Appendix A

Measure theory

In this appendix we collect some important notions from measure theory. The goal is not to present a self-contained presentation, but rather to establish the basic definitions and theorems from the theory for reference in the main text. As such, the presentation omits certain existence theorems and many of the proofs of other theorems (although references are given). The focus is strongly on finite (e.g. probability-) measures, in places at the expense of generality. Some background in elementary set-theory and analysis is required. As a comprehensive reference, we note Kingman and Taylor (1966) [52], alternatives being Dudley (1989) [29] and Billingsley (1986) [15].

A.1 Sets and sigma-algebras


Definition A.1.1. A measurable space \((\Omega, \mathcal{F})\) consists of a set \(\Omega\) and a \(\sigma\)-algebra \(\mathcal{F}\) of subsets of \(\Omega\).

A.2 Measures

Rough setup: set-functions, (signed) measures, probability measures, sigma-additivity, sigma-finiteness

Theorem A.2.1. Let \((\Omega, \mathcal{F})\) be a measurable space with measure \(\mu : \mathcal{F} \to [0, \infty]\). Then,

(i) for any monotone decreasing sequence \((F_n)_{n \geq 1}\) in \(\mathcal{F}\) such that \(\mu(F_n) < \infty\) for some \(n\),

\[
\lim_{n \to \infty} \mu(F_n) = \mu\left(\bigcap_{n=1}^{\infty} F_n\right),
\]  

(A.1)
(ii) for any monotone increasing sequence \((G_n)_{n \geq 1}\) in \(\mathcal{F}\),
\[
\lim_{n \to \infty} \mu(G_n) = \mu\left( \bigcup_{n=1}^{\infty} G_n \right),
\tag{A.2}
\]

Theorem A.2.1) is sometimes referred to as the continuity theorem for measures, because
if we view \(\cap_n F_n\) as the monotone limit \(\lim F_n\), (A.1) can be read as \(\lim_n \mu(F_n) = \mu(\lim_n F_n)\),
expressing continuity from below. Similarly, (A.2) expresses continuity from above. Note that
theorem A.2.1 does not guarantee continuity for arbitrary sequences in \(\mathcal{F}\). It should also be
noted that theorem A.2.1 is presented here in simplified form: the full theorem states that
continuity from below is equivalent to \(\sigma\)-additivity of \(\mu\) (for a more comprehensive formulation
and a proof of theorem A.2.1, see [VS], theorem TNSIN).

Example A.2.1. Let \(\Omega\) be a discrete set and let \(\mathcal{F}\) be the powerset \(2^{\Omega}\) of \(\Omega\), i.e. \(\mathcal{F}\) is the
collection of all subsets of \(\Omega\). The counting measure \(n : \mathcal{F} \to [0, \infty]\) on \((\Omega, \mathcal{F})\) is defined
simply to count the number \(n(F)\) of points in \(F \subset \Omega\). If \(\Omega\) contains a finite number of points,
n is a finite measure; if \(\Omega\) contains a (countably) infinite number of points, \(n\) is \(\sigma\)-finite. The
counting measure is \(\sigma\)-additive.

Example A.2.2. We consider \(\mathbb{R}\) with any \(\sigma\)-algebra \(\mathcal{F}\), let \(x \in \mathbb{R}\) be given and define the
measure \(\delta_x : \mathcal{F} \to [0, 1]\) by
\[
\delta_x(A) = 1\{x \in A\},
\]
for any \(A \in \mathcal{F}\). The probability measure \(\delta_x\) is called the Dirac measure (or delta measure, or
atomic measure) degenerate at \(x\) and it concentrates all its mass in the point \(x\). Clearly, \(\delta_x\)
is finite and \(\sigma\)-additive. Convex combinations of Dirac measures, i.e. measures of the form
\[
P = \sum_{j=1}^{m} \alpha_j \delta_{x_j},
\]
for some \(m \geq 1\) with \(\alpha_1, \ldots, \alpha_m\) such that \(\alpha_j \geq 0\) and \(\sum_{j=1}^{m} \alpha_j = 1\), can be used as a statistical
model for an observation \(X\) that take values in a discrete (but unknown) subset \(\{x_1, \ldots, x_m\}\)
of \(\mathbb{R}\). The resulting model (which we denote \(D(\mathbb{R}, \mathcal{B})\) for reference) is not dominated.

