Concepts of Measure and Probability Theory

Jens Christian Keil | j@polyhelion.com

September - November 2024
April - July 2025

London, UK

Table of contents

1. Introduction

These slides are the vastly enhanced version of a couple of talks I gave September to November 2024 and April to July 2025 in the course of two workshops on stochastic processes and machine learning. They are intended to give a structured overview of the measure and probability theory fundamentals to get a more 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 Ω and its power set $\mathcal{P}(\Omega)$. A subset $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is called a σ-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 ($\mathcal{A}$-) measurable sets.
- $\{\varnothing, \Omega\}$ is the smallest σ-algebra, the largest one is $\mathcal{P}(\Omega)$.
- A σ-algebra is either finite or uncountable (Axiom of choice needed to prove that).

Measure spaces

- For every system of subsets $\mathcal{T} \subseteq \mathcal{P}(\Omega)$ there is a smallest σ-algebra containing $\mathcal{T}$. Notation: $\sigma_\Omega(\mathcal{T})$.
- For a single $A \in \mathcal{A}$ is $\sigma_\Omega(\{A\}) = \{\varnothing, A, A^{\complement}, \Omega\}$.
- The ordered pair $(\Omega, \mathcal{A})$ is called a measurable space.
- A (countable) partition of an element $A \in \mathcal{A}$ is 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: $A = \bigsqcup_{i \in I} A_i$.

Measure spaces

- A function on a σ-algebra $\nu : \mathcal{A} \to \overline{\mathbb{R}} \equiv [-\infty, +\infty]$ is called a (countably additive) signed measure if
  - $\nu(\varnothing) = 0$
  - ν 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 $\mathcal{M}(\Omega, \mathcal{A})$ the set of signed measures on $(\Omega, \mathcal{A})$.
- Be $\mathcal{M}^+(\Omega, \mathcal{A}) \subseteq \mathcal{M}(\Omega, \mathcal{A})$ and $\mathcal{M}^-(\Omega, \mathcal{A}) \subseteq \mathcal{M}(\Omega, \mathcal{A})$ the subsets of nonnegative and nonpositive measures i.e. measures with the property $\nu(A) \in \overline{\mathbb{R}}_+ \equiv [0, +\infty]$ and $\nu(A) \in \overline{\mathbb{R}}_- \equiv [-\infty, 0]$ for all $A \in \mathcal{A}$ respectively. These sets are convex pointed cones in $\mathcal{M}(\Omega, \mathcal{A})$.
- Notation: A '+' and '-' as sub- or superscript will always mean 'nonnegative' i.e. ≥ 0 and 'nonpositive' i.e. ≤ 0 respectively.

Measure spaces

- 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 $\nu(A) \leq \nu(A')$
  σ - 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

- For a signed measure $\nu \in \mathcal{M}(\Omega, \mathcal{A})$ and $A \in \mathcal{A}$ with $A = \bigsqcup_{i \in I} A_i$ define

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

  the total variation measure of ν. The supremum in the definition runs across all partitions $\{A_i\}$ of A.
- $|\nu|$ is a nonnegative measure on $\mathcal{A}$, i.e. $|\nu| \in \mathcal{M}^+(\Omega, \mathcal{A})$. Obviously is $|\nu(A)| \leq |\nu|(A)$ and $|\nu|(A) = \nu(A)$ if $\nu \in \mathcal{M}^+(\Omega, \mathcal{A})$.

Measure spaces

- A signed measure ν is called a finite signed measure if $|\nu|(\Omega) < +\infty$ ($\nu(\Omega) < +\infty$ if it is nonnegative). Note that $|\nu|(\Omega) < +\infty$ implies $|\nu(A)| < +\infty$ for all $A \in \mathcal{A}$ while only requiring $|\nu(\Omega)| < +\infty$ does not.
- ν is called a probability measure if $\nu(\Omega) = 1$ and it has $[0, 1]$ as the range of values.
- ν is called σ-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 Ω is finite and σ-finite if Ω is countable.

Measure spaces

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

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

- $\mathcal{M}_\sigma(\Omega, \mathcal{A})$ and $\mathcal{M}_f(\Omega, \mathcal{A})$ are $\mathbb{R}$-vector spaces. $\mathcal{M}_f(\Omega, \mathcal{A})$ can even be made a Banach space with the total variation norm $\|\nu\| \equiv |\nu|(\Omega)$. $\mathcal{M}_1(\Omega, \mathcal{A})$ is not a vector space but can be studied as a metric space, e.g., under the total variation metric or Wasserstein metric.
- $\mathcal{M}_\sigma^+(\Omega, \mathcal{A})$ and $\mathcal{M}_f^+(\Omega, \mathcal{A})$ denote the corresponding pointed convex cones of the nonnegative measures.

Measure spaces

- 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.
- Ω is called the sample space of a probability space $(\Omega, \mathcal{A}, \nu)$ and $\mathcal{A}$ its event space.

Measure spaces

- Be E a normed $\mathbb{R}$-vector space. An E-valued (countably additive) vector measure on a measurable space $(\Omega, \mathcal{A})$ is a function $\nu : \mathcal{A} \to E$ s.t.
    - $\nu(\varnothing) = 0$
    - 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 total variation of a vector measure $\nu \in \mathcal{M}(\Omega, \mathcal{A}; E)$:

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

  Again, 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, ν 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

- Note that for an arbitrary normed space E and an E-valued vector measure ν, the notion 'finite/infinite measure of a set' usually doesn't make sense. $\nu(A)$ is an element of E and not of $[0, +\infty]$. We say instead that a set $A \in \mathcal{A}$ has finite variation if $|\nu|(A) < +\infty$.

Measure spaces

- If not explicitly stated otherwise, all measures are assumed nonnegative and real-valued.
- A measure ν on a measurable space $(\Omega, \mathcal{A})$ is called singular w.r.t a measure μ, notation $\nu \perp \mu$, if there exists a set $A \in \mathcal{A}$ such that $\nu(A^{\complement}) = 0$ and $\mu(A) = 0$.
- ν is called (singular) discrete (w.r.t. μ) if in addition A is at most countable and all $\{a\}$ for every $a \in A$ are measurable. singular continuous (w.r.t. μ) otherwise.
- Let δ_ω be the Dirac measure, assigning a measure of 1 to any set containing ω and a measure of 0 to any other set. Then a measure ν is singular discrete (w.r.t. μ) 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 $\mu(\{a_i\}) = 0$.

Measure spaces

- The Cantor measure is an example of a singular continuous measure w.r.t. the Lebesgue measure (see below).
- A measurable set $A \in \mathcal{A}$ is called a ν-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 ν is called purely atomic or simply atomic if every $A \in \mathcal{A}$ with $\nu(A) > 0$ contains a ν-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 σ-finite, this representation is unique.

Measure spaces

- A σ-finite atomic measure ν 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 ω_i are common points of the atoms (mod ν-zero). As intersections of atoms the $\{\omega_i\}$ are elements of $\mathcal{A}$.
- Assume all singletons $\{\omega\}, \omega \in \Omega$, are elements of $\mathcal{A}$ (This is the case for Borel σ-algebras $\mathcal{B}(\Omega)$ (see below) for example). A measure ν on $\mathcal{A}$ is called a continuous measure if $\nu(\{\omega\}) = 0$ for every ω. A nonatomic measure is necessarily continuous but not every continuous measure is nonatomic.

Measure spaces

- 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 Ω is a separable metric space and ν a σ-finite measure on $\mathcal{B}(\Omega)$, every atom is a union of a singleton with measure greater zero and a null set.
- A continuous Radon measure (definition see below) on a Borel space $(\Omega, \mathcal{B}(\Omega))$ is nonatomic. In particular λ^d on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.
- For a locally compact Hausdorff space Ω, a general Radon measure ν on $\mathcal{B}(\Omega)$ has a unique decomposition $\nu = \nu_d + \nu_c$ where ν_d and ν_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 ν-null set. The completion of $\mathcal{A}$ w.r.t. ν is the σ-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). ν is called complete if $\mathcal{A} = \widehat{\mathcal{A}}_\nu$ that is if $\mathcal{N}_\nu \subseteq \mathcal{A}$. The extension of ν 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 ν and μ on a measurable space $(\Omega, \mathcal{A})$ we say ν is absolutely continuous with respect to μ, 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 σ-finite measure ν 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 σ-algebra $\mathcal{B}(\Omega, \mathcal{T})$ is the σ-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 σ-algebra containing $\mathcal{T}$. The Hausdorff property not only ensures that all compact sets of Ω are closed and therefore part of $\mathcal{B}(\Omega, \mathcal{T})$, it prevents a whole number of other pathologies like failures of weak convergence of measures, too few measurable functions to produce conditional expectations (see below), other measure regularity issues and so on.
- Usually the topology of Ω 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

- 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.
- Intuition: $\mathcal{B}(\Omega, \mathcal{T})$ contains the subsets of Ω you reasonably want to be able to measure with regards to the topology $\mathcal{T}$.
- Is every σ-algebra on a space Ω identical to a $\mathcal{B}(\Omega, \mathcal{T})$ for some topology $\mathcal{T}$ on Ω ? 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.

Borel spaces

- An element of $\mathcal{M}(\Omega)$ is called a signed Borel measure, an element of $\mathcal{M}^+(\Omega)$ simply a Borel measure.
- The support of a Borel measure $\nu \in \mathcal{M}(\Omega)$ is defined as

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

- A Borel measure ν is called degenerate if $\dim \operatorname{supp}(\nu) < \dim \Omega$, nondegenerate otherwise.

Borel spaces

- Be $\nu \in \mathcal{M}^+(\Omega)$ a Borel measure.
- ν 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 $K \in \mathcal{B}(\Omega)$
- ν is called inner regular if for every $B \in \mathcal{B}(\Omega)$

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

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

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

- ν is called regular if it is both inner and outer regular.

Borel spaces

- If Ω is a locally compact Hausdorff space, then locally finite is equivalent to being finite on all compact sets from $\mathcal{B}(\Omega)$.
- A signed Borel measure $\nu \in \mathcal{M}(\Omega)$ is called a signed 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 (signed) Radon measures on a locally compact Hausdorff space are precisely the inner regular (signed) measures, finite on compact sets from $\mathcal{B}(\Omega)$.

Borel spaces

- If Ω is a complete separable metric space, every (signed) Borel measure is a (signed) Radon measure.
- The Dirac measure δ_ω is a Radon measure on any Borel space.
- The Borel measure μ^d is a Radon measure on $\mathcal{B}(\mathbb{R}^d)$ but the Lebesgue measure λ^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 Ω.
- 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.

Borel spaces

- On Polish spaces, all finite measures and therefore all probability measures are Radon measures.
- All Radon measures on Polish spaces are σ-finite.
- $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is the canonical Borel space on $\mathbb{R}^d$ aka $\mathbb{R}^d$ with the Borel σ-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 μ^d, 'the' Borel measure, assigning to every right half-open interval its volume. μ^d is not a probability measure but it is by definition locally finite and σ-finite.

Borel spaces

- All Borel measures on $\mathbb{R}^d$ are regular (since $\mathbb{R}^d$ with the euclidian metric is a Polish space).
- With a function $p \geq 0$ on $\mathbb{R}^d$, $p\,\mu^d$ is a measure equivalent to μ^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 μ^d, $\tilde{\mu}^d = \mu^d(U^d)\,\mu^d$ to be precise. μ^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 σ-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 $B \mapsto B \cup \varnothing$. λ^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 λ^d will be our weapons of choice for most of the time. We will write λ^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 σ-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 μ^d-measure-zero-set (see F_σ and G_δ sets).

Borel spaces

$$\begin{tikzcd}
\mathcal{B}(\mathbb{R}^d) \arrow[r, phantom, "\subset"] \arrow[d] & \mathcal{L}(\mathbb{R}^d) \arrow[r, phantom, "\subset"] \arrow[d] & \mathcal{P}(\mathbb{R}^d) \arrow[d] \\
|\mathcal{B}(\mathbb{R}^d)| \arrow[d, equal] & |\mathcal{L}(\mathbb{R}^d)| \arrow[d, equal] & |\mathcal{P}(\mathbb{R}^d)| \arrow[d, equal] \\
\beth_1 \equiv 2^{\beth_0} \arrow[r, phantom, "<"] & \beth_2 \equiv 2^{\beth_1} \arrow[r, phantom, "="] & \beth_2 \equiv 2^{\beth_1}
\end{tikzcd}$$

-

$$\begin{aligned} \beth_0 &= & &= |\mathbb{N}| \\ \beth_1 &= 2^{\beth_0} & &= |\mathbb{R}| \end{aligned}$$

- $\beth_i \geq \aleph_i$
- 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 σ-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 $f : (\Omega_1, \mathcal{A}_1) \to (\Omega_2, \mathcal{A}_2)$ 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 $f : (\Omega_1, \mathcal{A}_1) \to (\Omega_2, \mathcal{A}_2)$ 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 $f : (\Omega_1, \mathcal{T}_1) \to (\Omega_2, \mathcal{T}_2)$ is $(\mathcal{B}(\Omega_1, \mathcal{T}_1), \mathcal{B}(\Omega_2, \mathcal{T}_2))$-measurable. A $(\mathcal{B}(\Omega_1, \mathcal{T}_1), \mathcal{B}(\Omega_2, \mathcal{T}_2))$-measurable function $f : (\Omega_1, \mathcal{B}(\Omega_1, \mathcal{T}_1)) \to (\Omega_2, \mathcal{B}(\Omega_2, \mathcal{T}_2))$ 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).

Measurable Functions

- 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 $\min(f, g)$.
- A constant function $f : \Omega_1 \to \Omega_2$ with $f(\omega_1) = \omega \in \Omega_2$ is always measurable. For any two σ-algebras $\mathcal{A}_1, \mathcal{A}_2$ on Ω_1 and Ω_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 $\omega \notin A$.
- Notation: Denote $\mathcal{L}^0(\Omega_1, \mathcal{A}_1; \Omega_2, \mathcal{A}_2)$ the space of $(\mathcal{A}_1, \mathcal{A}_2)$-measurable functions mapping Ω_1 to Ω_2. $\mathcal{L}^0(\Omega, \mathcal{A}; \mathbb{R}, \mathcal{B}(\mathbb{R}))$ will be abbreviated as $\mathcal{L}^0(\Omega, \mathcal{A})$. (This notation will make sense in the next chapter.)

Measurable Functions

- Notation: For an arbitrary (not necessarily countable) index set I denote $\mathcal{F}(I)$ the set of nonempty finite subsets.
- Consider a set Ω and measurable spaces $(\Omega_i, \mathcal{A}_i)$ $(i \in I)$, I a not necessarily countable index set. Be Γ a family of functions $\{\gamma_i\}_{i \in I}$ with $\gamma_i : \Omega \to (\Omega_i, \mathcal{A}_i)$. For any $J \in \mathcal{F}(I)$, the sets

$$C_J \equiv \gamma_J^{-1}\left(\prod_{i \in J} \mathcal{A}_i\right)$$

  are called J-cylinder sets. The algebra

$$\mathcal{C}yl\,(\Omega, \Gamma) \equiv \bigcup_{J \in \mathcal{F}(I)} C_J \subseteq \mathcal{P}(\Omega)$$

  is called the cylinder algebra on Ω, generated by Γ.

Measurable Functions

- The σ-algebra

  $$\mathcal{E}(\Omega,\Gamma) \equiv \sigma_\Omega(\ \mathcal{C}yl\,(\Omega,\Gamma)\ ) \subseteq \mathcal{P}(\Omega)$$

  is called the cylindrical σ-algebra on Ω, generated by Γ. $\mathcal{E}(\Omega,\Gamma)$ is the smallest σ-algebra $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ for which the γ_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 σ-algebra $\mathcal{A}$, then $\mathcal{C}yl\,(\Omega,\Gamma)$ is already a σ-algebra and $\mathcal{E}(\Omega,\Gamma) = \mathcal{C}yl\,(\Omega,\Gamma) \subseteq \mathcal{A}$.
- If $i : U \hookrightarrow (\Omega,\mathcal{A})$, then $\mathcal{E}(U,\{i\}) = \mathcal{A}_{|U}$.

Measurable Functions

- If $\Omega = \prod_{i \in I} \Omega_i$ and Γ consists of the projections π_i, then $\mathcal{E}(\Omega, \Gamma)$ is often simply called the cylindrical σ-algebra on Ω and denoted by $\otimes_{i \in I} \mathcal{A}_i$ or simply $\mathcal{A}_I$ (see section on product measurable spaces for more details).
- Be Ω a topological vector space and Ω' its topological dual. We define $\mathcal{E}(\Omega) \equiv \mathcal{E}(\Omega, \Omega').$ By definition is $\mathcal{E}(\Omega) \subseteq \mathcal{B}(\Omega).$ This inclusion is in general strict but equality holds for separable Fréchet spaces (in particular separable Banach spaces).

Measurable Functions

- If Ω is a Hausdorff space $(\Omega, \mathcal{T}_\Omega)$, be $\mathcal{C} \equiv \mathcal{C}(\Omega, \mathcal{T}_\Omega)$ the space of real-valued functions on Ω, continuous w.r.t. the topology $\mathcal{T}_\Omega$, and $\mathcal{C}_c \equiv \mathcal{C}_c(\Omega, \mathcal{T}_\Omega)$ the subspace of continuous functions $f \in \mathcal{C}$ with compact support $\mathrm{supp}(f) \equiv \overline{\{\omega \in \Omega \,|\, f(\omega) \neq 0\}}$. Denote $\mathcal{C}_0 \equiv \mathcal{C}_0(\Omega, \mathcal{T}_\Omega)$ the space of continuous functions on Ω, vanishing at infinity. These are functions $f \in \mathcal{C}$, for which there exists a compact set $K \subset \Omega$ for every $\epsilon > 0$ s.t. $|f(\omega)| < \epsilon$ for all $\omega \in \Omega - K$.

Measurable Functions

- Be $\mathcal{C}_b \equiv \mathcal{C}_b(\Omega, \mathcal{T}_\Omega)$ the space of bounded functions from $\mathcal{C}$ and $\mathcal{L}^0 \equiv \mathcal{L}^0(\Omega, \mathcal{B}(\Omega, \mathcal{T}_\Omega))$.
- Obviously
$$\mathcal{C}_c \subseteq \mathcal{C}_0 \subseteq \mathcal{C}_b \subseteq \mathcal{C} \subseteq \mathcal{L}^0$$
If Ω is compact, the first four spaces coincide.
- It holds by definition that :
$$\mathcal{E}(\Omega, \mathcal{C}_c) \subseteq \mathcal{E}(\Omega, \mathcal{C}_0) \subseteq \mathcal{E}(\Omega, \mathcal{C}_b) \subseteq \mathcal{E}(\Omega, \mathcal{C}) \subseteq \mathcal{E}(\Omega, \mathcal{L}^0)$$
$$\mathcal{E}(\Omega, \mathcal{C}) = \mathcal{B}(\Omega), \quad \mathcal{E}(\Omega, \mathcal{L}^0) = \mathcal{P}(\Omega)$$

Measurable Functions

- If Ω is locally compact :

$$\mathcal{E}(\Omega, \mathcal{C}_c) \subseteq \mathcal{E}(\Omega, \mathcal{C}_0) = \mathcal{E}(\Omega, \mathcal{C}_b) = \mathcal{E}(\Omega, \mathcal{C}) = \mathcal{B}(\Omega)$$

  First equality: Since Ω is locally compact, $\mathcal{C}_0$ is dense in $\mathcal{C}_b$ with respect to the uniform topology i.e. every function in $\mathcal{C}_b$ can be approximated by functions in $\mathcal{C}_0$. But the cylindrical σ-algebra only depends on measurability (not topology).

  Second equality: The space $\mathcal{C}_b$ separates points and closed sets. That is, for every Borel set can its indicator function be approximated by bounded continuous functions. Since bounded continuous functions can distinguish open sets, they generate the full Borel σ-algebra $\mathcal{B}(\Omega)$.

  The inclusion is usually strict.

Measurable Functions

- If Ω is Polish :

$$\mathcal{E}(\Omega, \mathcal{C}_c) = \mathcal{E}(\Omega, \mathcal{C}_0) = \mathcal{E}(\Omega, \mathcal{C}_b) = \mathcal{E}(\Omega, \mathcal{C}) = \mathcal{B}(\Omega)$$

- $\mathcal{B}a(\Omega, \mathcal{T}_\Omega) \equiv \mathcal{E}(\Omega, \mathcal{C}(\Omega, \mathcal{T}_\Omega))$ is called the Baire σ-algebra.
- The inclusion $\mathcal{B}a(\Omega, \mathcal{T}_\Omega) \subseteq \mathcal{B}(\Omega, \mathcal{T}_\Omega)$ in general is strict. If (Ω, 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 $f : (\Omega, \mathcal{A}) \to \mathbb{R}$ 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 ν-a.e. against f. ν 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 ν 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 ν 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 ν-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 ν-quasi-integrable. If both integrals have an infinite value, the integral of f is not defined.

Integration

- $$\nu(A) = \int_\Omega d\nu \; 1_A = \int_A d\nu$$

  Notation:
  $$\nu(A) = \int_A \nu(d\omega)$$
- $f : \Omega \to \mathbb{R}$ 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 $f : \Omega \to \mathbb{R}$ on a measure space $(\Omega, \mathcal{A}, \nu)$ is called *p-th power ν-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.)

