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

Rough setup: set operations, monotony of sequences of subsets, set-limits, sigma-algebra's, measurable spaces, set-functions, product spaces.

Definition A.1.1. *A measurable space* $(\Omega, \mathscr{F})$ *consists of a set* Ω *and a* σ*-algebra* $\mathscr{F}$ *of subsets of* Ω*.*

A.2 Measures

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

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

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

$$\lim_{n\to\infty} \mu(F_n) = \mu\Big(\bigcap_{n=1}^{\infty} F_n\Big), \tag{A.1}$$

(ii) for any monotone increasing sequence $(G_n)_{n\geq 1}$ *in* $\mathscr{F}$*,*

$$\lim_{n\to\infty} \mu(G_n) = \mu\Big(\bigcup_{n=1}^{\infty} G_n\Big), \tag{A.2}$$

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 $\mathscr{F}$. 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 σ-additivity of μ (for a more comprehensive formulation and a proof of theorem A.2.1, see [52], theorem 3.2).

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

Example A.2.2. *We consider* $\mathbb{R}$ *with any* σ*-algebra* $\mathscr{F}$*, let* $x \in \mathbb{R}$ *be given and define the measure* $\delta_x : \mathscr{F} \to [0,1]$ *by*

$$\delta_x(A) = 1\{x \in A\},$$

for any $A \in \mathscr{F}$*. The probability measure* δ_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,* δ_x *is finite and* σ*-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* $D(\mathbb{R}, \mathscr{B})$ *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 $(\Omega, \mathscr{F}, P)$ *be a probability space and let* $(A_n)_{n\geq 1} \subset \mathscr{F}$ *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 $(\Omega, \mathscr{F}, P)$ be a probability space and let $(A_n)_{n\geq 1} \subset \mathscr{F}$ 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 $(\mathscr{Y}, \mathscr{B})$ be a measurable space. Given a set-function $\mu : \mathscr{B} \to [0, \infty]$, the total variation total-variation norm of μ is defined:*

$$\|\mu\|_{TV} = \sup_{B\in\mathscr{B}} |\mu(B)|. \tag{A.3}$$

Lemma A.2.3. *Let $(\mathscr{Y}, \mathscr{B})$ be a measurable space. The collection of all signed measures on $\mathscr{Y}$ 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 $(\Omega, \mathscr{F}, \mu)$ be a measure space. A real-valued measurable function $f : \Omega \to \mathbb{R}$ is said to be μ-integrable if*

$$\int_O mega|f|\, d\mu < \infty. \tag{A.4}$$

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 supermum-norm. This space forms a Banach space X' and $\mathscr{P}$ 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 Ξ on $\mathscr{P}$, if*

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

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 $(\Omega, \mathscr{F})$ be a measurable space and let $\mu, \nu : \mathscr{F} \to [0, \infty]$ be two σ-finite measures on $(\Omega, \mathscr{F})$. There exists a unique decomposition*

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

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

$$\mu_\parallel(F) = \int_F f \, d\nu. \tag{A.5}$$

The function f is ν-almost-everywhere unique.

The function $f : \Omega \to \mathbb{R}$ in the above theorem is called the *Radon-Nikodym derivative* of μ with respect to ν. If μ is a probability distribution, then f is called the (probability) density for μ with respect to ν. The assertion that f is "ν-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 ν-measure equal to zero. Through a construction involving increasing sequences of simple functions, we see that the Radon-Nikodym theorem has the following implication.

Corollary A.4.1. *Assume that the conditions for the Radon-Nikodym theorem are satisfied. Let $X : \Omega \to [0, \infty]$ be measurable and μ-integrable. Then the product Xf is ν-integrable and*

$$\int X \, d\mu = \int Xf \, d\nu.$$

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.

Definition A.5.1. *Let $(\Omega, \mathscr{F}, P)$ be a probability space, let T be an arbitrary set. A collection of $\mathscr{F}$-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 $(\Omega, \mathscr{F}, P)$ 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 $(\Omega, \mathscr{F}, P)$. Clearly, if the X_t take their values in a measurable space space $(\mathscr{X}, \mathscr{B})$, the obvious choice for Ω is the collection $\mathscr{X}^T$ in which the process takes its values. The question remains how to characterize P and its domain $\mathscr{F}$. 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 *consistency conditions*:

(K1) Let $k \geq 1$ and $\{t_1, \ldots, t_{k+1}\} \subset T$ be given. For any $C \in \sigma(\mathscr{B}^k)$,

$$P_{t_1 \ldots t_k}(C) = P_{t_1 \ldots t_{k+1}}(C \times \mathscr{X}),$$

(K2) Let$k \geq 1$, $\{t_1, \ldots, t_k\} \subset T$ and a permutation π of k elements be given. For any $A_1, \ldots, A_k \in \mathscr{B}$,

$$P_{t_{\pi(1)} \ldots t_{\pi(k)}}(A_1 \times \ldots \times A_k) = P_{t_1 \ldots t_k}(A_{\pi^{-1}(1)} \times \ldots \times A_{\pi^{-1}(k)}).$$

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

$$(X_{t_1}, \ldots, X_{t_k}) \sim P_{t_1 \ldots t_k}, \tag{A.6}$$

are defined and satisfy conditions (K1) and (K2). Then there exists a probability space $(\Omega, \mathscr{F}, P)$ and a stochastic process $\{ X_t : \Omega \to \mathscr{X} : t \in T \}$ 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 σ-algebra $\mathscr{F}$: $\mathscr{F}$ is the σ-algebra generated by the cylinder sets, which implies that measurability of events restricing an uncountable number of X_t's simultaneously can not be guaranteed! For instance, if $T = [0, \infty)$ and $\mathscr{X} = \mathbb{R}$, the probability that sample-paths of the process are coninuous,

$$P\big(\mathbb{R} \to \mathbb{R} : t \mapsto X_t \,\text{iscontinuous} \big),$$

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 $P : \mathscr{F} \to [0, 1]$.

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 σ-algebras and random variables, to arrive at the posterior distribution and Bayes' rule for densities.

Definition A.6.1. *Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $B \in \mathscr{F}$ be such that $P(B) > 0$. For any $A \in \mathscr{F}$, the conditional probability of the event A given event B is defined:*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{A.7}$$

Conditional probability given B describes a set-function on $\mathscr{F}$ and one easily checks that this set-function is a measure. The conditional probability measure $P(\,\cdot\,|B) : \mathscr{F} \to [0,1]$ can be viewed as the restriction of P to $\mathscr{F}$-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 $(\Omega, \mathscr{F}, P)$ be a probability space and let $A, B \in \mathscr{F}$ 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 σ-algebra.

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

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

The condition that X be P-integrable is sufficient for the existence of $E[X|\mathscr{C}]$; $E[X|\mathscr{C}]$ is unique P-almost-surely (see theorem 10.1.1 in Dudley (1989)). Often, the σ-algebra $\mathscr{C}$ is the σ-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 $(\mathscr{Y}, \mathscr{B})$ be a measurable space, let $(\Omega, \mathscr{F}, P)$ be a probability space and let $\mathscr{C}$ be a sub-σ-algebra of $\mathscr{F}$. Furthermore, let $Y : \Omega \to \mathscr{Y}$ be a random variable taking values in $\mathscr{Y}$. The conditional distribution of Y given $\mathscr{C}$ is P-almost-surely defined as follows:*

$$P_{Y|\mathscr{C}}(A, \omega) = E[1_A|\mathscr{C}](\omega). \tag{A.8}$$

Although seemingly innocuous, the fact that conditional expectations are defined only P-almost-surely poses a rather subtle problem: for every $A \in \mathscr{B}$ there exists an A-dependent null-set on which $P_{Y|\mathscr{C}}(A, \cdot)$ 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 $P_{Y|\mathscr{C}}$ as a probability measure, that is to say, it must be well-defined as a *map* on the σ-algebra $\mathscr{B}$ almost-surely. Since most σ-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 $\Pi_{Y|\mathscr{C}}$ is regular, if there exists a set $E \in \mathscr{F}$ such that $P(E) = 0$ and for all $\omega \in \Omega \setminus E$, $\Pi_{Y|\mathscr{C}}(\,\cdot\,, \omega)$ satisfies A.8 for all $A \in \mathscr{B}$.*

Definition A.6.5. *A topological space $(S, \mathscr{T})$ is said to be a Polish space if $\mathscr{T}$ 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 $\mathscr{Y}$ be a Polish space and denote its Borel σ-algebra by $\mathscr{B}$. Furthermore let $(\Omega, \mathscr{F}, P)$ be a probability space and $Y : \Omega \to \mathscr{Y}$ a random variable taking values in $\mathscr{Y}$. Let $\mathscr{C}$ be a sub-σ-algebra of $\mathscr{F}$. Then a conditional distribution [MORE MORE]*

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-σ-algebra $\mathscr{C}$: since $\Omega = \mathscr{X} \times \Theta$ (so that $\omega = (x, \theta)$) and we condition on θ, we choose $\mathscr{C} = \{\mathscr{X} \times G : G \in \mathscr{G}\}$.