Often, one has a sequence of events \((A_n)_{n \geq 1}\) and one is interested in the probability of a
limiting event \(A\), for example the event that \(A_n\) occurs infinitely often. The following three
lemmas pertain to this situation.

Lemma A.2.1. (First Borel-Cantelli lemma)
Let \((\Omega, \mathcal{F}, P)\) be a probability space and let \((A_n)_{n \geq 1} \subset \mathcal{F}\) be given and denote \(A = \limsup A_n\).
If
\[
\sum_{n \geq 1} P(A_n) < \infty,
\]
then \(P(A) = 0\).
In the above lemma, the sequence \((A_n)_{n \geq 1}\) is general. To draw the converse conclusion, the sequence needs to exist of independent events.

**Lemma A.2.2.** *(Second Borel-Cantelli lemma)*

Let \((\Omega, \mathcal{F}, P)\) be a probability space and let \((A_n)_{n \geq 1} \subset \mathcal{F}\) be independent and denote \(A = \limsup A_n\). If

\[
\sum_{n \geq 1} P(A_n) = \infty,
\]

then \(P(A) = 1\).

Together, the Borel-Cantelli lemmas assert that for a sequence of independent events \((A_n)_{n \geq 1}\), \(P(A)\) equals zero or one, according as \(\sum_n P(A_n)\) converges or diverges. As such, this corollary is known as a zero-one law, of which there are many in probability theory.

Exchangability, De Finetti’s theorem

**Theorem A.2.2.** *(De Finetti’s theorem)* State De Finetti’s theorem.

**Theorem A.2.3.** *(Ulam’s theorem)* State Ulam’s theorem.

**Definition A.2.1.** Let \((\mathcal{Y}, \mathcal{B})\) be a measurable space. Given a set-function \(\mu : \mathcal{B} \to [0, \infty]\), the total variation total-variation norm of \(\mu\) is defined:

\[
\|\mu\|_{TV} = \sup_{B \in \mathcal{B}} |\mu(B)|.
\]

**Lemma A.2.3.** Let \((\mathcal{Y}, \mathcal{B})\) be a measurable space. The collection of all signed measures on \(\mathcal{Y}\) forms a linear space and total variation is a norm on this space.

### A.3 Measurability and random variables

Rough setup: measurability, monotone class theorem, simple functions, random variables, approximating sequences.

### A.4 Integration

Rough setup: the definition of the integral, its basic properties, limit-theorems (Fatou, dominated convergence) and \(L_p\)-spaces.

**Definition A.4.1.** Let \((\Omega, \mathcal{F}, \mu)\) be a measure space. A real-valued measurable function \(f : \Omega \to \mathbb{R}\) is said to be \(\mu\)-integrable if

\[
\int_{\Omega} \max |f| \, d\mu < \infty.
\]
Remark A.4.1. If $f$ is a stochastic vector taking values in $\mathbb{R}^d$, the above definition of integrability is extended naturally by imposing (A.4) on each of the component functions. This extension is more problematic in infinite-dimensional spaces. However, various generalizations can be found in an approach motivated by functional analysis (see Megginson (1998) [67] for an introduction to functional analysis): suppose that $f : \Omega \to X$ takes its values in an infinite-dimensional space $X$. If $(X, \| \cdot \|)$ is a normed space, one can impose that

$$\int_{\Omega} \| f \| \, d\mu < \infty,$$

but this definition may be too strong, in the sense that too few functions $f$ satisfy it. If $X$ has a dual $X^*$, one may impose that for all $x^* \in X^*$,

$$\int_{\Omega} x^*(f) \, d\mu < \infty,$$

which is often a weaker condition than the one in the previous display. In case $X$ is itself (a subset of) the dual of a space $X'$, then $X' \subset X^*$ and we may impose that for all $x \in X'$,

$$\int_{\Omega} f(x) \, d\mu < \infty$$

which is weaker than both previous displays.