L^p spaces

- The set of $\mathbb{R}$-valued p-th power ν-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.
- Be $f^+ \equiv \max(f, 0)$ and $f^- \equiv \max(-f, 0)$. Obviously is $f = f^+ - f^-$. Since $|f| = f^+ + f^-$, f is ν-integrable iff $|f|$ is ν-integrable. I.e. f is ν-integrable iff $X \in \mathcal{L}^1(\Omega, \mathcal{A}, \nu; \mathbb{R})$.

L^p spaces

- 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 $(f\,g) \in \mathcal{L}^1(\Omega, \mathcal{A}, \nu)$ and

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

- 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 ν-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 ν- 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 $\|f + g\|_p^p \leq \|f\|_p^p \; + \; \|g\|_p^p$, 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 ν is a finite measure, then $L^p(\Omega, \mathcal{A}, \nu) \subseteq L^q(\Omega, \mathcal{A}, \nu)$ and $\nu(\omega)^{-1/q} \left\|f\right\|_q \;\leq\; \nu(\omega)^{-1/p} \left\|f\right\|_p$ whenever $0 < q \leq p \leq +\infty$. This is not true in general.
- Consider $L^p(\Omega, \mathcal{A}, \nu)' \equiv \mathcal{L}(L^p(\Omega, \mathcal{A}, \nu), \mathbb{R})$, the space of linear forms on $L^p(\Omega, \mathcal{A}, \nu)$, continuous w.r.t. the norm-induced topology on $L^p(\Omega, \mathcal{A}, \nu)$. $L^p(\Omega, \mathcal{A}, \nu)'$ is 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

$$\left\|T\right\| \equiv \sup_{\left\|f\right\|_p = 1} \left|T(f)\right| \quad , T \in L^p(\Omega, \mathcal{A}, \nu)'$$

L^p spaces

- If $L^p(\Omega, \mathcal{A}, \nu)$ is a Banach space, $L^p(\Omega, \mathcal{A}, \nu)'$ consists precisely of the bounded linear forms.
- $L^p(\Omega, \mathcal{A}, \nu)' \subseteq L^p(\Omega, \mathcal{A}, \nu)^* \equiv \mathrm{Hom}(L^p(\Omega, \mathcal{A}, \nu), \mathbb{R})$. $L^p(\Omega, \mathcal{A}, \nu)^*$ 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 $L^p(\Omega, \mathcal{A}, \nu)'$ is large enough. For $p \in (0, 1)$ $L^p(\Omega, \mathcal{A}, \nu)$ is not locally convex in general (check $L^p([0, 1], \mathcal{B}([0, 1]), \lambda)' = \varnothing$).

L^p spaces

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

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

  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.

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

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

L^p spaces

- $L^1(\Omega, \mathcal{A}, \nu)' = L^\infty(\Omega, \mathcal{A}, \nu)$ if ν is σ-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: $L^2(\Omega, \mathcal{A}, \nu) = L^2(\Omega, \mathcal{A}, \nu)'$ (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 ν- almost everywhere.

L^p spaces

- If ν is σ-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 ν 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. Ω consists of a finite number of ν-atoms.
- If the space $L^p(\Omega, \mathcal{A}, \nu)$ with $p \in [1, +\infty)$ is separable, there is a countable set Γ 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, Γ 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 σ-finite precisely because I is countable.
- $L^p(I, \mathcal{P}(I), \#) = \mathcal{L}^p(I, \mathcal{P}(I), \#) \quad$ (because $\mathcal{N} = \{0\}$).
- $\ell^0(I) = \mathcal{L}^0(I, \mathcal{P}(I), \#) \overset{!}{=} \mathsf{Fun}(I, \mathbb{R})$ (with $\mathcal{P}(I)$ as σ-algebra, every function $f : I \to \mathbb{R}$ 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 $\mathsf{Fun}(I,\mathbb{R}) = \mathbb{R}^I$, 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 $p \in (0,+\infty)$.
- $\|\{x_i\}_{i\in I}\|_\infty = \mathsf{sup}_{i\in I}|x_i|$.
- $\ell^p(I)$ 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 $\mathsf{sup}_{i\in I} |x_i| < +\infty$ 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 $(\ell^p)' = \ell^\infty$.
- For $p \in (1, +\infty)$ it holds that $(\ell^p)' = \ell^q$ 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 σ-finiteness does not suffice.
- ℓ^p is separable for $p \in (0, +\infty)$, ℓ^∞ is not.

Sequence spaces

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

Bochner spaces

- By definition it holds for vector-valued measurable maps that
$$L^0(\Omega, \mathcal{A}, \nu; \mathbb{R}^d) = \bigoplus_{i=1}^{d} L^0(\Omega, \mathcal{A}, \nu; \mathbb{R})$$
- The norm equivalence in finite-dimensional spaces implies that for $p \in (0, +\infty]$
$$L^p(\Omega, \mathcal{A}, \nu; \mathbb{R}^d) = \bigoplus_{i=1}^{d} L^p(\Omega, \mathcal{A}, \nu; \mathbb{R})$$
-
$$L^p(\Omega, \mathcal{A}, \nu; \mathbb{R}^d) = L^p(\Omega, \mathcal{A}, \nu; \mathbb{R}) \otimes \mathbb{R}^d \quad .$$

Bochner spaces

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

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

such that

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

Bochner spaces

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

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

  induces on it the topology of (local) convergence in measure. $L^0(\Omega, \mathcal{A}, \nu; \mathsf{E})$ 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 $f \in \mathsf{Fun}(\Omega, \mathsf{E})$ and $\|f(\cdot)\|_{\mathsf{E}} \in \mathsf{Fun}(\Omega, \mathbb{R})$ its norm function. Then $f \in L^0(\Omega, \mathcal{A}, \nu; \mathsf{E})$ implies that $\|f(\cdot)\|_{\mathsf{E}} \in L^0(\Omega, \mathcal{A}, \nu; \mathbb{R})$. (Note that $\| \cdot \|_{\mathsf{E}} : \mathsf{E} \to \mathbb{R}$ is $(\mathcal{B}(\mathsf{E}), \mathcal{B}(\mathbb{R}))$-measurable since as a norm it is continuous).
- A function $f : \Omega \to \mathsf{E}$ is called weakly measurable if $\phi \circ f : \Omega \to \mathbb{K}$ is $(\mathcal{A}, \mathcal{B}(\mathbb{K}))$-measurable for every $\phi \in \mathsf{E}' = \mathcal{L}(\mathsf{E}, \mathbb{K})$.
- If $\mathcal{T}_{\mathsf{E}}$ is the norm-induced topology on E, how does being $(\mathcal{A}, \mathcal{B}(\mathsf{E}, \mathcal{T}_{\mathsf{E}}))$-measurable compare to being strongly measurable or weakly measurable ?

Bochner spaces

- Pettis measurability theorem: For $f : \Omega \to \mathsf{E}$ the following assertions are equivalent
  - f is strongly measurable
  - f is weakly measurable and there exists a separable closed subspace $\mathsf{E}_0 \subseteq \mathsf{E}$ s.t. $f(\Omega) \subseteq \mathsf{E}_0$ (f is separably valued).
  - f is $(\mathcal{A}, \mathcal{B}(\mathsf{E}, \mathcal{T}_{\mathsf{E}}))$-measurable and separably valued.
- Consider the weak topology on E, denoted by $\sigma(\mathsf{E}, \mathsf{E}')$. $\sigma(\mathsf{E}, \mathsf{E}')$ is the weakest/coarsest topology on E for which all $\phi \in \mathsf{E}'$ are continuous. It is Hausdorff.

Bochner spaces

- By definition $\sigma(\mathsf{E},\mathsf{E}') \subseteq \mathcal{T}_{\mathsf{E}}$ and $\mathcal{B}(\mathsf{E},\sigma(\mathsf{E},\mathsf{E}')) \subseteq \mathcal{B}(\mathsf{E},\mathcal{T}_{\mathsf{E}})$. Denote

  $$\mathcal{B}a(\mathsf{E},\sigma(\mathsf{E},\mathsf{E}'))) \subseteq \mathcal{B}(\mathsf{E},\sigma(\mathsf{E},\mathsf{E}'))) \subseteq \mathcal{B}(\mathsf{E},\mathcal{T}_{\mathsf{E}})$$

  the Baire σ-algebra of E w.r.t. the weak topology. It can be shown that $f : \Omega \to \mathsf{E}$ is weakly measurable iff $f : (\Omega,\mathcal{A}) \to (\mathsf{E},\mathcal{B}a(\mathsf{E},\sigma(\mathsf{E},\mathsf{E}'))$ is $(\mathcal{A},\mathcal{B}a(\mathsf{E},\sigma(\mathsf{E},\mathsf{E}')))$-measurable.

Bochner spaces

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

Bochner spaces

- As expected, the Bochner integral of a simple function

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

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

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

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

Bochner spaces

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

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

  and

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

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

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 ν, that is $L^p(\Omega, \mathcal{A}, \nu; \mathsf{E}) \subseteq L^q(\Omega, \mathcal{A}, \nu; \mathsf{E})$ for $0 < q \leq p \leq +\infty$.
- When $p \in [1, +\infty]$ the spaces $L^p(\Omega, \mathcal{A}, \nu; \mathsf{E})$ 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 ν is finite then E' having the Radon-Nikodym property (see below) w.r.t. ν is equivalent to the canonical mapping $\phi : L^q(\Omega, \mathcal{A}, \nu; \mathsf{E}') \longrightarrow L^p(\Omega, \mathcal{A}, \nu; \mathsf{E})'$ ($p \in (1, +\infty)$ and $p^{-1} + q^{-1} = 1$)

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

being an isometric isomorphism i.e.

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

Bochner spaces

- If E is a reflexive Banach space it is separable iff E' is separable.
- If E′ is separable it has the Radon-Nikodym property.
- Hilbert spaces and reflexive spaces have the R-N property.
- $L^p(\Omega, \mathcal{A}, \mu; \mathrm{E})$ 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 $(f, e) \mapsto f(\cdot)\, e$)

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

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

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

  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

- It can be shown that Δ_1 corresponds to π, the projective norm. In particular there are isometric isomorphisms

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

- Δ_∞ corresponds to the injective norm ϵ. In particular

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

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

Bochner spaces

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

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

  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

- Many of the concepts from above carry over to Fréchet space-valued measurable functions. We can define function spaces $L^p(\Omega, \mathcal{A}, \nu; \mathsf{E}, \mathcal{E})$. But the situation is more complicated, ’weak measurability’ does not imply ’Borel measurability’ in case of a separable Fréchet space for example.

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 $\nu(\{\omega \in \Omega \mid w(\omega) \neq 0\}) < +\infty$.
- It can be shown (Dominated Convergence Theorem) that

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

- $$\overline{S(\Omega, \mathcal{A}, \nu)}^{\|\cdot\|_\infty} \subset L^\infty(\Omega, \mathcal{A}, \nu)$$

  Equality holds if ν is finite, a probability measure for example.

Density Theorems

- For a measurable space $(\Omega, \mathcal{B}(\Omega))$ with Ω 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 ℓ^∞ (not of the ℓ^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 $(\Omega, \mathcal{A})$ and denote $\mathcal{B}_b(\Omega, \mathcal{A}) \subseteq \mathcal{L}^0(\Omega, \mathcal{A}; \mathbb{R})$ the $\mathbb{R}$-vector subspace of $\| \cdot \|_\infty$- bounded $(\mathcal{A}, \mathcal{B}(\mathbb{R}))$-measurable real-valued functions on Ω. $(\mathcal{B}_b(\Omega, \mathcal{A}), \| \cdot \|_\infty)$ is a Banach space. Its strong topological dual $\mathcal{B}_b(\Omega, \mathcal{A})'$ is a Banach space w.r.t. the operator norm $\|\phi\| \equiv \sup_{\|f\|_\infty \leq 1} \|\phi(f)\|$.
- Be $ba(\Omega, \mathcal{A})$ the set of bounded finitely additive signed measures on $\mathcal{A}$. $ba(\Omega, \mathcal{A})$ is a Banach space w.r.t. the total variation norm $\|\nu\|_{TV} \equiv |\nu|(\Omega)$. Obviously is $\mathcal{M}_f(\Omega, \mathcal{A}) \subseteq ba(\Omega, \mathcal{A})$.

Spaces of Measures

- The mapping

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

induces an isomorphism

$$ba(\Omega, \mathcal{A}) = \mathcal{B}_b(\Omega, \mathcal{A})' \quad .$$

This isomorphism is isometric i.e. it translates the total variation norm on $ba(\Omega, \mathcal{A})$ into the operator norm on $\mathcal{B}_b(\Omega, \mathcal{A})'$.

Spaces of Measures

- The above isomorphism respects the decomposition of $ba(\Omega, \mathcal{A})$ and $\mathcal{B}_b(\Omega, \mathcal{A})'$ into positive and negative cones $ba^{\pm}(\Omega, \mathcal{A})$ and $\mathcal{B}_b(\Omega, \mathcal{A})'^{\pm}$ respectively. $ba^{\pm}(\Omega, \mathcal{A})$ is the convex cone of bounded finitely additive positive/negative measures on $(\Omega, \mathcal{A})$. $\mathcal{B}_b(\Omega, \mathcal{A})'^{\pm}$ is the convex cone of continuous linear forms ϕ on $\mathcal{B}_b(\Omega, \mathcal{A})$ for which $f \geq 0$ $(f \leq 0)$ implies $\langle \phi \mid f \rangle \geq 0$ $(\langle \phi \mid f \rangle \leq 0)$. In summary, we have isomorphisms of convex cones

$$ba^{+}(\Omega, \mathcal{A}) = \mathcal{B}_b(\Omega, \mathcal{A})'^{+} \quad , \quad ba^{-}(\Omega, \mathcal{A}) = \mathcal{B}_b(\Omega, \mathcal{A})'^{-} \quad .$$

Spaces of Measures

- Assume for now that Ω is a locally compact Hausdorff space with topology $\mathcal{T}_\Omega$ and that $\mathcal{A} \equiv \mathcal{B}(\Omega, \mathcal{T}_\Omega)$. Denote $\mathcal{M}_{R,f}(\Omega, \mathcal{A})$ the space of finite signed Radon measures on $(\Omega, \mathcal{A})$. It is a Banach space with the total variation norm. So is $\mathcal{C}_0(\Omega, \mathcal{T}_\Omega)$, the space of $\mathcal{T}_\Omega$-continuous functions on Ω, vanishing at infinity, equipped with the $\| \cdot \|_\infty$ norm.
- Riesz-Markov-Kakutani Representation Theorem: The mapping

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

induces an isometric isomorphism

$$\mathcal{M}_{R,f}(\Omega, \mathcal{A}) = \mathcal{C}_0(\Omega, \mathcal{T}_\Omega)'$$

Spaces of Measures

- Like above, there is a decomposition into convex cones:

$$\mathcal{M}^+_{R,f}(\Omega,\mathcal{A}) = \mathcal{C}_0(\Omega,\mathcal{T}_\Omega)'^+ \quad , \quad \mathcal{M}^-_{R,f}(\Omega,\mathcal{A}) = \mathcal{C}_0(\Omega,\mathcal{T}_\Omega)'^- \quad .$$

- The space $\mathcal{C}_c(\Omega,\mathcal{T}_\Omega)$ of continuous functions with compact support on Ω is not a Banach but an LF-space i.e. a direct limit of Fréchet spaces. $\mathcal{C}_c(\Omega,\mathcal{T}_\Omega)$ is dense in $\mathcal{C}_0(\Omega,\mathcal{T}_\Omega)$ and the above isomorphism induces not an isometric isomorphism but an isomorphism of tvs

$$\mathcal{M}_{R,f}(\Omega,\mathcal{A}) = \mathcal{C}_c(\Omega,\mathcal{T}_\Omega)' \quad .$$

Spaces of Measures

- The space of test functions $\mathcal{C}_c^\infty(\Omega, \mathcal{T}_\Omega)$ is a subset of $\mathcal{C}_c(\Omega, \mathcal{T}_\Omega)$ and thus leads to an embedding of the space of finite signed Radon measures into its strong topological dual $\mathcal{D}b(\Omega) \equiv \mathcal{C}_c^\infty(\Omega, \mathcal{T}_\Omega)'$, the space of distributions :

$$\mathcal{M}_{R,f}(\Omega, \mathcal{A}) = \mathcal{C}_c(\Omega, \mathcal{T}_\Omega)' \subset \mathcal{C}_c^\infty(\Omega, \mathcal{T}_\Omega)' = \mathcal{D}b(\Omega)$$

Distributions can be interpreted as a generalization of finite signed Radon measures. The distributions comming from a Radon measure are called regular distributions.

Radon-Nikodym Theorem

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

$$\nu(A) = \int_A d\nu \overset{!}{=} \int_A d\mu \, f_\nu \quad .$$

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

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

Radon-Nikodym Theorem

- Notation

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

- Note that $\nu \ll \mu$ implies that ν is σ-finite because μ is.
- The reqirement for the reference measure μ to be nonnegative and σ-finite can in general not be weakened.
- Intuition: The Radon-Nikodym derivative captures the degree of density change between two measures.

Radon-Nikodym Theorem

- For a σ-finite nonnegative measure μ denote $\mathcal{M}(\Omega, \mathcal{A}, \mu)$ the $\mathbb{R}$-vector space of signed measures on $(\Omega, \mathcal{A})$, absolutely continuous w.r.t. μ. $\mathcal{M}_f(\Omega, \mathcal{A}, \mu) \subset \mathcal{M}(\Omega, \mathcal{A}, \mu)$ denotes the subset of finite measures. It is an $\mathbb{R}$-vector space. Equipped with the total variation norm, is a Banach space.

Radon-Nikodym Theorem

- The Radon-Nikodym Theorem induces an isomorphism of $\mathbb{R}$-vector spaces

  $$\mathcal{M}(\Omega, \mathcal{A}, \mu) = L^0(\Omega, \mathcal{A}, \mu)$$

  and an isometric isomorphism of $\mathbb{R}$-Banach spaces

  $$\mathcal{M}_f(\Omega, \mathcal{A}, \mu) = L^1(\Omega, \mathcal{A}, \mu) \quad .$$

  Cone decompositions are preserved :

  $$\mathcal{M}^+(\Omega, \mathcal{A}, \mu) = L^0_+(\Omega, \mathcal{A}, \mu) \ \ , \ \ \mathcal{M}^-(\Omega, \mathcal{A}, \mu) = L^0_-(\Omega, \mathcal{A}, \mu)$$

  $$\mathcal{M}_f^+(\Omega, \mathcal{A}, \mu) = L^1_+(\Omega, \mathcal{A}, \mu) \ \ , \ \ \mathcal{M}_f^-(\Omega, \mathcal{A}, \mu) = L^1_-(\Omega, \mathcal{A}, \mu)$$

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 $g \in L^0(\Omega, \mathcal{A}, \nu)$ 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

- ν and μ 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 ν with bounded variation on a measure space $(\Omega, \mathcal{A}, \mu)$, the following statements are equivalent:
  - $\nu \ll \mu$
  - There exists a function $f_\nu \in L^0(\Omega, \mathcal{A}, \mu; E)$ such that for all $A \in \mathcal{A}$
$$\nu(A) = \int_A d\nu \overset{!}{=} \int_A d\mu\, f_\nu$$

Kullback-Leibler Divergence

- Two σ-finite measures ν, μ 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 μ is used to approximate ν.

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 ν and μ are absolutely continuous w.r.t a σ-finite measure λ, i.e. $\nu \ll \mu \ll \lambda$, the KL divergence can be calculated from the respective RN derivatives. With

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

  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 ν and μ 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 ν on a σ-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 ν is σ-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

- 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 Ω_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 π_K and π_i the main projections π_{IK} and $\pi_{I\{i\}}$ respectively. The partial projections obey a cocycle condition $\pi_{JL} \circ \pi_{KJ} = \pi_{KL}$ for $\varnothing \neq L \subseteq J \subseteq K \subseteq I$. Ω_I itself is equipped with the product topology, the coarsest topology for which the projections π_K are continuous.
- Define $\bigotimes_{i \in I} \mathcal{A}_i$, $\mathcal{A}_I$ for short, to be $\mathcal{E}(\Omega_I, \{\pi_i\}_{i \in I}) \subseteq \mathcal{P}(\Omega_I)$, the smallest σ-algebra $\mathcal{A}$ on Ω_I such that the π_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 $\varnothing \neq J \subseteq K \subseteq I$ 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

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

  The $\pi_K^{-1}(\mathcal{A}_K)$ as well as the $\pi_i^{-1}(\mathcal{A}_i)$ are σ-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 Ω_i are Hausdorff

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

Equality holds if the Ω_i are second countable and I is countable. In that case

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

Product Measurable Spaces

- If the $(\Omega_i, \mathcal{A}_i)$ are all identical to the same measurable space $(\Omega, \mathcal{A})$, we will use the notation $(\Omega^I, \mathcal{A}^I)$, i.e.

$$(\Omega^I, \mathcal{A}^I) = (\prod_{i \in I} \Omega, \bigotimes_{i \in I} \mathcal{A})$$

  with $\mathcal{A}^I$ understood to be the cylindrical σ-algebra.