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

Lemma A.6.2. (Bayes' Rule for densities)
State Bayes' rule for densities.

A.7 Convergence in spaces of probability measures

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

Definition A.7.1. *(topology of weak convergence)*
Let $(Q_n)_{n\geq 1}$ and Q in $M(\mathbb{R}, scrB)$ 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 serveral 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 $M(\mathbb{R}, scrB)$ be given. The following are equivalent:*

(i) Q_n converges weakly to Q.

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

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

(iv) For all non-negative, continuous $h : \mathbb{R} \to \mathbb{R}$, $\liminf_{n\to\infty} Q_n f \geq Qf$.

(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, δB 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* $M(\mathbb{R}, \mathscr{B})$ *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* $M(\mathbb{R}, scrB)$ *be given. We say that* Q_n *converges pointwise to* Q *if, for all* $B \in \mathscr{B}$, $Q_n(B) \to Q(B)$.

Definition A.7.3. *(topology of total variation)*
Let $(Q_n)_{n\geq 1}$ *and* Q *in* $M(\mathbb{R}, scrB)$ *be given. We say that* Q_n *converges in total variation to* Q *if,*

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

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

Bibliography

[1] M. Alpert and H. Raiffa, *A progress report on the training of probability assessors,* In *Judgement under uncertainty: heuristics and biases*, eds. D. Kahneman, P. Slovic and A. Tversky, Cambridge University Press, Cambridge (1982).

[2] S. Amari, *Differential-geometrical methods in statistics,* Lecture Notes in Statistics No. 28, Springer Verlag, Berlin (1990).

[3] M. Bayarri and J. Berger, *The interplay of Bayesian and frequentist analysis,* Preprint (2004).

[4] T. Bayes, *An essay towards solving a problem in the doctrine of chances,* Phil. Trans. Roy. Soc. **53** (1763), 370–418.

[5] A. Barron, L. Birgé and P. Massart, *Risk bounds for model selection via penalization*, Probability Theory and Related Fields **113** (1999), pp. 301–413.

[6] S. Bernstein, *Theory of probability,* (in Russian), Moskow (1917).

[7] A. Barron, M. Schervish and L. Wasserman, *The consistency of posterior distributions in nonparametric problems,* Ann. Statist. **27** (1999), 536–561.

[8] J. Berger, *Statistical decision theory and Bayesian analysis,* Springer, New York (1985).

[9] J. Berger and J. Bernardo, *On the development of reference priors,* Bayesian Statistics **4** (1992), 35–60.

[10] R. Berk, *Consistency of a posteriori,* Ann. Math. Statist. **41** (1970), 894–906.

[11] R. Berk and I. Savage, *Dirichlet processes produce discrete measures: an elementary proof,* Contributions to statistics, Reidel, Dordrecht (1979), 25–31.

[12] J. Bernardo, *Reference posterior distributions for Bayesian inference,* J. Roy. Statist. Soc. **B41** (1979), 113–147.

[13] J. Bernardo and A. Smith, *Bayesian theory,* John Wiley & Sons, Chichester (1993).

[14] P. Bickel and J. Yahav, *Some contributions to the asymptotic theory of Bayes solutions,* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **11** (1969), 257–276.

[15] P. Billingsley, *Probability and Measure, 2nd edition,* John Wiley & Sons, Chichester (1986).

[16] L. Birgé, *Approximation dans les espaces métriques et théorie de l'estimation,* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **65** (1983), 181–238.

[17] L. BIRGÉ, *Sur un théorème de minimax et son application aux tests,* Probability and Mathematical Statistics **3** (1984), 259–282.

[18] L. BIRGÉ and P. MASSART, *From model selection to adaptive estimation,* Festschrift for Lucien Le Cam, Springer, New York (1997), 55–87.

[19] L. BIRGÉ and P. MASSART, *Gaussian model selection*, J. Eur. Math. Soc. **3** (2001), 203–268.

[20] C. BISHOP, *Pattern Recognition and Machine Learning,* Springer, New York (2006).

[21] D. BLACKWELL and L. DUBINS, *Merging of opinions with increasing information,* Ann. Math. Statist. **33** (1962), 882–886.

[22] L. BOLTZMANN, *Vorlesungen über Gastheorie,* (2 Volumes), Leipzig (1895, 1898).

