly measurable functions on $(\Omega, \mathcal{A}, \nu)$ w.r.t the Banach space $E$ and $L^0(\Omega, \mathcal{A}, \nu; E) \equiv \mathcal{L}^0(\Omega, \mathcal{A}, \nu; E)/(= \nu \; a.e.)$ the factor space modulo $\nu$ equivalence (sometimes called Kolmogorov quotient).
- $L^0(\Omega, \mathcal{A}, \nu; E)$ is a $\mathbb{K}$-vector space and if $\nu$ is $\sigma$-finite w.r.t. a covering $\Omega = \bigcup_{i \in I} \Omega_i$ with $\Omega_i \subseteq \Omega_{i+1}$, the F-norm

$$\|f\|_0 \equiv \sum_{i \in I} 2^{-i} \int_{\Omega_i} \nu(d\omega) \, \frac{\|f(\omega)\|_E}{1 + \|f(\omega)\|_E}$$

  induces on it the topology of (local) convergence in measure. $L^0(\Omega, \mathcal{A}, \nu; E)$ can be shown to be an F-space but it is not a Frechet space. Convergence in the corresponding $d_0$ metric is equivalent to local convergence in measure (see below).

# Bochner spaces

- Be $f \in \text{Fun}(\Omega, E)$ and $\|f(\cdot)\|_E \in \text{Fun}(\Omega, \mathbb{R})$ its norm function. Then $f \in L^0(\Omega, \mathcal{A}, \nu; E)$ implies that $\|f(\cdot)\|_E \in L^0(\Omega, \mathcal{A}, \nu; \mathbb{R})$. (Note that $\|\cdot\|_E : E \to \mathbb{R}$ is $(\mathcal{B}(E), \mathcal{B}(\mathbb{R}))$-measurable since as a norm it is continuous).
- A function $f : \Omega \to E$ is called weakly measurable if $\phi \circ f : \Omega \to \mathbb{K}$ is $(\mathcal{A}, \mathcal{B}(\mathbb{K}))$-measurable for every $\phi \in E^\vee = \mathcal{L}(E, \mathbb{K})$.
- If $\mathcal{T}_E$ is the norm-induced topology on $E$, how does being $(\mathcal{A}, \mathcal{B}(E, \mathcal{T}_E))$-measurable compare to being strongly measurable or weakly measurable ?

# Bochner spaces

- Pettis measurability theorem: For $f\colon \Omega \to E$ the following assertions are equivalent
  - $f$ is strongly measurable
  - $f$ is weakly measurable and there exists a separable closed subspace $E_0 \subseteq E$ s.t. $f(\Omega) \subseteq E_0$ ($f$ is separably valued).
  - $f$ is $(\mathcal{A}, \mathcal{B}(E, \mathcal{T}_E))$-measurable and separably valued.
- Consider the weak topology on $E$, denoted by $\sigma(E, E^\vee)$. $\sigma(E, E^\vee)$ is the weakest/coarsest topology on $E$ for which all $\phi \in E^\vee$ are continuous. It is Hausdorff.
- By definition $\sigma(E, E^\vee) \subseteq \mathcal{T}_E$ and $\mathcal{B}(E, \sigma(E, E^\vee)) \subseteq \mathcal{B}(E, \mathcal{T}_E)$. Denote $\mathcal{B}a(E, \sigma(E, E^\vee))) \subseteq \mathcal{B}(E, \sigma(E, E^\vee))) \subseteq \mathcal{B}(E, \mathcal{T}_E)$ the Baire $\sigma$-algebra of $E$ w.r.t. the weak topology. It can be shown that $f\colon \Omega \to E$ is weakly measurable iff $f\colon (\Omega, \mathcal{A}) \to (E, \mathcal{B}a(E, \sigma(E, E^\vee))$ is $(\mathcal{A}, \mathcal{B}a(E, \sigma(E, E^\vee)))$-measurable.

# Bochner spaces

- If $E$ is a separable Banach space, $\mathcal{B}a(E, \sigma(E, E^\vee)) = \mathcal{B}(E, \mathcal{T}_E)$ and the properties 'weakly measurable', 'strongly measurable' and '$(\mathcal{A}, \mathcal{B}(E, \mathcal{T}_E))$-measurable' all coincide. Bochner measurable equals $(\mathcal{A}, \mathcal{B}(E, \mathcal{T}_E))$-measurable, meaning $L^0(\Omega, \mathcal{A}, \nu; E) = L^0(\Omega, \mathcal{A}, \nu; E, \mathcal{B}(\mathcal{T}_E))$.
- More generally $\mathcal{B}a(E, \sigma(E, E^\vee)) = \mathcal{B}(E, \mathcal{T}_E)$ holds for a separable metrizable space. Note that 'separable' is equivalent to 'second countable' for a metrizable space.
- Be $f\colon (\Omega, \mathcal{A}, P) \to E$ weakly measurable. $f$ is weakly equivalent to a Bochner-measurable function iff $f_* P$ is an inner regular probability measure on the measurable space $(E, \mathcal{B}(\sigma(E, E^\vee)))$.

# Bochner spaces

- As expected, the Bochner integral of a simple function

$$s(\omega) = \sum_{i=1}^{m} c_i \, 1_{A_i}(\omega) \text{ with } c_i \in E, \; A_i \in \mathcal{A}$$

on $(\Omega, \mathcal{A}, \nu)$ is defined by

$$\int_{\Omega} \nu(d\omega) \, s(\omega) \equiv \sum_{i=1}^{m} c_i \, \nu(A_i)$$

# Bochner spaces

- A strongly measurable function $f : \Omega \to E$ is called Bochner integrable if there exists a sequence of simple functions $(s_n)_{n\in\mathbb{N}}$ s.t.

$$\lim_{n\to+\infty} \int_\Omega \nu(d\omega)\, \|f(\omega) - s_n(\omega)\|_E = 0$$

  In that case, the Bochner integral of $f$ is defined by

$$\int_\Omega \nu(d\omega)\, f(\omega) \equiv \lim_{n\to+\infty} \int_\Omega \nu(d\omega)\, s_n(\omega)$$

- $f$ is Bochner-integrable iff $\|f(\cdot)\|_E \in L^1(\Omega, \mathcal{A}, \nu; \mathbb{R})$.
- -> $\nu$-Bochner measurable and integrable

# Bochner spaces

- With $p \in (0, +\infty]$, the spaces $L^p(\Omega, \mathcal{A}, \nu; E)$ of Bochner $p$-th power $\nu$-integrable functions are defined analogously to the Lebesgue spaces above as spaces of functions $f \in L^0(\Omega, \mathcal{A}, \nu; E)$ with $\|f\|_p < +\infty$ for the respective norms

$$\|f\|_p \equiv \left\{ \int_\Omega \nu(d\omega)\, \|f(\omega)\|_E^p \right\}^{\frac{1}{p}}$$

  and

$$\|f\|_\infty \equiv \operatorname*{ess\,sup}_{\omega \in \Omega} \|f(\omega)\|_E$$

- For $p \in (0, +\infty]$ there is $f \in L^p(\Omega, \mathcal{A}, \nu; E)$ iff $\|f(\cdot)\|_E \in L^p(\Omega, \mathcal{A}, \nu; \mathbb{R})$.

# Bochner spaces

- Many properties of the real valued case carry over to the Bochner $L^p$ spaces, like Hölder's inequality or the inclusion property for finite measures $\nu$, that is $L^p(\Omega, \mathcal{A}, \nu; E) \subseteq L^q(\Omega, \mathcal{A}, \nu; E)$ for $0 < q \leq p \leq +\infty$.
- When $p \in [1, +\infty]$ the spaces $L^p(\Omega, \mathcal{A}, \nu; E)$ are Banach spaces, reflexive for $p \in [1, +\infty)$ if $E$ is reflexive, separable for $p \in [1, +\infty)$ if $E$ is separable.
- For nonseparable $E$, Bochner integrability often is too restrictive. $->$ Pettis integral

# Bochner spaces

- If $E$ is a Banach space and $\nu$ is finite then $E^\vee$ having the Radon-Nikodym property (see below) w.r.t. $\nu$ is equivalent to the canonical mapping $\phi : L^q(\Omega, \mathcal{A}, \nu; E^\vee) \longrightarrow L^p(\Omega, \mathcal{A}, \nu; E)^\vee$ ($p \in (1, +\infty)$ and $p^{-1} + q^{-1} = 1$)

$$h \mapsto \phi_h,\ \phi_h(f) \equiv \int_\Omega \nu(d\omega) \langle h(\omega) | f(\omega) \rangle$$

being an isometric isomorphism i.e.

$$L^q(\Omega, \mathcal{A}, \nu; E^\vee) = L^p(\Omega, \mathcal{A}, \nu; E)^\vee$$

# Bochner spaces

- If $E$ is a reflexive Banach space it is separable iff $E^\vee$ is separable.
- If $E^\vee$ is separable it has the Radon-Nikodym property.
- Hilbert spaces and reflexive spaces have the R-N property.
- $L^p(\Omega, \mathcal{A}, \mu; E)$ for $p \in (1, +\infty)$ has the R-N property iff the Banach space $E$ has the R-N property.

# Bochner spaces

- Consider the Banach spaces $E$ and $L^p(\Omega, \mathcal{A}, \nu)$ (for $p \in [1, +\infty]$). The natural embedding (via $(f, e) \mapsto f(\cdot)\, e$)

$$L^p(\Omega, \mathcal{A}, \nu) \otimes E \hookrightarrow L^p(\Omega, \mathcal{A}, \nu; E)$$

  induces a norm $\Delta_p$ on $L^p(\Omega, \mathcal{A}, \nu) \otimes E$. Notation for $L^p(\Omega, \mathcal{A}, \nu) \otimes E$ with this norm is $L^p(\Omega, \mathcal{A}, \nu) \otimes_{\Delta_p} E$. The completion w.r.t $\Delta_p$ is denoted by $L^p(\Omega, \mathcal{A}, \nu) \hat{\otimes}_{\Delta_p} E$.
- For $p \in [1, +\infty)$ $L^p(\Omega, \mathcal{A}, \nu) \otimes E$ is dense in $L^p(\Omega, \mathcal{A}, \nu; E)$ and therefore

$$L^p(\Omega, \mathcal{A}, \nu) \hat{\otimes}_{\Delta_p} E = L^p(\Omega, \mathcal{A}, \nu; E)$$

  is an isometric isomorphism.

# Bochner spaces

- The canonical injection

$$L^p(\Omega_1, \mathcal{A}_1, \nu_1) \otimes_{\Delta_p} L^p(\Omega_2, \mathcal{A}_2, \nu_2) \\ \hookrightarrow L^p(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2, \nu_1 \otimes \nu_2)$$

is isometric. For $p \in [1, +\infty)$ it has dense range and therefore induces an isometric isomorphism

$$L^p(\Omega_1, \mathcal{A}_1, \nu_1) \hat{\otimes}_{\Delta_p} L^p(\Omega_2, \mathcal{A}_2, \nu_2) \\ = L^p(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2, \nu_1 \otimes \nu_2)$$

Note that

$$L^p(\Omega_1, \mathcal{A}_1, \nu_1) \hat{\otimes}_{\Delta_p} L^p(\Omega_2, \mathcal{A}_2, \nu_2) = L^p(\Omega_1, \mathcal{A}_1, \nu_1; L^p(\Omega_2, \mathcal{A}_2, \nu_2))$$

# Bochner spaces

- In [?defant1992tensor] it is shown that $\Delta_1$ corresponds to $\pi$, the projective norm. In particular there are isometric isomorphisms

$$L^1(\Omega, \mathcal{A}, \nu)\hat{\otimes}_\pi E = L^1(\Omega, \mathcal{A}, \nu)\hat{\otimes}_{\Delta_1} E \quad ( = L^1(\Omega, \mathcal{A}, \nu; E) \ )$$

- $\Delta_\infty$ corresponds to the injective norm $\epsilon$. In particular

$$L^\infty(\Omega, \mathcal{A}, \nu)\hat{\otimes}_\epsilon E = L^\infty(\Omega, \mathcal{A}, \nu)\hat{\otimes}_{\Delta_\infty} E \quad ( \hookrightarrow L^\infty(\Omega, \mathcal{A}, \nu; E) \ )$$

- $\epsilon \leq \Delta_p \leq \pi$ for $p \in (1, +\infty)$

# Bochner spaces

- Grothendieck: If $E$ is a nuclear space, it holds that

$$L^p(\Omega, \mathcal{A}, \nu) \,\hat{\otimes}_\pi\, E = L^p(\Omega, \mathcal{A}, \nu) \,\hat{\otimes}_\epsilon\, E$$

  where the projective and injective tensor products are taken in the category of lctvs.
- Banach spaces are only nuclear when finite dimensional.
- Rule of thumb: If an infinite dimensional space is not Banach it is probably nuclear.

# Bochner spaces

- $E = \mathbb{R}^d$: For $p \in [0, +\infty]$

$$L^p(\Omega, \mathcal{A}, \nu; \mathbb{R}^d) = \bigoplus_d L^p(\Omega, \mathcal{A}, \nu; \mathbb{R})$$

$$L^p(\Omega, \mathcal{A}, \nu; \mathbb{R}^d) = L^p(\Omega, \mathcal{A}, \nu; \mathbb{R}) \mathbin{\hat{\otimes}_{\Delta_p}} \mathbb{R}^d$$

- $\mathbb{R}^d$ is Banach and nuclear, i.e.

$$L^1(\Omega, \mathcal{A}, \nu; \mathbb{R}^d) = L^1(\Omega, \mathcal{A}, \nu) \mathbin{\hat{\otimes}_\pi} \mathbb{R}^d = L^1(\Omega, \mathcal{A}, \nu) \mathbin{\hat{\otimes}_\epsilon} \mathbb{R}^d$$

# Density Theorems

- Consider a measure space $(\Omega, \mathcal{A}, \nu)$ and the vector space $S(\Omega, \mathcal{A})$ of simple functions. Be $S(\Omega, \mathcal{A}, \nu) \subseteq S(\Omega, \mathcal{A})$ the subspace of simple functions $w$ with finite measure support i.e. the property $\nu(\{\omega \in \Omega \mid w(\omega) \neq 0\}) < +\infty$.
- $$\overline{S(\Omega, \mathcal{A})}^{\|\cdot\|_0} = L^0(\Omega, \mathcal{A}, \nu)$$
  Note that $S(\Omega, \mathcal{A})$ in general is not a subset of $L^p(\Omega, \mathcal{A}, \nu)$ when $p \in (0, +\infty]$.
- It can be shown (Dominated Convergence Theorem) that
  $$\overline{S(\Omega, \mathcal{A}, \nu)}^{\|\cdot\|_p^p} = L^p(\Omega, \mathcal{A}, \nu) \quad \text{for } p \in (0, 1)$$
  $$\overline{S(\Omega, \mathcal{A}, \nu)}^{\|\cdot\|_p} = L^p(\Omega, \mathcal{A}, \nu) \quad \text{for } p \in [1, +\infty)$$
- $$\overline{S(\Omega, \mathcal{A})}^{\|\cdot\|_\infty} = L^\infty(\Omega, \mathcal{A}, \nu)$$

# Density Theorems

- For a measurable space $(\Omega, \mathcal{B}(\Omega))$ with $\Omega$ locally compact Hausdorff, a Radon measure $\nu \in \mathcal{M}_R^+(\Omega)$ and $p \in (0, +\infty)$, the space $\mathcal{C}_c(\Omega)$ is dense in $L^p(\Omega, \mathcal{B}(\Omega), \nu)$:

$$\overline{\mathcal{C}_c(\Omega)}^{\|\cdot\|_p^p} = L^p(\Omega, \mathcal{B}(\Omega), \nu) \quad \text{for } p \in (0,1)$$
$$\overline{\mathcal{C}_c(\Omega)}^{\|\cdot\|_p} = L^p(\Omega, \mathcal{B}(\Omega), \nu) \quad \text{for } p \in [1,+\infty)$$

- For $p = +\infty$

$$\overline{\mathcal{C}_c(\Omega)}^{\|\cdot\|_\infty} = \mathcal{C}_0(\Omega) \subset L^\infty(\Omega, \mathcal{B}(\Omega), \nu)$$

- Proof uses the results for simple functions above, Lusin's theorem and Urysohn's lemma.

# Density Theorems

- Be $c$ the space of convergent sequences of real numbers, $c_0$ its subspace consisting of sequences with limit zero and $c_{00}$ its subspace consisting of sequences with only finite nonzero elements :

$$c_{00} \subset c_0 \subset c$$

- In analogous fashion to the above density results:

$$\overline{c_{00}}^{\|\cdot\|_p^p} = \ell^p \quad \text{for } p \in (0,1)$$
$$\overline{c_{00}}^{\|\cdot\|_p} = \ell^p \quad \text{for } p \in [1,+\infty)$$

- $c$ and $c_0$ are closed subspaces of $\ell^\infty$ (not of the $\ell^p$ for $p \in (0,+\infty)$) and therefore Banach spaces. $c_{00}$ is not closed in $c_0$ but it holds that

$$\overline{c_{00}}^{\|\cdot\|_\infty} = c_0 \subset \ell^\infty$$

# Spaces of Measures

- Consider a measurable space $(\Omega, \mathcal{B}(\Omega))$ with $\Omega$ locally compact Hausdorff and $\mathcal{M}^+_R(\Omega)$, $\mathcal{M}^+_{R,f}(\Omega)$, $\mathcal{M}^+_{R,1}(\Omega)$ the sets of positive Radon measures, positive finite Radon measures and Radon probability measures respectively. Obviously

$$\mathcal{M}^+_{R,1}(\Omega) \subseteq \mathcal{M}^+_{R,f}(\Omega) \subseteq \mathcal{M}^+_R(\Omega) \subseteq \mathcal{M}^+(\Omega)$$

- $\mathrm{Hom}^+(\mathcal{C}_c(\Omega, \mathbb{C}), \mathbb{C})$ the set of positive linear forms i.e. the space of linear forms $\phi$ on $\mathcal{C}_c(\Omega, \mathbb{C})$ for which $f \geq 0$ implies $\langle \phi \mid f \rangle \geq 0$.
  $\mathcal{L}^+(\mathcal{C}_c(\Omega, \mathbb{C}), \mathbb{C}) \subseteq \mathrm{Hom}^+(\mathcal{C}_c(\Omega, \mathbb{C}), \mathbb{C})$ is the subset of linear forms, continuous w.r.t. the inductive limit topology of $\mathcal{C}_c(\Omega)$.

# Spaces of Measures

- Riesz-Markov-Kakutani Representation Theorem: The mapping

$$\nu \mapsto \phi_\nu(f) \equiv \int_\Omega \nu(d\omega)\, f(\omega)$$

induces bijections

$$\mathcal{M}_R^+(\Omega) \longleftrightarrow \mathcal{L}^+(\mathcal{C}_c(\Omega), \mathbb{R})$$

and

$$\mathcal{M}_{R,f}^+(\Omega) \longleftrightarrow \mathcal{L}^+(\mathcal{C}_0(\Omega), \mathbb{R})$$

# Spaces of Measures

- $\mathcal{M}_{R,f}(\Omega)$, the vector space of signed finite Radon measures, can be made a Banach space with the total variation norm $\|\nu\|_R \equiv |\nu|(\Omega)$.
- $\nu \mapsto \phi_\nu$ induces an isometric isomorphism

$$\mathcal{M}_{R,f}(\Omega) = \mathcal{C}_0(\Omega)^\vee \ \ ( \ = \mathcal{L}\left(\mathcal{C}_0(\Omega), \mathbb{R}\right) \ )$$

# Radon-Nikodym Theorem

- Be $(\Omega, \mathcal{A}, \mu)$ a measure space where $\mu$ is $\sigma$-finite. For every measure $\nu$ with $\nu \ll \mu$ there exists a unique nonnegative function $f_\nu \in L^0_+(\Omega, \mathcal{A}, \mu)$ such that for all $A \in \mathcal{A}$

$$\nu(A) = \int_A d\nu = \int_A d\mu\, f_\nu \quad .$$

  Conversely, every nonnegative function from $L^0_+(\Omega, \mathcal{A}, \mu)$ defines this way a unique measure, absolutely continuous w.r.t. $\mu$.
  $f_\nu$ is called the Radon-Nikodym derivative, denoted by the symbol

$$\frac{d\nu}{d\mu}$$

- Notation

$$d\nu = \frac{d\nu}{d\mu} d\mu = f_\nu\, d\mu$$

# Radon-Nikodym Theorem

- If $\nu$ is $\sigma$-finite, it holds that $f_\nu \in L^1_+(\Omega, \mathcal{A}, \mu)$. In particular if $\nu$ is a probability measure.
- Intuition: The Radon-Nikodym derivative captures the degree of density change between two measures.
- For a $\sigma$-finite measure space $(\Omega, \mathcal{A}, \mu)$ denote by $ba\,(\Omega, \mathcal{A})$ the set of bounded finitely additive signed measures on $\mathcal{A}$, $ca\,(\Omega, \mathcal{A}) \subseteq ba\,(\Omega, \mathcal{A})$ the subset of (countably additive) measures $\mathcal{M}(\Omega, \mathcal{A})$ and $ca(\Omega, \mathcal{A}, \mu) \subseteq ca(\Omega, \mathcal{A})$ the subset of (countably additive) measures, absolutely continuous w.r.t. $\mu$. There should hold something like

$$\begin{aligned} ba(\Omega, \mathcal{A}) &= B(\Omega, \mathcal{A})^\vee \quad \text{(Hildebrand, Fichtenholtz, Kantorovich)} \\ ca(\Omega, \mathcal{A}, \mu) &= L^1(\Omega, \mathcal{A}, \mu) \quad \text{(R-N theorem)} \end{aligned}$$

  Isometries with $\|\cdot\|_\infty$-norms on the left and $L^p$-norms on the right (Check !).