- If $(\Omega, \mathcal{A}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and I is countable :

$$((\mathbb{R}^d)^I, \mathcal{B}(\mathbb{R}^d)^I) = (\ (\mathbb{R}^d)^I, \mathcal{B}((\mathbb{R}^d)^I)\ )\quad .$$

Product Measures

- If I is a nonempty and finite set, the product measure of σ-finite measures ν_i on measurable spaces $(\Omega_i, \mathcal{A}_i)$ is the unique measure $\bigotimes_{i\in I} \nu_i$ (ν_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 σ-finite since the ν_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:
  $$\bigotimes_{i \in I} (\Omega_i, \mathcal{A}_i, \nu_i) \equiv (\Omega_I, \mathcal{A}_I, \nu_I)$$
  $$\bigotimes_{i \in I} (\Omega_i, \mathcal{A}_i, P_i) \equiv (\Omega_I, \mathcal{A}_I, P_I)$$
- λ^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
  $$(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \lambda^d) = \bigotimes_{|I|=d} (\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda^1)$$

Product Measures

- Be ν a measure $(\Omega_I, \mathcal{A}_I)$. The measures ν_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 ν.

Product Measures

- Tonelli/Fubini I: Let $f : \Omega_1 \times \cdots \times \Omega_d \longrightarrow \mathbb{R}$ 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

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

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 ν_K need not be product measures). Then there exists a unique probability measure ν on $\mathcal{A}_I$ such that $\pi_{K*}\nu = \nu_K$ for every $K \in \mathcal{F}(I)$. ν is the projective limit of the ν_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 $\phi_s^d : \mathbb{R}^{d \times s} \longrightarrow \mathbb{R}^d$, $(x_1, \ldots, x_s) \mapsto x_1 + \cdots + x_s$ is a continuous and therefore $(\mathcal{B}(\mathbb{R}^{d \times s}), \mathcal{B}(\mathbb{R}^d))$-measurable mapping. Consider finite measures $\nu_1, \ldots, \nu_s$ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ (these are Borel measures). The ϕ_s^d -push forward measure of the product measure $\nu_1 \otimes \cdots \otimes \nu_s$ is called the convolution measure of the ν_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 λ^d. Convolution is not necessarily commutative.

Convolution

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

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

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

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

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: $L^0(\Omega, \mathcal{A}, P; \mathsf{E}, \mathcal{E})$
- Ω is called the sample space of X, $\mathcal{A}$ its event space and $(\mathsf{E}, \mathcal{E})$ its target space.
- If 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 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 $\mathcal{L}^0(\Omega, \mathcal{A}, P\,; \mathsf{E})$. If E is separable, 'weakly measurable', 'strongly measurable' and '$(\mathcal{A}, \mathcal{B}(\mathsf{E}, \mathcal{T}_{\mathsf{E}}))$-measurable' are equivalent. In that case being an E-valued random variable means being an $(\mathsf{E}, \mathcal{B}a(\mathsf{E}, \sigma(\mathsf{E}, \mathsf{E}')))$-valued random element or equivalently an E-valued Borel random element. In other words

$$\begin{aligned}&\mathcal{L}^0(\Omega, \mathcal{A}, P\,; \mathsf{E})\\ &\qquad\qquad = \mathcal{L}^0(\Omega, \mathcal{A}, P\,; \mathsf{E}, \mathcal{B}a(\mathsf{E}, \sigma(\mathsf{E}, \mathsf{E}')))\\ &\qquad\qquad = \mathcal{L}^0(\Omega, \mathcal{A}, P\,; \mathsf{E}, \mathcal{B}(\mathsf{E}, \mathcal{T}_{\mathsf{E}}))\end{aligned}$$

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 σ-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 Φ.

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 $g \geq 0$

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

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

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

  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 $U \subseteq \bigcup_i \mathbf{Im}(X_i)$, 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 $\{X_i\}$. $P_{g(X_1,\ldots,X_n)}$ 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 $Q \in \mathcal{M}_1(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ can be made the distribution of a random variable Y by taking as Y the identity element of $L^0(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), Q; \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

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. $\mathcal{B}(\mathbb{R}^d) = \sigma(\{(-\infty, x] \mid x \in \mathbb{R}^d\}) = \sigma(\{[x, +\infty) \mid x \in \mathbb{R}^d\})$.

-

$$P(X \leq x) \equiv P\,(\,X \in (-\infty, x]\,) = P_X((-\infty, x])$$

$$P(X \geq x) \equiv P\,(\,X \in [x, +\infty)\,) = P_X([x, +\infty))$$

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 $\mathcal{B}(\mathbb{R}^d)$.
- $F_X(x) = F_X(x_1, \ldots, x_d) = P(X_1 \leq x_1, \ldots, X_d \leq x_d)$
- $\overline{F}_X(x) \equiv P(X > x) = 1 - F_X(x)$ is called the complementary cumulative distribution function (ccdf) / tail distribution of X.
- F_X is monotone nondecreasing.
- $\lim_{x_1,\ldots,x_d \to +\infty} F_X(x_1, \ldots, x_d) = 1$
- $\lim_{x_i \to -\infty} F_X(x_1, \ldots, x_d) = 0 \quad \forall i \in \{1, \ldots, d\}$

Distribution Functions

- An important class of functions are the so called càdlàg functions (càdlàg = french "continue à droite, limite à gauche" = right-continuous with left limits). Accordingly a function from an interval $T \subseteq \mathbb{R}$ to a Polish space E is called a càdlàg function if it is right-continuous with left limits in every $t \in T$. These functions form an $\mathbb{R}$-vector space, denoted by $\mathcal{D}(T, \mathsf{E})$, the Skorokhod space. Obviously $\mathcal{C}(T, \mathsf{E}) \subset \mathcal{D}(T, \mathsf{E})$.
- Càdlàg functions will play a significant role later on as paths of Feller processes (see below).
- The cdf of a univariate random variable is a càdlàg function.

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} :

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

Distribution Functions

- The cdf $C(u_1, \ldots, u_d) = P(U_1 \leq u_1, \ldots, U_d \leq u_d) : [0,1]^d \to [0,1]$ of a random variable $U = (U_1, \ldots, U_d)^\top$ with $U_i \sim U([0,1])$ is called a d-dimensional copula.
- Be $X = (X_1, \ldots, X_d)$ a random variable with cdf F_X and univariate marginal cdfs $F_{X_i}(x) = P(X_i \leq x)$. If the F_{X_i} are continuous then the probability integral transform implies that $U_i \equiv F_{X_i}(X_i) \sim U([0,1])$. The cdf C_X of $U \equiv (U_1, \ldots, U_d)$ is called the copula of the random variable X .

Distribution Functions

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

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

  If the marginal distributions are continuous, C is uniquely defined i.e $C = C_X$. It then can be constructed: $C(u_1, \ldots, u_d) \equiv F_X\left(F_{X_1}^{-1}(u_1), \ldots, F_{X_d}^{-1}(u_d)\right)$ 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 describes the dependencies between the X_i, detached from 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. λ^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

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

  and that

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

Density Functions

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

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

  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

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

- 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:

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

- 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

- Be $Z = X + Y$ the sum of two random variables $X, Y \in L^0(\Omega, \mathcal{A}, P; \mathbb{R}^d)$ with pdfs p_X and p_Y. It holds that

  $$F_Z(x) = \int_{\mathbb{R}^d} dt\, F_X(x-t)\, p_Y(t) = \int_{\mathbb{R}^d} dt\, F_Y(t)\, p_X(x-t) \quad .$$

  The pdf of Z can be reconstructed like so:

  $$p_Z(x) = \int_{\mathbb{R}^d} dt\, p_{(X,Y)}(t, x-t)$$

  If X and Y are independent, p_Z reduces to the convolution of their pdfs :

  $$p_Z(x) = \int_{\mathbb{R}^d} dt\, p_X(t)\, p_Y(x-t) = (p_X * p_Y)(x)$$

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

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

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

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

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. λ^d. 'Absolutely continuous' implies 'continuous' because λ^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 $\mathbf{supp}(X)$). 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, $S_X \equiv \mathbf{supp}(X) \subset \mathsf{E}$, 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 $\mathsf{supp}(P_X)$ equals $\mathsf{supp}(X)$.
-

$$
\begin{aligned}
P_X(\mathsf{E}) &= \sum_{s\in S_X} m_s\,\delta_s(\mathsf{E}) = \sum_{s\in S_X} m_s\,\delta_s\left(\bigsqcup_{s\in S_X}\{s\}\sqcup \mathsf{E}\setminus \bigsqcup_{s\in S_X}\{s\}\right)\\
&= \sum_{s\in S_X} m_s\,\delta_s\left(\bigsqcup_{s\in S_X}\{s\}\right) = P_X\left(\bigsqcup_{s\in S_X}\{s\}\right)\\
&= \sum_{s\in S_X} P_X(\{s\}) = 1
\end{aligned}
$$

- In familiar notation:

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

Discrete Random Variables

- Denote $p_X(\epsilon) \equiv P\left(X = \epsilon\right) = P_X\left(\{\epsilon\}\right)$, $\epsilon \in \mathsf{E}$ 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 σ-finite on $\mathcal{B}(\mathbb{R}^d)$ and with $P_X \ll \#_{S_X}$, the Radon-Nikodym theorem gives that for $x \in \mathbb{R}^d$

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

- 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. λ^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 $\mathcal{B}(\mathbb{R})$

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

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

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

- For the cdf

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

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 $\mathsf{supp}(X) \subseteq \mathsf{supp}(Y)$:

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

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 $X \in \mathsf{Fun}(\Omega, \mathbb{R}^d)$ 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 $X \in \mathbf{Fun}(\Omega, \mathbb{R}^d)$ 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 $X \in \mathsf{Fun}(\Omega, \mathbb{R}^d)$:

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

  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 $X \in \mathsf{Fun}(\Omega, \mathbb{R}^d)$:

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

  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

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

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

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

  In particular

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

Moments

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

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

Moments

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

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

- Generalized to the setup of the Change-of-variables theorem:

$$\mathbb{E}_P\left[f(X)\right] = \int_\Omega f(X)\, dP = \int_\Omega f(X)\, \frac{dP}{dQ}(X)\, dQ = \mathbb{E}_Q\left[f(X)\, \frac{dP}{dQ}(X)\right]$$

  This identity is used by importance sampling, a variance reduction technique.

Moments

- If in addition X has a pdf, i.e. $P_X \ll \lambda^d$, the above reasoning in combination with the change-of-variables theorem leads to the Law Of The Unconscious Statistician (LOTUS):

$$\begin{aligned}\mathbb{E}_P\left[f(X)\right] &= \int_\Omega f(X)\,dP = \int_{\mathbb{R}^d} f(x)\,dP_X(x) \\ &= \int_{\mathbb{R}^d} f(x)\,\frac{dP_X}{d\lambda^d}(x)\,d\lambda^d(x) = \int_{\mathbb{R}^d} \lambda^d(dx)\,f(x)\,p(x)\end{aligned}$$

If X is a discrete random variable:

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

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 :

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

-

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

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, the following notations are common :

$$\mathbb{E}_{p(x)}\left[f\right] \equiv \mathbb{E}_{P_X}\left[f\right] = \mathbb{E}_P\left[f(X)\right] = \int_{\mathbb{R}^{2d}} dx\; f(x)\, p(x)$$

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

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

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

Moments

- Note that

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

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

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

- 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 ϕ: $\phi(x) = e^x$, $\phi(x) = |\cdot|^p$ 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

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

- $(L^1(\Omega, \mathcal{A}, P\,; \mathbb{R}^d), \|X\|_1)$ 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).
- Recall from above that

$$L^1(\Omega, \mathcal{A}, P; \mathbb{R}^d) = \bigoplus_{i=1}^{d} L^1(\Omega, \mathcal{A}, P; \mathbb{R})$$

  or equivalently that for $X = (X_1, \ldots, X_d)$

$$X \in L^1(\Omega, \mathcal{A}, P; \mathbb{R}^d) \;\text{ iff }\; X_i \in L^1(\Omega, \mathcal{A}, P; \mathbb{R}) \;\; \forall i \in \{1, \ldots, d\}\;.$$

Moments

- $(L^2(\Omega, \mathcal{A}, P\,; \mathbb{R}^d), \|X\|_2)$ 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 $\{\, X \in L^0(\Omega, \mathcal{A}, P\,; \mathbb{R}^d) \ \wedge \ \mathbb{E}[|X|^2] < +\infty \,\}$
- Again like above

$$X \in L^2(\Omega, \mathcal{A}, P; \mathbb{R}^d) \ \text{iff} \ X_i \in L^2(\Omega, \mathcal{A}, P; \mathbb{R}) \ \ \forall i \in [1, \ldots, d]$$

- $X \in L^2(\Omega, \mathcal{A}, P\,; \mathbb{R}^d)$ implies $X \in L^1(\Omega, \mathcal{A}, P\,; \mathbb{R}^d)$ and since $\mathsf{Cov}(X_i) = \mathbb{E}[|X_i|^2] - \mathbb{E}[X_i]^2$, it follows that $\mathsf{Cov}(X_i) < +\infty$ $\forall i \in \{1, \ldots, d\}$

Moments

- 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

- 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 ≤ 1.

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 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 $A = L\,L^\top$ 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 positive semidefinite matrix. It describes linear relations between the random vector components of X and Y.

Moments

- $\mathsf{Cov}(X) \equiv \mathsf{Cov}(X, X) \in \overline{\mathbb{R}}_+^{d \times d}$ is called the covariance matrix of the random variable X.
  Univariate:
  $$\mathsf{Var}(X) = \mathsf{Cov}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \in \overline{\mathbb{R}}_+$$
- The square root of $\mathsf{Cov}(X)$, denoted by $\sigma(X)$, is called the standard deviation of X: $\sigma(X) \equiv \mathsf{Cov}(X)^{\frac{1}{2}} \in \overline{\mathbb{R}}_+^{d \times d}$.
  Univariate: $\sigma(X) = \sqrt{\mathsf{Var}(X)} \in \overline{\mathbb{R}}_+$
- With $\mathbf{dim}(X) = 1$, 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 $\mathsf{Cov}(X, Y)_{ij} = \mathsf{Cov}(X_i, Y_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_j - \mathbb{E}[Y_j])]$. In particular $\mathsf{Cov}(X)_{ii} = \mathsf{Cov}(X_i) = \mathsf{Var}(X_i)$.
- The inverse $\mathsf{Cov}(X)^{-1}$ (if it exists) is called the precision matrix.
- $\mathbb{E}[|X|^2] = \mathsf{Tr}\{\mathsf{Cov}(X)\} + |\mathbb{E}[X]|^2$
- $Y \equiv AX + b$ an affine transformation of X with $A \in \mathbb{R}^{d \times d'}$: $E[Y] = A\,\mathbb{E}[X] + b$ and $\mathsf{Cov}(Y) = A\ \mathsf{Cov}(X)\,A^\top$
- $\mathbb{E}[X^\top AX] = \mathbb{E}[X]^\top A\,\mathbb{E}[X] + \mathsf{Tr}(A\ \mathsf{Cov}(X))$.

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.
- X and Y are independent iff

$$\mathbb{E}[f(X)\,g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)] \quad \forall f, g \in \mathcal{L}^0(\mathbb{R}^d, \mathbb{R}) \quad .$$

- Obviously the independence of X and Y implies they are uncorrelated. The other way around does not necessarily hold because there could be nonlinear relations.

Moments

- $\mathrm{Cor}(X,Y) \in \mathbb{R}^{d\times d}$ with $\sigma(X_i), \sigma(Y_j) \in (0,+\infty)$ and

  $$\mathrm{Cor}(X,Y)_{ij} \equiv \frac{\mathrm{Cov}(X_i, Y_j)}{\sigma(X_i)\,\sigma(Y_j)}$$

  is the correlation matrix of the random variables X and Y. Obviously is $\mathrm{Cor}(X,Y)_{ii} = 1$.
- $\mathrm{Cor}(X,Y) = D_X^{-1}\ \mathrm{Cov}(X,Y)\, D_Y^{-1}$ where D_X and D_Y are diagonal matrices with $\sigma(X_i)$ and $\sigma(Y_i)$ on the diagonal respectively.

Moments

- $\mathrm{Cor}(X,Y)$ is a positive semidefinite matrix. It measures the degree to which the random vector components of X and Y are linearly related.
- $$\begin{aligned}\mathrm{Cor}(X,Y)_{ij} &= \frac{\mathrm{Cov}(X_i,Y_j)}{\sigma(X_i)\,\sigma(Y_j)} \\ &= \frac{\mathbb{E}[(X_i-\mathbb{E}[X_i])(Y_j-\mathbb{E}[Y_j])]}{\sigma(X_i)\,\sigma(Y_j)} \\ &= \mathbb{E}\left[\frac{(X_i-\mathbb{E}[X_i])}{\sigma(X_i)}\frac{(Y_j-\mathbb{E}[Y_j])}{\sigma(Y_j)}\right]\end{aligned}$$

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

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

Moments

- $\mathrm{Cov}(1_A, 1_B) = \mathbb{E}[1_{A\cap B}] - P(A)P(B) = P(A \cap B) - P(A)P(B)$.
  In particular
  $\mathrm{Cov}(1_A) = P\left(A\right) - P\left(A\right)^2 = P\left(A\right)\left(1 - P\left(A\right)\right) = P\left(A\right)P\left(A^{\complement}\right)$.
-
  $$\mathrm{Cor}(1_A, 1_B) = \frac{P(A \cap B) - P(A)P(B)}{\sqrt{(P(A)(1 - P(A))P(B)(1 - P(B))}}$$
- IF events A and B are independent, this is equivalent to $P(A \cap B) = P(A)P(B)$ and therefore $\mathrm{Cor}(1_A, 1_B) = 0$.

Moments

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

Moments

- Moments encode quantitative information about the shape of probability distributions/densities of random variables.
- Be $k \in \mathbb{N}_0$. For a univariate random variable X, the expectation value $\mu_k(X) \equiv \mathbb{E}[X^k]$ is called the k-th raw moment.
- $\mathbb{E}[|X|^k]$ denotes the k-th absolute moment of X.
- The expectation values $\mu'_k(X) \equiv \mathbb{E}[(X - \mathbb{E}[X])^k]$ and

$$\tilde{\mu}_k(X) \equiv \mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sigma(X)}\right)^k\right] = \frac{\mu'_k(X)}{\sigma(X)^k}$$

  are called the k-th centralized moment and the k-th standardized moment of X respectively.

Moments

- $\mu_1(X) = \mathbb{E}[X], \mu_2(x) = \mathbb{E}[X^2] \quad \left(\text{i.e. } \mathsf{Var}(X) = \mu_2(X) - \mathbb{E}[X]^2\right).$
- $\mu_1'(X) = 0, \mu_2'(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathsf{Var}(X).$
- $\tilde{\mu}_1(X) = 0, \tilde{\mu}_2(X) = 1.$

Moments

- For a multivariate random variable X, the k-tensor $\mu_k(X) \equiv \mathbb{E}[X^{\otimes k}]$ is called the raw k-th moment.
- Correspondingly denote $\mu'_k(X) \equiv \mathbb{E}[(X - \mathbb{E}[X])^{\otimes k}]$ and

$$\tilde{\mu}_k(X) \equiv \mathbb{E}\left[\{\sigma(X)^{-1}(X - \mathbb{E}[X])\}^{\otimes k}\right] = \left(\sigma(X)^{-1}\right)^{\otimes k} \cdot \mu'_k(X)$$

  the k-th centralized moment and the k-th standardized moment of X respectively (the generalized dot product here is tensor contraction) .

Moments

- $\mu_1(X) = \mathbb{E}[X]$, $\mu_2(x) = \mathbb{E}[XX^\top]$ (i.e. $\mathrm{Cov}(X) = \mu_2(X) - \mathbb{E}[X]\mathbb{E}[X]^\top$) .
- $\mu_1'(X) = 0$, $\mu_2'(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] = \mathrm{Cov}(X)$.
- $\tilde{\mu}_1(X) = 0$, $\tilde{\mu}_2(X) = 1$.
- 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

- The skewness of a random variable X measures the asymmetry of the probability distribution/density of X around its mean. This is intuitive in the univariate case where skewness is defined by $\tilde{\mu}_3(X)$, the third standardized moment of X.
- Not so much in the multivariate case where joint behavior of the component variables plays in i.e. joint asymmetry.

Moments

- Mardia's theory generalizes the univariate notions of skewness and kurtosis to higher dimensions, using the Mahalanobis distance from X to $\mathbb{E}[X]$ for $\mathsf{Cov}(X) \succ 0$:

$$Q(X) \equiv (X - \mathbb{E}[X])^\top \mathsf{Cov}(X)^{-1}(X - \mathbb{E}[X])$$

- Mardia's skewness is defined as

$$\beta_{1,d}(X) \equiv \mathbb{E}\left[\left\{(X_1 - \mathbb{E}[X])^\top \mathsf{Cov}(X)^{-1}(X_2 - \mathbb{E}[X])\right\}^3\right]$$

  where X_1 and X_2 are two independent copies of X.

Moments

- Why not define as $\mathbb{E}\left[Q(X)^3\right]$? Because it is not invariant under affine transformations and rotations in $\mathbb{R}^d$. Another disadvantage, it doesn't relate to the univariate skewness : For $d = 1$ is $\mathbb{E}\left[Q(X)^3\right] = \tilde{\mu}_6(X)$.
- But

$$\begin{aligned}
\beta_{1,1}(X) &= \sigma(X)^{-6}\mathbb{E}\left[\left\{(X_1 - \mathbb{E}[X])^\top (X_2 - \mathbb{E}[X])\right\}^3\right] \\
&= \sigma(X)^{-6}\mathbb{E}\left[(X_1 - \mathbb{E}[X])^3\right]\mathbb{E}\left[(X_2 - \mathbb{E}[X])^3\right] \\
&= \sigma(X)^{-6}\left(\mathbb{E}\left[(X - \mathbb{E}[X])^3\right]\right)^2
\end{aligned}$$

and therefore $\beta_{1,1}(X) = (\tilde{\mu}_3(X))^2$.