[23] H. CRAMÉR, *Mathematical methods of statistics,* Princeton University Press, Princeton (1946).

[24] A. DAWID, *On the limiting normality of posterior distribution,* Proc. Canad. Phil. Soc. **B67** (1970), 625-633.

[25] P. DIACONIS and D. FREEDMAN, *On the consistency of Bayes estimates,* Ann. Statist. **14** (1986), 1–26.

[26] P. DIACONIS and D. FREEDMAN, *On inconsistent Bayes estimates of location,* Ann. Statist. **14** (1986), 68–87.

[27] P. DIACONIS and D. FREEDMAN, *Consistency of Bayes estimates for nonparametric regression: Normal theory,* Bernoulli, **4** (1998), 411-444.

[28] J. DOOB, *Applications of the theory of martingales,* Le calcul des Probabilités et ses Applications, Colloques Internationales du CNRS, Paris (1948), 22–28.

[29] R. DUDLEY, *Real analysis and probability,* Wadsworth & Brooks-Cole, Belmont (1989).

[30] A. DVORETZKY, J. KIEFER, and J. WOLFOWITZ, *Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator,* Ann. Math. Statist. **27** (1956), 642–669.

[31] B. EFRON, *Defining curvature on a statistical model,* Ann. Statist. **3** (1975), 1189–1242.

[32] B. EFRON and R. Tibshirani, An introduction to the Bootstrap, Chapman and Hall, London (1993).

[33] M. ESCOBAR and M. WEST, *Bayesian density estimation and inference with mixtures,* Journal of the American Statistical Association **90** (1995), 577–588.

[34] T. FERGUSON, *A Bayesian analysis of some non-parametric problems,* Ann. Statist. **1** (1973), 209–230.

[35] T. FERGUSON, *Prior distribution on the spaces of probability measures,* Ann. Statist. **2** (1974), 615–629.

[36] D. FREEDMAN, *On the asymptotic behavior of Bayes estimates in the discrete case I,* Ann. Math. Statist. **34** (1963), 1386–1403.

[37] D. FREEDMAN, *On the asymptotic behavior of Bayes estimates in the discrete case II,* Ann. Math. Statist. **36** (1965), 454–456.

[38] D. Freedman, *On the Bernstein-von Mises theorem with infinite dimensional parameters,* Ann. Statist. **27** (1999), 1119–1140.

[39] S. van de Geer, Empirical Processes in M-Estimation. Cambridge University Press, Cambridge (2000).

[40] S. Ghosal, J. Ghosh and R. Ramamoorthi, *Non-informative priors via sieves and packing numbers,* Advances in Statistical Decision theory and Applications (eds. S. Panchapakeshan, N. Balakrishnan), Birkhäuser, Boston (1997).

[41] S. Ghosal, J. Ghosh and A. van der Vaart, *Convergence rates of posterior distributions,* Ann. Statist. **28** (2000), 500–531.

[42] J. Ghosh and R. Ramamoorthi, *Bayesian Nonparametrics,* Springer Verlag, Berlin (2003).

[43] P. Green, *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, Biometrika **82** (1995), 711–732.

[44] T.-M. Huang, *Convergence rates for posterior distributions and adaptive estimation,* Carnegie Mellon University, preprint (accepted for publication in Ann. Statist.).

[45] I. Ibragimov and R. Has'minskii, *Statistical estimation: asymptotic theory,* Springer, New York (1981).

[46] H. Jeffreys, *An invariant form for the prior probability in estimation problems,* Proc. Roy. Soc. London **A186** (1946), 453–461.

[47] H. Jeffreys, *Theory of probability (3rd edition),* Oxford University Press, Oxford (1961).

[48] R. Kass and A. Raftery, *Bayes factors,* Journal of the American Statistical Association **90** (1995), 773–795.

[49] R. Kass and L. Wasserman, *A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion,* Journal of the American Statistical Association **90** (1995), 928–934.

[50] Yongdai Kim and Jaeyong Lee, *The Bernstein-von Mises theorem of survival models,* (accepted for publication in Ann. Statist.)

[51] Yongdai Kim and Jaeyong Lee, *The Bernstein-von Mises theorem of semiparametric Bayesian models for survival data,* (accepted for publication in Ann. Statist.)

[52] J. Kingman and S. Taylor, *Introduction to measure and probability,* Cambridge University Press, Cambridge (1966).

[53] B. Kleijn and A. van der Vaart, *Misspecification in Infinite-Dimensional Bayesian Statistics,* Ann. Statist. **34** (2006), 837–877.