Example A.4.1. Our primary interest here is in Bayesian statistics, where the prior and posterior can be measures on a non-parametric model, giving rise to a situation like that in remark A.4.1. Frequently, observations will lie in $\mathbb{R}^n$ and we consider the space of all bounded, measurable functions on $\mathbb{R}^n$, endowed with the supremum-norm. This space forms a Banach space $X'$ and $\mathcal{P}$ is a subset of the unit-sphere of the dual $X''$, since $X \to \mathbb{R} : f \mapsto Pf$ satisfies $|Pf| \leq \|f\|$, for all $f \in X$. Arguably, $P$ should be called integrable with respect to a measure $\Xi$ on $\mathcal{P}$, if

$$\left| \int_{\mathcal{P}} Pf \, d\Xi(P) \right| < \infty.$$

for all $f \in X$. Then, “suitable integrability” is not an issue in the definition of the posterior mean (2.2.1), since $P|f| \leq \sup_{\mathbb{R}^n} |f| = \|f\| < \infty$ for all $f \in X$ and the posterior is a probability measure.

Theorem A.4.1. (Fubini’s theorem) State Fubini’s theorem.

Theorem A.4.2. (Radon-Nikodym theorem) Let $(\Omega, \mathcal{F})$ be a measurable space and let $\mu, \nu : \mathcal{F} \to [0, \infty]$ be two $\sigma$-finite measures on $(\Omega, \mathcal{F})$. There exists a unique decomposition

$$\mu = \mu_\parallel + \mu_\perp,$$

such that $\nu_\parallel \ll \nu$ and $\mu_\perp$ and $\nu$ are mutually singular. Furthermore, there exists a finite-valued, $\mathcal{F}$-measurable function $f : \Omega \to \mathbb{R}$ such that for all $F \in \mathcal{F}$,

$$\mu_\parallel(F) = \int_F f \, d\nu.$$

(A.5)

The function $f$ is $\nu$-almost-everywhere unique.
The function \( f : \Omega \to \mathbb{R} \) in the above theorem is called the \textit{Radon-Nikodym derivative} of \( \mu \) with respect to \( \nu \). If \( \mu \) is a probability distribution, then \( f \) is called the (probability) density for \( \mu \) with respect to \( \nu \). The assertion that \( f \) is \( \nu\)-almost-everywhere unique” means that if there exists a measurable function \( g : \Omega \to \mathbb{R} \) such that (A.5) holds with \( g \) replacing \( f \), then \( f = g, \ (\nu - a.e.), \ i.e. \ f \) and \( g \) may differ only on a set of \( \nu \)-measure equal to zero. Through a construction involving increasing sequences of simple functions, we see that the Radon-Nikodym theorem has the following implication.

\textbf{Corollary A.4.1.} Assume that the conditions for the Radon-Nikodym theorem are satisfied. Let \( X : \Omega \to [0, \infty] \) be measurable and \( \mu \)-integrable. Then the product \( Xf \) is \( \nu \)-integrable and

\[
\int X \, d\mu = \int Xf \, d\nu.
\]

\textbf{Remark A.4.2.} Integrability is not a necessary condition here, but the statement of the corollary becomes rather less transparent if we indulge in generalization.

### A.5 Existence of stochastic processes

A stochastic processes have the following broad definition.