# Radon-Nikodym Theorem

- Reflexivity:

$$\frac{d\nu}{d\nu} = 1$$

- Linearity: For $\nu, \tau \ll \mu$

$$\frac{d(\nu + \tau)}{d\mu} = \frac{d\nu}{d\mu} + \frac{d\tau}{d\mu}$$

- Since $\nu \ll \mu$ implies $a\nu \ll \mu$ for $a \in \mathbb{R}$

$$\frac{d(a\nu)}{d\mu} = \frac{a d\nu}{d\mu}$$

# Radon-Nikodym Theorem

- Change of Variables: For $\nu \ll \mu$ and a $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-measurable function $g : \Omega \to \mathbb{R}$ with $g \in L^1(\Omega, \mathcal{A}, \nu)$ we get

$$\int_\Omega d\nu \, g = \int_\Omega d\mu \, g \, f_\nu$$

- Chain Rule: For $\nu \ll \tau \ll \mu$

$$\frac{d\nu}{d\mu} = \frac{d\nu}{d\tau} \frac{d\tau}{d\mu} \qquad \mu\text{-a.e.}$$

# Radon-Nikodym Theorem

- $\nu$ and $\mu$ are equivalent, i.e. $\nu \sim \mu$ iff $d\nu/d\mu$ and $d\mu/d\nu$ do exist. If in addition

$$\frac{d\nu}{d\mu} > 0 \quad \mu\text{-a.e.} \qquad \text{and} \qquad \frac{d\mu}{d\nu} > 0 \quad \nu\text{-a.e.} \quad ,$$

  it holds that

$$\frac{d\mu}{d\nu} = \left(\frac{d\nu}{d\mu}\right)^{-1} \qquad \nu\text{- and } \mu\text{-a.e.} \quad .$$

- Equivalent measures $\nu \sim \mu$ induce an isometric isomorphism

$$L^p(\Omega, \mathcal{A}, \nu) = L^p(\Omega, \mathcal{A}, \mu)$$

  via

$$\tau : f(\omega) \mapsto f(\omega)\frac{d\nu}{d\mu}(\omega)$$

# Radon-Nikodym Theorem

- A Banach space $E$ has the Radon-Nikodym property (R-N property) if for every $E$-valued measure $\nu$ with bounded variation on a measure space $(\Omega, \mathcal{A}, \mu)$ the following statements are equivalent:
  - $\nu \ll \mu$
  - There exists a nonnegative function $f_\nu \in L^0(\Omega, \mathcal{A}, \mu; E)$ such that for all $A \in \mathcal{A}$

$$\nu(A) = \int_A d\nu = \int_A d\mu \, f_\nu$$

# Kullback-Leibler Divergence

- Two $\sigma$-finite measures $\nu$, $\mu$ on a measurable space $(\Omega, \mathcal{A})$ can be compared via the Kullback-Leibler divergence or relative entropy, defined as

$$D_{KL}(\nu\|\mu) \equiv \begin{cases} \int_\Omega d\mu \log\left(\frac{d\nu}{d\mu}\right)\frac{d\nu}{d\mu} & , \nu \ll \mu \\ +\infty & , \text{ otherwise} \end{cases}$$

- Since in case $\nu \not\ll \mu$ the RN derivative does not exist, we define $D_{KL}(\nu\|\mu)$ to be $+\infty$. $D_{KL}(\nu\|\mu)$ is not necessarily finite if $\nu \ll \mu$ (see example). The definition includes the convention $\log(0)\,0 \equiv 0$ when $(d\nu/d\mu)(\omega) = 0$. This makes sense because $x\log(x) \to 0$ for $x \to +0$ by L'Hôpital's rule.
- Intuition: $D_{KL}(\nu\|\mu)$ is the amount of information lost when $\mu$ is used to approximate $\nu$.

## Kullback-Leibler Divergence

- $D_{KL}(\nu\|\mu) \geq 0$. $D_{KL}(\nu\|\mu) = 0$ iff $\nu = \mu$.
- $D_{KL}(\nu\|\mu) < \infty$ implies $\nu \ll \mu$.
- $D_{KL}$ is not symmetric and therefore not a distance measure, in particular not a metric. Minimising forward and reversed KL divergences puts quite different requirements on pairs of measures.
- If in addition $\nu$ and $\mu$ are absolutely continuous w.r.t a $\sigma$-finite measure $\lambda$, i.e. $\nu \ll \mu \ll \lambda$, the KL divergence can be calculated from the respective RN derivatives. With

$$p_\nu \equiv \frac{d\nu}{d\lambda} \quad \text{and} \quad p_\mu \equiv \frac{d\mu}{d\lambda} \quad ,$$

via chain rule and change of variables, there is

$$D_{KL}(p_\nu\|p_\mu) \equiv D_{KL}(\nu\|\mu) = \int_\Omega d\lambda \log\left(\frac{p_\nu}{p_\mu}\right) p_\nu \quad .$$

# Kullback-Leibler Divergence

- Let $\nu$ and $\mu$ be two probability measures on $([0,1], \mathcal{B}([0,1])$ with pdfs $p_\nu(x) = 1_{[0,1]}(x)$ and $p_\mu(x) = c \cdot \exp(-\frac{1}{x}) \cdot 1_{(0,1]}(x)$ where c is whatever constant will make $\mu([0,1]) = 1$. Despite $\nu \sim \mu$ and $\nu, \mu \ll \lambda$, we see that $D_{KL}(\nu|\mu)$ is infinite.

# Lebesgue Decomposition Theorem

- Every measure $\nu$ on a $\sigma$-finite measure space $(\Omega, \mathcal{A}, \mu)$ has a unique decomposition into a sum of measures

  $$\nu = \nu_{ac} + \nu_{sc} + \nu_{sd}$$

  where
  - $\nu_{ac} \ll \mu$ (absolutely continuous)
  - $\nu_{sc} \perp \mu$ (singular continuous)
  - $\nu_{sd} \perp \mu$ (singular discrete)

  If $\nu$ is $\sigma$-finite or finite, so are these measures.
- Additivity of support for Borel spaces

  $$\mathrm{supp}(\nu) = \mathrm{supp}(\nu_{ac}) \cup \mathrm{supp}(\nu_{sc}) \cup \mathrm{supp}(\nu_{sd})$$

# Product Measurable Spaces

- Consider an arbitrary nonempty index set $I$ (not necessarily countable and/or finite) and $\mathcal{F}(I)$ the set of nonempty finite subsets of $I$. For a family of measurable spaces $\{(\Omega_i, \mathcal{A}_i)\}_{i \in I}$ one can define a corresponding product measurable space $(\Omega_I, \mathcal{A}_I)$.
- Be $\Omega_I$ the Cartesian product $\prod_{i \in I} \Omega_i$ with partial projections $\pi_{KJ} : \Omega_K \to \Omega_J$ for $\varnothing \neq J \subseteq K \subseteq I$ and denote by $\pi_K$ and $\pi_i$ the main projections $\pi_{IK}$ and $\pi_{I\{i\}}$ respectively. The partial projections obey a cocycle condition $\pi_{JL} \circ \pi_{KJ} = \pi_{KL}$ for $\varnothing \neq L \subseteq J \subseteq K \subseteq I$. $\Omega_I$ itself is equipped with the product topology, the coarsest topology for which the projections $\pi_K$ are continuous.
- Define $\mathcal{A}_I$ to be $\bigotimes_{i \in I} \mathcal{A}_i \equiv \mathcal{C}yl(\Omega_I, \{\pi_i\}_{i \in I}) \subseteq \mathcal{P}(\Omega_I)$, the smallest $\sigma$-algebra $\mathcal{A}$ on $\Omega_I$ such that the $\pi_i$ are $(\mathcal{A}, \mathcal{A}_i)$-measurable.

# Product Measurable Spaces

- In general there is $\bigotimes_{i\in I}\mathcal{A}_i \subseteq \sigma(\prod_{i\in I}\mathcal{A}_i)$. If $I$ is finite or at least countable, it holds that $\bigotimes_{i\in I}\mathcal{A}_i = \sigma(\prod_{i\in I}\mathcal{A}_i)$. The inclusion is strict for uncountable $I$. $\bigotimes_{i\in I}\mathcal{A}_i$ has a couple of nice properties (measurability of functions ...).
- The partial projections $\pi_{KJ} : \Omega_K \to \Omega_J$ for $\emptyset \neq J \subseteq K \subseteq I$ are $(\mathcal{A}_K, \mathcal{A}_J)$-measurable if $K, J \in \mathcal{F}(I)$. Then $\pi_{KJ}^{-1}(\mathcal{A}_J) \subseteq \mathcal{A}_K$ and it holds that

$$\bigotimes_{i\in I}\mathcal{A}_i = \sigma_{\Omega_I}(\{\pi_K\}_{K\in\mathcal{F}(I)}) = \mathcal{C}yl(\Omega_I, \{\pi_K\}_{K\in\mathcal{F}(I)})$$

The $\pi_K^{-1}(\mathcal{A}_K)$ as well as the $\pi_i^{-1}(\mathcal{A}_i)$ are $\sigma$-algebras of cylinder sets.

# Product Measurable Spaces

- $(\Omega_I, \mathcal{A}_I) = (\prod_{i\in I}\Omega_i, \bigotimes_{i\in I}\mathcal{A}_i)$
- If the $\Omega_i$ are Hausdorff

$$\bigotimes_{i\in I}\mathcal{B}(\Omega_i,) \subseteq \mathcal{B}(\prod_{i\in I}\Omega_i,)$$

Equality holds if the $\Omega_i$ are second countable and $I$ is countable. In that case

$$(\prod_{i\in I}\Omega_i, \bigotimes_{i\in I}\mathcal{B}(\Omega_i,)) \stackrel{!}{=} (\prod_{i\in I}\Omega_i, \mathcal{B}(\prod_{i\in I}\Omega_i,)) = (\Omega_I, \mathcal{B}(\Omega_I))$$

n particular

$$(\prod_{i\in I}\mathbb{R}, \bigotimes_{i\in I}\mathcal{B}(\mathbb{R})) = (\mathbb{R}^I, \mathcal{B}(\mathbb{R}^I))$$

# Product Measures

- If $I$ is a nonempty and finite set, the product measure of $\sigma$-finite measures $\nu_i$ on measurable spaces $(\Omega_i, \mathcal{A}_i)$ is the unique measure $\bigotimes_{i \in I} \nu_i$ ($\nu_I$ for short) on $(\prod_{i \in I} \Omega_i, \bigotimes_{i \in I} \mathcal{A}_i)$, defined by

$$\bigotimes_{i \in I} \nu_i \left(\prod_{i \in I} A_i\right) \equiv \prod_{i \in I} \nu_i(A_i)$$

  The product measure is $\sigma$-finite since the $\nu_i$ are.
- This construction does not make sense for an arbitrary set $I$ but a unique product measure can be shown to exist for a family of probability measures.
- Andersen-Jessen theorem: For an arbitrary nonempty set $I$ and a family of probability spaces $(\Omega_i, \mathcal{A}_i, P_i)_{i \in I}$, there exists a unique probability measure $P_I$ on $(\Omega_I, \mathcal{A}_I)$ (see above) such that $\pi_{K*} P_I = \bigotimes_{i \in K} P_i$ for every $K \in \mathcal{F}(I)$.

# Product Measures

- $(\Omega_I, \mathcal{A}_I, \nu_I)$ and $(\Omega_I, \mathcal{A}_I, P_I)$ are called the product measure space and product probability space respectively.
  Notation:
  $$\bigotimes_{i \in I} (\Omega_i, \mathcal{A}_i, \nu_i) \equiv (\Omega_I, \mathcal{A}_I, \nu_I)$$
  $$\bigotimes_{i \in I} (\Omega_i, \mathcal{A}_i, P_i) \equiv (\Omega_I, \mathcal{A}_I, P_I)$$
- $\lambda^d$ is identical to the unique product measure $\lambda^1 \otimes \cdots \otimes \lambda^1$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and therefore
  $$(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda^d) = \bigotimes_{i \in I} (\mathbb{R}^1, \mathcal{B}(\mathbb{R}^1), \lambda^1)$$

# Product Measures

- Be $\nu$ a measure $(\Omega_I, \mathcal{A}_I)$. The measures $\nu_i$, defined by

$$\nu_i\,(A) \equiv \nu(A \times \prod_{j \neq i} \Omega_j) \quad , A \in \mathcal{A}_i \quad ,$$

are called the marginal measures of $\nu$.

# Product Measures

- Tonelli/Fubini I: Let $f\colon \Omega_1 \times \cdots \times \Omega_d \longrightarrow \mathbb{R}$ be a $(\bigotimes_{i=1}^d \mathcal{A}_i, \mathcal{B}(\mathbb{R}))$-measurable function. Then it holds that

$$\int_{\Omega_1 \times \cdots \times \Omega_d} d(\nu_1 \otimes \cdots \otimes \nu_d)\, |f| = \int_{\Omega_{i_1}} \cdots \left( \int_{\Omega_{i_d}} d\nu_{i_1}\, |f| \right) \cdots\, d\nu_{i_d}$$

- Tonelli/Fubini II: If any of the following integrals is finite

$$\int_{\Omega_1 \times \cdots \times \Omega_d} d(\nu_1 \otimes \cdots \otimes \nu_d)\, |f| \ , \ \int_{\Omega_{i_1}} \cdots \left( \int_{\Omega_{i_d}} d\nu_{i_1}\, |f| \right) \cdots\, d\nu_{i_d}$$

then there is

$$\int_{\Omega_1 \times \cdots \times \Omega_d} d(\nu_1 \otimes \cdots \otimes \nu_d)\, f = \int_{\Omega_{i_1}} \cdots \left( \int_{\Omega_{i_d}} d\nu_{i_1}\, f \right) \cdots\, d\nu_{i_d}$$

# Product Measures

- In particular if $f$ is $\nu_1 \otimes \cdots \otimes \nu_d$-integrable :

$$\int_{\Omega_1 \times \cdots \times \Omega_d} d(\nu_1 \otimes \cdots \otimes \nu_d)\, f = \int_{\Omega_{i_1}} \cdots \left( \int_{\Omega_{i_d}} f \, d\nu_{i_1} \right) \cdots \, d\nu_{i_d}$$

- As a consequence: Be $f, g$ integrable functions on finite measure spaces $(\Omega_1, \mathcal{A}_1, \nu_1)$ and $(\Omega_2, \mathcal{A}_2, \nu_2)$ respectively. Then

$$\int_{\Omega_1 \times \Omega_2} d(\nu_1 \otimes \nu_2)\, (fg) = \int_{\Omega_1} d\nu_1 \, f \cdot \int_{\Omega_2} d\nu_2 \, g$$

# Measures on Product Spaces

- A projective system of measures w.r.t. to the above family of measurable spaces $\{(\Omega_i, \mathcal{A}_i)\}_{i \in I}$ is a collection of probability measures $\{\nu_L : \mathcal{A}_L \to [0,1]\}_{L \in \mathcal{F}(I)}$, inner regular w.r.t. the corresponding product topologies (and therefore Radon), such that $\pi_{JK*}\,\nu_J = \nu_K$ for every pair $J, K \in \mathcal{F}(I)$ with $K \subseteq J$.
- Kolmogorov extension theorem: Consider an index set $I$ and a projective system of measures $\{\nu_K\}_{K \in \mathcal{F}(I)}$ for a family of Hausdorff measurable spaces $\{(\Omega_i, \mathcal{A}_i)\}_{i \in I}$ (the $\nu_K$ need not be product measures). Then there exists a unique probability measure $\nu$ on $\mathcal{A}_I$ such that $\pi_{K*}\,\nu = \nu_K$ for every $K \in \mathcal{F}(I)$. $\nu$ is the projective limit of the $\nu_K$. Notation: $\varprojlim_{K \in \mathcal{F}(I)} \nu_K = \nu$.

# Measures on Product Spaces

- This result is used in the construction of stochastic processes.
- Measures on a product space need not be product measures. (Joint distributions do not necessariliy come from independent random variables (see below)).
- The Kolmogorov extension theorem can nevertheless be used to create product probability measures. Check for the projective system $P_K = \bigotimes_{i \in K} P_i$ for $K \in \mathcal{F}(I)$. The result is weaker because Andersen-Jessen does not need the Hausdorff property.

# Convolution

- $\mathbb{R}^d$ is a topological group (via vector addition). In particular $\phi_s^d : \mathbb{R}^{d\times s} \longrightarrow \mathbb{R}^d$, $(x_1, \dots, x_s) \mapsto x_1 + \cdots + x_s$ is a continuous and therefore $(\mathcal{B}(\mathbb{R}^{d\times s}), \mathcal{B}(\mathbb{R}^d))$-measurable mapping. Consider finite measures $\nu_1, \dots, \nu_s$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ (these are Borel measures). The $\phi_s^d$-push forward measure of the product measure $\nu_1 \otimes \cdots \otimes \nu_s$ is called the convolution measure of the $\nu_i$ :

$$\nu_1 * \cdots * \nu_s : \mathcal{B}(\mathbb{R}^d) \longrightarrow \overline{\mathbb{R}}$$

$$\nu_1 * \cdots * \nu_s \equiv (\phi_s^d)_*(\nu_1 \otimes \cdots \otimes \nu_s)$$

- $\nu * \mu = \mu * \nu$, $\nu * (\mu_1 + \mu_2) = \nu * \mu_1 + \nu * \mu_2$
- Construction can be generalized to arbitrary locally compact Hausdorff topological groups $(G, \mathcal{B}(G))$ with a Haar measure replacing $\lambda^d$. Convolution is not necessarily commutative.

# Convolution

- As a consequence of the Tonelli theorem, for two finite measures $\nu$ and $\mu$ and a nonnegative $\mathcal{B}(\mathbb{R}^d)$-measurable function f, there is

$$\int_{\mathbb{R}^d} d(\nu * \mu)\, f = \int_{\mathbb{R}^{d\times 2}} d(\nu \otimes \mu)\, f \circ \phi_2^d$$
$$= \int_{\mathbb{R}^d} \mu(dy) \left( \int_{\mathbb{R}^d} \nu(dx)\, f(x+y) \right)$$

- With $f = 1_B$, $B \in \mathcal{B}(\mathbb{R}^d)$

$$(\nu * \mu)(B) = \int_{\mathbb{R}^{d\times 2}} \nu(dx)\, \mu(dy)\, 1_B(x+y)$$
$$= \int_{\mathbb{R}^d} \mu(dy) \left( \int_{\mathbb{R}^d} \nu(dx)\, 1_B(x+y) \right)$$

# 4. Random Variables

# Random Elements

- Consider a probability space $(\Omega, \mathcal{A}, P)$ and a measurable space $(\mathsf{E}, \mathcal{E})$. A $(\mathcal{A}, \mathcal{E})$-measurable function $X : (\Omega, \mathcal{A}, P) \to (\mathsf{E}, \mathcal{E})$ is called an $(\mathsf{E}, \mathcal{E})$-valued random element. Notation for the space of $(\mathsf{E}, \mathcal{E})$-valued random elements: $\mathcal{L}^0(\Omega, \mathcal{A}, P; \mathsf{E}, \mathcal{E})$. In many cases we will identify random elements, that differ only on a set of $P$-measure zero. Notation for the corresponding space of equivalence classes: $L^0(\Omega, \mathcal{A}, P; \mathsf{E}, \mathcal{E})$
- $\Omega$ is called the sample space of $X$, $\mathcal{A}$ its event space and $(\mathsf{E}, \mathcal{E})$ its target space.
- If $\mathsf{E}$ is a topological space $(\mathsf{E}, \mathcal{T}_\mathsf{E})$, an $(\mathsf{E}, \mathcal{B}(\mathsf{E}, \mathcal{T}_\mathsf{E}))$-valued random element is called an $\mathsf{E}$-valued Borel random element.

# Random Elements

- For an arbitrary Banach space E, an E-valued random variable is a strongly (or Bochner-Lebesgue) measurable function, i.e. an element of $\mathcal{L}^0(\Omega, \mathcal{A}, P; \mathrm{E})$. If E is separable, 'weakly measurable', 'strongly measurable' and '$(\mathcal{A}, \mathcal{B}(\mathrm{E}, \mathcal{T}_{\mathrm{E}}))$-measurable' are equivalent. In that case being an E-valued random variable means being an $(\mathrm{E}, \mathcal{B}a(\mathrm{E}, \sigma(\mathrm{E}, \mathrm{E}^{\vee}))$-valued random element or equivalently an E-valued Borel random element. In other words

$$
\begin{aligned}
&\mathcal{L}^0(\Omega, \mathcal{A}, P; \mathrm{E}) \\
&\qquad = \mathcal{L}^0(\Omega, \mathcal{A}, P; \mathrm{E}, \mathcal{B}a(\mathrm{E}, \sigma(\mathrm{E}, \mathrm{E}^{\vee}))) \\
&\qquad = \mathcal{L}^0(\Omega, \mathcal{A}, P; \mathrm{E}, \mathcal{B}(\mathrm{E}, \mathcal{T}_{\mathrm{E}}))
\end{aligned}
$$

# Random Elements

- A multivariate real-valued random variable is an $\mathbb{R}^d$-valued random variable. We will simply speek of random variables. $L^0(\Omega, \mathcal{A}, P\,; \mathbb{R}^d)$ is the space of (equivalence classes of) random variables .
- $L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ with the F-norm

$$\|f\|_0 \equiv \int_\Omega P(d\omega)\, \frac{\|f(\omega)\|}{1 + \|f(\omega)\|}$$

is an F-space. Convergence in the corresponding metric $d_0(f, g) \equiv \|f - g\|_0$ is equivalent to convergence in probability (see below).

# Random Elements

- Wherever possible and/or convenient, we will identify random variables with their equivalence classes and treat them as elements of $L^p$- instaed of $\mathcal{L}^p$-spaces.
- Note that random variables are not defined as $(\mathbb{R}^d, \mathcal{L}(\mathbb{R}^d))$-valued random elements. A Lebesgue- aka $(\mathcal{A}, \mathcal{L}(\mathbb{R}^d))$-measurable 'random variable' does not produce $(\mathcal{A}, \mathcal{L}(\mathbb{R}^d))$-measurable functions in composition with continuous functions and has other unwanted/impractical behaviour.