[54] B. Kleijn and A. van der Vaart, *The Bernstein-Von-Mises theorem under misspecification,* (submitted for publication in the Annals of Statistics).

[55] B. Kleijn and A. van der Vaart, *A Bayesian analysis of errors-in-variables regression,* (submitted for publication in the Annals of Statistics).

[56] A. Kolmogorov and V. Tikhomirov, *Epsilon-entropy and epsilon-capacity of sets in function spaces,* American Mathematical Society Translations (series 2), **17** (1961), 277–364.

[57] P. Laplace, *Mémoire sur la probabilité des causes par les évenements,* Mem. Acad. R. Sci. Presentés par Divers Savans **6** (1774), 621–656. (Translated in Statist. Sci. **1**, 359–378.)

[58] P. Laplace, *Théorie Analytique des Probabilités (3rd edition),* Courcier, Paris (1820).

[59] E. Lehmann and G. Casella, *Theory of point-estimation, (2nd ed.)* Springer, New York (1998).

[60] E. Lehmann and J. Romano, *Testing statistical hypothesis,* Pringer, New York (2005).

[61] L. Le Cam, *On some asymptotic properties of maximum-likelihood estimates and related Bayes estimates,* University of California Publications in Statistics, **1** (1953), 277–330.

[62] L. Le Cam, *On the assumptions used to prove asymptotic normality of maximum likelihood estimators,* Ann. Math. Statist. **41** (1970), 802–828.

[63] L. Le Cam, *Asymptotic methods in statistical decision theory,* Springer, New York (1986).

[64] L. Le Cam and G. Yang, *Asymptotics in Statistics: some basic concepts,* Springer, New York (1990).

[65] D. Lindley, *A measure of the information provided by an experiment,* Ann. Math. Statist. **27** (1956), 986–1005.

[66] D. Lindley and A. Smith, *Bayes estimates for the linear model,* J. Roy. Statist. Soc. **B43** (1972), 1–41.

[67] R. Megginson, *An introduction to Banach Space Theory,* Springer, New York (1998).

[68] R. Von Mises, *Wahrscheinlichkeitsrechnung,* Springer Verlag, Berlin (1931).

[69] J. Munkres, *Topology (2nd edition),* Prentice Hall, Upper Saddle River (2000).

[70] H. Raiffa, and R. Schlaifer, *Decision analysis: introductory lectures on choices under uncertainty,* Addison-Wesley, Reading (1961).

[71] C. Rao, *Information and the accuracy attainable in the estimation of statistical parameters,* Bull. Calcutta Math. Soc. **37** (1945), 81–91.

[72] C. Robert, *The Bayesian choice: from decision-theoretic foundations to computational implementation,* Springer, New York (2001).

[73] B. Ripley, *Pattern recognition and neural networks*, Cambridge University Press, Cambridge (1996).

[74] L. Savage, *The subjective basis of statistical practice,* Technical report, Dept. Statistics, University of Michigan (1961).

[75] M. Schervish, *Theory of statistics,* Springer, New York (1995).

[76] L. Schwartz, *On Bayes procedures,* Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **4** (1965), 10–26.

[77] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. **6** (1978), pp. 461–464.

[78] C. Shannon, *A Mathematical Theory of Communication,* Bell System Technical Journal **27** (1948), 379–423, 623–656.

[79] X. Shen and L. Wasserman, *Rates of convergence of posterior distributions,* Ann. Statist. **29** (2001), 687–714.

[80] X. Shen, *Asymptotic normality of semiparametric and nonparametric posterior distributions,* Journal of the American Statistical Association **97** (2002), 222–235.

[81] H. Strasser, *Mathematical Theory of Statistics,* de Gruyter, Amsterdam (1985).

[82] J. Tukey, *Exploratory data analysis,* Addison-Wesley, Reading (1977).

[83] A. van der Vaart, *Asymptotic Statistics,* Cambridge University Press, Cambridge (1998).

[84] A. Walker, *On the asymptotic behaviour of posterior distributions,* J. Roy. Statist. Soc. **B31** (1969), 80–88.

[85] L. Wasserman, *Bayesian model selection and model averaging*, J. Math. Psych. **44** (2000), 92–107.

[86] Y. Yang and A. Barron, *An asymptotic property of model selection criteria*, IEEE Transactions on Information Theory **44** (1998), 95–116.

Index

Cover Illustration

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:

$$P\big(\, k \;\big|\; X = x \,\big) = \frac{n!}{k!\,(n-k)!}\, x^k (1-x)^{n-k}.$$

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)