\textbf{Definition A.5.1.} Let \( (\Omega, \mathcal{F}, P) \) be a probability space, let \( T \) be an arbitrary set. A collection of \( \mathcal{F} \)-measurable random variables \( \{ X_t : \Omega \to \mathbb{R} : t \in T \} \) is called a stochastic process indexed by \( T \).

The problem with the above definition is the requirement that there exists an underlying probability space: typically, one approaches a problem that requires the use of stochastic processes by proposing a collection of random quantities \( \{ X_t : t \in T \} \). The guarantee that an underlying probability space \( (\Omega, \mathcal{F}, P) \) exists on which all \( X_t \) can be realised as random variables is then lacking so that we have not defined the stochastic process properly yet. Kolmogorov’s existence theorem provides an explicit construction of \( (\Omega, \mathcal{F}, P) \). Clearly, if the \( X_t \) take their values in a measurable space \( (\mathcal{X}, \mathcal{B}) \), the obvious choice for \( \Omega \) is the collection \( \mathcal{X}^T \) in which the process takes its values. The question remains how to characterize \( P \) and its domain \( \mathcal{F} \). Kolmogorov’s solution here is to assume that for any finite subset \( S = \{ t_1, \ldots, t_k \} \subset T \), the distribution \( P_{t_1 \ldots t_k} \) of the \( k \)-dimensional stochastic vector \( (X_{t_1}, \ldots, X_{t_k}) \) is given. Since the distributions \( P_{t_1 \ldots t_k} \) are as yet unrelated and given for all finite subsets of \( T \), consistency requirements are implicit if they are to serve as marginals to the probability distribution \( P \): if two finite subsets \( S_1, S_2 \subset T \) satisfy \( S_1 \subset S_2 \), then the distribution of \( \{ X_t : t \in S_1 \} \) should be marginal to that of \( \{ X_t : t \in S_2 \} \). Similarly, permutation of the components of the stochastic vector in the above display should be reflected in the respective distributions as well. The requirements for consistency are formulated in two requirements called Kolmogorov’s \textit{consistency conditions}:
(K1) Let \( k \geq 1 \) and \( \{t_1, \ldots, t_{k+1}\} \subset T \) be given. For any \( C \in \sigma(\mathcal{B}^k) \),
\[
P_{t_1 \ldots t_k}(C) = P_{t_1 \ldots t_{k+1}}(C \times \mathcal{B}),
\]
(K2) Let \( k \geq 1 \), \( \{t_1, \ldots, t_k\} \subset T \) and a permutation \( \pi \) of \( k \) elements be given. For any \( A_1, \ldots, A_k \in \mathcal{B} \),
\[
P_{t_{\pi(1)} \ldots t_{\pi(k)}}(A_1 \times \ldots \times A_k) = P_{t_1 \ldots t_k}(A_{\pi^{-1}(1)} \times \ldots \times A_{\pi^{-1}(k)}).
\]

**Theorem A.5.1. (Kolmogorov’s existence theorem)**

Let a collection of random quantities \( \{X_t : t \in T\} \) be given. Suppose that for any \( k \geq 1 \) and all \( t_1, \ldots, t_k \in T \), the finite-dimensional marginal distributions
\[
(X_{t_1}, \ldots, X_{t_k}) \sim P_{t_1 \ldots t_k},
\]
are defined and satisfy conditions (K1) and (K2). Then there exists a probability space \((\Omega, \mathcal{F}, P)\) and a stochastic process \( \{X_t : \Omega \to \mathcal{X} : t \in T\} \) such that all distributions of the form (A.6) are marginal to \( P \).