# Random Matrices

- Be $H$ a complex Hilbert space, $B(H)$ the space of bounded operators on $H$ and $A(H) \subseteq B(H)$ the subspace of self-adjoint operators. $B(H)$ is a Banach space. Consider an $A(H)$-valued random variable $X : \Omega \to A(H)$.
- Want invariance of $X_* P$ under unitary transformations.

# Random Variables

- For a random variable $X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$, the push-forward measure $X_*P(B) \equiv P(X^{-1}(B))$ $(B \in \mathcal{B}(\mathbb{R}^d))$ induces a probability measure on $\mathcal{B}(\mathbb{R}^d)$. This measure is called the (probability) distribution of $X$ and is denoted by $P_X$.
- With some notational shortcut $P(X \in B) \equiv P(\{\omega \mid X(\omega) \in B\})$

$$P_X(B) = P(X^{-1}(B)) = P(\{\omega \mid X(\omega) \in B\}) = P(X \in B)$$

- Note that the probability measure $P$ is only defined on the $\sigma$-algebra $\mathcal{A}$, which is not necessarily the entire $\mathcal{P}(\Omega)$. $P(X \in B)$ is valid because the $(\mathcal{A}, \mathcal{B}(\mathbb{R}^d))$-measurability of $X$ guarantees that $X^{-1}(B) \in \mathcal{A}$.
- Notation: $X \sim \mathcal{C}(\Phi)$ if $P_X$ has a special form $\mathcal{C}$ with parameters $\Phi$.

# Random Variables

- $$\int_\Omega dP\, 1_B(X) = P(X \in B) = P_X(B) = \int_{\mathbb{R}^d} P_X(dx)\, 1_B(x)$$
- The previous equation is a special case ($g = 1_B$) of the transformation formula :
  Be $X \in L^0(\Omega, \mathcal{A}, P\,; \mathbb{R}^d)$ a random variable with a function $g \in L^0(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda^d; \mathbb{K})$ (the latter condition ensures that $g(X) \in L^0(\Omega, \mathcal{A}, P; \mathbb{K})$).
  It holds that $g \in L^1(\mathbb{R}^d, P_X; \mathbb{K})$ iff $g(X) \in L^1(\Omega, \mathcal{A}, P; \mathbb{K})$.
  In that case or if $g \geq 0$

$$\int_\Omega dP\, g(X) \stackrel{!}{=} \int_{\mathbb{R}^d} P_X(dx)\, g(x)$$

# Random Variables

- The multivariate random variable $X$ can be seen as a vector of one-dimensional random variables $X_i : \Omega \to \mathbb{R}^1$

$$X(\omega) = (X_1(\omega), \ldots, X_n(\omega))^\top$$

  If $\pi_i : \mathbb{R}^d \to \mathbb{R}$ are the canonical projections, it holds that $X$ is an $\mathbb{R}^d$-valued random variable iff the $X_i \equiv \pi_i \circ X : \Omega \to \mathbb{R}$ are $\mathbb{R}$-valued random variables for every $i \in [1, \ldots, d]$.
- $P_X = P_{(X_1,\ldots,X_d)}$ is called the joint probability distribution.
- Random variables $X$ and $Y$ are called independent, notation $X \perp\!\!\!\perp Y$, if $P_{(X,Y)}$ is the product measure of the $P_X$ and $P_Y$ i.e. $P_{(X,Y)} = P_X \otimes P_Y$. In particular for a multivariate random variable $X(\omega) = (X_1(\omega), \ldots, X_n(\omega))^\top$ there is $P_X = P_{X_1} \otimes \cdots \otimes P_{X_n}$ if the $X_i$ are independent.

# Random Variables

- Random variables $X$ and $Y$ are called indentically distributed if $P_X(B) = P_Y(B)$ for all $B \in \mathcal{B}(\mathbb{R}^d)$.
- (Mutually) independent and identically distributed (i.i.d. for short) is a common assumption for data/random variables in the machine learning context.
- Random variables can have the same distribution but be different a.e. ($X \sim \mathcal{N}(0,1)$ and $Y \equiv -X$ for example).
- Random variables can even be i.i.d. and be different a.e.. For example $(X, Y) \sim U([0,1]^2)$ implies that $X$ and $Y$ are independent, both $\sim U([0,1])$ and therefore i.i.d., but $P(X = Y) = 0$.

# Random Variables

- Note that $P(X = Y) = 0$ does hold for arbitrary continuous random variables as long as $(X, Y)$ has a pdf, because

$$P\left(X = Y\right) = P\left((X, Y) \in \Delta\right) = P_{(X,Y)}\left(\Delta\right) = \int_{\Delta} p_{(X,Y)}(x, y) dx\, dy$$
$$= \int_{\mathbb{R}} \left( \int_{\mathbb{R}} 1_{\Delta}(x, y)\, p_{(X,Y)}(x, y)\, dx \right) dy = 0$$

  with $\Delta \equiv \{(x, y) \in \mathbb{R}^2 \mid x = y\}$ the diagonal. (This argument will make sense within the next slides.)
- If two random variables are equal a.e., they are identically distributed. $P((X \in B)\, \Delta\, (Y \in B)) = 0$ implies $P(X \in B) - P(Y \in B) = P(X \in B \setminus Y \in B) \leq P((X \in B)\, \Delta\, (Y \in B)) = 0$ and therefore $P(X \in B) = P(Y \in B)$.

# Random Variables

- A set of i.i.d random variables $X_1, \ldots, X_n \in L^0(\Omega, \mathcal{A}, P\,; \mathbb{R}^d)$ is called a random sample. For a function $g \in L^0(U; \mathbb{R}^{d'})$ with $U \subseteq \bigcup_i \mathrm{Im}(X_i)$ the corresponding random variable $g(X_1, \ldots, X_n) \in L^0(\Omega, \mathcal{A}, P\,; \mathbb{R}^{d'})$ is called a statistic of the random sample $\{X_i\}_{i \in \mathbb{N}}$. $P_{g(X_1,\ldots,X_n)}$ is called the sampling distribution.
- The sample mean
$$\frac{1}{n} \sum_i X_i$$
is obviously a statistic, so is covariance for example (see below).
- Every probability measure $Q \in \mathcal{M}_1^+(\mathbb{R}^d)$ can be made the distribution of a random variable $Y$ by taking as $Y$ the identity element of $L^0(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), Q; \mathbb{R})$. This holds in particular when $Q \equiv P_X$ for a random variable $X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ .

# Distribution Functions

- For a vector $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ define the $d$-dimensional intervals

$$(-\infty, x] \equiv \prod_{i=1}^{d} (-\infty, x_i] \text{ and } [x, +\infty) \equiv \prod_{i=1}^{d} [x_i, +\infty)$$

  These intervals are generators of $\mathcal{B}(\mathbb{R}^d)$ i.e. $\mathcal{B}(\mathbb{R}^d) = \sigma(\{(-\infty, x] \mid x \in \mathbb{R}^d\}) = \sigma(\{[x, +\infty) \mid x \in \mathbb{R}^d\})$.

- $$P(X \le x) \equiv P(\, X \in (-\infty, x]\,) = X_* P((-\infty, x])$$
  $$P(X \ge x) \equiv P(\, X \in [x, +\infty)\,) = X_* P([x, +\infty))$$

# Distribution Functions

- $F_X(x) \equiv P(X \leq x) : \mathbb{R}^d \to [0,1]$ is called the (joint) cumulative distribution function (cdf) of $X$.
- $F_X(x) = P_X((-\infty, x])$
  Intuition: $F_X$ is $P_X$ 'restricted' to the generators of $\mathcal{B}(\mathbb{R}^d)$.
- $F_X(x) = F_X(x_1, \dots, x_d) = P(X_1 \leq x_1, \dots, X_d \leq x_d)$
- $\overline{F}_X(x) \equiv P(X > x) = 1 - F_X(x)$ is called the complementary cumulative distribution function (ccdf) / tail distribution of $X$.
- $F_X$ is monotone nondecreasing and right-continuous in each variable $x_i$ (cadlag)
- $\lim_{x_1,\dots,x_d \to +\infty} \; F_X(x_1, \dots, x_d) = 1$
- $\lim_{x_i \to -\infty} \; F_X(x_1, \dots, x_d) = 0 \quad \forall i \in \{1, \dots, d\}$

# Distribution Functions

- Two random variables $X$ and $Y$ are independent iff $F_{(X,Y)}(x,y) = F_X(x)\, F_Y(y)$ for all $x, y \in \mathbb{R}^d$.
- Two random variables $X$ and $Y$ are identically distributed iff $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}^d$.
- $F_X$ determines its marginal distributions $F_{X_i}$ :

$$
\begin{aligned}
F_{X_i}(x_i) &= \lim_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d \to +\infty} F_X(x_1,\ldots,x_i,\ldots,x_d) \\
&= \lim_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d \to +\infty} P(X_1 \le x_1,\ldots,X_i \le x_i,\ldots,X_d \le x_d) \\
&= P(X_i \le x_i) = P_X\left((-\infty, x_i] \times \prod_{j \ne i} \mathbb{R}\right)
\end{aligned}
$$

- The cdf $C(u_1,\ldots,u_d) = P(U_1 \le u_1,\ldots,U_d \le u_d) : [0,1]^d \to [0,1]$ of a random variable $U = (U_1,\ldots,U_d)^\top$ with $U_i \sim U([0,1])$ is called a $d$-dimensional copula.

# Distribution Functions

- Sklar's theorem: For every random variable $X$ with cdf $F_X$ and univariate marginal distributions $F_{X_i}(x_i) = P(X_i \leq x_i)$, there exists a copula $C$ s.t. for all $x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$ :

$$F_X(x) = F_X\left(x_1, \ldots, x_d\right) = C\left(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\right)$$

If the marginal distributions are continuous, $C$ is uniquely defined. $C$ can be constructed:
$C\left(u_1, \ldots, u_d\right) \equiv F_X(F_{X_1}^{-1}(u_1), \ldots, F_{X_d}^{-1}(u_d))$ with
$F_{X_i}^{-1}(u_i) = \inf\{t \mid F_{X_i}(t) \geq u_i\}$.
Conversely, any arbitrary copula $C$ with univariate cdfs $F_{X_i}(x_i)$ as arguments defines a joint cdf for the random variable $X \equiv (X_1, \ldots, X_d)^\top$ with these cdfs as marginal distributions.

# Distribution Functions

- Intuition: The copula encodes the dependencies between the $X_i$ to separate the study of the marginal distributions and their dependencies.

# Density Functions

- For a random variable $X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$, $P_X \ll \lambda^d$, i.e. $P_X$ being absolutely continuous w.r.t. $\lambda^d$, is equivalent to the existence of a Radon-Nikodym-derivative

$$p_X \equiv \frac{dP_X}{d\lambda^d} \in L^1_+(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda^d; \mathbb{R}) \ .$$

$p_X$ is unique and called the probability density function (pdf) of $X$. It holds for every $B \in \mathcal{B}(\mathbb{R}^d)$ that

$$P(X \in B) = P_X(B) \stackrel{!}{=} \int_B \lambda^d(ds)\, p_X(s)$$

and that

$$\int_B dP_X = \int_{X^{-1}(B)} dP \stackrel{!}{=} \int_B \lambda^d(ds)\, p_X(s)$$

# Density Functions

- The cdf $F_X$ of $X$ as a special case. By the RN-theorem is

$$F_X(x) = P_X\left((-\infty, x]\right) \stackrel{!}{=} \int_{-\infty}^{x} \lambda^d(ds)\, p_X(s) \quad \forall x \in \mathbb{R}^d$$

and therefore

$$p_X(x) = \frac{\partial^d F_X}{\partial x_1 \cdots \partial x_d}(x) \qquad \lambda^d\text{-a.e.}$$

by the Lebesgue differentiation theorem.

# Density Functions

- Consider open balls $B(x,r) \equiv \{s \in \mathbb{R}^d \mid |s-x| < r\}$ for $x \in \mathbb{R}^d$ and $r > 0$. Then $p_X$ has a measure-theoretic characterization

$$p_X(x) \stackrel{!}{=} \lim_{r \to 0} \frac{P_X\left(B(x,r)\right)}{\lambda^d(B(x,r))} = \lim_{r \to 0} \frac{P\left(X \in B(x,r)\right)}{\lambda^d(B(x,r))}$$

- Casual way of formulating this:
  For $dx$ infinitesimally small: $p_X(x)\,dx = P\left(x < X < x + dx\right)$

## Density Functions

- The marginal distributions of $F_X$:

$$F_{X_i}(x_i) = P\left(X_i \leq x_i\right) =$$
$$= \int_{\mathbb{R}} \cdots \int_{-\infty}^{x_i} \cdots \int_{\mathbb{R}} \lambda^d(ds_1, \ldots, ds_d)\, p_X(s_1, \ldots, s_d)$$

- The marginal density functions:

$$p_{X_i}(x_i) = \int_{\mathbb{R}^{d-1}} \lambda^{d-1}(dx_1, \ldots, dx_{i-1}, dx_{i+1}, \ldots, dx_d)\, p_X(x_1, \ldots, x_d)$$

# Density Functions

- If $X$ and $Y$ are two random variables with existing pdfs, their joint random variable $(X, Y)$ does not necessarily have a pdf even if they are identically distributed (see the example $X, Y \sim U([0,1])$ and $Y \equiv X^2$). If it does, the marginal densities of $p_{(X,Y)}$ are the pdfs of $X$ and $Y$.
- This is different if the random variables are independent: Two random variables $X$ and $Y$ with pdfs $p_X$ and $p_Y$ respectively are independent iff $p_{(X,Y)}(x, y) = p_X(x) \cdot p_Y(y)$ with $p_{(X,Y)}(x, y)$ the joint probability density function of $(X, Y)$. The same holds for discrete random variables (see below) and their respective probability mass functions.

# Density Functions

- Change-of-variables theorem: Be $X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ a random variable and $p_X \in L^1_+(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda^d)$ its pdf. Be $g \equiv (g_1, \ldots, g_d) \in L^0(U, \mathcal{B}(\mathbb{R}^d), \lambda^d; \mathbb{R}^d)$ $(U \subseteq \mathrm{Im}(X))$ an injective measurable function with continuous first partial derivatives and a nonvanishing Jacobian, i.e. for $x \equiv (x_1, \ldots, x_d) \in \mathbb{R}^d$

$$J_g(x) = \det \left( \frac{\partial g_i}{\partial x_j}(x) \right)_{1 \leq i \leq d, 1 \leq j \leq d} \neq 0$$

  Then it holds for $Y \equiv g(X) \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ and $y \equiv (y_1, \ldots, y_d) \in \mathbb{R}^d$ that

$$p_Y(y) = \begin{cases} \left| J_g(g^{-1}(y)) \right|^{-1} \cdot \; p_X(g^{-1}(y)) & , y \in g(\mathbb{R}^d) \\ 0 & , y \in g(\mathbb{R}^d)^\complement \end{cases}$$

# Density Functions

- If $g$ is a linear transformation with an invertible $d \times d$ matrix $A$ and therefore $g(x) = A\,x$ and $g^{-1}(y) = A^{-1}\,y$ with $J_A(g^{-1}(y)) = \det(A)$ , it holds that

$$p_Y(y) = \begin{cases} |\det(A)|^{-1} \cdot p_X(A^{-1}\,y) & , y \in g(\mathbb{R}^d) \\ 0 & , y \in g(\mathbb{R}^d)^\complement \end{cases}$$

- If $g \in L^0(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda^d; \mathbb{R})$ with continuous first partial derivatives :

$$p_Y(y) = \int_{\mathbb{R}^d} \lambda^d(dx)\, p_X(x)\, \delta(y - g(x))$$

# Density Functions

- For the cdf of $Y = g(X)$ :

$$F_Y(y) = \int_{-\infty}^{y} \lambda^d(dt)\, p_Y(t)$$

Note that in particular

$$F_Y(y) = P\,(Y \leq y) = P\,(g(X) \leq y) =$$
$$= \begin{cases} P\,(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) & , g \text{ is monot. increasing} \\ P\,(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)) & , g \text{ is monot. decreasing} \end{cases}$$

# Density Functions

- If $X$ has a pdf $p_X$ and its cdf is given by $F_X(x_1, \ldots, x_d) = C_X(F_{X_1}(x_1), \ldots, F_{X_d}(x_d))$ with copula $C_X$, $p_X$ can be derived from $C_X$ via the derivative chain rule like so:

$$
\begin{aligned}
p_X(x) &= \frac{\partial^d C_X\left(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\right)}{\partial F_{X_1}(x_1) \cdots \partial F_{X_d}(x_d)} \prod_{i=1}^{d} p_{X_i}(x_i) \\
&= c_X\left(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\right) \prod_{i=1}^{d} p_{X_i}(x_i)
\end{aligned}
$$

$c_X$ is called the copula density.

# Notions of Continuity

- A random variable $X$ is called a continuous random variable if the induced probability measure $P_X$ is a continuous measure, i.e. $P_X(\{x\}) = 0$ (no point masses).
- $X$ is called an absolutely continuous random variable if $P_X$ is absolutely continuous w.r.t. $\lambda^d$. 'Absolutely continuous' implies 'continuous' because $\lambda^d$ is a continuous measure.
- For $d > 1$, the cdf $F_X$ can have discontinuities on $\mathbb{R}^d$ despite $P_X$ being a continuous measure. And a continuous $F_X$ does not imply a continuous measure $P_X$.

# Notations for Probability Density Functions

- If $X \sim \mathcal{C}(\Phi)$ and a pdf of $X$ exists, the notation $p_X(x) = \mathcal{C}(x\,;\Phi)$ is used. The notation $X \sim p(x)$ means that the random variable $X$ is distributed according to a pdf $p$.
- To escape notational clutter, $p(x)$ and $p(z)$ or $p(x)$ and $q(x)$ is often used instead of $p_X(x)$ and $p_Z(x)$ respectively.

# Discrete Random Variables

- Consider a probability space $(\Omega, \mathcal{A}, P)$ and a measurable space $(\mathsf{E}, \mathcal{E})$. A discrete $(\mathsf{E}, \mathcal{E})$-valued random element is an $(\mathsf{E}, \mathcal{E})$-valued random element $X : (\Omega, \mathcal{A}, P) \to (\mathsf{E}, \mathcal{E})$ where $P_X$ is a discrete measure on $\mathcal{E}$ (This should be equivalent to $X$ having a discrete image, which we call the support of the $X$, denoted by $\operatorname{supp}(X)$). If $(\Omega, \mathcal{A})$ and/or $(\mathsf{E}, \mathcal{E})$ are discrete measurable spaces of type $(S, \mathcal{P}(S))$ ($S$ at most countable), then every distribution $P_X$ is a discrete measure.
- $P_X$ is a discrete probability measure, concentrated on the support of $X$, $S_X \equiv \operatorname{supp}(X) \subset \mathsf{E}$, i.e.

$$P_X = \sum_{s \in S_X} m_s \, \delta_s \quad .$$

Obviously $m_s = P_X(\{s\})$.

# Discrete Random Variables

- Explain measure-theoretic $\text{supp}(P_X)$ equals $\text{supp}(X)$.
- $$\begin{aligned} P_X(\mathrm{E}) &= \sum_{s\in S_X} m_s\,\delta_s(\mathrm{E}) = \sum_{s\in S_X} m_s\,\delta_s\left(\bigsqcup_{s\in S_X}\{s\}\sqcup \mathrm{E}\setminus\bigsqcup_{s\in S_X}\{s\}\right) \\ &= \sum_{s\in S_X} m_s\,\delta_s\left(\bigsqcup_{s\in S_X}\{s\}\right) = P_X\left(\bigsqcup_{s\in S_X}\{s\}\right) \\ &= \sum_{s\in S_X} P_X(\{s\}) = 1 \end{aligned}$$
- In familiar notation:

$$P(X\in \mathrm{E}) = \sum_{s\in S_X} P(X=s) = 1$$

# Discrete Random Variables

- Denote $p_X(\epsilon) \equiv P\left(X = \epsilon\right) = P_X\left(\{\epsilon\}\right)$, $\epsilon \in \mathrm{E}$ the probability mass function (pmf) of a discrete random element.
- For an arbitrary $E \in \mathcal{E}$ there is

$$P_X\left(E\right) = \sum_{s \in S_X} p_X(s)\, \delta_s(E)$$

# Discrete Random Variables

- A $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$-valued discrete random element $X : (\Omega, \mathcal{A}, P) \to \mathbb{R}^d$ is called a discrete random variable. $P_X$ is a discrete probability measure, concentrated on the support $S_X \subset \mathbb{R}^d$, i.e.

  $$P_X = \sum_{s \in S_X} m_s \, \delta_s \quad .$$

  It is absolutely continuous w.r.t. to

  $$\#_{S_X} \equiv \#(\cdot \cap S_X) = \sum_{s \in S_X} \delta_s \quad ,$$

  the counting measure w.r.t. $S_X$. ($\#_{S_X}$ is the push-forward of the counting measure on $S_X$ via the inclusion map $i : (S_X, \mathcal{P}(S_X)) \hookrightarrow (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, i.e. $\#_{S_X} = i_* \#$. Observe that $\mathcal{P}(S_X) = \mathcal{B}(\mathbb{R}^d)_{|S_X}$ .)

# Discrete Random Variables

- $\#_{S_X}$ is $\sigma$-finite on $\mathcal{B}(\mathbb{R}^d)$ and with $P_X \ll \#_{S_X}$, the Radon-Nikodym theorem gives that for $x \in \mathbb{R}^d$

$$p_X(x) = P_X(\{x\}) = \int_{\{x\}} dP_X \stackrel{!}{=} \int_{\{x\}} \frac{dP_X}{d\#_{S_X}} \, d\#_{S_X}$$
$$= \frac{dP_X}{d\#_{S_X}}(x) \quad \in L^1_+(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \#_{S_X})$$

- As the choice of notation indicates, the probability mass function $p_X$ of $X$ is in fact an R-N derivative but w.r.t. the counting measure $\#_{S_X}$, not w.r.t. $\lambda^d$ as for a pdf. Note that such pdf $p_X \in L^1_+(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda^d)$ does not exist for a discrete $X$ because $P_X \perp \lambda^d$ (follows from $\delta_x \perp \lambda^d$ and therefore $\#_{S_X} \perp \lambda^d$, together with $P_X \ll \#_{S_X}$).