Moments

- The kurtosis of a univariate random variable X is $\tilde{\mu}_4(X)$, its fourth standardized moment. The kurtosis captures how likely it is for observations of X to fall far (in any direction) from its mean i.e. how much the probability distribution/ density of X is characterized through outliers. In the univariate case this translates into measuring the heaviness of the tails.
- For a multivariate random variable X, Mardia's kurtosis is defined as

  $$\beta_{2,d}(X) \equiv \mathbb{E}\left[Q(X)^2\right] \quad .$$

  This definition is analogous to the univariate case, i.e. $\beta_{2,1}(X) = \tilde{\mu}_4(X)$.

Moments

- Skewness of a d-multivariate normal distribution (or any other symmetric distribution) is 0.
- Kurtosis for a d-multivariate normal distribution is $d(d+2)$. Kurtosis $> d(d+2)$ indicates heavy tails, we speak of a leptokurtic distribution. Kurtosis $< d(d+2)$ suggests light tails, we speak of a platykurtic distribution.

Moments

- To complete the moment ladder of Mahalanobis distance, note that

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

If X is d-multivariate normal, then $Q(X) \sim \chi^2_d$, which confirms this result.

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 :

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

Moments

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

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

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

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

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

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

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 under the assumption $P(B) > 0$

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

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

Conditional Expectation

- Conditioning on a class of events or a zero probability event requires a broader notion of conditional expectation. As a generalization, the conditional expectation of Y w.r.t. a σ-subalgebra $\mathcal{G} \subseteq \mathcal{A}$, denoted by $\mathbb{E}[Y \mid \mathcal{G}]$ ($\mathbb{E}^{\mathcal{G}}[Y]$ a notational shortcut), is defined as 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: $\mathbb{E}^{\mathcal{G}}[Y] = (\mathbb{E}^{\mathcal{G}}[Y_1], \ldots, \mathbb{E}^{\mathcal{G}}[Y_d])$.
- 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 σ-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: $\mathbb{E}[\mathbb{E}^{\mathcal{G}}[Y]] = \mathbb{E}[Y]$

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:
  $\mathrm{Var}[X] = \mathbb{E}[\mathrm{Var}^{\mathcal{G}}[X]] + \mathrm{Var}[\mathbb{E}^{\mathcal{G}}[X]]$

Conditional Expectation

- $\mathbb{E}[\cdot \mid \mathcal{G}] : L^1(\Omega, \mathcal{A}, P\,; \mathbb{R}^d) \to L^1(\Omega, \mathcal{G}, P|_{\mathcal{G}}\,; \mathbb{R}^d)$ is a continuous linear operator.
- Restricted to $L^2(\Omega, \mathcal{A}, P\,; \mathbb{R}^d)$, $\mathbb{E}[\cdot \mid \mathcal{G}]$ is the orthogonal projection onto $L^2(\Omega, \mathcal{G}, P|_{\mathcal{G}}\,; \mathbb{R}^d)$ where

$$L^2(\Omega, \mathcal{A}, P\,; \mathbb{R}^d) = L^2(\Omega, \mathcal{G}, P|_{\mathcal{G}}\,; \mathbb{R}^d) \oplus L^2(\Omega, \mathcal{G}, P|_{\mathcal{G}}\,; \mathbb{R}^d)^{\perp} \quad .$$

Semigroups of Kernels

- A (transition) kernel from measurable space $(\Omega, \mathcal{A})$ to a measurable space $(\mathsf{E}, \mathcal{E})$ 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 $K(\omega, \mathsf{E}) \leq 1$, a probability or Markov kernel if $K(\omega, \mathsf{E}) = 1$. In these cases obviously is $K : \Omega \times \mathcal{E} \to [0, 1]$.
  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 σ-finite if $K(\omega, \cdot)$ is σ-finite for every $\omega \in \Omega$. K is called finite and bounded if $K(\omega, \mathsf{E})$ 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 measurable space $(\Omega, \mathcal{A})$ to a measurable 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 σ-finite, finite, bounded and of (sub-)Markov kernels respectively.

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

- $\mathcal{K}_\sigma(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$, $\mathcal{K}_f(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ 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 δ_ω. $1_\Omega \in \mathcal{K}_1(\Omega, \mathcal{A})$ is called the unit kernel.
- Any $(\mathcal{A}, \mathcal{E})$-measurable function $f : (\Omega, \mathcal{A}) \to (\mathsf{T}, \mathcal{E})$ induces a Markov kernel $K_f : \Omega \times \mathcal{E} \longrightarrow [0,1]$ from $(\Omega, \mathcal{A})$ to $(\mathsf{T}, \mathcal{E})$ with $K_f(\omega, E) \equiv f_*\delta_\omega(E) = \delta_{f(\omega)}(E)$.
- In general, image measures $f_*\nu$ with ν a measure on $\mathcal{A}$ are kernels, independent from ω (Markov kernels, if ν is a probability measure).

Semigroups of Kernels

- If $K, L \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ and $c \in \mathbb{R}_+$ then is $K + L \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ and $c\,K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$.
- $\mathcal{K}_b(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$, $\mathcal{K}_f(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ and $\mathcal{K}_\sigma(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ are closed under these two operations.
- Product of kernels: For $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ and $L \in \mathcal{K}(\Omega \times \mathsf{E}, \mathcal{A} \otimes \mathcal{E}; \Sigma, \mathcal{S})$ define a product $K \otimes L \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E} \times \Sigma, \mathcal{E} \otimes \mathcal{S})$ 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 $\mathcal{K}(\Omega \times \mathsf{E}, \mathcal{A} \otimes \mathcal{E}; \Sigma, \mathcal{S})$ without $(\Omega, \mathcal{A})$-dependency.

Semigroups of Kernels

- $K \in \mathcal{K}_f(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E}) \wedge L \in \mathcal{K}_f(\Omega \times \mathsf{E}, \mathcal{A} \otimes \mathcal{E}; \Sigma, \mathcal{S}) \implies K \otimes L \in \mathcal{K}_\sigma(\Omega, \mathcal{A}; \mathsf{E} \times \Sigma, \mathcal{E} \otimes \mathcal{S})$.
- The product preserves the Markov property 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 $K\,L \in \mathcal{K}(\Omega, \mathcal{A}; \Sigma, \mathcal{S})$

Semigroups of Kernels

- 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.
- Commutativity fails: $K, L \in \mathcal{K}(\Omega, \mathcal{A})$ does not imply $K\,L = L\,K$.
- $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E}) \wedge 1_\Omega \in \mathcal{K}_1(\Omega, \mathcal{A}) \implies 1_\Omega\,K = K$.
- $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E}) \wedge 1_\mathsf{E} \in \mathcal{K}_1(\mathsf{E}, \mathcal{E}) \implies K\,1_\mathsf{E} = K$.
- Notation: For $K \in \mathcal{K}(\Omega, \mathcal{A})$ define $K^n \equiv K\,K \cdots K$ (n times)

Semigroups of Kernels

- Associativity: For $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$, $L \in \mathcal{K}(\mathsf{E}, \mathcal{E}; \Sigma, \mathcal{S})$ and $M \in \mathcal{K}(\Sigma, \mathcal{S}; \mathsf{T}, \mathcal{T})$:

$$(K\,L)\,M = K\,(L\,M)$$

- Distributivity: If $K, L \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$, $M, N \in \mathcal{K}(\mathsf{E}, \mathcal{E}; \Sigma, \mathcal{S})$, it holds that :

$$(K + L)\,M = K\,M + K\,L \quad , \; K\,(M + N) = K\,M + K\,N$$

Semigroups of Kernels

- Translation-invariance : Be E a topological vector space. Then $K \in \mathcal{K}(\mathsf{E}, \mathcal{E}; \mathsf{E}, \mathcal{E})$ is called translation-invariant if

$$K(\epsilon + h, E + h) = K(\epsilon, E) \quad \forall\, \epsilon, h \in \mathsf{E}, E \in \mathcal{E} \quad .$$

Semigroups of Kernels

- A kernel $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ induces two natural integral operators $(\cdot\, K) : \mathcal{M}(\Omega, \mathcal{A}) \longrightarrow \mathcal{M}(\mathsf{E}, \mathcal{E})$, 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 $(K\,\cdot) : \mathcal{L}^0(\mathsf{E}, \mathcal{E}; \mathbb{R}) \longrightarrow \mathcal{L}^0(\Omega, \mathcal{A}; \mathbb{R})$, defined by

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

- $L^0(\mathsf{E}, \mathcal{E}, Q)$ -> lift if $K\,(\omega, \cdot) \ll Q \quad \forall \omega \in \Omega$
- 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 $f \in \mathcal{L}^0(\Omega, \mathcal{A}; \mathbb{R})$.

Semigroups of Kernels

- For $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$, $c \in \mathbb{R}$, $\nu, \mu \in \mathcal{M}(\Omega, \mathcal{A})$ it holds that $c\,(\nu\, K) = (c\,\nu)\, K$ and that $(\nu + \mu)\, K = \nu\, K + \mu\, K$ i.e. that $(\cdot\, K) : \mathcal{M}(\Omega, \mathcal{A}) \longrightarrow \mathcal{M}(\mathsf{E}, \mathcal{E})$ is an $\mathbb{R}$-linear operator. $(\cdot\, K)$ respects the cone structures of $\mathcal{M}(\Omega, \mathcal{A})$ and $\mathcal{M}(\mathsf{E}, \mathcal{E})$ i.e. $(\cdot\, K)_{|\mathcal{M}^+(\Omega,\mathcal{A})}$ is an $\mathcal{M}^+(\mathsf{E}, \mathcal{E})$-valued additive $\mathbb{R}_+$- homogenous operator and $(\cdot\, K)_{|\mathcal{M}^-(\Omega,\mathcal{A})}$ is an $\mathcal{M}^-(\mathsf{E}, \mathcal{E})$-valued additive $\mathbb{R}_-$-homogenous operator
- Recover the kernel from its operator: $(\delta_\omega\, K)\,(E) = K(\omega, E)$.
- $K \in \mathcal{K}_1(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E}) \implies (\cdot\, K) : \mathcal{M}_1(\Omega, \mathcal{A}) \longrightarrow \mathcal{M}_1(\mathsf{E}, \mathcal{E})$.
- $K \in \mathcal{K}_{\leq 1}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E}) \implies (\cdot\, K) : \mathcal{M}_{\leq 1}(\Omega, \mathcal{A}) \longrightarrow \mathcal{M}_{\leq 1}(\mathsf{E}, \mathcal{E})$.

Semigroups of Kernels

- For $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$, $c \in \mathbb{R}, f, g \in \mathcal{L}^0(\mathsf{E}, \mathcal{E}; \mathbb{R})$, it holds that $(Kf)\, c = K(f c)$ and that $K(f + g) = Kf + Kg$ i.e. that $(K\,\cdot) : \mathcal{L}^0(\mathsf{E}, \mathcal{E}; \mathbb{R}) \longrightarrow \mathcal{L}^0(\Omega, \mathcal{A}; \mathbb{R})$ is an $\mathbb{R}$-linear operator. It respects the cone structures of $\mathcal{M}(\Omega, \mathcal{A})$ and $\mathcal{M}(\mathsf{E}, \mathcal{E})$.
- Recover the kernel from its operator: $(K1_E)(\omega) = K(\omega, E)$
- If $K \in \mathcal{K}_{\leq 1}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$, then the left operation can be restricted to bounded measurable functions $\mathcal{B}_b(\mathsf{E}, \mathcal{E}) \subseteq \mathcal{L}^0(\mathsf{E}, \mathcal{E})$. $(K\,\cdot)_{|\mathcal{B}_b(\mathsf{E}, \mathcal{E})}$ is a $\mathcal{B}_b(\Omega, \mathcal{A})$-valued $\mathbb{R}$-linear operator. It respects the cone structures of $\mathcal{B}_b(\mathsf{E}, \mathcal{E})$ and $\mathcal{B}_b(\Omega, \mathcal{A})$.

Semigroups of Kernels

- Consider a pairing between $\mathcal{M}(\mathsf{E}, \mathcal{E})$ and $\mathcal{L}^0(\mathsf{E}, \mathcal{E}; \mathbb{R})$, defined by:

  $$\langle \nu \,|\, f \rangle \equiv \int_{\mathsf{E}} \nu(d\epsilon)\, f(\epsilon)$$

  Applying Tonelli's theorem results in an adjoint relation between the two integral operators :

  $$\begin{aligned} \langle \nu\, K \,|\, f \rangle &= \int_{\mathsf{E}} (\nu K)(d\epsilon)\, f(\epsilon) = \int_{\mathsf{E}} \left(\int_{\Omega} \nu(d\omega)\, K(\omega, d\epsilon) \right) f(\epsilon) \\ &= \int_{\Omega} \nu(d\omega) \left(\int_{\mathsf{E}} K(\omega, d\epsilon)\, f(\epsilon) \right) = \langle \nu \,|\, K f \rangle \end{aligned}$$

  The pairing respects the cone structure in both measure and function space.

Semigroups of Kernels

- $(\cdot\, K) = (K\, \cdot)^*$
- Note that the kernel operators $(K\, \cdot)$ and $(\cdot\, K)$ are treated here as 'adjoint' in informal language but they are not adjoint operators in a strict functional analytic sense. $\mathcal{L}^0(\mathsf{E}, \mathcal{E}; \mathbb{R})$ is not dual to $\mathcal{M}(\mathsf{E}, \mathcal{E})$ ($\mathcal{L}^0(\mathsf{E}, \mathcal{E}; \mathbb{R})$ and $\mathcal{M}(\mathsf{E}, \mathcal{E})$ are not even Banach spaces). But if we, for example, replace both spaces with $\mathcal{B}_b(\mathsf{E}, \mathcal{E})$ and $ba(\mathsf{E}, \mathcal{E})$ respectively, (with $ba(\mathsf{E}, \mathcal{E}) = \mathcal{B}_b(\mathsf{E}, \mathcal{E})'$ (see above)), the operators are adjoint.

Semigroups of Kernels

- Assume a measurable space $(\Omega, \mathcal{A})$, a measure space $(\mathsf{E}, \mathcal{E}, \lambda)$ and a kernel $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$. A measurable function $k \in \mathcal{L}^0_+(\Omega \times \mathsf{E}, \mathcal{A} \otimes \mathcal{E}; \mathbb{R})$ s.t.

$$K(\omega, E) = \int_E \lambda(d\epsilon)\, k(\omega, \epsilon)$$

  is called a kernel density function (kdf) of K (w.r.t. λ).
- K has a kdf k w.r.t λ iff $K(\omega, \cdot) \ll \lambda$ for all $\omega \in \Omega$. $k(\omega, \_)$ is then the R-N derivative of $K(\omega, \cdot)$ w.r.t. λ :

$$k(\omega, \_) = \frac{dK(\omega, \cdot)}{d\lambda} \;\in \mathcal{L}^0_+(\mathsf{E}, \mathcal{E}; \mathbb{R})$$

Semigroups of Kernels

- A kdf $k \in \mathcal{L}^0_+(\Omega \times \mathsf{E}, \mathcal{A} \otimes \mathcal{E}; \mathbb{R})$ is called a Markov kernel density function if it defines a Markov kernel, i.e. if it holds that
$$K(\omega, \mathsf{E}) = \int_{\mathsf{E}} \lambda(d\epsilon)\, k(\omega, \epsilon) = 1 \;\; \forall \omega \in \Omega \quad .$$
- If in addition to being a Markov kdf, there is
$$\int_{\Omega} \nu(d\omega)\, k(\omega, \epsilon) = 1 \;\; \forall \epsilon \in \mathsf{E} \quad ,$$
then k is called a double kernel density function.

Semigroups of Kernels

- A kdf $k \in \mathcal{L}^0_+(\mathsf{E} \times \mathsf{E}, \mathcal{E} \otimes \mathcal{E}; \mathbb{R})$ is called symmetric if $k(\epsilon, \epsilon') = k(\epsilon', \epsilon)$ for all $(\epsilon, \epsilon') \in \mathsf{E} \times \mathsf{E}$.
- If k is a symmetric Markov kdf, then it is a double kdf but not vice versa.
- For kernels $K \in \mathcal{K}(\Omega, \mathcal{A}; \mathsf{E}, \mathcal{E})$ and $L \in \mathcal{K}(\mathsf{E}, \mathcal{E}; \Sigma, \mathcal{S})$ with kdfs $k \in \mathcal{L}^0_+(\Omega \times \mathsf{E}, \mathcal{A} \otimes \mathcal{E}; \mathbb{R})$ and $l \in \mathcal{L}^0_+(\mathsf{E} \times \Sigma, \mathcal{E} \otimes \mathcal{S}; \mathbb{R})$ respectively, the kdf $k\,l \in \mathcal{L}^0_+(\Omega \times \Sigma, \mathcal{A} \otimes \mathcal{S}; \mathbb{R})$ of their composition kernel $K\,L \in \mathcal{K}(\Omega, \mathcal{A}; \Sigma, \mathcal{S})$ is calculated like so

$$k\,l(\omega, \sigma) = \int_{\mathsf{E}} \lambda(d\epsilon)\, k(\omega, \epsilon)\, l(\epsilon, \sigma) \quad .$$

Semigroups of Kernels

- For a $f \in \mathcal{L}^0(\mathsf{E}, \mathcal{E}, \lambda; \mathbb{R})$ is

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

- The adjoint operation (see below) for a $f \in \mathcal{L}^0(\Omega, \mathcal{A}, \nu; \mathbb{R})$:

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

Semigroups of Kernels

- The common use case is $(\Omega, \mathcal{A}, \nu) = (\mathsf{E}, \mathcal{E}, \lambda)$. $\mathcal{K}(\mathsf{E}, \mathcal{E})$, $\mathcal{K}_f(\mathsf{E}, \mathcal{E})$, $\mathcal{K}_b(\mathsf{E}, \mathcal{E})$, $\mathcal{K}_{\leq 1}(\mathsf{E}, \mathcal{E})$ and $\mathcal{K}_1(\mathsf{E}, \mathcal{E})$ are monoids w.r.t. the composition i.e. they are semigroups w.r.t. the composition with 1_E as their identity element.
- There is a monoid structure w.r.t. the addition for $\mathcal{K}(\mathsf{E}, \mathcal{E})$, $\mathcal{K}_\sigma(\mathsf{E}, \mathcal{E})$, $\mathcal{K}_f(\mathsf{E}, \mathcal{E})$ and $\mathcal{K}_b(\mathsf{E}, \mathcal{E})$. But when we speak of monoids or semigroups of kernels, we will always understand these to be w.r.t. the composition operation.
- An arbitrary semigroup of kernels $K \subset \mathcal{K}(\mathsf{E}, \mathcal{E})$ is called finite, bounded or (sub-) Markovian if all its members have this property.

Semigroups of Kernels

- Consider a family of kernels $K = (K_{t,s})_{\{t,s \in \mathbb{R}_+ | s \leq t\}}$ on $(\mathsf{E}, \mathcal{E})$. K fullfills the Chapman-Kolmogorv equations (C-K equations) if it holds $\forall\ s \leq t \leq u \in \mathbb{R}_+$ that

$$K_{u,s} = K_{t,s}\, K_{u,t} \quad ,$$

that is if

$$K_{u,s}(\epsilon, E) = \int_{\mathsf{E}} K_{t,s}(\epsilon, d\epsilon')\, K_{u,t}(\epsilon', E) \quad (\forall\ (\epsilon, E) \in \mathsf{E} \times \mathcal{E}) \ .$$

- A family of kernels satisfying the C-K equations constitutes a semigroup. This can be generalized to n-parameter families and n-parameter C-K equations.

Semigroups of Kernels

- If the kernels have kdfs w.r.t. a reference measure λ, these satisfy corresponding C-K equations

$$k_{u,s}(\epsilon, \epsilon') = \int_{\mathsf{E}} \lambda(d\rho)\, k_{t,s}(\epsilon, \rho)\, k_{u,t}(\rho, \epsilon') \quad .$$

Semigroups of Kernels

- If $K = (K_{t,s})_{\{t,s \in \mathbb{R}_+ | s \leq t\}} \subset \mathcal{K}(\mathsf{E}, \mathcal{E})$ is a family of translation-invariant kernels, defining

  $$\mu_{t,s}(E - \epsilon) \equiv K_{t,s}(\epsilon, E) \quad \text{for} \;\; \epsilon \in \mathsf{E}, E \in \mathcal{E} \quad ,$$

  results in a family of measures $(\mu_{t,s})_{\{t,s \in \mathbb{R}_+ | s \leq t\}}$ on $\mathcal{E}$. Note that

  $$K_{t,s}(\epsilon, E) = \delta_\epsilon * \mu_{t,s}(E) \quad .$$

  Such a kernel is called a convolution kernel .
- Obeying the C-K equations translates into

  $$\mu_{u,s} = \mu_{t,s} * \mu_{u,t} \quad .$$

Semigroups of Kernels

- If the $K_{s,t}$ are time-homogeneous, i.e. depend only on the time-difference $t - s$, the Chapman-Kolmogorv equations turn into

  $$K_{s+t} = K_s\, K_t \quad .$$

  This obviously implies commutativity $(s + t = t + s)$
- For the sake of completeness, the kdf version :

  $$k_{s+t}(\epsilon, \epsilon') = \int_{\mathsf{E}} \lambda(d\rho)\, k_s(\epsilon, \rho)\, k_t(\rho, \epsilon') \quad .$$

Semigroups of Kernels

- Be $K \subset \mathcal{K}(\mathsf{E}, \mathcal{E})$ a semigroup of kernels. Since $(\nu\, K')\, L' = \nu\, (K'\, L')$ and $K'\, (L'\, g) = (K'\, L')\, g$ for $K', L' \in K$, the associated integral operators (see above) induce semigroup representations

$$\begin{aligned} K &\longrightarrow \mathbf{End}(\mathcal{M}(\mathsf{E}, \mathcal{E})) \text{ via } K' \mapsto (K' \,\cdot) \\ K &\longrightarrow \mathbf{End}(\mathcal{L}^0(\mathsf{E}, \mathcal{E})) \text{ via } K, \mapsto (\cdot\, K') \quad . \end{aligned}$$

- The images of these representations are semigroups of operators i.e. semigroups of transition kernels translate into semigroups of operators on measure and/or function spaces.