Kolmogorov’s approach to the definition and characterization of stochastic processes in terms of finite-dimensional marginals turns out to be of great practical value: it allows one to restrict attention to finite-dimensional marginal distributions when characterising the process. The drawback of the construction becomes apparent only upon closer inspection of the \( \sigma \)-algebra \( \mathcal{F} \): \( \mathcal{F} \) is the \( \sigma \)-algebra generated by the cylinder sets, which implies that measurability of events restricting an uncountable number of \( X_t \)'s simultaneously can not be guaranteed! For instance, if \( T = [0, \infty) \) and \( \mathcal{X} = \mathbb{R} \), the probability that sample-paths of the process are continuous,
\[
P(\mathbb{R} \to \mathbb{R} : t \mapsto X_t \text{ is continuous}),
\]
may be ill-defined because it involves an uncountable number of \( t \)'s. This is the ever-recurring trade-off between generality and strength of a mathematical result: Kolmogorov’s existence theorem always works but it does not give rise to a comfortably ‘large’ domain for the resulting \( P : \mathcal{F} \to [0, 1] \).

### A.6 Conditional distributions

In this section, we consider conditioning of probability measures. In first instance, we consider straightforward conditioning on events and illustrate Bayes’ rule, but we also cover conditioning on \( \sigma \)-algebras and random variables, to arrive at the posterior distribution and Bayes’ rule for densities.

**Definition A.6.1.** Let \((\Omega, \mathcal{F}, P)\) be a probability space and let \( B \in \mathcal{F} \) be such that \( P(B) > 0 \). For any \( A \in \mathcal{F} \), the conditional probability of the event \( A \) given event \( B \) is defined:
\[
P(A | B) = \frac{P(A \cap B)}{P(B)},
\]
Conditional probability given $B$ describes a set-function on $\mathcal{F}$ and one easily checks that this set-function is a measure. The conditional probability measure $P(\cdot|B) : \mathcal{F} \to [0,1]$ can be viewed as the restriction of $P$ to $\mathcal{F}$-measurable subsets of $B$, normalized to be a probability measure. Definition (A.7) gives rise to a relation between $P(A|B)$ and $P(B|A)$ (in case both $P(A) > 0$ and $P(B) > 0$, of course), which is called Bayes’ Rule.

**Lemma A.6.1.** (Bayes’ Rule)
Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $A, B \in \mathcal{F}$ be such that $P(A) > 0$, $P(B) > 0$. Then

$$P(A|B) P(B) = P(B|A) P(A).$$

However, being able to condition on events $B$ of non-zero probability only is too restrictive. Furthermore, $B$ above is a definite event; it is desirable also to be able to discuss probabilities conditional on events that have not been measured yet, i.e. to condition on a $\sigma$-algebra.

**Definition A.6.2.** Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{C}$ be a sub-$\sigma$-algebra of $\mathcal{F}$ and let $X$ be a $P$-integrable random variable. The conditional expectation of $X$ given $\mathcal{C}$, denoted $E[X|\mathcal{C}]$, is a $\mathcal{C}$-measurable random variable such that for all $C \in \mathcal{C}$,

$$\int_C X \, dP = \int_C E[X|\mathcal{C}] \, dP.$$

The condition that $X$ be $P$-integrable is sufficient for the existence of $E[X|\mathcal{C}]$; $E[X|\mathcal{C}]$ is unique $P$-almost-surely (see theorem 10.1.1 in Dudley (1989)). Often, the $\sigma$-algebra $\mathcal{C}$ is the $\sigma$-algebra $\sigma(Z)$ generated by another random variable $Z$. In that case we denote the conditional expectation by $E[X|Z]$. Note that conditional expectations are random themselves; realisation occurs only when we impose $Z = z$.