# Discrete Random Variables

- The cdf of a discrete random variable:

$$F_X(x) = P\left(X \leq x\right) = \sum_{s \leq x} P\left(X = s\right) = \sum_{s \leq x} p_X(s)$$

- A random variable $X : \Omega \to \mathbb{R}^d$ is almost surely constant with value $x \in \mathbb{R}^d$ iff $P_X = \delta_x$.

# Discrete Random Variables

- For an $A \in \mathcal{A}$, the indicator function $1_A : \Omega \to \{0, 1\} \subset \mathbb{R}$ is a discrete random variable with pmf

$$p_{1_A}(x) = \begin{cases} P(A) & , x = 1 \\ P(A^\complement) = (1 - P(A)) & , x = 0 \\ 0 & , x \in \mathbb{R} \setminus \{0, 1\} \end{cases}$$

and the induced probability measure on $\mathcal{B}(\mathbb{R})$

$$P_{1_A} = (1 - P(A))\, \delta_0 + P(A)\, \delta_1 \quad .$$

# Discrete Random Variables

- Obviously $P_{1_A} = \mathrm{Ber}_{P(A)}$ and $1_A \sim \mathrm{Ber}(P(A))$, i.e. $1_A$ is Bernoulli distributed. Note that

$$P_{1_A}(\mathbb{R}) = P(1_A \in \mathbb{R}) = P(\Omega) = 1 \quad .$$

- For the cdf

$$F_{1_A}(x) = P(1_A \leq x) = \begin{cases} 0 & , x < 0 \\ (1 - P(A)) & , x \in [0, 1) \\ 1 & , x \geq 1 \end{cases}$$

# Kullback-Leibler Divergence

- For two random variables $X, Y \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ with pdfs $p(x)$ and $q(x)$ respectively, we can consider the corresponding KL-divergence $D_{KL}(p\|q)$.
- If $X$ and $Y$ are continuous and $P_X \ll P_Y \ll \lambda^d$ :

$$D_{KL}(p\|q) = \int_{\mathbb{R}^d} \lambda^d(dx) \log\left(\frac{p(x)}{q(x)}\right) p(x)$$

with the conventions

$$\log\left(\frac{0}{0}\right) 0 \equiv 0, \; \log\left(\frac{0}{q(x)}\right) 0 \equiv 0, \; \log\left(\frac{p(x)}{0}\right) p(x) \equiv +\infty \quad .$$

# Kullback-Leibler Divergence

- If $X$ and $Y$ are discrete with $\text{supp}(X) \subseteq \text{supp}(Y)$ :

$$D_{KL}(p \| q) = \sum_{x \in \text{supp}(X)} \log\left(\frac{p(x)}{q(x)}\right) p(x)$$

# Convergence

- Consider sequences of random variables $X_1, X_2, \ldots$ with $X_n \in L^0(\Omega_n, \mathcal{A}_n, P_n; \mathbb{R}^d)$ (the probability spaces not necessarily identical). There are various notions of convergence associated with such sequences.
- Pointwise convergence Be $X \in \mathrm{Fun}(\Omega, \mathbb{R}^d)$ the pointwise limit of random variables $X_n \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ i.e.

$$\lim_{n \to +\infty} \|X(\omega) - X_n(\omega)\| = 0 \quad \forall \omega \in \Omega$$

Then it holds that $X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$.

# Convergence

- Almost sure convergence of a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ with $X_n \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ towards a random variable $X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$:

$$P\left(\lim_{n\to+\infty} X_n = X\right) = 1$$

  Notation: $X_n \xrightarrow{a.s.} X$

- Note that a.s. convergence towards a function $X \in \mathrm{Fun}(\Omega, \mathbb{R}^d)$ does not imply it to be a random variable. The completeness of $(\Omega, \mathcal{A}, P)$ is needed for that.

# Convergence

- Convergence in distribution or weak convergence of a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ with $X_n \in L^0(\Omega_n, \mathcal{A}_n, P_n; \mathbb{R}^d)$ against a random variable $X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ :

$$\lim_{n\to+\infty} P_{X_n}(B) = P_X(B) \quad \forall B \in \mathcal{B}'(\mathbb{R}^d)$$

  or equivalently

$$\lim_{n\to+\infty} P_n(X_n \in B) = P(X \in B) \quad \forall B \in \mathcal{B}'(\mathbb{R}^d)$$

  $\mathcal{B}'(\mathbb{R}^d) \equiv \{B \in \mathcal{B}(\mathbb{R}^d) \mid \lambda^d(\partial B) = 0\}$ ('continuity sets').
  Notation: $X_n \Rightarrow X$

- The $P_{X_n}$ are weakly converging towards a probability measure $Q \in \mathcal{M}_1^+(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. As the random variable with that distribution we can pick $X \equiv id \in L^0(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), Q; \mathbb{R}^d)$ .

# Convergence

- Convergence in probability of a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ with $X_n \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ towards a function $X \in \mathrm{Fun}(\Omega, \mathbb{R}^d)$:

$$\lim_{n\to+\infty} P\left(|X - X_n| > \epsilon\right) = 0 \quad \forall \epsilon > 0$$

  Notation: $X_n \xrightarrow{P} X$
- $(L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d), d_0)$ is an F-space (see above). Since convergence in the $d_0$-metric is equivalent to convergence in probability, $X_n \xrightarrow{P} X$ implies that $X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$.
- Generalization to arbitrary measures: convergence in measure

# Convergence

- Convergence in $L^p$-norm of a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ with $X_n \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ towards a function $X \in \mathrm{Fun}(\Omega, \mathbb{R}^d)$:

$$\lim_{n\to+\infty} \|X - X_n\|_p = 0$$

  Notation: $X_n \xrightarrow{L^p} X$
- $X_n \xrightarrow{L^p} X$ implies $X \in L^p(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ if $p \in [1, +\infty]$ (Banach space property).
- For $r > s \geq 1$ :

$$X_n \xrightarrow{L^r} X \quad \text{implies} \quad X_n \xrightarrow{L^s} X$$

# Convergence

- $$X_n \xrightarrow{L^p} X \quad \text{implies} \quad \lim_{n \to +\infty} \mathbb{E}[|X_n|^p] = \mathbb{E}[|X|^p]$$
- $L^1$-convergence is called convergence in mean
  $L^2$-convergence is called convergence in quadratic mean
- a.s. convergence does not imply $L^1$-convergence (but see dominated convergence theorem).
- Convergence in probability, almost sure convergence and $L^p$-convergence for $X_n \to X$ is linear i.e. equivalent to the same form of convergence on the component level, i.e. for $X_n[i] \to X[i]$.

# Convergence

- Weak convergence is not linear or multiplicative in general, but if $X_n \Rightarrow X$ and $Y_n \Rightarrow c \in \mathbb{R}^d$ then $X_n + Y_n \Rightarrow X + c$ (Slutsky).
- Continuous mapping theorem: If $g : \mathbb{R}^d \to \mathbb{R}^s$ is continuous, it holds that:

$$X_n \Rightarrow X \text{ implies } g(X_n) \Rightarrow g(X)$$

$$X_n \xrightarrow{P} X \text{ implies } g(X_n) \xrightarrow{P} g(X)$$

$$X_n \xrightarrow{a.s.} X \text{ implies } g(X_n) \xrightarrow{a.s.} g(X)$$

# Convergence

- $$X_n \xrightarrow{a.s.} X \quad \text{implies} \quad X_n \xrightarrow{P} X$$
  $$X_n \xrightarrow{L^2} X \quad \text{implies} \quad X_n \xrightarrow{L^1} X$$
  $$X_n \xrightarrow{L^1} X \quad \text{implies} \quad X_n \xrightarrow{P} X$$
  $$X_n \xrightarrow{P} X \quad \text{implies} \quad X_n \Rightarrow X$$
- Weak convergence does not imply convergence of the corresponding pdfs but convergence of the pdfs implies weak convergence (Scheffé theorem).

# 5. Moments

# Moments

- For a random variable $X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$

$$\mathbb{E}[X] \equiv \mathbb{E}_P[X] \equiv \int_\Omega dP\, X \quad \in \overline{\mathbb{R}}^{\,d}$$

denotes the expectation or the first moment.
For $X \equiv (X_1, \dots, X_d)^\top \in L^0(\Omega, \mathcal{A}, P)$ this amounts to

$$\mathbb{E}[X] \equiv \mathbb{E}_P[X] \equiv (\mathbb{E}[X_1], \dots, \mathbb{E}[X_d])^\top \in \overline{\mathbb{R}}^{\,d}$$

For a random matrix $X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^{d \times d'})$:

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_{11}] & \cdots & \mathbb{E}[X_{1d'}] \\ \vdots & & \vdots \\ \mathbb{E}[X_{d1}] & \cdots & \mathbb{E}[X_{dd'}] \end{pmatrix} \in \overline{\mathbb{R}}^{\,d \times d'}$$

# Moments

- For $X \in \mathcal{L}^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ denote $X^+$ and $X^-$ the obvious random vectors $(X_1^+, \ldots, X_d^+)^\top$ and $(X_1^-, \ldots, X_d^-)^\top$ respectively with $X = X^+ - X^-$ and $|X| = X^+ + X^-$.
- $X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ is called $P$-integrable if $|\mathbb{E}[X]| < +\infty$ which is equivalent to the $P$-integrability of the univariate components i.e. the condition $|\mathbb{E}[X_i]| < +\infty$ for all $i \in \{1, \ldots, d\}$. The $P$-integrability of the $X_i$ is equivalent to the $P$-integrability of the $|X_i|$ and therefore of $|X|$.
- To summarize: $X$ is $P$-integrable iff $|X|$ is $P$-integrable. I.e. $X$ is $P$-integrable iff $X \in L^1(\Omega, \mathcal{A}, P; \mathbb{R})$.
- Nonnegativity: $X \geq 0 \;\Rightarrow\; \mathbb{E}[X] \geq 0$
- Linearity: $\mathbb{E}[cX] = c\mathbb{E}[X]$ for any $c \in \mathbb{R}$ and $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $X = Y \;\; P$ a.s. $\;\Rightarrow\; \mathbb{E}[X] = \mathbb{E}[Y]$
- $X = c \in \mathbb{R}^d \;\; P$ a.s. $\;\Rightarrow\; \mathbb{E}[X] = c$

# Moments

- Be $A \in \mathcal{A}$ an event and $1_A : \Omega \to \mathbb{R}$ the associated indicator random variable. Then it holds that

$$\mathbb{E}[1_A] = \int_A dP = P(A)$$

- For an event $B \in \mathcal{B}(\mathbb{R}^d)$ and the associated indicator random variable $1_{\{X \in B\}} : \Omega \to \mathbb{R}^1$

$$\mathbb{E}[1_{\{X \in B\}}] = \int_{\{X \in B\}} dP = P_X(B) = P(X \in B)$$

In particular

$$\mathbb{E}[1_{\{X \leq x\}}] = F_X(x)$$

# Moments

- From Radon-Nikodym theorem: If $Q$ is a $\sigma$-finite measure on $\mathcal{A}$ with $Q \ll P$:

$$\mathbb{E}_Q[1_A] = \int_A dQ = \int_A \frac{dQ}{dP}\, dP = \mathbb{E}_P[1_A \frac{dQ}{dP}]$$

- Change-of-variables theorem: Consider a random variable $X \in L^1(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ and a function $g \in L^0(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda^d; \mathbb{K})$ (the latter condition ensures that $g(X) \in L^0(\Omega, \mathcal{A}, P; \mathbb{K})$). It holds that $g \in L^1(\mathbb{R}^d, P_X; \mathbb{K})$ iff $g(X) \in L^1(\Omega, \mathcal{A}, P; \mathbb{K})$. In that case or if $g \geq 0$

$$\mathbb{E}_P[g(X)] = \int_\Omega dP\, g(X) \stackrel{!}{=} \int_{\mathbb{R}^d} P_X(dx)\, g(x)$$

# Moments

- Law of the unconscious statistician (LOTUS):
  Under the conditions of the change-of-variables theorem.
  If in addition $X$ has a pdf, i.e. $P_X \ll \lambda^d$, it holds that

$$\mathbb{E}_P\left[g(X)\right] = \int_{\mathbb{R}^d} P_X(dx)\, g(x) = \int_{\mathbb{R}^d} \lambda^d(dx)\, g(x)\, \frac{dP_X}{d\lambda^d}$$
$$= \int_{\mathbb{R}^d} \lambda^d(dx)\, g(x)\, p(x)$$

  If $X$ is a discrete random variable:

$$\mathbb{E}_P\left[g(X)\right] = \sum_{x \in \mathbb{R}^d} g(x)\, p_X(x)$$

# Moments

- With $\mathbb{E}[X] \equiv \mathbb{E}_P[X] \equiv (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])^\top$, there is

$$\mathbb{E}[X_i] = \int_{\mathbb{R}} \lambda(x_i)\, x_i\, p_{X_i}(x_i) \quad .$$

Since $p_{X_i}$ is the marginal density function of $p_X$ w.r.t. $X_i$, we get that

$$\begin{aligned}
&\int_{\mathbb{R}} \lambda(dx_i)\, x_i\, p_{X_i}(x_i) \\
&= \int_{\mathbb{R}} \lambda(dx_i)\, x_i \left( \int_{\mathbb{R}^{d-1}} \lambda^{d-1}(dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_d)\, p_X(x) \right) \\
&= \int_{\mathbb{R}^d} \lambda^d(dx)\, x_i\, p_X(x)
\end{aligned}$$

# Moments

- As a consequence, it holds that :

$$\mathbb{E}_P[X] = \begin{cases} \int_{\mathbb{R}^d} \lambda(dx)\, x\, p(x) & , X \text{ absolutely continuous} \\ \sum_{x \in \mathbb{R}^d} x\, p(x) & , X \text{ discrete} \end{cases}$$

- 

$$\|X\|_p = \mathbb{E}[\,|X|^p\,]^{\frac{1}{p}}$$

$$= \begin{cases} \left( \int_{\mathbb{R}^d} \lambda^d(dx)\, |x|^p\, p(x) \right)^{\frac{1}{p}} & , X \text{ absolutely continuous} \\ \left( \sum_{x \in \mathbb{R}^d} |x|^p\, p(x) \right)^{\frac{1}{p}} & , X \text{ discrete} \end{cases}$$

# Moments

- Notations: For random variables $X, Y \in L^1(\Omega, \mathcal{A}, P\,; \mathbb{R}^d)$ and $(X, Y) \in L^1(\Omega, \mathcal{A}, P\,; \mathbb{R}^{2d})$ with existing pdfs $p(x), q(x), p(x, y)$ respectively, we introduce the following notations to specify the pdf (other that the joint one) under the integral:

$$\mathbb{E}_{p(x)}\left[g(X, Y)\right] \equiv \int_{\mathbb{R}^{2d}} dx\, dy\; g(x, y)\, p(x)$$

$$\mathbb{E}_{Y \sim q(x)}\left[g(X, Y)\right] \equiv \int_{\mathbb{R}^{2d}} dx\, dy\; g(x, y)\, q(x)$$

$$\mathbb{E}_{q(x)}\left[g(p(x), q(y))\right] \equiv \int_{\mathbb{R}^{2d}} dx\, dy\; g(p(x), q(y))\, q(x)$$

# Moments

- Note that

$$\mathbb{E}\left[X\right] = \begin{cases} \int_{\mathbb{R}^d} dP_X(dx)\, x & , X \text{ absolutely continuous} \\ \sum_{x \in \mathbb{R}^d} x\, P_X(\{x\}) & , X \text{ discrete} \end{cases}$$

- Even without $X$ having a pdf, $\mathbb{E}[g(X)]$ can be written as an integral over $\mathbb{R}^d$, as a Lebesgue-Stieltjes (LS-) integral with $F_X$ as the integrator:

$$\mathbb{E}\left[g(X)\right] = \int_{\mathbb{R}^d} dF_X(x)\, g(x)$$

- As we will later see, the stochastic integrals from the Itô calculus are a generalization of this construction with stochastic processes as integrators.

# Moments

- Jensen inequality: For a random variable $X \in L^1(\Omega, \mathcal{A}, P\,; \mathbb{R}^d)$ and a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ there is

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$$

  Examples for $\phi$: $\phi(x) = e^x$, $\phi(x) = |\cdot|^p$ for $p \geq 1$
- $\mathbb{E}$ is a linear operator on $L^1(\Omega, \mathcal{A}, P\,; \mathbb{R}^d)$ (see above).
- $\mathbb{E}$ is bounded and therefore continuous because $|\mathbb{E}[X]| = \|\mathbb{E}[X]\|_1 \leq \mathbb{E}[\|X\|_1] = \|X\|_1$ (Jensen inequality).
- Multiplication theorem: Be $X_1, \ldots, X_d \in L^1(\Omega, \mathcal{A}, P; \mathbb{R})$ independent univariate random variables. Then it holds that

$$\mathbb{E}\left[\prod_{i=1}^{d} X_i\right] = \prod_{i=1}^{d} \mathbb{E}[X_i]$$

  (From Change-of-variables and Tonelli/Fubini)

# Moments + Spaces of Random Variables

- $L^0(\Omega, \mathcal{A}, P\,; \mathbb{R}^d)$ is the space of (multivariate) random variables on the probability space $(\Omega, \mathcal{A}, P)$, i.e. we consider random variables as equivalence classes. We dont't write "P a.s." for example. Everything is modulo behaviour on a set of measure zero.
- Since $P$ is a probability measure and therefore finite:

$$L^\infty(\Omega, \mathcal{A}, P\,; \mathbb{R}^d) \subset \cdots \subset L^0(\Omega, \mathcal{A}, P\,; \mathbb{R}^d)$$

  These inclusions are strict. A random variable with a Pareto type distribution for example can be an element of $L^1$ while not being in $L^2$.

# Moments + Spaces of Random Variables

- $(L^1(\Omega, \mathcal{A}, P; \mathbb{R}^d), \|X\|_1)$ is a Banach space (see section on Lebesgue spaces) with $\|X\|_1 = \mathbb{E}[|X|]$.
- $X \in L^1(\Omega, \mathcal{A}, P; \mathbb{R}^d)$
  iff $\{\ X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)\ \wedge\ \mathbb{E}[|X|] < +\infty\ \}$
  iff $\{\ X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)\ \wedge\ |\mathbb{E}[X]| < +\infty\ \}$
  i.e. iff X is $P$-integrable (see above).

# Moments + Spaces of Random Variables

- $(L^2(\Omega, \mathcal{A}, P; \mathbb{R}^d), \|X\|_2)$ is even a Hilbert space with $\|X\|_2 = \sqrt{\mathbb{E}[|X|^2]}$ and inner product $\langle X \mid Y \rangle = \mathbb{E}[X^\top Y]$.
  Cauch-Schwarz inequality: $|\mathbb{E}[X^\top Y]| \leq \|X\|_2 \, \|Y\|_2$
- $X \in L^2(\Omega, \mathcal{A}, P; \mathbb{R}^d)$
  iff $\{ \; X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d) \; \wedge \; \mathbb{E}[|X|^2] < +\infty \; \}$
- Since $\mathrm{Var}(X_i) = \mathbb{E}[|X_i|^2] - \mathbb{E}[X_i]^2$ : $X \in L^2(\Omega, \mathcal{A}, P; \mathbb{R}^d)$
  iff $\{ \; |\mathbb{E}[X_i]| < +\infty \; \wedge \; \mathrm{Var}(X_i) < +\infty \; \forall i \in \{1, \ldots, d\} \; \}$
  iff $\{ \; X \in L^1(\Omega, \mathcal{A}, P; \mathbb{R}^d) \; \wedge \; \mathrm{Var}(X_i) < +\infty \; \forall i \in \{1, \ldots, d\} \; \}$

# Moments + Spaces of Random Variables

- In general, the $(L^p(\Omega, \mathcal{A}, P; \mathbb{R}^d), \|X\|_p)$ are quasi Banach spaces for $p \in (0,1)$ and Banach spaces for $p \in [1, +\infty)$ with $\|X\|_p = \mathbb{E}[\,|X|^p\,]^{\frac{1}{p}}$.
- $X \in L^p(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ iff $\{\, X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ and $\mathbb{E}[|X|^p] < +\infty \,\}$
- For $p \in [1, +\infty)$ and $d_p(X, Y) \equiv \|X - Y\|_p$ the $L^p(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ are complete metric spaces. For $p \in (0,1)$ the same holds for the metric ${d_p}^p(X, Y) \equiv \|X - Y\|_p^p$.
- $d_2$ is the root mean square distance and for a fixed $X \in L^2(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ is

$$\mathrm{RMSE}(\hat{X}) \equiv d_2(X, \hat{X}) = \|X - \hat{X}\|_2$$

  the root mean square error. $\mathbb{E}[X]$ is the constant minimizer of RMSE with $\mathrm{RMSE}(\mathbb{E}[X]) = \sigma(X)$, the standard deviation (see below).

# Moments + Spaces of Random Variables

- Almost sure equality: $X$ and $Y$ are considered as equal almost surely (a.s.) if $d_\infty(X, Y) = 0$ or equivalently if $P(X = Y) = 1$ / $P(X \neq Y) = 0$
- $L^\infty(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ is the space of essentially bounded random variables.
- For $A \in \mathcal{A}$ it holds that $1_A \in L^p(\Omega, \mathcal{A}, P; \mathbb{R})$ for all $p \in [0, +\infty]$. Because

$$\|1_A\|_p = \begin{cases} P(A)^{\frac{1}{p}} & p \in (0, +\infty) \\ 1 & p = +\infty \end{cases}$$

is always $\leq 1$.

# Moments

- Moments encode quantitative information about $X$.
- $\mathbb{E}[X]$, $\mathbb{E}[XX^\top]$ and $\mathbb{E}[X_1^{k_1} \cdots X_d^{k_d}]$ are called raw moments.
- $\mathbb{E}[(X - \mathbb{E}[X]) (X - \mathbb{E}[X])^\top]$ are called centralized and $\mathbb{E}[\{(X - \mathbb{E}[X])^\top \mathrm{Cov}(X)^{-1}(X - \mathbb{E}[X])\}^k]$ standardized moments.
- The $\mathbb{E}[|X|^k]$ are called absolute moments.
- Moments not necessarily exist. A Cauchy distributed (see below) random variable $X \sim \mathit{Cauchy}\,(\mu, \Sigma)$ for example does have absolute moments with finite value but its raw moments are either undefined or have infinite value.