Conditional Probability

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

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

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

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

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

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 $P(B) > 0$
  $$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \int_B dP\, \mathbb{P}^{\mathcal{G}}[A] \,/\, \int_\Omega dP\, \mathbb{P}^{\mathcal{G}}[A]$$

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 restrictions on 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

- For $Y, X \in \mathcal{L}^0(\Omega, \mathcal{A}, P; \mathsf{E}, \mathcal{E})$ and $\mathcal{G} \equiv \sigma(X) \subset \mathcal{A}$ examine $\mathbb{E}^{\mathcal{G}}[Y] = \mathbb{E}[Y \mid X]$. The Doob-Dynkin lemma guarantees that there exists a measurable function $h \in \mathcal{L}^0(\mathsf{E}, \mathcal{E}; \mathsf{E}, \mathcal{E})$ s.t. $\mathbb{E}[Y \mid X] = h(X)$. h is unique P_X-a.e. . We define

  $$\mathbb{E}[Y \mid X = \epsilon] \equiv h(\epsilon) \text{ for any } \epsilon \in \mathsf{E} \quad .$$

  (Notation mimics the discrete situation.)
- $\mathbb{E}\,[Y \mid X](\omega) = \mathbb{E}[Y \mid X = X(\omega)]$ P-a.e.
- Conditioning on probability zero events.

Conditional Probability

- Be $Y \in \mathcal{L}^0(\Omega, \mathcal{A}, P; \mathsf{E}, \mathcal{E})$ an $(\mathsf{E}, \mathcal{E})$-valued random element. A regular conditional distribution (rcd) 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 $(\mathsf{E}, \mathcal{E})$ s.t. for all $E \in \mathcal{E}$

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

  i.e. such that $K_{Y|\mathcal{G}}(\cdot, E)$ is a version of $\mathbb{P}^{\mathcal{G}}[Y \in E]$.
- By this definition, $\mathbb{P}^{\mathcal{G}}[Y \in \cdot](\omega)$ is a probability measure on $(\mathsf{E}, \mathcal{E})$ or equivalently $\mathbb{P}^{\mathcal{G}}[\cdot](\omega)$ is a probability measure for events of type $\{Y \in E\} \in \mathcal{A}, (E \in \mathcal{E}) \quad P$-a.e. .

Conditional Probability

- If $(\mathsf{E}, \mathcal{E}) = (\Omega, \mathcal{A})$ and $Y(\omega) = \omega$, a regular conditional distribution of Y (if it exists) is called a regular conditional probability (rcp) because $K_{Y|\mathcal{G}}(\cdot, E) = \mathbb{P}^{\mathcal{G}}[Y \in E] = \mathbb{P}^{\mathcal{G}}[E]$ P-a.e. (see slides above).
- An rcd 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.
- If $X \in \mathcal{L}^0(\Omega, \mathcal{A}, P; \mathsf{E}, \mathcal{E})$ and $\mathcal{G} = \sigma(X)$ then (by Doob-Dynkin)

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

Conditional Probability

- Note that $K(\epsilon, E) \equiv \mathbb{P}[Y \in E \mid X = \epsilon]$ constitutes a Markov kernel from $(\mathsf{E}, \mathcal{E})$ to $(\mathsf{E}, \mathcal{E})$. Obviously is

  $$K_{Y|X}(\omega, E) = K(X(\omega), E) \quad P\text{-a.e. } .$$

  Since K is unique (again by the Doob-Dynkin lemma), $K_{Y|X}$ and K are used interchangeably in some textbooks. K is then also called an rcd of Y w.r.t. X and denoted by $K_{Y|X}$.

Conditional Probability

- If E is a Polish space, the existence of an rcd allows a LOTUS-style representation of generalized conditional expectation: Be $Y \in \mathcal{L}^0(\Omega, \mathcal{A}, P; \mathsf{E}, \mathcal{E})$, $\mathcal{G} \subseteq \mathcal{A}$ a σ-subalgebra and $K_{Y|\mathcal{G}}$ a corresponding rcd. By definition is

  $$\mathbb{E}\left[1_{\{Y \in E\}} \mid \mathcal{G}\right](\omega) = \mathbb{P}\left[Y \in E \mid \mathcal{G}\right](\omega) = K_{Y|\mathcal{G}}(\omega, E) \quad P\text{-a.e.} \ .$$

  This can be generalized to the identity

  $$\mathbb{E}^{\mathcal{G}}\left[f(Y)\right](\omega) = \int_{\mathsf{E}} K_{Y|\mathcal{G}}(\omega, d\epsilon)\, f(\epsilon) \quad P\text{-a.e.} \ .$$

  where $f \in \mathcal{L}^0(\mathsf{E}, \mathcal{E}; \mathbb{R})$ with $f(Y) \in L^1(\Omega, \mathcal{A}, P; \mathbb{R})$.
- $\mathcal{F} = \sigma(X)$:

  $$\mathbb{E}\left[f(Y) \mid X\right](\omega) = \int_{\mathsf{E}} K_{Y|X}(\omega, d\epsilon)\, f(\epsilon) \quad P\text{-a.e.}$$

Conditional Probability Density

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

  $$p_X(x) \equiv \int_{\mathbb{R}^d} \lambda^d(dy)\, p_{(X,Y)}(x, y) \quad ,$$

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

  $$p_{Y|X=x}(y \mid x) \equiv \frac{p_{(X,Y)}(x, y)}{p_X(x)} \quad .$$

- Notation : As with pdfs, we simply write $p(y \mid x)$ instead of $p_{Y|X=x}(y \mid x)$ if the environment and its actors are known.

Conditional Probability

- If X and Y are independent, it holds that $p(x, y) = p(x)\, p(y)$, hence
$$p(y \mid x) = p_Y(y) \quad .$$
- For any $B \in \mathcal{B}(\mathbb{R}^d)$:
$$\mathbb{P}[Y \in B \mid X = x] = \int_B \lambda^d(dy)\, p(y \mid x)$$
$$\mathbb{P}[Y \in B \mid X](\omega) = \int_B \lambda^d(dy)\, p(y \mid X(\omega))\, g(y)$$

Conditional Probability

- For a general $g \in \mathcal{L}^0(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d); \mathbb{R})$ with $g(Y) \in L^1(\Omega, \mathcal{A}, P; \mathbb{R})$:

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

$$\mathbb{E}[f(Y) \mid X](\omega) = \int_{\mathbb{R}^d} \lambda^d(dy)\, p(y \mid X(\omega))\, f(y)$$

Conditional Probability

- If the joint pdf of X and Y exists :

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

- If the joint pdf of X and Y does not exist but the conditional probability distribution $\mathbb{P}[Y \in B \mid X = x]$ is well-defined (in general probability spaces without pdfs for example) :

$$\mathbb{E}[Y \mid X = x] = \int_{\mathbb{R}^d} \mathbb{P}[Y \in dy \mid X = x]\, y$$

Conditional Probability Density

- Bayes theorem:

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

’Mathematically easy, conceptually powerful.’

7. Gaussian Measures

Characteristic Functions and Moment generation

- The characteristic function $\Phi_\nu : \mathbb{R}^d \to \mathbb{C}$ of a measure ν 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 Φ_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 $p(x)$.
- Φ_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:

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

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

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

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

- For two random variables X and Y

$$\Phi_X = \Phi_Y \text{ iff } P_X = P_Y$$

or equivalently

$$\Phi_X = \Phi_Y \text{ iff } F_X = F_Y \quad .$$

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

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

- If $\boldsymbol{\Sigma} = 0$, we put

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

Gaussian Measures

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

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

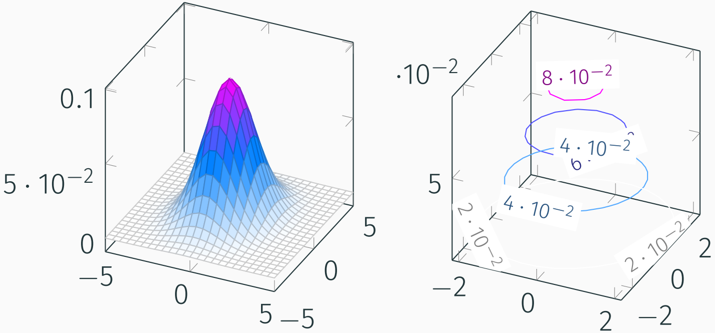
- In case $\gamma^d_{\mu,\boldsymbol{\Sigma}}$ is degenerate, there is $\gamma^d_{\mu,\boldsymbol{\Sigma}} \perp \lambda^d$.

Gaussian Measures

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

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 positiv semidefinite matrix Σ, its characteristic function is. Allows for a more general definition of a Gaussian measure:

- A measure γ 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 ϕ_γ is of the form

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

  and Σ is positive semidefinite.

Gaussian Measures

- If Σ is positive definite, this is the nondegenerate Gaussian measure.
- Note that for $\Sigma = 0$ we get the characteristic function of the Dirac measure, which is defined to be a Gaussian measure.
- Gaussian measures with mean zero are called centered Gaussian measures.
- Gaussian measures γ^d_{0,I_d} are called standard Gaussian measures, denoted by γ^d.
- $\mathrm{supp}(\gamma^d_{\mu,\Sigma}) = \{\mu + \Sigma x : x \in \mathbb{R}^d\}$

Gaussian Measures

- Every general Gaussian measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ can be written as the pushforward of a standard Gaussian measure via an affine transformation.
- As a consequence, every 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, Σ_γ the covariance operator with $q_\gamma(t) = \langle \Sigma_\gamma t, t \rangle$.

Gaussian Measures

- If γ is a Gaussian measure on $\mathbb{R}^d$, we get its mean μ_γ via

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

and its covariance matrix Σ_γ via

$$\begin{aligned}\Sigma_\gamma &= \int_{\mathbb{R}^d} \gamma(dx)\,(x - \mu_\gamma)(x - \mu_\gamma)^\top \\ &= \int_{\mathbb{R}^d} \gamma(dx) x\, x^\top - \left(\int_{\mathbb{R}^d} \gamma(dx)\, x\right)\left(\int_{\mathbb{R}^d} \gamma(dx)\, x^\top\right) \quad .\end{aligned}$$

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 P_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

- $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ 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. Even if they are i.i.d. . If (X, Y) is Gaussian, its marginals X and Y are.
- A vector of (univariate) random variables $(X_1, \dots, X_d)$ is a (multivariate) Gaussian random variable if, for any $\alpha_1, \dots, \alpha_d \in \mathbb{R}$, the random variable $\sum \alpha_i X_i$ is Gaussian.
- If $X = (X_1, \dots, X_d)^\top$ is a Gaussian random variable, the X_i are independent from each other iff they are uncorrelated. If for example $X \sim \mathcal{N}(\mu, E_d)$, then the $X_i \sim \mathcal{N}(0, 1)$ are uncorrelated and therefore i.i.d. .

Gaussian Random Variables

- A different characterization of Gaussian random variables:

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

iff there exist a vector $\mu \in \mathbb{R}^d$, a Cholesky decomposition $\Sigma = A\,A^\top$ with $A \in \mathbb{R}^{d \times d}$ and a standard Gaussian random variable $Z \sim \mathcal{N}(0, E_d)$ such that

$$X = \mu + A\,Z \quad .$$

Remember that the Cholesky decomposition is unique iff Σ is not only positive semidefinite but positive definite.

Gaussian Random Variables

- Every Gaussian random variable $X \sim \mathcal{N}(\mu, \Sigma)$ can be produced from a standard Gaussian random variable Z via an affine transformation $T(Z) \equiv \mu + A\,Z$.
  Put differently, Gaussian random variables are closed under affine transformations.
- If Σ is singular i.e. has $r \equiv rank(\Sigma) < d = \dim \operatorname{supp}(P_X)$, choose A to be a $d \times r$ matrix and $Z \sim \mathcal{N}(0, E_r)$.
  X is here effectively an r-dimensional random variable embedded in $\mathbb{R}^d$.
- In particular $\Phi_{T(Z)}(t) = e^{i\,t^\top \mu}\,\Phi_Z(A\,Z)$

Gaussian Random Variables

- The above characterisation of Gaussian random variables is used by the so called reparameterization trick in a number of machine learning papers. Usually X there has a diagonal covariance matrix, that is $\Sigma = \sigma^{\odot 2} E_d$ with a variance vector $\sigma = (\sigma_1, \ldots, \sigma_d)^\top \in \mathbb{R}^d$ ($\odot$ is the elementwise product).

  As a consequence, X has a representation

$$X = \mu + \sigma \odot Z \quad .$$

  with $\mu \in \mathbb{R}^d$ and $Z \sim \mathcal{N}(0, E_d)$.

Gaussian Random Elements

- Gaussian measures can be defined for general cylindrical σ-algebras.
- Let E be a vector space and let $\Gamma \subseteq \mathsf{Hom}(\mathsf{E}, \mathbb{R})$ be some linear space, separating the points in E. A measure γ on $\mathcal{E}(\mathsf{E}, \Gamma)$ is called Gaussian if the push-forward measure $f_*\gamma$ is Gaussian on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for any $f \in \Gamma$.
- If E is a tvs, a measure γ on $\mathcal{E}(\mathsf{E})$ is called Gaussian if $f_*\gamma$ is Gaussian on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for every $f \in \mathsf{E}'$.
- An immediate consequence of the definition: Gaussian measures are probability measures.
- An $(\mathsf{E}, \mathcal{E})$-valued random element is called Gaussian if the induced measure on $\mathcal{E}$ is Gaussian.

Gaussian Random Elements

- The concept of ’characteristic function’ does exist for nuclear spaces. If E is a nuclear space, a function $\Phi : \mathsf{E} \to \mathbb{C}$ is called a characteristic function if

  (i) It is continuous on E,

  (ii) It is positiv definite i.e. for every $I \in \mathcal{F}(\mathbb{N})$ and tuples $(\alpha_{i_1}, \dots, \alpha_{i_{|I|}}) \in \mathbb{C}^{|I|}$, $(\epsilon_{i_1}, \dots, \epsilon_{i_{|I|}}) \in \mathsf{E}^{|I|}$ there is

$$\sum_{j,\,k=1}^{|I|} \alpha_j \, \overline{\alpha_k} \, \Phi(\epsilon_j - \epsilon_k) \geq 0$$

  (iii) $\Phi(0) = 1$

Gaussian Random Elements

- Bochner-Minlos Theorem. Be E a nuclear space, E' its strong dual and $\mathcal{E}(\mathsf{E}')$ the cylindrical σ-algebra on E'. If $\Phi : \mathsf{E} \to \mathbb{C}$ is a characteristic function on E like above, then there exists a unique probability measure μ_Φ on $(\mathsf{E}', \mathcal{E}(\mathsf{E}'))$ s.t.

$$\Phi(\epsilon) = \mathcal{F}[\mu_\Phi](\epsilon) = \int_{\mathsf{E}'} d\mu_\Phi \; \exp(i \langle t, \epsilon \rangle)$$

  that is, s.t. that Φ is the characteristic function of μ_Φ.

Law of Large Numbers (LLN)

- Be $X_1, X_2, \dots, X_n, \dots$ 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 μ 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 $\mathsf{Cov}(X_i) = \Sigma$, it holds that

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

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

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

Stochastic Processes

- A function $X : \Omega \to \mathsf{E}^T$ with the above property is called a stochastic process on $(\Omega, \mathcal{A}, P)$ with state space $(\mathsf{E}, \mathcal{E})$ and index set T. Equivalently such a stochastic process is a collection of $(\mathsf{E}, \mathcal{E})$-valued random elements $(X_t)_{t \in T}$. Notation: $X = (\Omega, \mathcal{A}, P;\ (X_t)_{t \in T}; \mathsf{E}, \mathcal{E})$ or simply $(X_t)_{t \in T}$ if the probability space and state space is understood.
- If 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 ω, the mapping $X(\cdot, \omega) : T \to \mathsf{E}, t \mapsto X_t(\omega)$ is called the sample function, realization, path or time series of the process X at ω.

Stochastic Processes

- The probability measure $P_X = X_*P \in \mathcal{M}_1(\mathsf{E}^T, \mathcal{E}^T)$ is called the distribution of the process. For finite T-subsets $\{s_1, \dots, s_n\} = S \in \mathcal{F}(T)$, the marginal distributions $P_{X_S} = X_{S*}P = (X_{s_1}, \dots, X_{s_n})_* P$ are called the finite-dimensional distributions of the process.
- If E is a topological space and $\mathcal{E} = \mathcal{B}(\mathsf{E})$, P_X is a probability measure on the cylindrical σ-algebra $\mathcal{B}(\mathsf{E})^T$ but in general not a Borel measure on $\mathcal{B}(\mathsf{E}^T) \supseteq \mathcal{B}(\mathsf{E})^T$. The inclusion

$$\mathcal{B}(\mathsf{E})^T = \bigotimes_T \mathcal{B}(\mathsf{E}) \subset \mathcal{B}(\prod_T \mathsf{E}) = \mathcal{B}(\mathsf{E}^T)$$

  is usually strict, especially if T is uncountable. If E is Polish ($\mathbb{R}^d$ for example) and T is finite or at least countable, then $\mathcal{B}(\mathsf{E})^T = \mathcal{B}(\mathsf{E}^T)$ does hold.

Stochastic Processes

- A $((\mathbb{R}^d)^T, \mathcal{B}(\mathbb{R}^d)^T)$-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

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

  X is $(\mathcal{A}, \mathcal{B}(\mathbb{R}^d)^T)$-measurable iff each X_t is $(\mathcal{A}, \mathcal{B}(\mathbb{R}^d))$-measurable.

Stochastic Processes

- Stochastic processes with the same state space and index set (not necessarily the same sample space) are called equivalent if they have the same finite-dimensional distributions.
- Stochastic processes with the same state space and index set (not necessarily the same sample space) are equivalent if they have the same overall distributions. Not necessarily the other way around.
- The same event can have different probabilities for equivalent processes. Take $(\Omega, \mathcal{A}, P) = ([0,1], \mathcal{B}([0,1]), \lambda)$ and $X_t(\omega) \equiv 0$, $Y_t(\omega) \equiv \{1 \text{ if } \omega = t; 0 \text{ else}\}$.

Stochastic Processes

- A stochastic process $X = (\Omega, \mathcal{A}, P;\ (X_t)_{t\in T}; \mathsf{E}, \mathcal{E})$ with state space $(\mathsf{E}, \mathcal{E})$ is obviously equivalent to the process

  $$Y = (\mathsf{E}^T, \mathcal{E}^T, P_X;\ (Y_t)_{t\in T} \equiv (\pi_t)_{t\in T}; \mathsf{E}, \mathcal{E})$$

  where the π_t are the projections from the t coordinate position. If we interpret E^T as the function space $\mathsf{Fun}(T, \mathsf{E})$, Y is the identity morphism

  $$Y = id : (\mathsf{E}^T, \mathcal{E}^T, P_X) \to (\mathsf{E}^T, \mathcal{E}^T)$$

  i.e. $Y(f) = f$ for $f \in \mathsf{Fun}(T, \mathsf{E})$. Y is called the first canonical process associated to X.

Stochastic Processes

- Note that by construction, the cylindrical σ-algebra $\mathcal{E}^T$ consists solely of events that depend on a countable number of the Y_t. If T is uncountable, say an interval on the real line, this is usually too coarse for practical purposes.
- Be E a Polish space, X a process with state space $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$ and Y the first canonical process

  $$Y = id : (\mathsf{E}^T, \mathcal{B}(\mathsf{E})^T, P_X) \to (\mathsf{E}^T, \mathcal{B}(\mathsf{E})^T) \quad .$$

  Can we extend the distribution P_X on $\mathcal{B}(\mathsf{E})^T$ to a distribution on $\mathcal{B}(\mathsf{E}^T)$? Tightness of P_X necessary at least.