**Definition A.6.3.** Let $(\mathcal{Y}, \mathcal{B})$ be a measurable space, let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{C}$ be a sub-$\sigma$-algebra of $\mathcal{F}$. Furthermore, let $Y : \Omega \to \mathcal{Y}$ be a random variable taking values in $\mathcal{Y}$. The conditional distribution of $Y$ given $\mathcal{C}$ is $P$-almost-surely defined as follows:

$$P_{Y|\mathcal{C}}(A, \omega) = E[1_A|\mathcal{C}](\omega).$$

(A.8)

Although seemingly innocuous, the fact that conditional expectations are defined only $P$-almost-surely poses a rather subtle problem: for every $A \in \mathcal{B}$ there exists an $A$-dependent null-set on which $P_{Y|\mathcal{C}}(A, \cdot)$ is not defined. This is not a problem if we are interested only in $A$ (or in a countable number of sets). But usually, we wish to view $P_{Y|\mathcal{C}}$ as a probability measure, that is to say, it must be well-defined as a map on the $\sigma$-algebra $\mathcal{B}$ almost-surely. Since most $\sigma$-algebras are uncountable, there is no guarantee that the corresponding union of exceptional null-sets has measure zero as well. This means that definition (A.8) is not sufficient for our purposes: the property that the conditional distribution is well-defined as a map is called regularity.
Definition A.6.4. Under the conditions of definition A.6.3, we say that the conditional distribution \( \Pi_{Y|C} \) is regular, if there exists a set \( E \in \mathcal{F} \) such that \( P(E) = 0 \) and for all \( \omega \in \Omega \setminus E \), \( \Pi_{Y|\omega}(\cdot, \omega) \) satisfies A.8 for all \( A \in \mathcal{B} \).

Definition A.6.5. A topological space \((S, \mathcal{T})\) is said to be a Polish space if \( \mathcal{T} \) is metrizable with metric \( d \) and \((S, d)\) is complete and separable.

Polish spaces appear in many subjects in probability theory, most notably in a theorem that guarantees that conditional distributions are regular.

Theorem A.6.1. (regular conditional distributions) Let \( Y \) be a Polish space and denote its Borel \( \sigma \)-algebra by \( \mathcal{B} \). Furthermore let \((\Omega, \mathcal{F}, P)\) be a probability space and \( Y : \Omega \to \mathcal{Y} \) a random variable taking values in \( \mathcal{Y} \). Let \( \mathcal{C} \) be a sub-\( \sigma \)-algebra of \( \mathcal{F} \). Then a conditional distribution \( \Pi_{Y|C} \)

Proof For a proof of this theorem, the reader is referred to Dudley (1989) [29], theorem 10.2.2).

In Bayesian context we can be more specific regarding the sub-\( \sigma \)-algebra \( \mathcal{C} \): since \( \Omega = \mathcal{X} \times \Theta \) (so that \( \omega = (x, \theta) \)) and we condition on \( \theta \), we choose \( \mathcal{C} = \{ \mathcal{X} \times G : G \in \mathcal{G} \} \).

Note also that due to the special choice for \( \mathcal{C} \), \( \mathcal{C} \)-measurability implies that \( \Pi_{Y|\theta}(\cdot, (y, \theta)) \) depends on \( \theta \) alone. Hence we denote it \( \Pi_{Y|\theta} : \mathcal{B} \times \Theta \to [0, 1] \).

Lemma A.6.2. (Bayes’ Rule for densities)

State Bayes’ rule for densities.

A.7 Convergence in spaces of probability measures

Let \( M(\mathbb{R}, \mathcal{B}) \) denote the space of all probability measures on \( \mathbb{R} \) with Borel \( \sigma \)-algebra \( \mathcal{B} \).

Definition A.7.1. (topology of weak convergence)

Let \((Q_n)_{n \geq 1}\) and \( Q \) in \( M(\mathbb{R}, \text{scr}\mathcal{B}) \) be given. Denote the set of points in \( \mathbb{R} \) where \( \mathbb{R} \to [0, 1] : t \mapsto Q(-\infty, t] \) is continuous by \( C \). We say that \( Q_n \) converges weakly to \( Q \) if, for all \( t \in C \), \( Q_n(-\infty, t] \to Q(-\infty, t] \).