# Moments

- A matrix $A \in \mathbb{R}^{d \times d}$ is symmetric if $A = A^\top$
- A symmetric matrix $A$ is positive semidefinite, notation $A \succeq 0$, if one of the following equivalent conditions is met:
  - $x^\top A x \geq 0$ for all $x \in \mathbb{R}^d$
  - All principal minors of $A$ are nonnegative.
  - Cholesky decomposition: There exists a lower triangular matrix with nonnegative diagonal $L$ such that $A = L\,L^\top$.
  - Square root decomposition: There is a unique positive semidefinite matrix $B$ such that $A = B\,B$. Notation: $B = A^{\frac{1}{2}}$.
- A symmetric positive semidefinite matrix $A$ is positive definite, notation $A \succ 0$, if one of the following equivalent conditions is met:
  - $x^\top A x > 0$ for all $x \neq 0$
  - $A$ is nonsingular i.e. $\det(A) > 0$.
  - The Cholesky decomposition $A = L^\top L$ is unique, i.e. the diagonal of $L$ is positive.
  - The square root $A^{\frac{1}{2}}$ of $A$ is positive definite.

# Moments

- The covariance matrix of two random variables $X$ and $Y$:

$$\begin{aligned}\mathrm{Cov}(X, Y) &\equiv \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top] \\ &= \mathbb{E}[XY^\top] - \mathbb{E}[X]\mathbb{E}[Y]^\top \in \overline{\mathbb{R}}^{d \times d}\end{aligned}$$

  Univariate:

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \in \overline{\mathbb{R}}\end{aligned}$$

- $\mathrm{Cov}(X, Y)$ is a symmetric positive semidefinite matrix. It describes linear relations between the random vector components of $X$ and $Y$.

# Moments

- $\mathrm{Cov}(X) \equiv \mathrm{Cov}(X, X) \in \overline{\mathbb{R}}_{+}^{d\times d}$ is called the covariance matrix of the random variable $X$.
  Univariate:
  $$\mathrm{Var}(X) = \mathrm{Cov}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \in \overline{\mathbb{R}}_{+}$$
- The square root of $\mathrm{Cov}(X)$, denoted by $\sigma(X)$, is called the standard deviation of $X$: $\sigma(X) \equiv \mathrm{Cov}(X)^{\frac{1}{2}} \in \overline{\mathbb{R}}_{+}^{d\times d}$.
  Univariate: $\sigma(X) = \sqrt{\mathrm{Var}(X)} \in \overline{\mathbb{R}}_{+}$
- With $\dim(X) = 1$, the random variables $X - \mathbb{E}[X]$ and $(X - \mathbb{E}[X])/\sigma(X)$ (defined if $\sigma(X) \in (0, +\infty)$) are called the centered and standardized forms of $X$ respectively.

# Moments

- With $X = (X_1, \ldots, X_d)^\top$ and $Y = (Y_1, \ldots, Y_d)^\top$ it holds that $\mathrm{Cov}(X, Y)_{ij} = \mathrm{Cov}(X_i, Y_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_j - \mathbb{E}[Y_j])]$. In particular $\mathrm{Cov}(X)_{ii} = \mathrm{Cov}(X_i) = \mathrm{Var}(X_i)$.
- The inverse $\mathrm{Cov}(X)^{-1}$ (if it exists) is called the precision matrix.
- $\mathrm{Cor}(X, Y) \equiv \mathbb{E}[XY^\top] \in \mathbb{R}^{d\times d}$ is the correlation matrix of the random variables $X$ and $Y$.
  Univariate: $\mathrm{Cor}(X, Y) = \mathbb{E}[XY]$
- $\mathrm{Cor}(X, Y)$ is a symmetric positive semidefinite matrix. It measures the degree to which the random vector components of $X$ and $Y$ are linearly related.
- $\mathrm{Cor}(X) \equiv \mathbb{E}[XX^\top]$ is the correlation matrix of $X$ or its second moment matrix.
  Univariate: $\mathrm{Cor}(X) = \mathbb{E}[X^2]$

# Moments

- $\mathrm{Cov}(X, Y) = \mathrm{Cor}(X, Y) - \mathbb{E}[X]\mathbb{E}[Y]^\top$.
- $\mathrm{Cor}(X, Y)_{ij} = \mathbb{E}[X_i Y_j]$ and in particular $\mathrm{Cor}(X)_{ii} = \mathbb{E}[X_i^2]$. $\mathrm{Cor}(X)_{ii} = 1$ if it exists.
- $\mathbb{E}[|X|^2] = \mathrm{Tr}\{\mathbb{E}[X^\top X]\} = \mathbb{E}[\mathrm{Tr}\{X^\top X\}] = \mathbb{E}[\mathrm{Tr}\{XX^\top\}] = \mathrm{Tr}\{\mathbb{E}[XX^\top]\} = \mathrm{Tr}\{\mathrm{Cor}(X)\}$
- $\mathrm{Tr}\{\mathrm{Cov}(X)\} = \mathrm{Tr}\{\mathbb{E}[XX^\top]\} - \mathrm{Tr}\{|\mathbb{E}[X]|^2\}$
- $Y \equiv AX + b$ an affine transformation of $X$ with $A \in \mathbb{R}^{d\times d'}$: $E[Y] = A\,\mathbb{E}[X] + b$ and $\mathrm{Cov}(Y) = A\ \mathrm{Cov}(X)\ A^\top$
- $\mathbb{E}[X^\top AX] = \mathbb{E}[X]^\top A\,\mathbb{E}[X] + \mathrm{Tr}(A\ \mathrm{Cov}(X))$.

# Moments

- If $\mathbb{E}[XY^\top] = \mathbb{E}[X]\,\mathbb{E}[Y]^\top$, or equivalently $\mathrm{Cov}(X, Y) = 0$, the random variables are called uncorrelated.
- Independence of $X$ and $Y$ implies they are uncorrelated :

$$\begin{aligned}&\mathbb{E}[X_i Y_j] - \mathbb{E}[X_i]\,\mathbb{E}[Y_j] \\ &\qquad = \int_{\mathbb{R}} \lambda(dy) \int_{\mathbb{R}} \lambda(dx)\, xy\,(\, p_{(X,Y)}(x, y) - p_X(x)\; p_Y(y)\,)\end{aligned}$$

The other way around does not necessarily hold because there could be nonlinear relations.

# Moments

- Pearson correlation: If $\sigma(X_i), \sigma(Y_j) \in (0, +\infty)$ for every $i, j \in [1, \ldots, d]$ consider the correlation matrix of the standardized forms :

$$\begin{aligned}
\mathrm{Cor}(X, Y)_{ij} &\equiv \mathbb{E}\left[\frac{(X_i - \mathbb{E}[X_i])}{\sigma(X_i)} \frac{(Y_j - \mathbb{E}[Y_j])}{\sigma(Y_j)}\right] \\
&= \frac{\mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_j - \mathbb{E}[Y_j])]}{\sigma(X_i)\,\sigma(Y_j)} \\
&= \frac{\mathrm{Cov}(X_i, Y_j)}{\sigma(X_i)\,\sigma(Y_j)}
\end{aligned}$$

If $X, Y \in L^2(\Omega, \mathcal{A}, P; \mathbb{R}^d)$, the Schwarz inequality implies that:

$$|\mathrm{Cor}(X, Y)_{ij}| \leq \left\| \frac{X_i - \mathbb{E}[X_i]}{\sigma(X_i)} \right\|_2 \left\| \frac{Y_j - \mathbb{E}[Y_j]}{\sigma(Y_j)} \right\|_2 = 1$$

# Moments

- $\text{Cov}(1_A) = \mathbb{E}[(1_A)^2] - \mathbb{E}[1_A]^2 = \mathbb{E}[1_A] - \mathbb{E}[1_A]^2 = P(A) - P(A)^2 = P(A)(1 - P(A)) = P(A)\,P(A^{\complement})$
  $\text{Cor}(1_A) = \text{Cov}(1_A) + \mathbb{E}[1_A]\,\mathbb{E}[1_A]^\top = P(A)(P(A^{\complement}) + P(A)) = P(A)$
- A random variable $X$ with $\text{Cov}(X) = \Sigma$ and $\mathbb{E}[X] = \mu$ can be constructed for any pair $(\Sigma, \mu)$ where $\Sigma \in \mathbb{R}^{d\times d}$ is symmetric positive semidefinite and $\mu \in \mathbb{R}^d$ a random vector. Take a Cholesky decomposition $\Sigma = L^\top L$ and put $X \equiv LY + \mu$. $Y$ is a standard Gaussian random variable (see below) with $\mathbb{E}[Y] = 0$ and $\text{Cov}[Y] = E_d$. Therefore $\mathbb{E}[X] = L\,\mathbb{E}[Y] + \mu = \mu$ and $\text{Cov}(X) = L\ \text{Cov}(Y)L^\top = \Sigma$.

# Moments

- If $\mathrm{Cov}(X) \succ 0$, the kurtosis of $X$ (the third standardized moment) is defined by:

$$\mathbb{E}\left[\left\{(X-\mathbb{E}[X])^\top \mathrm{Cov}(X)^{-1}(X-\mathbb{E}[X])\right\}^2\right]$$

- If $\mathrm{Cov}(X) \succ 0$, the skewness of $X$ (the fourth standardized moment) is defined by:

$$\mathbb{E}\left[\left\{(X-\mathbb{E}[X])^\top \mathrm{Cov}(X)^{-1}(X-\mathbb{E}[X])\right\}^3\right]$$

- Note that

$$\begin{aligned}
&\mathbb{E}\left[(X-\mathbb{E}[X])^\top \mathrm{Cov}(X)^{-1}(X-\mathbb{E}[X])\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left\{(X-\mathbb{E}[X])^\top \mathrm{Cov}(X)^{-1}(X-\mathbb{E}[X])\right\}\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left\{\mathrm{Cov}(X)^{-1}(X-\mathbb{E}[X])(X-\mathbb{E}[X])^\top\right\}\right] \\
&= \mathrm{Tr}\{\mathrm{Cov}(X)^{-1}\,\mathrm{Cov}(X)\} \\
&= \dim(X) = n
\end{aligned}$$

# Moments

- Estimation of tail probabilities of random variables.
- Markov inequality: If $Y : \Omega \to \mathbb{R}$ is a nonnegative random variable, it holds for every $t \in \mathbb{R}_{>0}$ that :

$$P(Y \geq t) \leq \min\left(1, \frac{\mathbb{E}[Y]}{t}\right)$$

# Moments

- Chebyshev inequality: If $X \in L^1(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ with $\mathrm{Cov}(X) \succ 0$ i.e. $\mathrm{Cov}(X)^{-1}$ exists, it holds for every $t \in \mathbb{R}_{>0}$ that :

$$P\left(\sqrt{(X - \mathbb{E}[X])^\top \mathrm{Cov}(X)^{-1}(X - \mathbb{E}[X])} > t\right) \le \min\left(1, \frac{\dim(X)}{t^2}\right)$$

- In case $d = 1$ we retain the classical Chebyshev inequality:

$$P(\,|X - \mathbb{E}[X]| > t\,\sigma(X)\,) \le \min\left(1, \frac{1}{t^2}\right)$$

# Moments

- Principal component analysis (PCA): Consider a random variable $X$ with $\mathbb{E}[X] = 0$ and $\mathrm{Cov}(X) = \mathbb{E}[X\,X^\top]$. $\mathrm{Cov}(X)$ is real symmetric and therefore has an eigendecomposition $\mathrm{Cov}(X) = Q\,\Lambda Q^\top$. $\Lambda = \{\lambda_1, \dots, \lambda_d\}$ is a diagonal matrix with entries the eigenvalues of $\mathrm{Cov}(X)$ ordered top down according to size (i.e. $\lambda_1 \geq \cdots \geq \lambda_d$). $Q = \{q_1, \dots, q_d\}$ is an orthogonal matrix with the corresponding (orthonormal chosen) eigenvectors as columns. Since

$$\begin{aligned}\mathrm{Cov}(Q^\top X) =& \mathbb{E}[Q^\top X\,(Q^\top X)^\top] = \mathbb{E}[Q^\top XX^\top Q] \\ =& Q^\top \mathbb{E}[XX^\top]Q = Q^\top\,\mathrm{Cov}(X)Q = \Lambda\end{aligned}$$

it holds that

$$\mathrm{Var}(q_1^\top X) = \lambda_1 \geq \cdots \geq \mathrm{Var}(q_d^\top X) = \lambda_d.$$

# Moments

- Times series regression analysis: Autocorrelation models ARIMA, VAR. Problem: Autocorrelation of the errors.

# 6. Conditional Expectation and Probability

# Conditional Expectation

- Consider classical conditional probability on a probability space $(\Omega, \mathcal{A}, P)$ :

$$P(A \mid B) \equiv \frac{P(A \cap B)}{P(B)} \quad (A, B \in \mathcal{A}; P(B) > 0)$$

denotes the conditional probability of event $A$ given event $B$.

- In general not true that $P(A \mid B) = P(B \mid A)$.
Bayes theorem:

$$P(A \mid B) \equiv \frac{P(B \mid A)\, P(A)}{P(B)} \quad (A, B \in \mathcal{A}; P(B) > 0)$$

- $P^B \equiv P(\cdot \mid B) : A \mapsto P(A \mid B)$ is again a probability measure on $(\Omega, \mathcal{A})$.

# Conditional Expectation

- $$P^B(A) = \frac{P(B \mid A)\, P(A)}{P(B)} = \frac{1}{P(B)} \int_\Omega dP\, 1_A\, 1_B = \frac{1}{P(B)} (1_B\, P)(A)$$

  Therefore
  $$P^B = \frac{1}{P(B)} (1_B\, P)$$

- Be $Y \in L^1(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ an integrable random variable. The conditional expectation of $Y$ w.r.t. an event $B \in \mathcal{A}$, notation $\mathbb{E}^B[Y] \equiv \mathbb{E}[Y \mid B]$, is

  $$\mathbb{E}[Y \mid B] \equiv \int_\Omega dP^B\, Y = \frac{1}{P(B)} \int_B dP\, Y = \frac{1}{P(B)} \mathbb{E}[1_B\, Y] \quad .$$

  In particular is $P^B(A) = \mathbb{E}[1_A \mid B]$.

# Conditional Expectation

- As a generalization, the conditional expectation of $Y$ w.r.t. a $\sigma$-subalgebra $\mathcal{G} \subseteq \mathcal{A}$, denoted by $\mathbb{E}[Y \mid \mathcal{G}]$ ($\mathbb{E}^{\mathcal{G}}[Y]$ a notational shortcut), is a unique random variable from $L^1(\Omega, \mathcal{G}, P|_{\mathcal{G}}; \mathbb{R}^d)$ satisfying

$$\mathbb{E}[1_G \cdot \mathbb{E}[Y \mid \mathcal{G}]] = \mathbb{E}[1_G \cdot Y]$$

  for every $G \in \mathcal{G}$.
- Intuition: In the above condition, $\mathbb{E}[Y \mid \mathcal{G}]$ is averaging out $Y$ over all sets of $\mathcal{G}$. $\mathcal{G}$ encodes pre-existing information. $\mathbb{E}[Y \mid \mathcal{G}]$ allows to fine tune the expected value aka the degree of information loss about $Y$. Extremes: $\mathbb{E}[Y \mid \{\varnothing, \Omega\}] = E[Y]$ with no information given and $\mathbb{E}[Y \mid \mathcal{A}] = Y$ with full information.

# Conditional Expectation

- Existence: If $d = 1$ and $Y \geq 0$, $Q_Y(A) \equiv \mathbb{E}[1_A\, Y] \geq 0$ is a finite measure on $(\Omega, \mathcal{A})$ with $Q_Y \ll P$. It follows that $Q_Y|_{\mathcal{G}} \ll P|_{\mathcal{G}}$ and the existence of

$$\mathbb{E}[Y \mid \mathcal{G}] = \frac{dQ_Y|_{\mathcal{G}}}{dP|_{\mathcal{G}}} \in L^1_+(\Omega, \mathcal{G}, P|_{\mathcal{G}}\,; \mathbb{R})$$

  is a consequence of the Radon-Nikodym theorem. For arbitrary univariate $Y$, put $\mathbb{E}^{\mathcal{G}}[Y] \equiv \mathbb{E}^{\mathcal{G}}[Y^+] - \mathbb{E}^{\mathcal{G}}[Y^-]$. For general (multivariate) $Y$ : $\mathbb{E}^{\mathcal{G}}[Y] = (\mathbb{E}^{\mathcal{G}}[Y_1], \dots, \mathbb{E}^{\mathcal{G}}[Y_d])$.
- Uniqueness: In general, a lot of candidates for $\mathbb{E}[Y \mid \mathcal{G}]$, obeying the above conditions, can exist. Consider candidates $E, E' \in \mathcal{L}^1(\Omega, \mathcal{G}, P|_{\mathcal{G}}\,; \mathbb{R}^d)$. Since $\mathbb{E}[1_G \cdot E] = \mathbb{E}[1_G \cdot Y] = \mathbb{E}[1_G \cdot E'\,]$ for every $G \in \mathcal{G}$ implies $P|_{\mathcal{G}}(E \neq E') = 0$, both $E$ and $E'$ belong to the same class in $L^1(\Omega, \mathcal{G}, P|_{\mathcal{G}}\,; \mathbb{R}^d)$.

# Conditional Expectation

- Be $X : \Omega \to \mathbb{R}^{d'}$ a random variable and $\sigma(X) \equiv X^{-1}(\mathcal{B}(\mathbb{R}^d)) \subset \mathcal{A}$ the $\sigma$-algebra, generated by $X$. Then there is $\mathbb{E}[Y \mid X] \equiv \mathbb{E}[Y \mid \sigma(X)]$. In particular if $Y \geq 0$ and $d = d' = 1$:

$$\mathbb{E}[Y \mid X] = \frac{d(Q_Y \circ X^{-1})}{d(P \circ X^{-1})} = \frac{dX_* Q_Y}{dP_X}$$

- If $Y$ is independent of $\mathcal{G}$, then $\mathbb{E}^{\mathcal{G}}[Y](\omega) = \mathbb{E}[Y] \quad \forall \omega \in \Omega$, i.e. $\mathbb{E}^{\mathcal{G}}[Y]$ is the constant random variable $\mathbb{E}[Y]$.
- Stability: If $Y$ is $(\mathcal{G}, \mathcal{B}(\mathbb{R}^d))$-measurable, i.e. $\sigma(Y) \subseteq \mathcal{G}$, it holds that $\mathbb{E}[Y \mid \mathcal{G}] = Y$. In particular $\mathbb{E}[Y \mid \mathcal{A}] = Y$ since $Y$ is $(\mathcal{A}, \mathcal{B}(\mathbb{R}^d))$-measurable.
- Law of total expectation: $\mathbb{E}[\mathbb{E}^{\mathcal{G}}[Y]] = \mathbb{E}[Y]$
- $\mathbb{E}[\cdot \mid \mathcal{G}] : L^1(\Omega, \mathcal{A}, P; \mathbb{R}^d) \to L^1(\Omega, \mathcal{G}, P|_{\mathcal{G}}; \mathbb{R}^d)$ is a continuous (see below) linear operator.

# Conditional Expectation

- Tower rule: For $\mathcal{H} \subset \mathcal{G} \subset \mathcal{A}$ there is $\mathbb{E}[\mathbb{E}[Y \mid \mathcal{G}] \mid \mathcal{H}] = \mathbb{E}[Y \mid \mathcal{H}]$. $\mathbb{E}[Y \mid \mathcal{H}] = \mathbb{E}[\mathbb{E}[Y \mid \mathcal{H}] \mid \mathcal{G}]$ because of the stability property.
- Generalized Jensen inequality:

$$\phi\left(\mathbb{E}^{\mathcal{G}}[Y]\right) \leq \mathbb{E}^{\mathcal{G}}[\phi(Y)]$$

- For $p \in [1, +\infty]$:
$Y \in L^p(\Omega, \mathcal{A}, P; \mathbb{R}^d) \Rightarrow \mathbb{E}^{\mathcal{G}}[Y] \in L^p(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ because $|\mathbb{E}^{\mathcal{G}}[Y]|^p \leq \mathbb{E}^{\mathcal{G}}[|Y|^p]$ implies $\mathbb{E}[|\mathbb{E}^{\mathcal{G}}[Y]|^p] \leq \mathbb{E}[\mathbb{E}^{\mathcal{G}}[|Y|^p]]$ $= \mathbb{E}[|Y|^p]$ and therefore $\|\mathbb{E}^{\mathcal{G}}[Y]\|_p \leq \|Y\|_p$.
- Since $\|\mathbb{E}^{\mathcal{G}}[1_A]\|_p \leq \|1_A\|_p \leq 1$ with $A \in \mathcal{A}$, $\mathbb{E}^{\mathcal{G}}[1_A]$ is a random variable in $L^p(\Omega, \mathcal{A}, P; [0, 1])$ for all $p \in (0, +\infty]$.