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 $(\mathsf{E}, \mathcal{E})$ and index set T are said to be modifications of each other if it holds that $P\left(X_t = Y_t\right) = 1$ for every $t \in T.$ Modifications of each other have the same finite-dimensional distributions and are therefore equivalent.
- 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 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 $\mathcal{F}$ of σ-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 $\mathcal{F} = (\mathcal{F}_t)_{t\in T}$, 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 $(\mathsf{E}, \mathcal{E})$-valued stochastic process on a probability space $(\Omega, \mathcal{A}, P)$. The σ-algebras $\mathcal{F}_t^X \equiv \sigma(X_s : s \leq t) \subseteq \mathcal{A}$ constitute the so called natural filtration $\mathcal{F}^X$ 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 $\mathcal{F} = (\mathcal{F}_t)_{t \in T}$ 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 $\mathcal{F}^X$ 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 $(\Omega, \mathcal{A}, P; (X_t)_{t \in T}; \mathsf{E}, \mathcal{B}(\mathsf{E}))$, E Polish, and finite T-subsets $S \in \mathcal{F}(T)$, the random variables $X_S : \Omega \to \mathsf{E}^S$,$\omega \mapsto (X_s(\omega))_{s \in S}$ induce probability measures P_{X_S} on $\mathcal{B}(\mathsf{E})^S$, the finite-dimensional distributions of X. ($\mathcal{B}(\mathsf{E}^S) = \mathcal{B}(\mathsf{E})^S$ here, since E is Polish)
- Consider T-subsets $S \subseteq R \subseteq T$ with the corresponding projections $\pi_{RS} : (\mathsf{E}^R, \mathcal{B}(\mathsf{E})^R) \to (\mathsf{E}^S, \mathcal{B}(\mathsf{E})^S)$. The π_{RS} are $(\mathcal{B}(\mathsf{E})^R, \mathcal{B}(\mathsf{E})^S)$-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 $\{P_{X_S} : \mathcal{B}(\mathsf{E})^S \to [0,1]\}_{S \in \mathcal{F}(T)}$ 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 $Q \equiv \{Q_S : \mathcal{B}(\mathsf{E})^S \to [0,1]\}_{S \in \mathcal{F}(T)}$ there is a unique probability measure $Q_T \equiv \varprojlim_{S \in \mathcal{F}(T)} Q_S$ on $\mathcal{B}(\mathsf{E})^T$ 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:
  $X^Q \equiv (\mathsf{E}^T, \mathcal{B}(\mathsf{E})^T, Q_T; \mathsf{E}^T, \mathcal{B}(\mathsf{E})^T)$ with $X^Q(\omega) \equiv id$.

Construction of Stochastic Processes

- In many situations, the regularity conditions of the Kolmogorov extension theorem cannot be met. A more general approach is needed. Be $\Gamma = \{\gamma_i\}_{i \in I}$ a family of functions $\gamma_i : \mathsf{E} \to (\mathsf{E}_i, \mathcal{E}_i)$ and $\mathcal{C}yl(\mathsf{E}, \Gamma)$ the corresponding cylinder algabra on E, generated by Γ. Be further $\{Q_S : \sigma_{\mathsf{E}}(\mathcal{C}yl(\mathsf{E}, S) \to [0,1]\}_{S \in \mathcal{F}(\Gamma)}$ a projective system of measures denoted by Q. Its projective limit is a finitely additive set function

$$Q_\Gamma : \mathcal{C}yl(\mathsf{E}, \Gamma) \to [0, +\infty] \quad ,$$

defined on the cylindrical algebra $\mathcal{C}yl(\mathsf{E}, \Gamma)$. Q_Γ is not required to (and in general does not) extend to a full measure on the cylindrical σ-algebra $\sigma_{\mathsf{E}}(\mathcal{C}yl(\mathsf{E}, \Gamma)$. (Q, Q_Γ) is called a cylindrical measure on E .

Construction of Stochastic Processes

- If $\mathsf{E} = \prod_{i \in I} \mathsf{E}_i$ and the γ_i are the projections : π_i, $Q = \{Q_S : \mathcal{E}^S \to [0,1]\}_{S \in \mathcal{F}(I)}$ and $Q_I : \mathcal{C}yl(\mathsf{E}, \{\pi_i\}) \to [0, +\infty]$
- Common scenario : E is a tvs and $\Gamma = \mathsf{E}'$, its topological dual.
- The canonical Gaussian cylindrical measure on the Hilbert space $L^2(\mathbb{R}, \mathbb{R}^d)$ induces a cylindrical Gaussian measure Q on $\mathcal{S}' \equiv \mathcal{S}'(\mathbb{R}, \mathbb{R}^d)$, the space of vector-valued tempered distributions (see below). This measure Q extends uniquely to a full Radon Gaussian measure on $\mathcal{E}(\mathcal{S}', \mathcal{S})$, the canonical white noise measure (Check this -> Minlos theorem). See continuous-time white noise construction below.

Construction of Stochastic Processes

- The set of paths of X is E^T. In particular it may include discontinuous ones which causes some unwanted behaviour (nonmeasurability of certain events, ...). Is there an equivalent process with $\mathcal{C}(T, \mathsf{E}) \subset \mathsf{Fun}(T, \mathsf{E}) = \mathsf{E}^T$ its set of paths ?
- If the index set T is a topological space, a stochastic process $X = (\Omega, \mathcal{A}, P; (X_t)_{t \in T}; \mathsf{E}, \mathcal{B}(\mathsf{E}))$, E Polish, is called sample- continuous if $X(\cdot, \omega) \in \mathcal{C}(T, \mathsf{E}) \subseteq \mathsf{Fun}(T, \mathsf{E})$ for P-almost all $\omega \in \Omega$. It is called continuous if $X(\cdot, \omega) \in \mathcal{C}(T, \mathsf{E})$ for all $\omega \in \Omega$ i.e. if it is a stochastic process with paths in $\mathcal{C}(T, \mathsf{E})$.
- A modification Y of X is called a (sample-) continuous modification if Y is (sample-) continuous.

Construction of Stochastic Processes

- A sample continuous process is indistinguishable from a continuous process, so in particuar has a continuous modification.
- For a sample-continuous process, the finite distributions and the overall distribution determine each other.
- Kolmogorov continuity theorem: 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 $(\mathcal{L}^2 \equiv \mathcal{L}^2(\mathbb{R}_+, \lambda; \mathrm{E}), L^2 \equiv \mathcal{L}^2/\sim)$:

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

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

Construction of Stochastic Processes

- If the set T is countable, $\ell_\infty(T)$ is a separable Banach space and we can identify bounded stochastic processes, indexed by T, with $\ell_\infty(T)$-valued random variables.
- If T is a compact metric space, $\mathcal{C}(T, \mathbb{R})$ is a separable Banach space and and a sample-continuous process, indexed by T, corresponds to a $(\mathcal{C}(T, \mathbb{R}), \mathcal{B}(\mathcal{C}(T, \mathbb{R}))$-valued random variable with a Radon measure distribution.
- Be E a Polish space. The spaces of E-valued càdlàg functions on $T = \mathbb{R}_+$, denoted by $\mathcal{D}(\mathbb{R}_+, E)$, can be metrized s.t. they are a Polish spaces and the relative topology on the $\mathcal{C}(\mathbb{R}_+, E) \subset \mathcal{D}(\mathbb{R}_+, E)$ is the uniform topology. The spaces $\mathcal{D}(\mathbb{R}_+, E)$ with the metric-induced topology are called Skorokhod spaces.

Construction of Stochastic Processes

- A stochastic process $X = (\Omega, \mathcal{A}, P; (X_t)_{t \in \mathbb{R}_+}; \mathsf{E}, \mathcal{B}(\mathsf{E}))$, E Polish, is called a càdlàg process if $X(\cdot, \omega) \in \mathcal{D}(\mathbb{R}_+, \mathsf{E})$ for all $\omega \in \Omega$ i.e. if it is a stochastic process with paths in the Skorokhod space $\mathcal{D}(\mathbb{R}_+, \mathsf{E})$.
- Pattern: Take a stochastic process X with index set T and state space $(\mathsf{E}, \mathcal{E})$. Find a modification of X that factors over an interesting path space $(U, \mathcal{E}^T{}_{|U}) \subset (\mathsf{E}^T, \mathcal{E}^T)$ which is a separable Banach space or more general a Polish space. We then have modeled X as $(U, \mathcal{E}^T{}_{|U})$-valued random element. Examples: $\mathcal{C}(T, \mathsf{E})$ and Skorokhod spaces $\mathcal{D}(T, \mathsf{E})$.

Construction of Stochastic Processes

- Properties of Polish spaces that come into play:
  - $\mathcal{B}(\mathsf{E}^T) = \mathcal{B}(\mathsf{E})^T$ if T is countable
  - $\mathcal{B}(U) = \mathcal{B}(\mathsf{E})^T \cap U$ (This holds for $U = \mathcal{C}(T, \mathsf{E})$ and $T = \mathbb{R}_+$. Holds if T is any locally compact set. Note that $\mathcal{C}(T, \mathsf{E})$ is still a Polish space then. Does it hold for arbitrary Polish spaces U, E with $U \subset \mathsf{E}^T$, arbitrary index sets T ?)
- Dealing with nonseparable (Banach) spaces as path spaces requires additional precautions.

Construction of Stochastic Processes

- A stochastic process $(\Omega, \mathcal{A}, P; X = (X_t)_{t \in \mathbb{R}_+}; \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is called (strictly) 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 $s \in \mathbb{R}_+$.
- Define $\mu_X(t) \equiv \mathbb{E}[X_t]$ and

  $$K_{XX}(t, s) \equiv \mathrm{Cov}(X_t, X_s) \quad t, s \in \mathbb{R}_+$$

  the process mean and auto-covariance of X respectively.
- If $X_t \in L^1(\Omega, \mathcal{A}, P; \mathbb{R}^d) \;\; \forall t \in T$, stationarity implies that $\mu_X(t)$ is shift-invariant

  $$\mu_X(t) = \mu_X(t + r) \quad \forall t, r \in \mathbb{R}_+$$

  and therefore constant.

Construction of Stochastic Processes

- If $X_t \in L^2(\Omega, \mathcal{A}, P; \mathbb{R}^d) \ \ \forall t \in T$, stationarity implies shift-invariant covariance i.e.

$$K_{XX}(t,s) = K_{XX}(t-s,0) \quad \forall t,s \in \mathbb{R}_+ \quad .$$

- Weak-sense stationarity or wide-sense stationarity (WSS) is a more general concept. It requires the shift-invariance of process mean and autocovariance from above and finite second moments i.e. $X_t \in L^2(\Omega, \mathcal{A}, P; \mathbb{R}^d) \ \ \forall t \in T$. (Finite second moment are essential here to have finite variances $\mathbf{Var}((X_t)_i)$ and cross-covariances $\mathbf{Cov}((X_t)_i, (X_s)_j)$ ensuring well-defined auto-covariance matrices $K_{XX}(t,s)$.)

Gaussian Processes

- A Gaussian process is a stochastic process $X = (\Omega, \mathcal{A}, P;\ (X_t)_{t \in \mathbb{R}_+})$ 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 $\mu_X(t) = 0 \ \ \forall t \in \mathbb{R}_+$ 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

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

  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 $B = (B_t)_{t\in\mathbb{R}_+} = (B_t^{(1)}, \dots, B_t^{(d)})^\top_{t\in\mathbb{R}_+}$ 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. $B_t^{(1)}, \dots, B_t^{(d)}$ 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 ν. 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.
- $$Q^{\nu}(A) = \int_{\mathbb{R}^d} \nu(dx)\, Q^x(A)$$

White Noise

- A discrete-time white noise process with mean $\mu \in \mathbb{R}^d$ and positive semidefinite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is a stochastic process $W = (W_n)_{n \in \mathbb{N}_0}$, $W_n \in L^2(\Omega, \mathcal{A}, P; \mathbb{R}^d)$, with constant mean $\mathbb{E}[W_n] = \mu$, constant covariance matrix $\mathrm{Cov}\,(W_n) = \boldsymbol{\Sigma}$ and serially uncorrelated random variables i.e.
$$\mathrm{Cov}\,(W_n, W_m) = \delta_{nm} \boldsymbol{\Sigma} \quad , \forall n, m \in \mathbb{N}_0 \quad .$$
- W is by definition identically distributed but not i.i.d. (uncorrelated does not imply independent). In many use cases, being i.i.d. or even being Gaussian (which implies i.i.d.) is required of the W_n. We then speak of an i.i.d. or Gaussian discrete-time white noise process.
- A discrete-time white noise process is by definition strict stationary and WSS.

White Noise

- An intuitive definition of a continuous-time white noise process with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ is a stochastic process $W = (W_t)_{t \in \mathbb{R}_+}$, $W_t \in L^2(\Omega, \mathcal{A}, P; \mathbb{R}^d)$, s.t.

$$\mathbb{E}[W_t] = \mu, \quad \forall t \in \mathbb{R}_+$$

and

$$\operatorname{Cov}(W_t, W_s) = \delta(t-s)\,\Sigma \quad , \forall s, t \in \mathbb{R}_+ \quad ,$$

where δ is the Dirac delta distribution. This does not work out. A more general approach is needed.

White Noise

- Mathematically rigorous definition of continuous-time white noise:
  Be $\mathcal{S}(\mathbb{R}) \equiv \mathcal{S}(\mathbb{R}, \mathbb{R})$ the Schwartz space on $\mathbb{R}$ i.e. the space of smooth functions $f \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ that decay rapidly at infinity along with all their derivatives. Denote

  $$p_{k,l}(\phi) \equiv \sup_{t \in \mathbb{R}} |t^k \frac{d^l}{dt^l} \phi(t)|$$

  a family of semi-norms on $\mathcal{S}$, parameterized by indices $k, l \in \mathbb{N}_0$. Then $\phi \in \mathcal{S}(\mathbb{R})$ is equivalent to $p_{k,l}(\phi) < +\infty$ for all $k, l \in \mathbb{N}_0$. $\mathcal{S}(\mathbb{R})$ is a Fréchet space i.e. it is a complete metrizable Hausdorff lctvs (note that metrizability of a tvs implies the Hausdorff property).

White Noise

- $\mathcal{S}(\mathbb{R})$ is separable (consider functions with rational Taylor expansion around rational points) and therefore a Polish space. Note that because of the metrizability, $\mathcal{S}(\mathbb{R})$ being separable is equivalent to $\mathcal{S}(\mathbb{R})$ being second countable.
- Now consider $\mathcal{S} \equiv \mathcal{S}(\mathbb{R}, \mathbb{R}^d) = \mathcal{S}(\mathbb{R}) \otimes \mathbb{R}^d$, the space of vector-valued Schwartz functions. Obviously we can naturally identify $\mathcal{S}$ with the direct sum

$$\mathcal{S} = \bigoplus_{i=1}^{d} \mathcal{S}(\mathbb{R}) \quad .$$

Semi-norms on $\mathcal{S}$: For $\phi = (\phi_1, \ldots, \phi_d) \in \mathcal{S}$, $\phi_i \in \mathcal{S}(\mathbb{R})$, be

$$p_{k,l}(\phi) \equiv \max_{1 \leq i \leq d} p_{k,l}(\phi_i) \quad .$$

White Noise

- $\mathcal{S}$ with the above family of semi-norms is a Fréchet space. The topology remains the product topology of d copies of $\mathcal{S}(\mathbb{R})$. $\mathcal{S}$ is separable or equivalently second countable because $\mathcal{S}(\mathbb{R})$ is.
- Be $\mathcal{S}' \equiv \mathcal{S}'(\mathbb{R}, \mathbb{R}^d)$ the topological dual of $\mathcal{S}$ w.r.t. the Fréchet topology. $\mathcal{S}'$ is called the space of vector-valued tempered distributions on $\mathbb{R}$. $\mathcal{S}'$ with the $\sigma(\mathcal{S}', \mathcal{S})$- aka weak-* topology is a complete Hausdorff lctvs. It is separable but not second countable (not even first countable). Since $\mathcal{S}'$ is nonmetrizable, this is no contradiction.

White Noise

- $\mathcal{S}'$ does have an LF-space structure. The topology induced by that is the strong dual topology, which is finer than $\sigma(\mathcal{S}', \mathcal{S})$. $\mathcal{S}'$ with the strong dual topology is still a complete Hausdorff lctvs. The strong dual topology is nonmetrizable and non-separable. But it is nuclear, which $\sigma(\mathcal{S}', \mathcal{S})$ is not.
- The mapping $\mathcal{S} \ni \chi \longmapsto T_\chi \in \mathcal{S}'$ with

  $$\langle T_\chi \,|\, \phi \rangle \equiv \int_{\mathbb{R}} \chi(t)^\top \phi \, dt \quad \forall \phi \in \mathcal{S}$$

  induces a natural injective homomorphism of $\mathcal{S}$ into its dual space $\mathcal{S}'$. This embedding does not preserve the Fréchet topology on $\mathcal{S}$. $\mathcal{S}$ inherits the weak-* subspace topology of $\mathcal{S}'$.

White Noise

- $\mathcal{S}$ is sequentially dense in $\mathcal{S}'$, i.e. for every tempered distribution $T \in \mathcal{S}'$ there exists a sequence of test functions $\chi_n \in \mathcal{S}$ which converges to T in the weak-* topology :

$$\langle T \,|\, \phi \rangle = \lim_{n \to +\infty} \langle T_{\chi_n} \,|\, \phi \rangle \quad \forall \phi \in \mathcal{S}$$

- Every $\phi \in \mathcal{S}$ defines a unique evaluation functional $T \longmapsto \langle T \,|\, \phi \rangle$, which is an element of $\mathcal{S}''$, the topological dual of $\mathcal{S}'$ w.r.t. the weak-* topology. The induced natural embedding $\mathcal{S} \hookrightarrow \mathcal{S}''$ is in fact an isomorphism of vector spaces, but not of tvs. The Fréchet topology of $\mathcal{S}$ is not preserved. Note that this isomorphism factorizes over $\mathcal{S} \hookrightarrow \mathcal{S}'$.

White Noise

- $\mathcal{E}(\mathcal{S}') \equiv \mathcal{E}(\mathcal{S}', \mathcal{S})$ is the cylindrical σ-algebra on $\mathcal{S}'$, generated by the topological dual of $\mathcal{S}'$ which is $\mathcal{S}$. $\mathcal{E}(\mathcal{S}')$ is the smallest σ-algebra making all evaluation functionals $(\mathcal{E}(\mathcal{S}'), \mathcal{B}(\mathbb{R}))$-measurable.
- Consider an $(\mathcal{S}', \mathcal{E}(\mathcal{S}'))$-valued random element

$$W : (\Omega, \mathcal{A}, P) \to (\mathcal{S}', \mathcal{E}(\mathcal{S}')) \quad .$$

  W is not parameterized by an index set T like a usual stochastic process but by the function space $\mathcal{S}$. This is sometimes called a generalized stochastic process.

  Notation : $W = (W_\phi)_{\phi \in \mathcal{S}}$ with $W_\phi(\omega) \equiv \langle\, W(\omega) \,|\, \phi \,\rangle$ for a $\phi \in \mathcal{S}$ pointwise

White Noise

- A continuous-time white noise process with mean $\mu \in \mathbb{R}^d$ and positive semidefinite covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ is a random element $W \in \mathcal{L}^0(\Omega, \mathcal{A}, P; \mathcal{S}', \mathcal{E}(\mathcal{S}'))$, fullfilling the following two conditions:
- Expectation :

$$W_\phi \in L^1(\Omega, \mathcal{A}, P; \mathbb{R}^d) \ \wedge \ \mathbb{E}[W_\phi] = \mu^\top \int_{\mathbb{R}} \phi(t)\, dt \quad \forall \phi \in \mathcal{S}$$

- Covariance structure : $W_\phi \in L^2(\Omega, \mathcal{A}, P; \mathbb{R}^d) \wedge$

$$\mathrm{Cov}\,(W_\phi, W_\chi) = \mathbf{\Sigma}\,\langle\, \phi\,,\, \chi\,\rangle \quad \forall\, \phi, \chi \in \mathcal{S}$$

  $\langle\,,\,\rangle$ is the L^2 inner product

$$\langle\, \phi\,,\, \chi\,\rangle \equiv \int_{\mathbb{R}} \phi(t)^\top \chi(t)\, dt$$

  where

$$\mathcal{S} \subset L^2(\mathbb{R}, \mathbb{R}^d) \subset \mathcal{S}' \quad .$$

White Noise

- The covariance property implies that W is a generalized Gaussian process or equivalently that W_*P a Gaussian measure on $(\mathcal{S}', \mathcal{E}(\mathcal{S}'))$.
- W_*P is a Gaussian and therefore probability measure on the cylindrical σ-algebra $\mathcal{E}(\mathcal{S}')$. But is in general not even a measure on the Borel σ-algebra $\mathcal{B}(\mathcal{S}') \supseteq \mathcal{E}(\mathcal{S}')$. Fails with the countable additivity property.
- The characteristic function of W_*P is

$$\Phi(\chi) = \mathbb{E}[e^{i\,\langle W,\chi\rangle}] = \exp\left(i\,\langle \mu, \chi\rangle - \frac{1}{2}\langle \chi, \Sigma\chi\rangle\right) \quad .$$

- Usually Gaussian fields, like continuous-time white noise, are constructed by creating a Gaussian measure from a characteristic function via the Minlos theorem (see above).