Weak convergence has several equivalent definitions. The following lemma, known as the Portmanteau lemma (from the French word for coat-rack),

Lemma A.7.1. Let \((Q_n)_{n \geq 1}\) and \( Q \) in \( M(\mathbb{R}, \text{scr}\mathcal{B}) \) be given. The following are equivalent:

(i) \( Q_n \) converges weakly to \( Q \).

(ii) For every bounded, continuous \( f : \mathbb{R} \to \mathbb{R} \), \( Q_n f \to Q f \).

(iii) For every bounded, Lipschitz \( g : \mathbb{R} \to \mathbb{R} \), \( Q_n g \to Q g \).

(iv) For all non-negative, continuous \( h : \mathbb{R} \to \mathbb{R} \), \( \lim \inf_{n \to \infty} Q_n f \geq Q f \).
(v) For every open set $F \subset \mathbb{R}$, $\liminf_{n \to \infty} Q_n(F) \geq Q(F)$.

(vi) For every closed set $G \subset \mathbb{R}$, $\limsup_{n \to \infty} Q_n(G) \leq Q(G)$.

(vii) For every Borel set $B$ such that $Q(\delta B) = 0$, $Q_n(B) \to Q(B)$.

In (vii) above, $\delta B$ denotes the boundary of $B$, which is defined as the closure of $B$ minus the interior of $B$.

**Lemma A.7.2.** When endowed with the topology of weak convergence, the space $M(\mathbb{R}, \mathcal{B})$ is Polish, i.e. complete, separable and metric.

**Definition A.7.2.** (topology of pointwise convergence)

Let $(Q_n)_{n \geq 1}$ and $Q$ in $M(\mathbb{R}, s\sigma B)$ be given. We say that $Q_n$ converges pointwise to $Q$ if, for all $B \in \mathcal{B}$, $Q_n(B) \to Q(B)$.

**Definition A.7.3.** (topology of total variation)

Let $(Q_n)_{n \geq 1}$ and $Q$ in $M(\mathbb{R}, s\sigma B)$ be given. We say that $Q_n$ converges in total variation to $Q$ if,

$$\sup_{B \in \mathcal{B}} |Q_n(B) - Q(B)| \to 0.$$ 

**Lemma A.7.3.** When endowed with the topology of total variation, the space $M(\mathbb{R}, \mathcal{B})$ is a Polish subspace of the Banach space of all signed measures on $(\mathbb{R}, \mathcal{B})$.
Bibliography


Index

R-better, 39
data, 1
categorical, 1
interval, 1
nominal, 1
ordinal, 1
ranked, 1
ratio, 1
decision, 38
decision principle
minimax, 39
decision rule, 38
Bayes, 42
minimax, 40
randomized, 40
decision-space, 38
density, see probability density, 95
Dirichlet distribution, 69
Dirichlet family, 70
Dirichlet process, 71
distribution
unimodal, 23

empirical Bayes, 66
empirical expectation, 11
empirical process, 11
entropy
Lindley, 58
Shannon, 58