# Conditional Expectation

- Conditional covariance :
  $\mathrm{Cov}^{\mathcal{G}}[X, Y] \equiv \mathbb{E}^{\mathcal{G}}[(X - \mathbb{E}^{\mathcal{G}}[X])(Y - \mathbb{E}^{\mathcal{G}}[Y])^{\top}]$.
- Conditional variance :
  $\mathrm{Var}^{\mathcal{G}}[X] \equiv \mathbb{E}^{\mathcal{G}}[(X - E^{\mathcal{G}}[X])(X - E^{\mathcal{G}}[X])^{\top}]$.
- Law of total covariance:
  $\mathrm{Cov}[X, Y] = \mathbb{E}[\mathrm{Cov}^{\mathcal{G}}[X, Y]] + \mathrm{Cov}[\mathbb{E}^{\mathcal{G}}[X], \mathbb{E}^{\mathcal{G}}[Y]]$
- Law of total variance:
  $\mathrm{Var}[X] = \mathbb{E}[\mathrm{Var}^{\mathcal{G}}[X]] + \mathrm{Var}[\mathbb{E}^{\mathcal{G}}[X]]$

# Semigroups of Kernels

- A (transition) kernel from measure space $(\Omega, \mathcal{A})$ to a measure space $(\mathrm{E}, \mathcal{E})$ is a mapping $K : \Omega \times \mathcal{E} \to \overline{\mathbb{R}}_+$ s.t. $K(\omega, \cdot) : E \mapsto K(\omega, E)$ is a measure on $\mathcal{E}$ for every $\omega \in \Omega$ and $K(\cdot, E) : \omega \mapsto K(\omega, E)$ is $(\mathcal{A}, \mathcal{B}(\overline{\mathbb{R}}_+))$-measurable for every $E \in \mathcal{E}$.
- $K$ is called a sub-Markov or sub-probability kernel if $K(\omega, \mathrm{E}) \leq 1$, a probability or Markov kernel if $K(\omega, \mathrm{E}) = 1$. In these cases obviously $K : \Omega \times \mathcal{E} \to [0, 1]$.
  A Markov kernel is a generalisation of the transition matrix of finite state space Markov processes to arbitrary Markov processes (see section on stochastic processes below).
- $K$ is called $\sigma$-finite if $K(\omega, \cdot)$ is $\sigma$-finite for every $\omega \in \Omega$. $K$ is called finite and bounded if $K(\omega, \mathrm{E})$ is so for every $\omega \in \Omega$.

# Semigroups of Kernels

- Denote $\mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ the set of kernels from measure space $(\Omega, \mathcal{A})$ to a measure space $(\mathsf{E}, \mathcal{E})$. Denote $\mathcal{K}_\sigma(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$, $\mathcal{K}_f(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$, $\mathcal{K}_b(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ and $\mathcal{K}_{\leq 1}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$, $\mathcal{K}_1(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ the subsets of $\sigma$-finite, finite, bounded and of (sub-)Markov kernels respectively.

$$\mathcal{K}_1(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E}) \subseteq \mathcal{K}_{\leq 1}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E}) \subseteq \mathcal{K}_b(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E}) \subseteq \\ \subseteq \mathcal{K}_f(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E}) \subseteq \mathcal{K}_\sigma(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$$

- $\mathcal{K}_\sigma(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$, $\mathcal{K}_f(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ and $\mathcal{K}_b(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ are convex cones.

# Semigroups of Kernels

- $1_\Omega : \Omega \times \mathcal{A} \longrightarrow [0,1]$ with $1_\Omega(\omega, A) \equiv 1_A(\omega)$, is a Markov kernel from $(\Omega, \mathcal{A})$ to itself where $1_\Omega(\cdot, A) : \omega \mapsto 1_A(\omega)$ is $(\mathcal{A}, \mathcal{B}([0,1]))$-measurable for every $A \in \mathcal{A}$ and where $1_\Omega(\omega, \cdot) : A \mapsto 1_A(\omega)$ is the Dirac measure $\delta_\omega$. $1_\Omega \in \mathcal{K}_1(\Omega, \mathcal{A})$ is called the unit kernel.
- Any $(\mathcal{A}, \mathcal{E})$-measurable function $f \colon (\Omega, \mathcal{A}) \to (\mathrm{T}, \mathcal{E})$ induces a Markov kernel $K_f : \Omega \times \mathcal{E} \longrightarrow [0,1]$ from $(\Omega, \mathcal{A})$ to $(\mathrm{T}, \mathcal{E})$ with $K_f(\omega, E) \equiv f_* \delta_\omega(E) = \delta_{f(\omega)}(E)$.
- In general, image measures $f_* \nu$ with $\nu$ a measure on $\mathcal{A}$ are kernels, independent from $\omega$ (Markov kernels, if $\nu$ is a probability measure).

# Semigroups of Kernels

- A kernel $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathrm{E}, \mathcal{E})$ induces two natural integral operators $(\cdot\, K) : \mathcal{M}^+(\Omega, \mathcal{A}) \longrightarrow \mathcal{M}^+(\mathrm{E}, \mathcal{E})$, defined by

  $$(\nu\, K)\,(E) \equiv \int_\Omega \nu(d\omega)\, K(\omega, E) \quad , \nu \in \mathcal{M}^+(\Omega, \mathcal{A}), \forall E \in \mathcal{E}$$

  and $(K\, \cdot) : \mathcal{L}^0_+(\mathrm{E}, \mathcal{E}; \mathbb{R}) \longrightarrow \mathcal{L}^0_+(\Omega, \mathcal{A}; \mathbb{R})$, defined by

  $$(K\, f)\,(\omega) \equiv \int_{\mathrm{E}} K\,(\omega, d\epsilon)\, f(\epsilon) \quad , f \in \mathcal{L}^0(\mathrm{E}, \mathcal{E}; \mathbb{R}), \forall \omega \in \Omega$$

- $L^0_+(\mathrm{E}, \mathcal{E}, Q)$ -> lift if $K\,(\omega, \cdot) \ll Q$
- Unit kernel $1_\Omega \in \mathcal{K}_1(\Omega, \mathcal{A})$: $(\nu\, 1_\Omega) = \nu$ for $\nu \in \mathcal{M}^+(\Omega, \mathcal{A})$ and $(1_\Omega\, f) = f$ for $f \in \mathcal{L}^0_+(\Omega, \mathcal{A}; \mathbb{R})$.
- $K \in \mathcal{K}_1(\Omega, \mathcal{A}; \mathrm{E}, \mathcal{E}) \implies (\cdot\, K) : \mathcal{M}^+_1(\Omega, \mathcal{A}) \longrightarrow \mathcal{M}^+_1(\mathrm{E}, \mathcal{E})$.

# Semigroups of Kernels

- Product of kernels: For $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ and $L \in \mathcal{K}(\Omega \times \mathsf{E}, \mathcal{A} \times \mathcal{E}; \Sigma, \mathcal{S})$ define a product $K \otimes L \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E} \times \Sigma, \mathcal{E} \times \mathcal{S})$ like so :

$$K \otimes L\,(\omega, S) \equiv \int_{\Omega} K(\omega, d\epsilon) \int_{\Sigma} L((\omega, \epsilon), d\sigma)\, 1_S((\epsilon, \sigma))$$

  for all $\omega \in \Omega, S \in \mathcal{E} \otimes \mathcal{S}$.
- Note that $L \in \mathcal{K}(\mathsf{E}, \mathcal{E}; \Sigma, \mathcal{S})$ is an element of $\mathcal{K}(\Omega \times \mathsf{E}, \mathcal{A} \times \mathcal{E}; \Sigma, \mathcal{S})$ without $(\Omega, \mathcal{A})$-dependency.
- $K \in \mathcal{K}_f(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E}) \wedge L \in \mathcal{K}_f(\Omega \times \mathsf{E}, \mathcal{A} \times \mathcal{E}; \Sigma, \mathcal{S}) \implies K \otimes L \in \mathcal{K}_\sigma(\Omega, \mathcal{A}; \mathsf{E} \times \Sigma, \mathcal{E} \times \mathcal{S})$.
- The Markov property of kernels is preserved.

# Semigroups of Kernels

- Consider kernels $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ and $L \in \mathcal{K}(\mathsf{E}, \mathcal{E}; \Sigma, \mathcal{S})$. Their composition $K\,L$ is defined as

$$K\,L(\omega, S) \equiv \int_{\mathsf{E}} K(\omega, d\epsilon)\, L(\epsilon, S) \quad \omega \in \Omega, S \in \mathcal{S}$$

  $K\,L$ is again a kernel, in fact $K\,L \in \mathcal{K}(\Omega, \mathcal{A}; \Sigma, \mathcal{S})$
- In particular: $K \in \mathcal{K}_f(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E}) \wedge L \in \mathcal{K}_f(\mathsf{E}, \mathcal{E}; \Sigma, \mathcal{S}) \implies K\,L \in \mathcal{K}_f(\Omega, \mathcal{A}; \Sigma, \mathcal{S})$
- Same holds for bounded and (sub-) Markov kernels.

# Semigroups of Kernels

- $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E}) \wedge 1_\Omega \in \mathcal{K}_1(\Omega, \mathcal{A}) \implies 1_\Omega K = K$.
- $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E}) \wedge 1_\mathsf{E} \in \mathcal{K}_1(\mathsf{E}, \mathcal{E}) \implies K 1_\mathsf{E} = K$.
- Notation: For $K \in \mathcal{K}(\Omega, \mathcal{A})$ define $K^n \equiv K K \cdots K$ ($n$ times)
- Associativity: For $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$, $L \in \mathcal{K}(\mathsf{E}, \mathcal{E}; \Sigma, \mathcal{S})$, $M \in \mathcal{K}(\Sigma, \mathcal{S}; \mathsf{T}, \mathcal{T})$, $c \in \mathbb{R}_+$, $\nu \in \mathcal{M}^+(\Omega, \mathcal{A})$, $f \in \mathcal{L}^0(\mathsf{E}, \mathcal{E}; \mathbb{R})$ and $g \in \mathcal{L}^0(\Sigma, \mathcal{S}; \mathbb{R})$

$$c(\nu K) = (c\nu) K \ , \ c(K f) = (c K) f \ , \ (\nu K) f = \nu (K f)$$
$$(\nu K) L = \nu (K L) \ , \ K(L g) = (K L) g \ , \ c(K L) = (c K) L$$
$$(K L) M = K(L M)$$

# Semigroups of Kernels

- Distributivity: For $K, L \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$, $M, N \in \mathcal{K}(\mathsf{E}, \mathcal{E}; \Sigma, \mathcal{S})$, $c \in \mathbb{R}_+$, $\nu \in \mathcal{M}^+(\Omega, \mathcal{A})$ and $f, g \in \mathcal{L}^0(\mathsf{E}, \mathcal{E}; \mathbb{R})$

$$(K + L)\, M = K\, M + K\, L \ \ , \ K\,(M + N) = K\, M + K\, N$$
$$\nu\,(K + L) = \nu\, K + \nu\, L \ \ , \ (K + L)\, f = K\, f + L\, f$$

$$(\nu + \mu)\, K = \nu\, K + \mu\, K \ \ , \ K\,(f + g) = K\, f + K\, g$$
$$\nu\,(f + g) = \nu\, f + \nu\, g \ \ , \ (\nu + \mu)\, f = \nu\, f + \mu\, f$$

- From the above we see that $(\cdot\, K) : \mathcal{M}^+(\Omega, \mathcal{A}) \longrightarrow \mathcal{M}^+(\mathsf{E}, \mathcal{E})$ and $(K\, \cdot) : \mathcal{L}^0(\mathsf{E}, \mathcal{E}; \mathbb{R}) \longrightarrow \mathcal{L}^0(\Omega, \mathcal{A}; \mathbb{R})$ are linear operators.
- left/right eigenvectors ? $\nu\, K = \nu$ and $K\, f = f$ the invariants from $\mathcal{M}^+(\Omega, \mathcal{A})$ and $\mathcal{L}^0(\mathsf{E}, \mathcal{E}; \mathbb{R})$ respectively.

# Semigroups of Kernels

- Two probability spaces $(\Omega, \mathcal{A}, P)$ and $(\mathrm{E}, \mathcal{E}, Q)$ and a measurable function $k \in \mathcal{L}^0(\Omega \times \mathrm{E}, \mathcal{A} \otimes \mathcal{E}; \mathbb{R}_+)$. $k$ defines a kernel $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathrm{E}, \mathcal{E})$ like so

$$K(\omega, E) \equiv \int_E Q(d\epsilon)\, k(\omega, \epsilon)$$

  and is called a kernel (density) function (w.r.t to $Q$).
- For a $f \in \mathcal{L}^0(\mathrm{E}, \mathcal{E}, Q; \mathbb{R})$ is

$$(K f) = \int_{\mathrm{E}} Q(d\epsilon)\, k(\cdot, \epsilon)\, f(\epsilon) \in \mathcal{L}^0(\Omega, \mathcal{A}; \mathbb{R}) \quad .$$

- The analogous left operation :

$$(f K) = \int_{\Omega} P(d\omega)\, k(\omega, \cdot)\, f(\omega) \in \mathcal{L}^0(\mathrm{E}, \mathcal{E}; \mathbb{R}) \quad .$$

# Semigroups of Kernels

- Write $f \in \mathcal{L}^0_+(\Omega, \mathcal{A}, P; \mathbb{R})$ as a R-N derivative $d\mu/dP$ with a $\mu \in \mathcal{M}^+(\Omega, \mathcal{A})$. Then is $fK = d(\mu K)/dQ$.
- Be $k \in \mathcal{L}^0(\Omega \times \mathrm{E}, \mathcal{A} \otimes \mathcal{E}; \mathbb{R}_+)$ a kernel function with probability spaces $(\Omega, \mathcal{A}, P)$ and $(\mathrm{E}, \mathcal{E}, Q)$. Then $k$ is called a double stochastic kernel function if both

$$\int_\Omega P(d\omega)\, k(\omega, \epsilon) = 1 \quad \forall \epsilon \in \mathrm{E}$$
$$\int_\mathrm{E} Q(d\epsilon)\, k(\omega, \epsilon) = 1 \quad \forall \omega \in \Omega$$

- A kernel function $k \in \mathcal{L}^0(\Omega \times \Omega, \mathcal{A} \otimes \mathcal{A}; \mathbb{R}_+)$ is called symmetric if $k(\omega, \omega') = k(\omega', \omega)$ for all $(\omega, \omega') \in \Omega \times \Omega$.

# Semigroups of Kernels

- Recover the kernel from its operators: $(K1_E)(\omega) = K(\omega, E)$ and $(\delta_\omega K)(E) = K(\omega, E)$.
- The common use case is $(\Omega, \mathcal{A}, P) = (\mathsf{E}, \mathcal{E}, Q)$. $\mathcal{K}_f(\Omega, \mathcal{A})$, $\mathcal{K}_b(\Omega, \mathcal{A})$, $\mathcal{K}_{\leq 1}(\Omega, \mathcal{A})$, $\mathcal{K}_{\leq 1}(\Omega, \mathcal{A})$ and $\mathcal{K}_1(\Omega, \mathcal{A})$ are semigroups w.r.t. the composition.

# Semigroups of Kernels

- Consider a family of kernels $K = (K_t)_{t \in \mathbb{R}_+}$ on $(\mathrm{E}, \mathcal{E})$. If $K$ fullfills the Chapman-Kolmogorv equations, i.e. if it holds that

  $$K_{s+t} = K_s K_t \equiv K_s \circ K_t \ \ (\forall\ s, t \in \mathbb{R}_+) \ ,$$

  that is

  $$K_{s+t}(\epsilon, E) = \int_{\mathrm{E}} K_s(\epsilon, d\epsilon')\, K_t(\epsilon', E) \ \ (\forall\ (\epsilon, E) \in \mathrm{E} \times \mathcal{E}, s, t \in \mathbb{R}_+) \ ,$$

  then we speak of a semigroup of kernels. It is obviously commutative $(s + t = t + s)$.
- A semigroup of kernels is called (sub-) Markovian if all its members have this property. Normal, if $K_0 = 1$, the unit kernel.

# Conditional Probability

- Conditional probability of $A \in \mathcal{A}$ given a $\sigma$-algebra $\mathcal{G} \subset \mathcal{A}$ (generalizing $P(A) = \mathbb{E}[1_A]$) :

$$\mathbb{P}^{\mathcal{G}}[A] \equiv \mathbb{P}[A \mid \mathcal{G}] \equiv \mathbb{E}^{\mathcal{G}}[1_A]$$

  In particular for $B \in \mathcal{G}$ :

$$\mathbb{P}[A \mid B] \equiv \mathbb{P}[A \mid \sigma(B)]$$

- $$\mathbb{P}[A \mid B](\omega) = \begin{cases} P(A \mid B) & , \omega \in B \\ P(A \mid B^{\complement}) & , \omega \notin B \end{cases}$$

# Conditional Probability

- For $B \in \mathcal{G}$
$$\int_\Omega dP\, 1_B\, \mathbb{P}^{\mathcal{G}}[A] = \int_\Omega dP\, 1_B\, 1_A$$
and therefore
$$\int_B dP\, \mathbb{P}^{\mathcal{G}}[A] = P(A \cap B)$$
- For $A \in \mathcal{A}$, $B \in \mathcal{G} \subseteq \mathcal{A}$ with $P(B) > 0$
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \int_B dP\, \mathbb{P}^{\mathcal{G}}[A] \Big/ \int_\Omega dP\, \mathbb{P}^{\mathcal{G}}[A]$$

# Conditional Probability

- $\mathbb{P}^{\mathcal{G}} : \mathcal{A} \to L^1(\Omega, \mathcal{G}, P|_{\mathcal{G}}; \mathbb{R})$ is an $L^1(\Omega, \mathcal{G}, P|_{\mathcal{G}}; \mathbb{R})$-valued vector measure.
- Note that $A \mapsto \mathbb{P}^{\mathcal{G}}[A](\omega)$, for a fixed $\omega \in \Omega$, does not constitute a probability measure because $\mathbb{P}^{\mathcal{G}}[A]$ is only defined a.s., which leads into trouble if we deal with uncountable families of events. Some regularity conditions on $(\Omega, \mathcal{A})$ are needed. See below.
- Conditional probability of a random variable $Y$ given $\mathcal{G}$ (generalizing $P_Y(B) = P(Y \in B) = \mathbb{E}[1_{\{Y \in B\}}]$) :

$$\mathbb{P}^{\mathcal{G}}[Y \in B] = \mathbb{E}^{\mathcal{G}}[1_{\{Y \in B\}}]$$

for $B \in \mathcal{B}(\mathbb{R}^d)$.

# Conditional Probability

- Be $Y \in \mathcal{L}^0(\Omega, \mathcal{A}, P; \mathrm{E}, \mathcal{E})$ an $(\mathrm{E}, \mathcal{E})$-valued random element. A regular conditional distribution of $Y$, given $\mathcal{G}$, is a Markov kernel $K_{Y|\mathcal{G}} : \Omega \times \mathcal{E} \to \overline{\mathbb{R}}_+$ from $(\Omega, \mathcal{A})$ to $(\mathrm{E}, \mathcal{E})$ s.t. for all $E \in \mathcal{E}$

$$K_{Y|\mathcal{G}}(\omega, E) = \mathbb{P}^{\mathcal{G}}[Y \in E](\omega) \quad \text{for a.e. } \omega \in \Omega.$$

- If $(\mathrm{E}, \mathcal{E}) = (\Omega, \mathcal{A})$ and $Y(\omega) = \omega$, the regular conditional distribution of $Y$ (if it exists) is called the regular conditional probability because $K_{Y|\mathcal{G}}(\cdot, E) = \mathbb{P}^{\mathcal{G}}[Y \in E] = \mathbb{P}^{\mathcal{G}}[E]$ a.e. (see slides above).
- $K_{Y|\mathcal{G}}$ is obviously not unique and does not exist for an arbitrary target space of the random element $Y$. It can be shown to exist for Polish spaces and $\mathbb{R}^d$ in particular.

# Conditional Probability

- LOTUS: Be $Y \in \mathcal{L}^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ a random variable and $g \in \mathcal{L}^0(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d); \mathbb{R})$ with $g(Y) \in L^1(\Omega, \mathcal{A}, P; \mathbb{R}^d)$. Then is

$$\int_{\mathbb{R}^d} K_{Y|\mathcal{G}}(\cdot, dx)\, g(x) \in \mathcal{L}^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$$

and

$$\mathbb{E}^{\mathcal{G}}[g(Y)](\omega) = \int_{\mathbb{R}^d} K_{Y|\mathcal{G}}(\omega, dx)\, g(x) \quad P\text{-a.e.} \quad .$$

In particular

$$\mathbb{E}[g(Y) \mid X](\omega) = \int_{\mathbb{R}^d} K_{Y|X}(\omega, dx)\, g(x) \quad P\text{-a.e.} \quad .$$

# Conditional Probability

- For $Y, X \in \mathcal{L}^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ and $\mathcal{G} \equiv \sigma(X) \subset \mathcal{A}$ examine $\mathbb{E}^{\mathcal{G}}[Y] = \mathbb{E}[Y \mid X]$. It can be shown (Doob-Dynkin lemma) that there exists a measurable function $h \in \mathcal{L}^0(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d); \mathbb{R}, \mathcal{B}(\mathbb{R}))$ s.t. $\mathbb{E}[Y \mid X] = h(X)$. $h$ is unique $P_X$-a.e. . For $B \in \mathcal{B}(\mathbb{R}^d)$, it holds that

$$\begin{aligned}\int_{\mathbb{R}^d} dP_X(dx)\, 1_B(x)\, h(x) &= \int_{\{X \in B\}} dP\, h \circ X \\ &= \int_{\{X \in B\}} dP\, \mathbb{E}[Y \mid X] = \int_{\{X \in B\}} dP\, X\end{aligned}$$

If $B = \{x\}, x \in \mathbb{R}^d$, there is

$$h(x)\, P(X = x) = \int_{\{X = x\}} dP\, X = \mathbb{E}\left[1_{\{X=x\}}\, Y\right]$$

# Conditional Probability

- Note that the above construction makes sense even with $P(X = x) = 0$ (the usual situation). We define

$$\mathbb{E}[Y \mid X = x] \equiv h(x) \text{ for any } x \in \mathbb{R}^d \quad .$$

(Notation mimics the discrete situation.)