White Noise

- W_ϕ is a Gaussian random variable with

  $$W_\phi \sim \mathcal{N}(\mu^\top \langle 1, \phi \rangle, \Sigma \langle \phi, \phi \rangle)$$

  and characteristic function

  $$\Phi_{W_\phi}(t) = \mathbb{E}[e^{i W_\phi^\top t}] = \exp\left(it^\top \mu - \frac{1}{2} t^\top \Sigma t \right) \quad .$$

- The W_ϕ s covariance fully determines their dependency structure. The inner product $\langle \, \phi \, , \, \chi \, \rangle$ acts as a weight for how much W_ϕ and W_χ are correlated: If ϕ and χ have disjoint supports, then $\langle \, \phi \, , \, \chi \, \rangle = 0$, making them uncorrelated. If ϕ and χ overlap significantly, then their inner product is nonzero, meaning W_ϕ and W_χ are correlated. This is how white noise behaves: uncorrelated at distinct times, but when tested against overlapping functions, correlation appears due to their functional overlap.

White Noise

- Continuous-time white noise is not uncorrelated for all distinct test functions $\phi, \chi \in \mathcal{S}$. Instead, its covariance depends on the inner product $\langle \phi , \chi \rangle$ which is nonzero unless ϕ and χ are orthogonal.
- Uncorrelatedness arises naturally for functions with disjoint supports, but not necessarily for all distinct functions.
- Note that continuous-time white noise is modelled as a generalized Gaussian process while discrete-time white noise need not be Gaussian. Discrete-time white noise is by definition identically distributed, continuous-time white noise is not.

White Noise

- Intuition: Continuous-time white noise W is the formal time derivative of Brownian motion:

  $$W_t = \frac{d}{dt} B_t$$

  W is everywhere continuous a.s. but nowhere-differentiable a.s. so we understand derivatives in the distributional sense.
- Sampling W at discrete points $t_n \equiv n\Delta$ (aka testing against Dirac delta functions δ_{t_n}) yields the discrete-time white noise process (W_n) with the same variance. Consider a sequence of test functions $\phi_n \in \mathcal{S}$, each ϕ_n centered in t_n, converging against the delta distribution $\delta \in \mathcal{S}'$ in the weak-* topology.

Markov Processes

- Consider a stochastic process, adapted to a filtration $\mathcal{F}$, $(\Omega, \mathcal{A}, P; X = (X_t, \mathcal{F}_t)_{t \in \mathbb{R}_+}; \mathsf{E}, \mathcal{E})$. X is called an $\mathcal{F}$-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 $f \in \mathcal{B}_b(\mathsf{E}, \mathcal{E})$.

Markov Processes

- Any $\mathcal{F}$-Markov process is always an $\mathcal{F}^X$-Markov process. $\mathcal{F}^X$-Markov processes are simply called Markov processes.
- Intuition: A Markov process is a process without memory. $\sigma(X_t)$ represents all the information there is at time t about the process. $\sigma(X_t, (X_s)_{s<t})$ does not provide any more information.

Markov Processes

- Every family of independent $(\mathsf{E}, \mathcal{E})$-valued random variables $X = (X_t)_{t \in T}$ represents a Markov process since $\sigma(X_t)$ is independent of $\mathcal{F}_s^X$ for $s < t$.
- We associate to X a family of Markov kernels $K = (K_{t,s})_{\{t,s \in \mathbb{R}_+ | t \geq s\}}$ with $K_{t,s} : \mathsf{E} \times \mathcal{E} \longrightarrow [0, 1]$ defined by

$$K_{t,s}(\epsilon, E) \equiv \mathbb{P}[X_t \in E \mid X_s = \epsilon] \ .$$

- The $K_{t,s}$ are called the transition probability kernels of the Markov process X.
- Note that by definition $K_{t,t}(\epsilon, E) = 1_{\mathsf{E}}(\epsilon, E)$, the unit kernel.

Markov Processes

- Intuition: $K_{t,s}(\epsilon, E)$ is the probability of the process to have a state in E at time t given it was in state ϵ at time s.
- If we consider states along a sample path $(X_t(\omega))_{t\in\mathbb{R}_+}$, $\omega \in \Omega$, we find that

  $$K_{t,s}(X_s(\omega), E) = \mathbb{P}[X_t \in E \mid X_s](\omega) \quad .$$

  $K_{t,s}(X_s(\omega), E)$ is a realization of the transition kernel where the current state is $X_s(\omega)$.
- If E is a Polish space, an rcd of of X_t given X_s exists and

  $$K_{t,s}(X_s(\omega), E) = K_{X_t|X_s}(\omega, E) \quad P\text{-a.e.}$$

  i.e. the transition probability kernel evaluated at the state $X_s(\omega)$ is an rcd of X_t given X_s.

Markov Processes

- The Markov property from above implies that

$$K_{t,s}(X_s(\omega), E) = K_{X_t|X_s}(\omega, E) = K_{X_t|\mathcal{F}_s}(\omega, E) \quad P\text{-a.e.} \quad .$$

- The $K_{t,s}$ obey the Chapman-Kolmogorov equations

$$K_{u,s} = K_{t,s}\ K_{u,t} \quad (0 \leq s \leq t \leq u < +\infty)$$

- Note that because the $K_{t,s}(\epsilon, \cdot)$ are measures by definition, a Markov kernel $K_{t,s}$ is already an integral kernel :

$$K_{t,s}(\epsilon, E) = \int_E K_{t,s}(\epsilon, d\epsilon') \quad \epsilon \in \mathsf{E}, E \in \mathcal{E}$$

Markov Processes

- Be $\lambda \in \mathcal{M}^+(\mathsf{E}, \mathcal{E})$ a reference measure on $\mathcal{E}$. If for the transition measures $K_{t,s}(\epsilon, \cdot)$ it does hold that $K_{t,s}(\epsilon, \cdot) \ll \lambda$ for all $\epsilon \in \mathsf{E}, t, s \in \mathbb{R}_+$, then the corresponding R-N derivatives

$$k_{t,s}(\epsilon, \_) = \frac{dK_{t,s}(\epsilon, \cdot)}{d\lambda} \quad \in L^1_+(\mathsf{E}, \mathcal{E}, \lambda; \mathbb{R})$$

do exist and are the kernel density functions of the $K_{t,s}$:

$$K_{t,s}(\epsilon, E) = \int_E \lambda(d\epsilon')\, k_{t,s}(\epsilon, \epsilon')$$

Markov Processes

- $K_{t,s}(\epsilon, \cdot) \ll \lambda \;\; \forall \epsilon \in \mathsf{E}, t, s \in \mathbb{R}_+$ is not in general true. Even for a $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$-valued Markov process and the Lebesgue-measure on $\mathcal{B}(\mathbb{R}^d)$, additional regularity assumptions are required.
- If X is translation-invariant (for example because it has stationary increments), then it has convolution kernels i.e. there exists a family of pdfs $(p_{t,s})_{\{t,s \in \mathbb{R}_+ | t \geq s\}}$ with $p_{t,s} \in L^1(\mathsf{E}, \mathcal{E}, \lambda; \mathbb{R})$ s.t.

  $$K_{t,s}(\epsilon, E) = \int_E \lambda(d\epsilon')\, p_{t,s}(\epsilon' - \epsilon)$$

  Obviously $K_{t,s}(\epsilon, \cdot) \ll \lambda$ does hold here and $k_{t,s}(\epsilon, \epsilon') = p_{t,s}(\epsilon' - \epsilon)$.

Markov Processes

- Jump processes with absolutely continuous Lévy measures and diffusions with smooth coefficients (elliptic SDEs) usually do have $K_{t,s}(\epsilon, \cdot) \ll \lambda$.
- If $\lambda \in \mathcal{M}^+(\mathsf{E}, \mathcal{E})$ is a reference measure on $\mathcal{E}$ and $P_{(X_t, X_s)} \ll \lambda \otimes \lambda$ does hold for all $t, s \in \mathbb{R}_+$, then

  $$K_{t,s}(\epsilon, E) = \mathbb{P}[X_t \in E \mid X_s = \epsilon] = \int_E \lambda(d\epsilon')\, p_{X_t|X_s}(\epsilon \mid \epsilon')$$

  that is, $p_{X_t|X_s}$ is the kernel density function of $K_{t,s}$.
- $P_{(X_t, X_s)} \ll \lambda \otimes \lambda$ does not hold for singular initial distributions of X, δ_ϵ for example. It does, if $P_{X_0} \ll \lambda$ and $K_{t,s}(\epsilon, \cdot) \ll \lambda$ for all $t, s \in \mathbb{R}_+$

Markov Processes

- If X is time-homogeneous :

$$K_{t,s}(\epsilon, E) = \mathbb{P}[X_t \in E \mid X_s = \epsilon] = \mathbb{P}[X_{t-s} \in E \mid X_0 = \epsilon] = K_{t-s,0}(\epsilon, E)$$

- We identify for $t = v - u \quad (0 \leq u \leq v \leq +\infty)$

$$K_t(\epsilon, E) \equiv K_{t,0}(\epsilon, E) = K_{v-u,0}(\epsilon, E) = K_{v,u}(\epsilon, E)$$

- The kernels

$$K_t(\epsilon, E) \equiv \mathbb{P}[X_t \in E \mid X_0 = \epsilon]$$

form a (commutative) Markov semigroup $K = (K_t)_{t \in \mathbb{R}_+}$.
Define for this Markov semigroup

$$K_{t,s}(\epsilon, E) \equiv K_{t-s}(\epsilon, E)$$

Markov Processes

- $K = (K_t)_{t \in \mathbb{R}_+}$ satisfies the C-K-equations

$$K_{s+t} = K_s \; K_t \quad .$$

- Note that $K_0(\epsilon, E) = 1_{\mathsf{E}}(\epsilon, E)$ with $K_0(\epsilon, \cdot) = \delta_\epsilon$.
- Every process can be turned into a time-homogeneous one by enlarging its state space.
- We introduce some notational shortcuts:

$$\mathbb{E}^{s,\epsilon}[f(X_t)] \equiv \mathbb{E}[f(X_t) \mid X_s = \epsilon] \;\; , \; \mathbb{E}^{\epsilon}[f(X_t)] \equiv \mathbb{E}^{0,\epsilon}[f(X_t)]$$

$$\mathbb{P}^{s,\epsilon}[X_t \in E] \equiv \mathbb{E}^{s,\epsilon}[1_E(X_t)] = \mathbb{P}[X_t \in E \mid X_s = \epsilon]$$

$$\mathbb{P}^{\epsilon}[X_t \in E] \equiv \mathbb{P}^{0,\epsilon}[X_t \in E] \qquad , \; s < t \; , \; \epsilon \in \mathsf{E}$$

Markov Processes

- From here on, we assume all processes to be time-homogeneous and understand all Markov semigroups to follow the C-K-equations.

Markov Processes

- K induces two types of integral operators (see chapter on kernels above). The left operation $(K_t\,\cdot) : \mathcal{B}_b(\mathsf{E},\mathcal{E}) \to \mathcal{B}_b(\mathsf{E},\mathcal{E})$:

$$(K_t\,f) = \int_{\mathsf{E}} K_t(\,\cdot\,, d\epsilon')\,f\,(\epsilon') \quad \text{for } f \in \mathcal{B}_b(\mathsf{E},\mathcal{E}) \quad .$$

- Note that

$$(K_t\,f)(\epsilon) = \mathbb{E}^{\epsilon}[f\,(X_t)]$$

since

$$(K_t\,1_E)(\epsilon) = K_t(\epsilon, E) = \mathbb{P}^{\epsilon}[X_t \in E] \quad ,$$

and that

$$(K_0\,f)(\epsilon) = f(\epsilon) \quad .$$

Markov Processes

- Intuition: Given that the process is at state ϵ now (at time 0), what is the expected value of test function f at time t ? Propagating the function/'observable' backward to time 0, we are evaluating the expectation now, based on future behavior. The left operation $(K_t \cdot)$ is called the backward operator. (-> "Heisenberg Picture")
- Recall that $\mathcal{B}_b(\mathsf{E}, \mathcal{E})$ with the supremum norm is a Banach space. It holds that $\|K_t f\| \leq \|f\|$ for $f \in \mathcal{B}_b(\mathsf{E}, \mathcal{E})$ i.e. that $(K_t \cdot)$ is a contractive and therefore continuous operator.

Markov Processes

- The right operation $(\cdot\, K_t) : \mathcal{M}_f(\mathsf{E}, \mathcal{E}) \to \mathcal{M}_f(\mathsf{E}, \mathcal{E})$:

$$(\nu\, K_t) = \int_{\mathsf{E}} \nu(d\epsilon')\, K_t(\epsilon', \cdot) \quad \text{for } \nu \in \mathcal{M}_f(\mathsf{E}, \mathcal{E}) \quad .$$

- Note that $(\nu\, K_0) = \nu$.
- Operation on distributions :

$$\begin{aligned}(P_{X_s}\, K_t) &= \int_{\mathsf{E}} P_{X_s}(d\epsilon')\, K_t(\epsilon', \cdot) \\ &= \int_{\mathsf{E}} P_{X_s}(d\epsilon')\, \mathbb{P}^{s,\epsilon'}[X_{s+t} \in \cdot] \\ &= \mathbb{E}[\, \mathbb{P}[X_{s+t} \in \cdot \mid X_s]\,] = P\,(X_{s+t} \in \cdot) \\ &= P_{X_{s+t}}\end{aligned}$$

Markov Processes

- If P_{X_0} is the initial distribution of X, $P_{X_t} = (P_{X_0} K_t)$.
- If $P_{X_0} = \delta_\epsilon$ i.e. $X_0 = \epsilon$:

$$P_{X_t} = (P_{X_0} K_t) = (\delta_\epsilon K_t) = K_t(\epsilon, \cdot) = \mathbb{P}^\epsilon[X_t \in \cdot]$$

- Intuition: The right operation $(\cdot K_t)$ is called the forward operator on the distributions. It drives the evolution of the distributions of the process forward in time. (-> ”Schrödinger picture”).

Markov Processes

- Be $\lambda \in \mathcal{M}_\sigma^+(\mathsf{E}, \mathcal{E})$ a σ-finite reference measure on $(\mathsf{E}, \mathcal{E})$ and assume the X_t all to have pdfs $p_t \in L_+^1(\mathsf{E}, \mathcal{E}, \lambda)$. The above operations reduce to $(\cdot\, K_t) : \mathcal{M}_f(\mathsf{E}, \mathcal{E}, \lambda) \to \mathcal{M}_f(\mathsf{E}, \mathcal{E}, \lambda)$ and $(K_t \,\cdot) : L^\infty(\mathsf{E}, \mathcal{E}, \lambda) \to L^\infty(\mathsf{E}, \mathcal{E}, \lambda)$.
- The Radon-Nikodym isomorphism $\mathcal{M}_f(\mathsf{E}, \mathcal{E}, \lambda) = L^1(\mathsf{E}, \mathcal{E}, \lambda)$ allows to lift the forward operator to an operator $L_t : L^1(\mathsf{E}, \mathcal{E}, \lambda) \to L^1(\mathsf{E}, \mathcal{E}, \lambda)$. Explicitly :

$$L_t\, p \;\equiv\; \frac{d((p\,\lambda)\,K_t)}{d\lambda} \qquad \text{for } p \in L^1(\mathsf{E}, \mathcal{E}, \lambda)$$

Markov Processes

- L_t is the adjoint of the backward operator $(K_t\,\cdot)$ under the duality pairing of $L^1(\mathsf{E},\mathcal{E},\lambda) = L^\infty(\mathsf{E},\mathcal{E},\lambda)'$ i.e.

  $$\langle\, K_t f \,|\, p \,\rangle = \langle\, f \,|\, L_t\, p \,\rangle \quad .$$

  (The dualilty requires λ to be σ-finite. In general $L^1(\mathsf{E},\mathcal{E},\lambda) \subset L^\infty(\mathsf{E},\mathcal{E},\lambda)'$)
- Notation : From now on we simply write $K_t f$ for the backward operator, operating on a test function f, and $K_t^* p$ or $K_t^* \nu$ for the forward operator, operating on a density function p or a distribution ν respectively.
- $K^* \equiv (K_t^*)_{t\in\mathbb{R}_+}$ is called the adjoint Markov semigroup of a Markov process.

Markov Processes

- Let the K_t have kernel densities $k_t \in L^0_+(\mathsf{E} \times \mathsf{E}, \mathcal{E} \otimes \mathcal{E}, \lambda \otimes \lambda)$. It then holds that

  $$(K_t f) = \int_{\mathsf{E}} \lambda(d\epsilon')\, k_t(\cdot, \epsilon')\, f(\epsilon') \in L^\infty(\mathsf{E}, \mathcal{E}, \lambda) \quad \text{for } f \in L^\infty(\mathsf{E}, \mathcal{E}, \lambda)$$

  and that

  $$(K_t^* p) = \int_{\mathsf{E}} \lambda(d\epsilon')\, k_t(\epsilon', \cdot)\, p(\epsilon') \in L^1(\mathsf{E}, \mathcal{E}, \lambda) \quad \text{for } p \in L^1(\mathsf{E}, \mathcal{E}, \lambda)$$

- If the k_t are symmetric, these operators coincide pointwise on $L^1(\mathsf{E}, \mathcal{E}, \lambda)$ and on $L^\infty(\mathsf{E}, \mathcal{E}, \lambda)$ i.e. $(K_t f)(\epsilon) = (K_t^* f)(\epsilon)$ for all $\epsilon \in \mathsf{E}$ if $f \in L^1(\mathsf{E}, \mathcal{E}, \lambda)$ or $f \in L^\infty(\mathsf{E}, \mathcal{E}, \lambda)$.

Markov Processes

- If $k_t \in L^\infty_+(\mathsf{E} \times \mathsf{E}, \mathcal{E} \otimes \mathcal{E}, \lambda \otimes \lambda)$, or more generally if

  $$\sup_{\epsilon \in \mathsf{E}} \int_{\mathsf{E}} \lambda(d\epsilon')\, |k_t(\epsilon, \epsilon')| < +\infty$$

  and

  $$\sup_{\epsilon \in \mathsf{E}} \int_{\mathsf{E}} \lambda(d\epsilon')\, |k_t(\epsilon', \epsilon)| < +\infty \quad ,$$

  then both K_t and K_t^* define bounded linear operators on $L^p(\mathsf{E}, \mathcal{E}, \lambda)$ for every $p \in [1, +\infty]$ by a generalized Schur's test.

  In particular, K_t and K_t^* both map $L^p(\mathsf{E}, \mathcal{E}, \lambda)$ to $L^p(\mathsf{E}, \mathcal{E}, \lambda)$ for every $p \in [1, +\infty]$ and as a consequence coincide pointwise on each of these spaces.

Markov Processes

- Under the above conditions, $K_t : L^2(\mathsf{E}, \mathcal{E}, \lambda) \to L^2(\mathsf{E}, \mathcal{E}, \lambda)$ and $K_t^* : L^2(\mathsf{E}, \mathcal{E}, \lambda) \to L^2(\mathsf{E}, \mathcal{E}, \lambda)$ are bounded linear operators that coincide pointwise. On a Hilbert space, this implies they are identical i.e. that K_t is self-adjoint on $L^2(\mathsf{E}, \mathcal{E}, \lambda)$.

Markov Processes

- Consider an abstract Markov semigroup of kernels $K = (K_t)_{t\in\mathbb{R}_+}$ and a probability measure $\nu \in \mathcal{M}_1(\mathsf{E},\mathcal{E})$. Together they induce a projective family of probability measures $P^{K,\nu} = (P_J^{K,\nu})_{J\in\mathcal{F}(\mathbb{R}_+)}$ on $(\mathsf{E}^{\mathbb{R}_+},\mathcal{E}^{\mathbb{R}_+})$, defined by

$$P_J^{K,\nu} = \left(K_{t_0}^{*}\nu \otimes \bigotimes_{i=0}^{|J|-2} K_{t_{i+1}-t_i}\right) \quad \in \mathcal{M}_1(\mathsf{E}^{\,|J|},\mathcal{E}^{\,\otimes|J|})$$

  with $J \equiv \{t_0,\ldots,t_n\} \in \mathcal{F}(\mathbb{R}_+)$ and $t_0 \le t_1 \le \cdots \le t_n$.
- If $\nu = \delta_\epsilon$ for $\epsilon \in \mathsf{E}$ (denoting P_J^{K,δ_ϵ} by $P_J^{K,\epsilon}$) :

$$P_J^{K,\epsilon} = \left(K_{t_0}^{*}\delta_\epsilon \otimes \bigotimes_{i=0}^{|J|-2} K_{t_{i+1}-t_i}\right) = \left(K_{t_0} \otimes \bigotimes_{i=0}^{|J|-2} K_{t_{i+1}-t_i}\right)(\epsilon,\cdot)$$

Markov Processes

- If 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.
- The above process is the canonical process associated to K and ν :

$$X^{K,\nu} \equiv (\mathsf{E}^{\mathbb{R}_+}, \mathcal{B}(\mathsf{E})^{\mathbb{R}_+}, P^{K,\nu}; (X_t^{K,\nu})_{t\in\mathbb{R}^+}; \mathsf{E}^{\mathbb{R}_+}, \mathcal{B}(\mathsf{E})^{\mathbb{R}_+})$$

where $P^{K,\nu} = \varprojlim_{J\in\mathcal{F}(\mathbb{R}_+)} P_J^{K,\nu}$ and $X^{K,\nu}(\epsilon) \equiv \epsilon$. Note that

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

Markov Processes

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

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

- $P^{K,\nu}(X_t^{K,\nu} \in E) = \mathbb{E}^{K,\nu}[1_E \circ X_t^{K,\nu}]$ ($\mathbb{E}^{K,\nu}$ denotes the expectation w.r.t. the measure $P^{K,\nu}$)
- From

$$\mathbb{E}^{K,\nu}[1_E \circ X_t^{K,\nu}] = \int_{\mathsf{E}} \nu(d\epsilon) K_t(\epsilon, E) = \int_{\mathsf{E}} \nu(d\epsilon)(K_t\, 1_E)(\epsilon)$$

we can generalize (usual simple function machinery) to

$$\mathbb{E}^{K,\nu}[f(X_t^{K,\nu})] = \int_{\mathsf{E}} \nu(d\epsilon)\, (K_t f)(\epsilon)$$

where $f \in \mathcal{L}^0_+(\mathsf{E}, \mathcal{E}; \mathbb{R})$.