estimator, 4

M-, 26
MAP, 25
maximum-a-priori, 25

Bayes factor, 35
Bayes’ billiard, 17
belief, 10
bootstrap, 53

classification, 28, 43
classifier, 44
conditional distribution, 14, 97
regular, 14, 98
conditional expectation, 97
conditional independence, 18
conditional probability, 96
confidence level, 32, see level, confidence
confidence region, 32
credible interval, see credible set
credible region, see credible set
credible set, 33
HPD-, 34
critical region, 28
maximum-likelihood, 6, 11
minimax, 41
non-parametric ML, 76
penalized maximum-likelihood, 27
small-ball, 25
exchangeability, 20
expectation
empirical, 5
exponential family, 61
canonical representation, 61
of full rank, 61
feature vector, 44
hyperparameter, 63
hyperprior, 63
hypothesis, 27
identifiability, 2
inadmissible, 39
inference, 38
infinite divisibility, 70
integrability, 93
lemma
First Borel-Cantelli, 92
Second Borel-Cantelli, 93
level, 28, 33
confidence, 32
significance, 28
likelihood, 7
likelihood principle, 6
limit distribution, 5
location, 22
loss, see loss-function
loss-function, 25, 38
$L_2^\ast$, 41
measure
atomic, 92
delta, 92
Dirac, 92
misclassification, 44
ML-II estimator, 68
MLE, see estimator, maximum-likelihood model, 2
dimension, 3
dominated, 2
full non-parametric, 4
hierarchical Bayes, 63
identifiable, 2
mis-specified, 3
non-parametric, 4
normal, 3
parameterized, 2
parametric, 3
well-specified, 3
model selection, 66, 67
norm
total-variation, 6, 93
NPMLE, see non-parametric MLE
null hypothesis, 27
odds ratio
posterior, 35
prior, 35
optimality criteria, 6
over-fitting, 67
parameter space, 2
point-estimator, see estimator
pointwise convergence, 99
Polish space, 98
Portmanteau lemma, 98
posterior, 8, 15
posterior expectation, 23
posterior mean, 23
parametric, 23
posterior median, 25
posterior mode, 25
power function, 28

107
sequence, 31
power-set, 4
powerset, 92
predictive distribution
  posterior, 19
  prior, 19, 75
preferred
  Bayes, 42
  minimax, 39
prior, 8, 20
  conjugate, 60
  Dirichlet process, 21
  improper, 54
  informative, 50
  Jeffreys, 56
  non-informative, 53
  objective, 53
  reference, 58
  subjective, 50
  subjectivist, 15
probability density, 95
probability density function, 2
Radon-Nikodym derivative, 95
rate of convergence, 5
regularity, 14, 20, 97
risk
  Bayes, 42
  minimax, 39
risk function
  Bayesian, 42
sample-average, 5, 11
sample-size, 5
samplespace, 1, 38
significance level, see level, see level, sign-
  significance
  asymptotic, 29
simple
  hypothesis, 28
simplex, 4
state, 38
state-space, 38
statistic, 5, 32
statistical decision theory, 38
statistics
  inferential, 38
stochastic process, 95
support, 16
test
  asymptotic, 30
  more powerful, 31
  uniformly more powerful, 31
  uniformly most powerful, 29, 31
test sequence, 30
test-statistic, 28
theorem
  central limit, 5
  De Finetti’s, 93
  Fubini’s, 94
  Glivenko-Cantelli, 6
  Minimax, 40
  Radon-Nikodym, 94
  Ulam’s, 93
type-I error, 28
type-II error, 28
utility, see utility-function
utility-function, 38
Weak convergence, 98
zero-one law, 93
The figure on the front cover originates from Bayes (1763), *An essay towards solving a problem in the doctrine of chances*, (see [4] in the bibliography), and depicts what is nowadays known as Bayes’ Billiard. To demonstrate the uses of conditional probabilities and Bayes’ Rule, Bayes came up with the following example: one white ball and \( n \) red balls are placed on a billiard table of length normalized to 1, at independent, uniformly distributed positions. Conditional on the distance \( X \) of the white ball to one end of the table, the probability of finding exactly \( k \) of the \( n \) red balls closer to that end, is easily seen to be:

\[
P(k \mid X = x) = \frac{n!}{k!(n-k)!} x^k (1-x)^{n-k}.
\]

One finds the probability that \( k \) red balls are closer than the white, by integrating with respect to the position of the white ball:

\[
P(k) = \frac{1}{n+1}.
\]

Application of Bayes’ Rule then gives rise to a Beta-distribution \( B(k+1, n-k+1) \) for the position of the white ball conditional on the number \( k \) of red balls that are closer. The density:

\[
\beta_{k+1,n-k+1}(x) = \frac{(n+1)!}{k!(n-k)!} x^k (1-x)^{n-k},
\]

for this Beta-distribution is the curve drawn at the bottom of the billiard in the illustration. (See example 2.1.2)