- $$\mathbb{E}\,[Y \mid X](\omega) = \mathbb{E}[Y \mid X = X(\omega)\,] \quad P\text{-a.e.}$$

and

$$K_{Y|X}(\omega, E) = \mathbb{P}[Y \in E \mid X = X(\omega)\,] \quad P\text{-a.e.} \quad .$$

# Conditional Probability Density

- Consider two random variables $Y, X \in \mathcal{L}^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ with their joint pdf $p(x,y)$. $p(x)$ is the marginal pdf w.r.t. $X$, i.e.

$$p(x) \equiv \int_{\mathbb{R}^d} \lambda^d(dy)\, p(x,y) \quad ,$$

which we assume to be strictly positive. Define the conditional probability density function of $Y$, given $X(\omega) = x$, as

$$p(y \mid x) \equiv \frac{p(x,y)}{p(x)} \quad .$$

If $X$ and $Y$ are independent, it holds that $p(x,y) = p(x)\, p(y)$, hence

$$p(y \mid x) = p(y) \quad .$$

# Conditional Probability

- Under the above prerequisites it can be shown that

$$\mathbb{E}[Y \mid X = x] = \int_{\mathbb{R}^d} \lambda^d(dy)\, p(y \mid x)\, y \quad P_X\text{-a.e.}$$

  and that

$$\mathbb{E}[Y \mid X](\omega) = \int_{\mathbb{R}^d} \lambda^d(dy)\, p(y \mid X(\omega))\, y \quad P\text{-a.e.} \quad .$$

- $$\mathbb{E}[g(Y) \mid X = x] = \int_{\mathbb{R}^d} \lambda^d(dx)\, p(y \mid x)\, g(x)$$

$$\mathbb{E}[g(Y) \mid X](\omega) = \int_{\mathbb{R}^d} K_{Y|X}(\omega, dx)\, g(x)$$

# Conditional Probability Density

- Bayes theorem: 'Mathematically easy, conceptually powerful.'

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

# 7. Gaussian Measures

# Characteristic Functions and Moment generation

- The characteristic function $\phi_\nu : \mathbb{R}^d \to \mathbb{C}$ of a measure $\nu$ on $\mathcal{B}(\mathbb{R}^d)$ is defined as its Fourier transform

$$\phi_\nu(t) \equiv \mathcal{F}[\nu](t) = \int_{\mathbb{R}^d} \nu(dx)\, e^{i\,t^\top x} \text{ for } t \in \mathbb{R}^d \quad .$$

- The characteristic function $\phi_X$ of a random variable $X$ is defined as the characteristic function of its distribution

$$\phi_X(t) \equiv \phi_{P_X}(t) = \int_{\mathbb{R}^d} P_X(dx)\, e^{i\,t^\top x} \text{ for } t \in \mathbb{R}^d \quad .$$

- Change-of-variables means

$$\phi_X(t) = \int_\Omega dP\, e^{i\,t^\top X}$$

and therefore

$$\phi_X(t) = \mathbb{E}[e^{i\,t^\top X}] \quad .$$

# Characteristic Functions and Moment generation

- If $X$ has a pdf $p(x)$ with $P_X \ll \lambda^d$, it holds that

  $$\phi_X(t) = \mathbb{E}[e^{i\,t^\top X}] = \int_{\mathbb{R}^d} \lambda^d(dx)\, e^{i\,t^\top x}\, p(x)$$

  i.e. $\phi_X(t) = \mathcal{F}[p](t)$, meaning the characteristic function is the Fourier transform of $p(x)$.
- $\phi_X$ always exists and is uniformly continuous.
- $|\phi_X(t)| = |\mathbb{E}_{p(x)}[e^{i\,t^\top X}]| \leq \mathbb{E}_{p(x)}[|e^{i\,t^\top X}|] = 1$ (Jensen inequality)
- $\overline{\phi_X(t)} = \phi_X(-t)$

# Characteristic Functions and Moment generation

- If $X = X_1 + \dots X_n$ is a sum of independent random variables

$$\phi_X(t) = \mathbb{E}_{p(x)}[e^{i\,t^\top X}] = \mathbb{E}_{p(x)}[e^{i\,t^\top (X_1 + \dots X_n)}] = \mathbb{E}_{p(x)}[\prod_{i=1}^{n} e^{i\,t^\top X_i}]$$

$$= \prod_{i=1}^{n} \mathbb{E}_{p(x_i)}[e^{i\,t^\top X_i}] = \prod_{i=1}^{n} \phi_{X_i}(t)$$

- Continuity theorem (Lévy): For a random variable $X \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ and a sequence of random variables $X_1, X_2, \dots$ with $X_i \in L^0(\Omega_i, \mathcal{A}_i, P_i; \mathbb{R}^d)$ (the probability spaces not necessarily identical) it holds that:

$$X_i \Rightarrow X \quad \text{iff} \quad \lim_{i \to +\infty} \phi_{X_i}(x) = \phi_X(x) \;\; \forall x \in \mathbb{R}^d$$

# Characteristic Functions and Moment generation

- Inversion formula: Be $\nu \in \mathcal{M}_1^+(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $(x, y) \in \mathcal{B}(\mathbb{R}^d)$ with $x, y \in \mathbb{R}^d$ and $T \in \mathbb{R}^d$. Then it holds that

$$\nu((x,y)) + 2^{-d}\nu(\{x,y\}) =$$

$$(2\pi)^{-d} \lim_{T\uparrow+\infty} \int_{[-T,T]} \lambda^d(ds) \left[\prod_{j=1}^{d} \frac{e^{-is_j x_j} - e^{-is_j y_j}}{is_j}\right] \phi_\nu(s)$$

- For two random variables $X$ and $Y$

$$\phi_X = \phi_Y \text{ iff } F_X = F_Y$$

# Characteristic Functions and Moment generation

- If $\phi_\nu \in L^1(\mathbb{R}^d)$ then $\nu \ll \lambda^d$ and

$$\frac{d\nu}{d\lambda^d} = \mathcal{F}^{-1}[\phi_\nu](x) \equiv (2\pi)^{-d} \int_{\mathbb{R}^d} \lambda^d(dt)\, \phi_\nu(t)\, e^{-i\,t^\top x}$$

- If $\phi_X \in L^1(\mathbb{R}^d)$, the pdf of $X$ is the inverse Fourier transformation of its characteristic function

$$p(x) = \mathcal{F}^{-1}[\phi_X](x) \equiv (2\pi)^{-d} \int_{\mathbb{R}^d} \lambda^d(dt)\, \phi_X(t)\, e^{-i\,t^\top x}$$

# Gaussian Measures

- Be $\lambda^d$ the Lebesgue-Borel measure on $\mathcal{B}(\mathbb{R}^d)$, $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ an arbitrary vector and symmetric and positive semidefinite matrix respectively.
- If $\Sigma$ in addition is positiv definite (equivalent to $\Sigma$ being nonsingular), the nondegenerate Gaussian measure $\gamma^n_{\mu,\Sigma} : \mathcal{B}(\mathbb{R}^d) \to [0,1]$ is defined as

$$\gamma^d_{\mu,\Sigma}(B) \equiv \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \int_B \lambda^d(dx) \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

- If $\Sigma = 0$, the degenerate Gaussian measure is defined as

$$\gamma^d_{\mu,\Sigma}(B) \equiv \delta_\mu(B)$$

# Gaussian Measures

- $\gamma^d_{\mu,\Sigma}$ is a probability measure (since $\gamma^d_{\mu,\Sigma}(\mathbb{R}^d) = 1$) and therefore finite (locally finite and $\sigma$-finite in particular). Since $\mathbb{R}^d$ is locally compact Hausdorff and it is locally finite, $\gamma^d_{\mu,\Sigma}$ is a Borel measure and therefore regular.
- If $\gamma^d_{\mu,\Sigma}$ is nondegenerate, obviously $\lambda^d \sim \gamma^d_{\mu,\Sigma}$ with RN derivative

$$\frac{d\gamma^d_{\mu,\Sigma}}{d\lambda^d}(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

- Notation

$$\mathcal{N}(x; \mu, \Sigma) \equiv \frac{d\gamma^d_{\mu,\Sigma}}{d\lambda^d}(x)$$

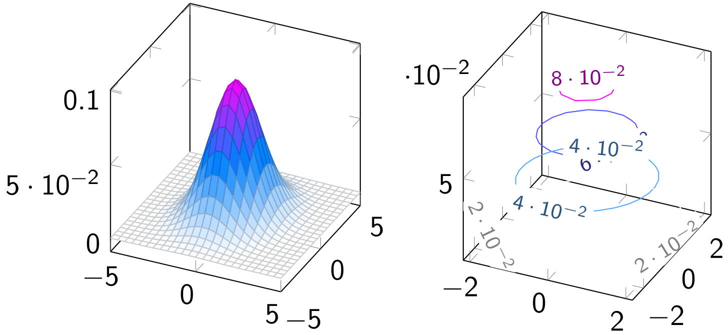
- In case $\gamma^d_{\mu,\Sigma}$ is degenerate, there is $\gamma^d_{\mu,\Sigma} \perp \lambda^d$.

# Gaussian Measures

- The pictures below show the situation in lower dimensions, where probability of $\gamma^d_{\mu,\Sigma}$ is concentrated around $\mu$. In higher dimensions the mass is concentrated near the boundary of a sphere around $\mu$.

# Gaussian Measures

The bivariate Gaussian pdf $\mathcal{N}(0, \Sigma)$ with $\Sigma = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$.
The figure on the right shows contours of constant density.



# Gaussian Measures

- The characteristic function of $\gamma^d_{\mu,\Sigma}$ :

$$\phi_{\gamma^d_{\mu,\Sigma}}(t) = \exp\left(it^\top \mu - \frac{1}{2}t^\top \Sigma t\right) .$$

  While $\gamma^d_{\mu,\Sigma}$ is not defined for a symmetric positiv semidefinite matrix $\Sigma$, its characteristic function is. Allows for a more general definition of a Gaussian measure:

- A measure $\gamma$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is called a Gaussian measure with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ if its characteristic function $\phi_\gamma$ is of the form

$$\phi_\gamma(t) = \exp\left(it^\top \mu - \frac{1}{2}t^\top \Sigma t\right)$$

  and $\Sigma$ is symmetric positive semidefinite.

# Gaussian Measures

- A general Gaussian measure is a probability measure.
- Since two probability measures are identical iff their characteristic functions are identical, there is for every $d \in \mathbb{N}$, $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ a unique Gaussian measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ with the respective characteristic function. Notation: $\gamma^d_{\mu,\Sigma}$
- $q_\gamma(t) \equiv t^\top \Sigma t$ is the associated quadratic form, $\Sigma_\gamma$ the covariance operator with $q_\gamma(t) = \langle \Sigma_\gamma t, t \rangle$.
- If $\gamma$ is a Gaussian measure on $\mathbb{R}^d$, it holds that

$$\mu_\gamma = \int_{\mathbb{R}^d} \gamma(dx)\, x$$

and

$$\langle \Sigma_\gamma u, v \rangle = \int_{\mathbb{R}^d} \gamma(dx)\, \langle u, x - \mu \rangle \, \langle v, x - \mu \rangle \quad \forall u, v \in \mathbb{R}^d$$

# Gaussian Measures

- Gaussian measures with mean zero are called centered Gaussian measures.
- Gaussian measures $\gamma^d_{0,I_d}$ are called standard Gaussian measures, denoted by $\gamma^d$.
- Note that for $\Sigma = 0$ we get the characteristic function of the degenerate Gaussian measure aka the dirac measure, which is considered a Gaussian measure under this definition.
- If $\Sigma$ is symmetric positive definite, this is the nondegenerate Gaussian measure.
- $\text{supp}(\gamma^d_{\mu,\Sigma}) = \{\mu + \Sigma x : x \in \mathbb{R}^d\}$

# Gaussian Random Variables

- A random variable $X$ on a probability space $(\Omega, \mathcal{A}, P)$ is called Gaussian or normally distributed with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ if $\mu = \mathbb{E}[X]$, $\Sigma = \mathrm{Cov}(X)$ and if $P_X = \gamma^d_{\mu,\Sigma}$, i.e.

$$\phi_X(t) = \exp\left(it^\top \mu - \frac{1}{2} t^\top \Sigma t\right)$$

- For a normally distributed $X$ it holds that $P_X \ll \lambda^d$ iff $X$ is nondegenerate. In that case

$$p_X(x) = \frac{dP_X}{d\lambda^d}(x) = \frac{d\gamma^d_{\mu,\Sigma}}{d\lambda^d}(x)$$

- Notation: $X \sim \mathcal{N}(\mu, \Sigma)$ and $p(x) = \mathcal{N}(x; \mu, \Sigma)$ if a pdf for $X$ exists.

# Gaussian Random Variables

- $X\colon (\Omega, \mathcal{A}, P) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $X \sim \mathcal{N}(\mu, \Sigma)$ implies $X \in L^2(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ (Note that $\mathbb{E}[|X|^2] = \mathrm{Tr}\{\mathrm{Cov}(X)\}$).
- If $X$ and $Y$ are Gaussian random variables, their joint random variable $(X, Y)$ does not need to be Gaussian. If $(X, Y)$ is Gaussian, its marginals are.
- A vector of (univariate) random variables $(X_1, \ldots, X_d)$ is a (multivariate) Gaussian random variable if, for any $\alpha_1, \ldots, \alpha_d \in \mathbb{R}$, the random variable $\sum \alpha_i X_i$ is Gaussian.
- If $X = (X_1, \ldots, X_d)^\top$ is a Gaussian random variable the $X_i$ are independent from each other iff they are uncorrelated.

# Gaussian Random Variables

- A different characterization of Gaussian random variables:

$$X \sim \mathcal{N}(\mu, \Sigma) \text{ iff}$$

  $\exists\, \mu \in \mathbb{R}^d, A \in \mathbb{R}^{d\times d}$ and a random variable $Z$ with $Z_i \sim \mathcal{N}(0,1)$ i.i.d. such that $X = A\,Z + \mu$ with $A^\top A = \Sigma$, i.e. $A = \sqrt{\Sigma}$ is the "root" of a Cholesky decomposition $\Sigma = A^\top A$ (decomposition is unique iff $\Sigma$ is symmetric positive definite)
- Every Gaussian random variable $X \sim \mathcal{N}(\mu, \Sigma)$ can be produced from a standard Gaussian random variable $Z$ via an affine transformation $T(Y) := \sqrt{\Sigma}\; Y + \mu$ .
- If $\Sigma$ is singular $rank(A) < n = \dim \operatorname{supp}(X_* P)$
- In particular $\phi_{T(Z)}(t) = e^{it^\top \mu}\, \phi_Z(AZ)$
- We will use this characterisation later with the so called reparameterization trick.

# Law of Large Numbers (LLN)

- Be $X_1, X_2, \ldots, X_n, \ldots$ an i.i.d. sequence of random variables with $X_i \in L^1(\Omega, \mathcal{A}, P; \mathbb{R}^d)$, $\mathbb{E}[X_i] = \mu$ and $S_n \equiv \sum_{i=1}^n X_i$. The weak LLN states that under these conditions the sample average
$$\overline{X_n} \equiv \frac{1}{n} S_n$$
converges in probability to the expected value $\mu$ for $n \to +\infty$ :
$$\overline{X_n} \xrightarrow{P} \mu$$

- The strong LLN states that under the above conditions it even holds that the convergence is a.s. :
$$\overline{X_n} \xrightarrow{a.s.} \mu$$

# Law of Large Numbers (LLN)

- Of course the strong LLN implies the weak LLN (since a.s. convergence implies convergence in probability). Examples with nonfinite expectation can be constructed where the weak LLN holds but not the strong LLN.
- $$\mathbb{E}[\overline{X_n}] = \frac{1}{n}\mathbb{E}[S_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mu$$
- Note that finite covariance of the $X_i$ is not required. If we assume $\mathrm{Cov}(X_i) = \Sigma$, it holds that

  $$\mathrm{Cov}(\overline{X_n}) = \frac{\mathrm{Cov}(S_n)}{n^2} = \frac{1}{n}\Sigma \quad .$$

  Large or infinite variance slows down the rate of convergence.

# Central Limit Theorem (CLT)

- Be $X_1, X_2, \ldots, X_n, \ldots$ an i.i.d. sequence of random variables with $X_i \in L^2(\Omega, \mathcal{A}, P; \mathbb{R}^d)$, $\mathbb{E}[X_i] = \mu$ and $\mathrm{Cov}(X_i) = \Sigma$. The Central Limit Theorem (CLT) states that under the above conditions the standardized sample average

$$\widetilde{X_n} \equiv \sqrt{n}\,(\overline{X_n} - \mu)$$

converges in distribution to a random variable $X \in L^2(\Omega', \mathcal{A}', P'; \mathbb{R}^d)$ with $X \sim \mathcal{N}(0, \Sigma)$ :

$$\widetilde{X_n} \Rightarrow X$$

# Central Limit Theorem (CLT)

- Note that $P_{S_n} = P_{X_1} * \cdots * P_{X_n}$.
- Intuition: The CLT underscores the exceptional status of the normal distribution. It serves as the blueprint for many ideas and concepts in probability theory/stochastic analysis.
- Tell something about rate of convergence here.

# 8. Stochastic Processes

# Stochastic Processes

- Consider a probability space $(\Omega, \mathcal{A}, P)$, a measurable space $(\mathrm{E}, \mathcal{E})$, an index set $T$ and a subset $U \subseteq \mathrm{E}^T$. Be $\pi_t : \mathrm{E}^T \to \mathrm{E}$ the projections. (If we understand $E^T$ as the space of functions $\mathrm{Fun}(T, \mathrm{E})$, the projections $\pi_t$, generators of $\mathcal{E}^T$, are the evaluation maps, i.e. $\pi_t f = f(t)$ for $f \in \mathrm{Fun}(T, \mathrm{E})$.) Note that $(U, (\mathcal{E}^T)_{|U} = \mathcal{E}^T \cap U)$ is again a measurable space.
  It now holds for a function $X \in \mathrm{Fun}(\Omega, U)$ that $X$ is an $(U, (\mathcal{E}^T)_{|U})$-valued random element iff $X_t = \pi_t \circ X$ is an $(\mathrm{E}, \mathcal{E})$-valued random element for every $t \in T$. That is

  $$X \in \mathcal{L}^0(\Omega, \mathcal{A}, P; U, (\mathcal{E}^T)_{|U}) \iff X_t \in \mathcal{L}^0(\Omega, \mathcal{A}, P; \mathrm{E}, \mathcal{E}) \;\; \forall t \in T$$

# Stochastic Processes

- A function $X : \Omega \to \mathsf{E}^T$ with the above property is called an $(\mathsf{E}, \mathcal{E})$-valued stochastic process on $T$ with paths in $U$. Equivalently an $(\mathsf{E}, \mathcal{E})$-valued stochastic process is a collection of $(\mathsf{E}, \mathcal{E})$-valued random elements $(X_t)_{t \in T}$.
  Notation: $(\Omega, \mathcal{A}, P; X = (X_t)_{t \in T}; \mathsf{E}, \mathcal{E})$ or simply $(X_t)_{t \in T}$ if the probability space and state space is understood.
- If $\mathsf{E}^T$ is a metric space: $(U, (\mathcal{E}^T)_{|U}) = (U, \mathcal{B}(U))$

# Stochastic Processes

- Depending on $T$, the process is either called a discrete-time stochastic process or a continuous-time stochastic process.
- If the $X_t$ are all discrete or all continuous random elements, the corresponding process is called a discrete-valued stochastic process or a continuous-valued stochastic process respectively.
- For a fixed $\omega$, the mapping $X(\cdot, \omega) : T \to \mathrm{E}, t \mapsto X_t(\omega)$ is called the sample function, realization, path or time series of the process $X$ at $\omega$.
- The probability measure $X_* P$ is called the distribution of the process.

# Stochastic Processes

- An $((\mathbb{R}^d)^T, \mathcal{B}(\mathbb{R}^d)^T)$-valued stochastic process is simply called a stochastic process. It is a $((\mathbb{R}^d)^T, \mathcal{B}(\mathbb{R}^d)^T)$-valued random variable

$$X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}^d)^T \text{ with } \omega \mapsto (X_t(\omega))_{t \in T} \ .$$

  $X$ is $(\mathcal{A}, \mathcal{B}(\mathbb{R}^d)^T)$-measurable iff each $X_t$ is $(\mathcal{A}, \mathcal{B}(\mathbb{R}^d))$-measurable.
- Note that for a Polish space $E$, $\mathbb{R}^d$ in particular, $\mathcal{B}(E)^T \equiv \bigotimes_T \mathcal{B}(E) \subseteq \mathcal{B}(\prod_{t \in T} E)$. $\mathcal{B}(E)^T = \mathcal{B}(\prod_{t \in T} E)$ holds if $T$ is countable. The inclusion is strict if $T$ is uncountable.

# Stochastic Processes

- Stochastic processes with the same state space and index set (not necessarily the same sample space) are called (stochastically) equivalent if they have the same finite-dimensional distributions.
- Two processes $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ with the same sample probability space $(\Omega, \mathcal{A}, P)$, the same state space $(\mathsf{E}, \mathcal{E})$ and index set $T$ are said to be modifications of each other if it holds that $P(X_t = Y_t) = 1$ for every $t \in T$. Modifications of each other have the same finite-dimensional distributions and are therefore equivalent.

# Stochastic Processes

- Two processes $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ with the same sample probability space $(\Omega, \mathcal{A}, P)$, the same state space $(\mathrm{E}, \mathcal{E})$ and index $T$ set are said to be indistinguishable if $P\left(\forall t \in T : X_t = Y_t\right) = 1$. Indistinguishable processes are modifications of each other but not the other way around.
- If the index set $T$ is countable, two processes are indistinguishable iff they are modifications of each other.

# Stochastic Processes

- A family of $\sigma$-algebras $(\mathcal{F}_t)_{t\in T'}$ where $\mathcal{F}_t \in \mathcal{P}(\Omega)$ and $T'$ is a set with total order $\leq$, is called a filtration if $\mathcal{F}_s \subseteq \mathcal{F}_t$ for $s \leq t$.
- Given a filtration $(\mathcal{F}_t)_{t\in T}$, its corresponding right- and left-continuous filtrations are defined by

  $$\mathcal{F}_t^+ \equiv \bigcap_{s>t} \mathcal{F}_s \ , \ \ \mathcal{F}_t^- \equiv \sigma\left(\bigcup_{s<t} \mathcal{F}_s\right)$$

  respectively. A filtration is called right-/left-continuous if it is equal to its corresponding right-/left-continuous filtration, continuous if it is both left- and right-continuous.