Markov Processes

- If $\nu = \delta_\epsilon$:

$$\mathbb{E}^{K,\epsilon}[f(X_t^{K,\epsilon})] = (K_t f)(\epsilon) = \mathbb{E}[f(X_t^{K,\epsilon}) \mid X_0^{K,\epsilon} = \epsilon]$$

- If $t = 0$:

$$P^{K,\nu}(X_0^{K,\nu} \in E) = \int_{\mathsf{E}} \nu(d\epsilon)\, K_0(\epsilon, E)$$

  If K is normal, i.e. $K_0 = 1_{\mathsf{E}}$:

$$P^{K,\nu}_{X_0^{K,\nu}}(E) = P^{K,\nu}(X_0^{K,\nu} \in E) = \nu(E)$$

Markov Processes

- It can be shown, that the $(\mathsf{E},\mathcal{E})$-valued stochastic process $X \equiv X^{K,\nu}$, induced by a Markov semigroup $K = (K_t)_{t\in\mathbb{R}_+}$ and a (initial) probability measure $\nu \in \mathcal{M}_1(\mathsf{E},\mathcal{E})$, is in fact a Markov process w.r.t. to its natural filtration $(\mathcal{F}_t^X)_{t\in\mathbb{R}_+}$.
It holds P-a.e. that for all $E \in \mathcal{E}$ and $0 \leq s \leq t \in \mathbb{R}_+$

$$\mathbb{P}[X_t \in B \mid \mathcal{F}_s^X](\omega) = K_{t-s}(X_s(\omega), E) \quad .$$

- If E is Polish and $\mathcal{E} \equiv \mathcal{B}(\mathsf{E})$, the Markov process $X^{K,\nu}$ is uniquely determined by K and ν.

Markov Processes

- Recall that

$$P^{K,\epsilon}(X_t^{K,\nu} \in E) = K_t(\epsilon, E) = \mathbb{P}^{\epsilon}(X_t^{K,\nu} \in E)$$

and that

$$P^{K,\nu}(X_t^{K,\nu} \in E) = \int_{\mathsf{E}} \nu(d\epsilon)\, K_t(\epsilon, E) \quad .$$

We will denote the distribution of $X^{K,\epsilon}$ by $\mathbb{P}^{\epsilon}$ and that of X^{ν} for a process with initial distribution ν by $\mathbb{P}^{\nu}$ (Whenever possible, the superscript for the Markov semigroup is ommitted for readability). Consequently is

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

Markov Processes

- Expectation w.r.t. $\mathbb{P}^\nu$ will be denoted by $\mathbb{E}^\nu$. Since

$$\begin{aligned}\mathbb{E}^\epsilon[f(X_t^{K,\nu})] &= \int_{\mathsf{E}^{\mathbb{R}_+}} f(X_t^{K,\nu}(\epsilon'))\,\mathbb{P}^\epsilon(d\epsilon') \\ &= \int_{\mathsf{E}} f(\epsilon')\,K_t(\epsilon, d\epsilon') = (K_t f)(\epsilon)\end{aligned}$$

  by the transformation formula, we can derive

$$\mathbb{E}^\nu[f(X_t^{K,\nu})] = \int_{\mathsf{E}} \nu(d\epsilon)\,(K_t f)(\epsilon) \quad .$$

- Note that we have justified our notational shortcuts from above.

Markov Processes

- A measure $\pi \in \mathcal{M}(\mathsf{E}, \mathcal{E})$ is called invariant w.r.t. to a Markov semigroup $K \equiv (K_t)_{t \in \mathbb{R}_+}$ if $K_t^* \pi = \pi$ for all $t \in \mathbb{R}_+$.
- The Markov process X^ν is said to have a stationary distribution π if $\pi = P_{X_s}$ for some $s \in \mathbb{R}_+$ and π is invariant w.r.t. to the Markov semigroup of X^ν.
- By definition, a stationary Markov process X^ν does have a stationary distribution, the initial distribution ν.
- If the process starts with the stationary distribution, it will remain in that distribution at all future times.

Markov Processes

- If the process is started from any other initial distribution, it may converge to the stationary distribution over time or not. If it is also ergodic (i.e. irreducible and aperiodic), it will converge in distribution to the stationary distribution, the unique stationary distribution of the process.
- Intuition: A stationary distribution represents a long-term equilibrium state of the process.
- Markov Chain Monte Carlo (MCMC) methods: Want to sample from a pdf $\pi(x)$. Produce an ergodic Markov Chain whose stationary distribution π has pdf $\pi(x)$.

Markov Processes

- A process is called time-reversible if $P_{X_K} = P_{X_{(s-K)}}$ for any fixed $s \in \mathbb{R}_+$ and all $K \in \mathcal{F}(\mathbb{R}_+)$.
- Be K a Markov kernel with kernel densities k_t and let the associated Markov process X have pdfs p_t. Then X is time-reversible if there exists a stationary distribution π such that the detailed balance condition holds:

$$\pi(d\epsilon)\, K_t(\epsilon, d\epsilon') = \pi(d\epsilon')\, K_t(\epsilon', d\epsilon)$$

If $\pi \ll \lambda$ with pdf $\pi(\epsilon)$:

$$\pi(\epsilon)\, k_t(\epsilon, \epsilon') = \pi(\epsilon')\, k_t(\epsilon', \epsilon)$$

Markov Processes

- The detailed balance condition ensures that the process run backward in time (under stationarity) has the same finite-dimensional distributions as when run forward.
- Symmetric kernels are neither sufficient nor necessary for time-reversibility. With nonsymmetric k_t, a process can still be time-reversible under certain conditions.
- Metropolis-Hastings, an MCMC method, often uses nonsymmetric transition probabilities but still satisfies detailed balance, ensuring time-reversibility.

Markov Processes

- Every $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$-valued stochastic process with stationary and independent increments is a Markov process, Brownian motion for example. (Do we really need convolution semigroups to prove this ?)
- Varying the initial distribution while keeping the transition kernels aka the Markov semigroup leads to the notion of a universal Markov process.

Markov Processes

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

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

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

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

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

Markov Processes

- A normal Markov semigroup of kernels $K = (K_t)_{t \in \mathbb{R}_+}$ on a Polish space $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$ obviously induces 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.

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

  This is the canonical construction from above.
- Vice versa does a universal Markov process $(\Omega, \mathcal{A}, (P^\epsilon)_{\epsilon \in \mathsf{E}}; (X_t)_{t \in \mathbb{R}_+}; \mathsf{E}, \mathcal{B}(\mathsf{E}))$ with a Polish state space $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$ induce a normal Markov semigroup of kernels on $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$ via

  $$K_t(\epsilon, E) \equiv P^\epsilon(X_t \in E) \quad .$$
- 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 E Polish and initial distribution $\nu = P_{X_0}$. It then holds that the family of finite-dimensional distributions of X is comming from a normal Markov semigroup $(K_t)_{t\in\mathbb{R}_+}$ and the 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 $(\Omega\,', \mathcal{A}\,', P^{\,\nu}; (X_t')_{t\in\mathbb{R}_+}; \mathsf{E}, \mathcal{B}(\mathsf{E}))$ with

$$P^{\,\nu}(A) \equiv \int_{\mathsf{E}} \nu(d\epsilon)\, P^{\,\epsilon}(A) \quad \text{for } A \in \mathcal{A}'$$

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

Markov Processes

- Intuition: Stochastic processes generated by normal Markov semigroups from an initial distribution are (modulo equivalency) the stochastic processes created by averaging that distribution against the family of probability measures of a universal Markov process.

Markov Processes

- If $\mathcal{C} \equiv \mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ and $x \in \mathbb{R}^d$, then is $(\mathcal{C}, \mathcal{B}(\mathcal{C}), Q^x; (X_t)_{t \in \mathbb{R}_+})$ the universal Markov process of Brownian motion (see above). $K_t(x, B) = Q^x(X_t \in B)$ for $x \in \mathbb{R}^d$ and $B \in \mathcal{B}(\mathbb{R}^d)$. Explicit description :

  $$K_t(x, B) = \int_B \lambda^d(dy)\, k_t(x, y)$$

  with kernel density function

  $$k_t(x, y) = \frac{1}{(2\pi t)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2t}(y - x - \mu t)^T \Sigma^{-1} (y - x - \mu t) \right)$$

  Parameters: $\mu \in \mathbb{R}^d$, the drift vector, and $\Sigma \in \mathbb{R}^{d \times d}$ the covariance matrix.

Markov Processes

- Note that $k_t(x, y) = p_{X_t}(y - x)$ applies here since X has stationary increments and therefore is translation-invariant.

Markov Processes

- The universal Ornstein-Uhlenbeck (OU) process $(\mathcal{C}, \mathcal{B}(\mathcal{C}), P^x; (X_t)_{t \in \mathbb{R}_+})$ is not translation-invariant. It has transition density

$$k_t(x, y) = \frac{1}{(2\pi)^{d/2} |\Sigma_t|^{1/2}} \exp\left(-\frac{1}{2} \left(y - e^{-\theta t} x - \mu(1 - e^{-\theta t})\right)^T \right. \\ \left. \Sigma_t^{-1} (y - e^{-\theta t} x - \mu(1 - e^{-\theta t})) \right) \quad .$$

Parameters: $\mu \in \mathbb{R}^d$ is the long-term mean, $\theta \in \mathbb{R}^{d \times d}$ is the rate matrix, Σ_t is the covariance matrix at time t, given by:

$$\Sigma_t = \int_0^t e^{-\theta s} \, Q \, e^{-\theta^T s} ds,$$

where Q is the diffusion matrix.

Markov Processes

- The OU process is (modulo linear transformations in space and time variables) the only nontrivial process that is stationary, Gaussian and Markov.
- The OU process is mean reverting, that is it tends to drift towards its mean function $t \mapsto \mathbb{E}[X_t]$ over time.

Markov Chains

- A Markov process $(\Omega, \mathcal{A}, P; X = (X_n, \mathcal{F}_n)_{t \in N}; \mathsf{E}, \mathcal{E})$ with an at most countable index set N is called a Markov chain.
- The C-K equation $K_n = K_{n-1} K_1$ results in an inductively generated Markov semigroup $(K_n)_{n \in N}$.
- For every $E \in \mathcal{B}(\mathsf{E})$ and $n \in N$ there is

$$K_{n+1}(X_n, E) = \mathbb{P}[X_{n+1} \in E \mid \mathcal{F}_n] \ .$$

Markov Chains

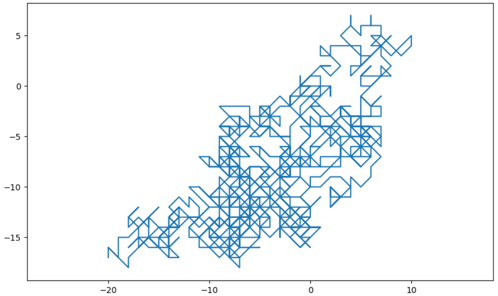
- Continuous-time Markov processes produce Markov chains :
- Sampling : Consider a continuous-time Markov process $X = (X_t)_{t \in T}$ with a discrete sequence $\{t_0, t_1, t_2, \dots\} \subset T$ s.t. $t_0 < t_1 < t_2 < \cdots$. Then is $Y = (Y_n)_{n \in N}$, defined by $Y_n \equiv X_{t_n}$, a discrete-time Markov process aka a Markov chain.
- Discretization :
- The Euler-Maruyama discretization of a Brownian motion and of a OU process generates Markov chains, a random walk and a discrete-time AR(1) process respectively.
- Milstein discretization (stochastic volatility scenario).

Markov Chains

- The standard random walk process on the integers $\mathbb{Z}$ is a Markov chain $(X_n)_{n\in\mathbb{N}}$ where

$$X_n \equiv \sum_{j=1}^{n} Z_j \quad , \quad Z_j = \begin{cases} +1 \\ -1 \end{cases} \quad , \; Z_j \sim \mathrm{Ber}(0.5)$$

- The 2-dimensional random walk ('drunkard's walk'):



Markov Chains

- Random Walk with Gaussian Noise :

$$X_n = X_{n-1} + \epsilon_n \quad , \epsilon_n \sim \mathcal{N}(\mu, \sigma^2)$$

- AR(1) Process :

$$X_n = \phi\, X_{n-1} + \epsilon_n \quad , (\epsilon_n) \sim WN(\mu, \sigma^2)$$

$$\begin{aligned} K_n(x, B) &= \mathbb{P}\left[X_n \in B \mid X_{n-1} = x\right] = P\left(\phi\, x + \epsilon_n \in B\right) \\ &= P(\epsilon_n \in B - \phi\, x) = P_{\epsilon_n}(B - \phi\, x) \end{aligned}$$

If $(\epsilon_n) \sim GWN(\mu, \sigma^2)$, then

$$K_n(x, B) = \mathcal{N}(\phi\, x + \mu, \sigma^2) \quad .$$

Markov Chains

- Markov chains induce continuous time Markov processes :
- Scaling limit :
- Brownian motion is the scaling limit of a Markov chain, the random walk.
- The OU process is a scaling limit (–> Kac process).
- In fact, every continuous-time Markov process is the scaling limit of a Markov chain, but not every scaling of a Markov chain leads to a well-defined continuous-time Markov process.

Martingales

- Nearly all Markov processes are semi-martingales. Semi-martingales are the largest class of processes, satisfying stochastic differential equations. Most reasonable diffusion processes are semi-martingales (regular SDE coefficients).

Feller Processes

- Be $(\mathsf{E}, \mathcal{E})$ a measurable space where E is a locally compact, separable metric space equipped with its Borel σ-algebra $\mathcal{E} \equiv \mathcal{B}(\mathsf{E})$. Remember that $\mathcal{C}_0(\mathsf{E})$, the space of real-valued continuous functions on E vanishing at infinity, is a Banach space w.r.t. the $\| \cdot \|_\infty$ norm.
- A Feller (kernel) semigroup is a (time-homogenous) Markov (kernel) semigroup $K \equiv (K_t)_{t \in \mathbb{R}_+}$ with $(K_t \, \cdot) : \mathcal{C}_0(\mathsf{E}) \to \mathcal{C}_0(\mathsf{E})$ and strong continuity property for $t = 0$:
$$\lim_{t \to 0} \|K_t f - f\|_\infty = 0 \quad \text{for all } f \in \mathcal{C}_0(\mathsf{E})$$

Feller Processes

- A Feller semigroup is a strong continuous semigroup aka a $\mathcal{C}_0$-semigroup. Because of the Markov property, it is a contraction $\mathcal{C}_0$-semigroup.
- A Feller process is a Markov process whose transition semigroup is a Feller semigroup.
- Every Feller process has a modification where all sample paths are càdlàg.

Kolmogorov equations

- Consider an $(\mathsf{E}, \mathcal{B}(\mathsf{E}))$-valued Feller process $X \equiv (X_t)_{t\in\mathbb{R}_+}$ with transition semigroup $K \equiv (K_t)_{t\in\mathbb{R}_+}$ operating on $\mathcal{C}_0(\mathsf{E})$. The infinitesimal generator of X (also of K) is the operator $A : D(A) \to \mathcal{C}_0(\mathsf{E})$, defined by (note that $K_0 f = f$)

$$A f \equiv \lim_{t\to 0} \frac{K_t f - f}{t} \quad \text{for} \;\; t \in \mathbb{R}_+, \; f \in \mathcal{C}_0(\mathsf{E})$$

with domain

$$D(A) \equiv \left\{ f \in \mathcal{C}_0(\mathsf{E}) \mid \lim_{t\to 0} \frac{K_t f - f}{t} \;\text{ exists } \right\} \subset \mathcal{C}_0(\mathsf{E}) \quad .$$

Kolmogorov equations

- $D(A) \subset \mathcal{C}_0(\mathsf{E})$ is a linear subspace.
- A is a closed operator and densely defined i.e. $\overline{D(A)}^{\|\cdot\|_\infty} = \mathcal{C}_0(\mathsf{E})$.
- Notational shorthand : $K_t = e^{tA}$ on $D(A)$. (This is formal. The operator Taylor series does not converge under these conditions.)
- The Hille-Yosida theorem gives conditions under which a closed and densely defined linear operator on a Banach space generates a contraction $\mathcal{C}_0$-semigroup. Used to construct Feller processes for given differential operators (see below).

Kolmogorov equations

- For a function $f \in \mathcal{C}_0(\mathsf{E})$, put

$$u(\epsilon, t, f) \equiv K_t f(\epsilon) = \mathbb{E}[f(X_t) \mid X_0 = \epsilon] \ .$$

  As a function of t is $u(\epsilon, \cdot, f) \in L^0(\mathbb{R}_+, \mathcal{C}_0(\mathsf{E}))$. The Hille-Yosida theorem implies that $u(\epsilon, \cdot, f) \in L^0(\mathbb{R}_+, D(A))$ if $f \in D(A)$.
- Therefore calculating the time-derivative with $f \in D(A)$ results in :

$$\frac{\partial}{\partial t} u(\cdot, t, f) = \frac{d}{dt}\,(K_t f) = A\, K_t\, f = A\, u(\cdot, t, f)$$

Kolmogorov equations

- The Kolmogorov backward equation :

$$\frac{\partial}{\partial t} u(\cdot, t, f) = A\, u(\cdot, t, f) \quad , \quad u(\cdot, 0, f) = f$$

- Relax the assumptions and start with a Markov process whose Markov kernel semigroup shows strong continuity. If the associated operator semigroup maps an arbitrary Banach space $\mathcal{S}$, not necessarily $\mathcal{C}_0(\mathsf{E}, \mathcal{E})$, to itself, it is a $\mathcal{C}_0$-semigroup and we still get an infinitesimal generator with the above properties and the Kolmogorov backwards equation (-> heat equation scenario).
- The Kolmogorov backward equation is the Feynman-Kac formula modulo a potential term.

Kolmogorov equations

- The space of finite, signed Radon measures $\mathcal{M}_{R,f}(\mathsf{E},\mathcal{E})$ on $(\mathsf{E},\mathcal{E}\equiv\mathcal{B}(\mathsf{E}))$ from above is the topological dual of $\mathcal{C}_0(\mathsf{E},\mathcal{E})$. The adjoint semigroup $K^* \equiv (K_t^*)_{t\in\mathbb{R}_+}$ acts on $\mathcal{M}_{R,f}(\mathsf{E},\mathcal{E})$ via

  $$\langle\, K_t^*\,\mu \,|\, f\,\rangle = \langle\, \mu \,|\, K_t\, f\,\rangle \quad .$$

  With $\mu_t \equiv K_t^*\,\nu$, ν the initial distribution, we get

  $$\begin{aligned}\frac{d}{dt}\langle\, \mu_t \,|\, f\,\rangle &= \frac{d}{dt}\langle\, \mu_0 \,|\, K_t\, f\,\rangle = \langle\, \mu_0 \,|\, \frac{d}{dt}\,K_t\, f\,\rangle \\ &= \langle\, \mu_0 \,|\, A\,K_t\, f\,\rangle = \langle\, \mu_0 \,|\, K_t\, A\, f\,\rangle \\ &= \langle\, \mu_t \,|\, A\, f\,\rangle \quad \text{for every } f \in D(A) \quad .\end{aligned}$$

Kolmogorov equations

- This is a weak formulation of the Kolmogorov forward equation (or Fokker-Planck equation)

  $$\frac{\partial}{\partial t}\mu_t = A^*\,\mu_t \quad , \quad \mu_0 \equiv \nu$$

  where A^* is the adjoint infinitesimal generator, defined on a suitable domain in $\mathcal{M}_{R,f}(\mathsf{E},\mathcal{E})$.
- The Kolmogorov forward equation is usually lifted to an equation describing the evolution of the pdf of the process :

  $$\frac{\partial}{\partial t}p_t = A^*\,p_t \quad , \quad p_0 \equiv \frac{d\nu}{d\lambda}$$

  where λ is a σ-finite reference measure on $\mathcal{E}$ and $p_t \in L^1(\mathsf{E},\mathcal{E},\lambda)$.

Kolmogorov equations

- If the Markov process is given by a stochastic differential equation, say a (time-homogenous) diffusion process

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dW_t \ \ , \ a(x) \equiv \frac{1}{2}\sigma(x)^2$$

(suitable regularity conditions assumed), then it has infinitesimal generator

$$A f = b(x)\,\partial_x f + a(x)\,\partial_{xx} f \quad (f \in \mathcal{C}^2) \ \ .$$

Kolmogorov equations

- The adjoint infinitesimal operator is

$$A^* p(x) = -\partial_x \{b(x)p(x)\} + \partial_{xx} \{a(x)p(x)\} \; .$$

9. References

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

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

Billingsley, P.: Convergence of probability measures (Second Edition) (Wiley Series in Probability and Statistics) (1999)

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