# Stochastic Processes

- Be $X = (X_t)_{t \in T}$ an $(\mathrm{E}, \mathcal{E})$-valued stochastic process on a probability space $(\Omega, \mathcal{A}, P)$. The $\sigma$-algebras $\mathcal{F}_t^X \equiv \sigma(X_s : s \leq t) \subseteq \mathcal{A}$ constitute the so called natural filtration of the process $X$. Intuition: $\mathcal{F}_t^X$ represents the information obtained observing the process $X$ up to point in time $t$.
- A process $X$ is adapted to a filtration $(\mathcal{F}_t)_{t \in T}$ if every $X_t$ is $(\mathcal{F}_t, \mathcal{B}(\mathbb{R}^d))$-measurable i.e. $\mathcal{F}_t^X \subseteq \mathcal{F}_t$. A process is always adapted to its natural filtration $(\mathcal{F}_t^X)_{t \in T}$ which is the smallest filtration it can be adapted to. Notation: $(X_t, \mathcal{F}_t)_{t \in T}$
- Full information filtration

# Construction of Stochastic Processes

- For a random process $(\Omega, \mathcal{A}, P; (x_t)_{t\in T}; \mathrm{E}, \mathcal{E})$, E Polish, and finite $T$-subsets $S \in \mathcal{F}(T)$, the random variables $X_S : \Omega \to \mathrm{E}^S, \omega \mapsto (X_s(\omega))_{s\in S}$ induce probability measures $P_{X_S}$ on $\mathcal{B}(\mathrm{E})^S$, the finite-dimensional distributions of $X$. ($\mathcal{B}(\mathrm{E}^S) = \mathcal{B}(\mathrm{E})^S$ here, since E is Polish)
- Consider $T$-subsets $S \subseteq R \subseteq T$ with the corresponding projections $\pi_{RS} : (\mathrm{E}^R, \mathcal{B}(\mathrm{E})^R) \to (\mathrm{E}^S, \mathcal{B}(\mathrm{E})^S)$. The $\pi_{RS}$ are $(\mathcal{B}(\mathrm{E})^R, \mathcal{B}(\mathrm{E})^S)$-measurable because they are continuous. For finite $T$-subsets $R, S \in \mathcal{F}(T)$ it holds that $\pi_{RS*}(P_{X_R}) = P_{X_S}$ i.e. the finite-dimensional distributions of $X$ form a projective system of probability measures $\{P_{X_S} : \mathcal{B}(\mathrm{E})^S \to [0,1]\}_{S\in\mathcal{F}(T)}$ with projective limit $\varprojlim_{S\in\mathcal{F}(T)} P_{X_S} = P_X$.

# Construction of Stochastic Processes

- Conversely the Kolmogorov extension theorem ensures that for a projective system of probability measures $Q \equiv \{Q_S : \mathcal{B}(\mathrm{E})^S \to [0,1]\}_{S \in \mathcal{F}(T)}$ there is a unique probability measure $Q_T \equiv \varprojlim_{S \in \mathcal{F}(T)} Q_S$ on $\mathcal{B}(\mathrm{E})^T$ s.t. $\pi_{TS*} Q_T = Q_S$ for every $S \in \mathcal{F}(T)$.
  In particular, there exists a canonical stochastic process $X^Q$ which has this projective system as its finite-dimensional distributions: $(\mathrm{E}^T, \mathcal{B}(\mathrm{E})^T, Q_T; X^Q)$ with $X^Q(\omega) \equiv \omega$.
- The set of paths of $X^Q$ is $\mathrm{E}^T$. In particular it may include discontinuous ones which causes some unwanted behaviour (nonmeasurability of certain events, ...). Is there an equivalent process with $\mathcal{C}(T, \mathrm{E}) \subset \mathrm{Fun}(T, \mathrm{E}) = \mathrm{E}^T$ the set of paths ?

# Construction of Stochastic Processes

- If the index set $T$ is a topological space, a stochastic process $(\Omega, \mathcal{A}, P; X = (X_t)_{t\in T}; \mathrm{E}, \mathcal{B}(\mathrm{E}))$, E Polish, is called sample-continuous if $X(\cdot, \omega) \in \mathcal{C}(T, \mathrm{E}) \subseteq \mathrm{Fun}(T, \mathrm{E})$ for $P$-almost all $\omega \in \Omega$. It is called continuous if $X(\cdot, \omega) \in \mathcal{C}(T, \mathrm{E})$ for all $\omega \in \Omega$ i.e. if it is a stochastic process with paths in $\mathcal{C}(T, \mathrm{E})$.
- A modification $Y$ of $X$ is called a (sample-) continuous modification if $Y$ is (sample-) continuous.
- A sample continuous process is indistinguishable from a continuous process, so in particuar has a continuous modification.

# Construction of Stochastic Processes

- For a sample-continuous process, the finite distributions and the overall distribution determine each other.
- Kolmogorov continuity theorem: For $T = \mathbb{R}_+$ be $(X_t)_{t\in\mathbb{R}_+}$ a stochastic process, meeting the condition

$$\mathbb{E}[\,|X_t - X_s|^{\alpha}\,] \leq K\,|t-s|^{1+\beta}$$

for some fixed $\alpha, \beta, \gamma \in \mathbb{R}_{>0}$ and any pair of $t, s \in \mathbb{R}_+$. Then there exists a continuous modification of $X$.

# Construction of Stochastic Processes

- A process $X : (\Omega, \mathcal{A}, P) \longrightarrow (\mathsf{E}^{\mathbb{R}_+}, \mathcal{B}(\mathsf{E})^{\mathbb{R}_+})$ with a Polish state space E and a continuous modification ($\mathcal{C} \equiv \mathcal{C}(\mathbb{R}_+, \mathsf{E})$)

$$
\begin{tikzcd}
\tilde{X} : (\Omega, \mathcal{A}, P) \arrow[rr] \arrow[dr] & & (\mathsf{E}^{\mathbb{R}_+}, \mathcal{B}(\mathsf{E})^{\mathbb{R}_+}) \\
& (\mathcal{C}, \mathcal{B}(\mathcal{C})) \arrow[ur, hook] &
\end{tikzcd}
$$

  is equivalent to the so-called $\mathcal{C}$-canonical process of its distribution

$$Y : (\mathcal{C}, \mathcal{B}(\mathcal{C}), Q^X) \longrightarrow (\mathcal{C}, \mathcal{B}(\mathcal{C}))$$

  where $Q^X(B \cap C) \equiv P_X(B)$ and $Y(f) \equiv f$.
- $\mathcal{C}$ with the topology of uniform convergence on compact $\mathbb{R}_+$-subsets is a Polish space with $\mathcal{B}(\mathcal{C}) = \mathcal{B}(\mathsf{E})^{\mathbb{R}_+} \cap \mathcal{C}$.

## Construction of Stochastic Processes

- $L^2$-modification of X ($\mathcal{L}^2 \equiv \mathcal{L}^2(\mathbb{R}_+, \lambda; \mathsf{E}), L^2 \equiv \mathcal{L}^2/\sim$) :

$$\begin{tikzcd}
\tilde{X} : (\Omega, \mathcal{A}, P) \arrow[rr] \arrow[dr] & & (\mathsf{E}^{\mathbb{R}_+}, \mathcal{B}(\mathsf{E})^{\mathbb{R}_+}) \\
& (\mathcal{L}^2, \mathcal{B}(\mathcal{L}^2)) \arrow[ur, hook] \arrow[d, "\pi"] & \\
& (L^2, \mathcal{B}(L^2)) &
\end{tikzcd}$$

- $\mathcal{L}^2$ is not a Banach space, but $L^2$ is, a Hilbert space even. Should consider processes $Z : (\Omega, \mathcal{A}, P) \to L^2(\mathbb{R}_+, \lambda; \mathsf{E})$.

# Construction of Stochastic Processes

- A stochastic process $X = (X_t)_{t \in \mathbb{R}_+}$ is called strictly stationary or strong stationary if its finite distributions $P_{X_K}$ $(K \in \mathcal{F}(\mathbb{R}_+))$, are shift invariant i.e. if $P_{X_K} = P_{X_{K+s}}$ for every $s \in \mathbb{R}_+$.
- If $X_t \in L^1(\Omega, \mathcal{A}, P; \mathbb{R}^d) \, \forall t \in T$, strictly stationary implies that $t \mapsto \mathbb{E}[X_t]$ is shift-invariant and therefore constant. If $X_t \in L^2(\Omega, \mathcal{A}, P; \mathbb{R}^d) \, \forall t \in T$, strictly stationary implies shift-invariant covariance.
- weak-sense stationarity or wide-sense stationarity requires shift-invariance of mean and covariance.

# Gaussian Processes

- A Gaussian process is a stochastic process $(\Omega, \mathcal{A}, P;\ X = (X_t)_{t \in \mathbb{R}_+})$ whose finite-dimensional distributions are Gaussian measures. In other words, every finite vector $(X_{t_1}, \ldots, X_{t_n})$ is a (multivariate) Gaussian random variable.
- A a stochastic process $X$ is Gaussian iff its distribution $P_X$ is a Gaussian measure (explain Gaussian measures on infinite dimensional spaces here).
- A Gaussian process with $\mathbb{E}[X_t] = 0\ \forall t \in \mathbb{R}_+$ is called a centered Gaussian process.

# Brownian Motion

- The physical phenomenon of Brownian motion can be modelled by a stochastic process $(\Omega, \mathcal{A}, P; (B_t, \mathcal{F}_t)_{t\in\mathbb{R}_+}; \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, adapted to a filtration $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$ and characterized by the following properties:
- Stationary Gaussian increments:

$$(B_t - B_s)_* P = \bigotimes_{[d]} \gamma^1_{0,t-s} \quad (0 \leq s < t)$$

- Independent increments:

$$(B_t - B_s) \perp\!\!\!\perp \mathcal{F}_s \qquad (0 \leq s < t)$$

- Sample-Continuity

# Brownian Motion

- $(B_t, \mathcal{F}_t)_{t \in \mathbb{R}_+}$ is called an $(\mathcal{F}_t)$-Brownian motion, simply Brownian motion if the filtration is the natural of $B$, denoted by $(\mathcal{F}_t^B)$.
- If $B_0 = 0$ $P$-a.e., $B$ is called standard or normal Brownian motion.
- For any $K \in \mathcal{F}(\mathbb{R}_+)$, a finite-dimensional distribution $P_{B_K}$ looks like

$$\gamma_{0,\Sigma_K}^{|K|} \quad \text{where } (\Sigma_K)_{ij} = k_i \wedge k_j, \forall k_i.k_j \in K \quad .$$

That is, a standard Brownian motion is a centered Gaussian process. This is not true for a general Brownian motion with an arbitrary initial distribution.

# Brownian Motion

- Be $B = (B_t)_{t\in\mathbb{R}_+} = (B_t^{(1)}, \ldots, B_t^{(d)})^\top_{t\in\mathbb{R}_+}$ a $d$-dimensional (standard) Brownian motion. Then its 1-dimensional component processes $(B_t^{(i)})_{t\in\mathbb{R}_+}$ are independent (standard) 1-dimensional Brownian motion processes. $B_t^{(1)}, \ldots, B_t^{(d)}$ are independent for every $t \in \mathbb{R}_+$ if $B$ is a standard Brownian motion.
- Any two $d$-dimensional Brownian motion processes $(\Omega, \mathcal{A}, P;\ (B_t)_{t\in\mathbb{R}_+})$ and $(\Omega', \mathcal{A}', P';\ (B'_t)_{t\in\mathbb{R}_+})$ with identical initial distributions, i.e. $P_{B_0} = P_{B'_0}$, are equivalent. Conversely any sample-continuous $d$-dimensional process, equivalent to a Brownian motion process, is itself a Brownian motion process with the same initial distribution.

# Brownian Motion

- As a consequence, every $d$-dimensional Brownian motion $B$ is equivalent to a second canonical process $(\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d), \mathcal{B}(\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)), Q^B\,;\, (Y_t)_{t\in\mathbb{R}_+})$ where $Y_t(f) \equiv f(t)$.
- For every $\nu \in \mathcal{M}_1^+(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ there is an equivalence class of Brownian motions with the initial distribution $\nu$. Notation for the representative of that class: $(\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d), \mathcal{B}(\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)), Q^\nu\,;\, (Y_t)_{t\in\mathbb{R}_+})$
- If $\nu \equiv \delta_x$ for some $x \in \mathbb{R}^d$, we define $Q^x \equiv Q^{\delta_x}$.
  $Y_{0*}Q^x = \delta_x$ i.e. $Y_0 = x \quad Q^x$-a.e. .
  $Q^0$ is called the ($d$-dimensional) Wiener measure.
- $$Q^\nu(A) = \int_{\mathbb{R}^d} \nu(dx)\, Q^x(A)$$

# Markov Processes

- Consider a stochastic process, adapted to a filtration, $(\Omega, \mathcal{A}, P; X = (X_t, \mathcal{F}_t)_{t \in T}; \mathrm{E}, \mathcal{E})$. $X$ is called a Markov process if

$$\mathbb{P}[X_t \in E \mid \mathcal{F}_s] = \mathbb{P}[X_t \in E \mid X_s] \quad P\text{-a.e.}$$

  for every pair $(s, t) \in \mathbb{R}_+ \times \mathbb{R}_+$ with $s \leq t$ and every $E \in \mathcal{E}$.
- Equivalently:

$$\mathbb{E}[f(X_t) \mid \mathcal{F}_s] = \mathbb{E}[f(X_t) \mid X_s] \quad P\text{-a.e.}$$

  for every pair $(s, t) \in \mathbb{R}_+ \times \mathbb{R}_+$ with $s \leq t$ and every $f \in B(\mathrm{E})$.
- Intuition: A process without memory.

# Markov Processes

- If E is Polish, the above condition implies (equality is $P$-a.e.)

$$K_{s,t}(\cdot, E) \equiv K_{X_t|\mathcal{F}_s}(\cdot, E) = K_{X_t|X_s}(\cdot, E) \quad .$$

- Time homogenous: $K_{s,t}(\cdot, E) = K_{t-s}(\cdot, E)$

# Markov Processes

- A Markov semigroup of kernels $K = (K_t)_{t \in \mathbb{R}_+}$ and a probability measure $\nu$ on a measure space $(\mathsf{E}, \mathcal{E})$ together induce a projective family of probability measures $(Q_J)_{J \in \mathcal{F}(\mathbb{R}_+)}$ on $(\mathsf{E}^{\mathbb{R}_+}, \mathcal{E}^{\mathbb{R}_+})$, defined by

$$Q_J^{K,\nu} = \nu \otimes \bigotimes_{i=1}^{n-1} K_{t_{i+1}-t_i}$$

  with $J \equiv \{t_1, \ldots, t_n\} \in \mathcal{F}(\mathbb{R}_+)$ and $t_1 < \cdots < t_n$.
- If $\mathsf{E}$ is Polish and $\mathcal{E} \equiv \mathcal{B}(\mathsf{E})$, the Kolmogorov extension theorem implies that this projective family of probability measures is the family of finite distributions of a unique $(\mathsf{E}, \mathcal{E} \equiv \mathcal{B}(\mathsf{E}))$-valued stochastic process.

# Markov Processes

- The above process is the canonical process associated to $(K_J)_{J\in\mathcal{F}(\mathbb{R}_+)}$ : $(\mathsf{E}^{\mathbb{R}_+}, \mathcal{B}(\mathsf{E})^{\mathbb{R}_+}, Q^{K,\nu}; X^K)$, $Q^{K,\nu} = \varprojlim_{J\in\mathcal{F}(\mathbb{R}_+)} Q_J^{K,\nu}$ and $X^K(\epsilon) \equiv \epsilon$. If $\nu \equiv \delta_\epsilon$ for $\epsilon \in \mathsf{E}$, we use the notation $Q^{K,\epsilon}$. Note that

$$Q^{K,\nu}(E') = \int_{\mathsf{E}} Q^{K,\epsilon}(E')\,\nu(d\epsilon) \quad \forall E' \in \mathcal{B}(\mathsf{E})^{\mathbb{R}_+} \quad .$$

For $n = 1$ and $E \in \mathcal{B}(\mathsf{E})$, there is $Q^{K,\epsilon}(X_t \in E) = K_t(\epsilon, E)$ and

$$\begin{aligned} Q^{K,\nu}(X_t \in E) &= \int_{\mathsf{E}} K_t(\epsilon, E)\,\nu(d\epsilon) \\ &= (\nu\, K_t)\,(E) \quad\quad \forall E \in \mathcal{B}(\mathsf{E}) \quad . \end{aligned}$$

# Markov Processes

- Consequently, an $(\mathrm{E}, \mathcal{E})$-valued stochastic process $(\Omega, \mathcal{A}, P; X = (X_t)_{t\in\mathbb{R}_+})$, comming from a Markov semigroup $(K_t)_{t\in\mathbb{R}_+}$ and an (initial) probability measure $\nu \in \mathcal{M}_1^+(\mathrm{E}, \mathcal{E})$, represents a Markov process w.r.t. to its natural filtration $(\mathcal{F}_t^X)_{t\in\mathbb{R}_+}$. It holds $P$-a.e. that

$$\mathbb{P}[X_t \in B \mid \mathcal{F}_s^X] = K_{t-s}(X_s, E) \quad \forall E \in \mathcal{E} \text{ and } s \leq t \in \mathbb{R}_+ \quad .$$

# Markov Processes

- A universal Markov process (with state space $(\mathrm{E}, \mathcal{E})$), denoted by $(\Omega, \mathcal{A}, (P^\epsilon)_{\epsilon \in \mathrm{E}}; X = (X_t)_{t \in \mathbb{R}_+}; \mathrm{E}, \mathcal{E})$, is a family of $(\mathrm{E}, \mathcal{E})$-valued stochastic processes $(\Omega, \mathcal{A}, P^\epsilon; X = (X_t)_{t \in \mathbb{R}_+}; \mathrm{E}, \mathcal{E})$, parametrized by E s.t.

  (i) $P^\epsilon(X_0 = \epsilon) = 1$ for all $\epsilon \in \mathrm{E}$,

  (ii) $\epsilon \mapsto P^\epsilon(A)$ is $\mathcal{E}$-measurable for every $A \in \mathcal{A}$,

  (iii) $\mathbb{P}^\epsilon[X_{s+t} \in E \mid \mathcal{F}_s^X](\omega) = \mathbb{P}^\epsilon[X_t \in E \mid X_s(\omega)]$
  $P^\epsilon$-a.e. for all $E \in \mathcal{E}$, $\epsilon \in E$ and $s, t \in \mathbb{R}_+$ .

  Notation: $\mathbb{P}^\epsilon, \mathbb{E}^\epsilon$ mean conditional probability and (conditional) expectation w.r.t. $P^\epsilon$.

# Markov Processes

- A normal Markov semigroup of kernels $K = (K_t)_{t\in\mathbb{R}_+}$ on a Polish space $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$ induces a universal Markov process $(\Omega, \mathcal{A}, (P^\epsilon)_{\epsilon\in\mathsf{E}}; (X_t)_{t\in\mathbb{R}_+}; \mathsf{E}, \mathcal{E})$ via

$$K_t(\epsilon, E) = P^\epsilon(X_t \in E) \ .$$

  and vice versa.
- So, for a chosen $\epsilon \in \mathsf{E}$, $K_t(\epsilon, E) \equiv P^\epsilon(X_t \in E)$ is the kernel of transition probabilities of $X$ with time difference $t$.

# Markov Processes

- Be $(\Omega, \mathcal{A}, P; X = (X_t)_{t \in \mathbb{R}_+}; \mathsf{E}, \mathcal{B}(\mathsf{E}))$ a stochastic process with $\mathsf{E}$ Polish and initial distribution $\nu \equiv P_{X_0}$. It then holds that the family of finite-dimensional distributions of $X$ is comming from a normal Markov semigroup and an initial distribution $\nu \in \mathcal{M}_1^+(\mathsf{E}, \mathcal{B}(\mathsf{E}))$ iff there exists a universal Markov process $(\Omega, \mathcal{A}, (P^\epsilon)_{\epsilon \in \mathsf{E}}; (X_t')_{t \in \mathbb{R}_+}; \mathsf{E}, \mathcal{B}(\mathsf{E}))$ s.t. $X$ is equivalent to $(\Omega, \mathcal{A}, Q^\nu; (X_t')_{t \in \mathbb{R}_+}; \mathsf{E}, \mathcal{B}(\mathsf{E}))$ with

$$Q^\nu(E) \equiv \int_{\mathsf{E}} Q^\epsilon(E)\, \nu(d\epsilon) \quad .$$

- Similar mechanism for other semigroups and initial distribution: If $\mathcal{C} \equiv \mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$, $(\mathcal{C}, \mathcal{B}(\mathcal{C}), (Q^x)_{x \in \mathbb{R}^d}; (X_t)_{t \in \mathbb{R}_+})$ is the translation-invariant universal Markov process of Brownian motion (see above).

# Markov Processes

- A (universal) Markov process with an at most countable index set $\mathbb{Z}_+$ is called a (universal) Markov chain.
- For a universal Markov chain $(\Omega, \mathcal{A}, (P^\epsilon)_{\epsilon \in \mathsf{E}}; (X_t)_{t \in \mathbb{N}_0}; \mathsf{E}, \mathcal{E})$ there is for every $\epsilon \in \mathsf{E}$, $E \in \mathcal{B}(\mathsf{E})$ and $t \in \mathbb{Z}_+$

$$K_1(X_t, E) = \mathbb{P}^\epsilon[X_{t+1} \in E \mid \mathcal{F}_t] \ .$$

Obviously we have $K_n = K_{n-1} \circ K_1$, resulting in the inductively generated Markov semigroup $(K_n)_{n \in \mathbb{Z}_+}$ .

# 10. Stochastic Differential Equations

# Stochastic Differential Equations

- Itô calculus coming soon ...

# References

Billingsley, P.: Probability and Measure (Wiley Series in Probability and Statistics) (2012)

Bauer, H.: Probability Theory (De Gruyter Studies in Mathematics, 23) (1995)

Kallenberg, O.: Foundations of Modern Probability Theory (SpringerNature) (2021)