Appendix A

Probabilistic Concepts

Modern probability is based on measure theory (Appendix A.1). Measure theory will provide the means to formally describe random variables, random processes, and most generally, random measures. A model represented by random measures can be fitted to the data using Bayesian inference (Appendix A.2). We give three typical examples of Bayesian model compositions, among which an infinite mixture model (Appendix A.3). A number of processes are described that can be used with (for example as prior distribution) infinite mixture models (Appendix A.4). We introduce plate notation which visualizes infinite models particularly well (Appendix A.5). Then we investigate completely random measures and Lévy measures (Appendix A.6), exchangebility (Appendix A.7), and stick-breaking processes (Appendix A.8). For mathematically more thorough approaches we refer to the literature (Halmos, 1974; Rosenthal, 2006; Cohn, 2013).

A.1 Measure Theory

A random variable is a *function* that assigns values to a *set* of possible outcomes. The formal definition requires concepts such as "measurable function" and "probability space" from *measure theory* (Feller, 1950). Measure theory is used to generalize the notion of a random variable to that of a "random process".

Informally, a measure generalizes the concepts of length, area, and volume of an Euclidean object to a concept of size for sets and subsets. The definition of a measure is based on the definition of a σ-algebra. A σ-algebra ascribes a value to a sum of individual disjoint sets, even if they are infinite in number.

▼ **Definition A.1** — ***σ-algebra***

A **σ-algebra** is a *subset* $\Sigma \in 2^X$, with X a set and 2^X its powerset, with three requirements:

- Σ is non-empty: at least one $A \in X$ is in Σ;

- Σ is closed under complementation: if A in Σ, so is its complement A^c;
- Σ is closed under countable unions: if $A_1, A_2, \ldots$ in Σ, so is $A = A_1 \cup A_2 \cup \ldots$.

The members of a σ-algebra are called *measurable sets*. Let $X = \{1,2,3,4\}$ and let us define a σ-algebra $\Sigma = \{\varnothing, \{1\}, \{4\}, \{2,3\}, \{1,4\}, \{1,2,3\}, \{2,3,4\}, \{1,2,3,4\}\}$. Here $\varnothing$ denotes the empty set. The complement of A is defined with respect to X: $A \cup A^c = X$. An example of closure under complementation: let $A_1 = \{1\}$, then $A_1^c = \{2,3,4\}$ and A_1^c is indeed a member of Σ: $A_1^c \in \Sigma$. An example of closure under countable unions: let $A_1 = \{1\}$ and $A_2 = \{2,3\}$, then $A_1 \cup A_2 = \{1,2,3\}$ and $A_1 \cup A_2 \in \Sigma$.

> The notion of a σ-algebra (Fremlin, 2000) can be applied to solve the so-called Banach-Tarski paradox (Banach and Tarski, 1924). This paradox describes how a unit-ball in $\mathbf{R}^3$ can be partitioned into a finite number of disjoint infinite sets (scattering of points) and then can be reassembled into two unit-balls again. This violates the intuitive notion of preservation of volume. If the measure μ of the union of two disjoint sets is equal to the sum of the measures of the two sets, this is called *finite additivity*: $\mu(\bigcup_{i=1}^{N} A_i) = \sum_{i=1}^{N} \mu(A_i)$. In probability theory *σ-additivity* extends this to infinite disjoint sets: $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$. Measure theory solves the Banach-Tarski paradox by only assigning a measure to subsets that are measurable sets (Tao, 2011).

A *measure* assigns values to measurable sets (as stated before, measurable sets are members or subsets of Σ).

▼ Definition A.2 — *measure*

A **measure** μ is a function from Σ to $[-\infty, +\infty]$, with three requirements:

- μ is non-negative: $\mu(A) \geq 0$ for $\forall A \in \Sigma$;
- μ has a null empty set: $\mu(\varnothing) = 0$;
- μ is σ-additive: $\mu(\bigcup_{i \in I_\Sigma} A_i) = \sum_{i \in I_\Sigma} \mu(A_i)$ for A_i disjoint.

The first statement defines that a measure μ only assigns non-negative values to sets in Σ. The second statement equals the measure of the empty set $\varnothing$ to 0. The third statement defines that σ-additivity is required. For any two sets in Σ the measure of the union of the sets equals the sum of the measures of the individual sets. Here I_Σ defines an index over sets in Σ.

Informally, a measure relates the concepts of *sets* and *subsets* to notions of size. A measure can be seen as a *monotonically* increasing function. Let the set A in X be the interval $[0,1)$, an uncountable (infinite) set of real numbers. Define the σ-algebra $\{\varnothing, A\}$. The empty set has measure 0, the set A has measure 1. Let us define the σ-algebra $\{\varnothing, A_{0,0.5}, A_{0.5,1}, A\}$. The set $A_{0,0.5}$ corresponds to the interval $[0,0.5)$ and $A_{0.5,1}$ to $[0.5,1)$. Both sets are assigned measure 0.5 and their union has measure 1. This examples shows that with σ-additive unions, measures can be assigned to sets that are uncountable.

A *measurable space* (X, Σ) is defined as a pair.

▼ Definition A.3 — *measurable space*

A **measurable space** (X, Σ) is a pair with:

- X a set;
- Σ a σ-algebra over X.

A *measure space* (X, Σ, μ) is defined as a triple.

▼ Definition A.4 — *measure space*

A **measure space** (X, Σ, μ) is a triple with:

- X a set;
- Σ a σ-algebra over X;
- μ a measure from Σ to $[-\infty, \infty]$.

A finite measure μ assigns a finite real number to all A.

▼ Definition A.5 — *finite measure*

A **finite measure** μ is a measure from Σ to $[0, \infty)$:

- μ is non-negative: $\mu(A) \geq 0$ for $\forall A \in \Sigma$;
- μ has a null empty set: $\mu(\varnothing) = 0$;
- μ is σ-additive: $\mu(\bigcup_{i \in I_\Sigma} A_i) = \sum_{i \in I_\Sigma} \mu(A_i)$ for A_i disjoint;
- μ for the whole sample space, X, is finite: $\mu(X) = N$.

A σ-finite measure allows A to be a countable union of sets with finite measure.

▼ **Definition A.6 — σ-*finite measure***

A **σ-finite measure** μ is a finite measure with:

- X is a countable union of sets with finite measures.

We will now define five measures: (A.1.1) the *probability measure* (Definition A.7), (A.1.2) the *counting measure* (Definition A.9), (A.1.3) the *Borel measure* (Definition A.11), (A.1.4) the *Lebesgue measure* (Definition A.16), and (A.1.5) the *random measure* (Definition A.17). These measures are important because they are fundamental to different branches of mathematics. In probability theory a σ-algebra is interpreted as a collection of events to which probabilities are assigned. Counting measures play a fundamental role in discrete probability distributions. In integration theory a σ-algebra corresponding to the Borel and Lebesgue measures are relevant for integration in the Euclidean space $\mathscr{R}^n$. In statistics a σ-algebra formally defines the concept of sufficient statistics and generalizes random variables to random functions and measures.

A.1.1 Probability Measure

A *probability measure*, $\mathbb{P}$, is a finite measure that assigns non-negative values $\mathbb{P}$, called probabilities, to sets A, called events (see Definition A.7).

▼ **Definition A.7 — *probability measure***

A **probability measure** $\mathbb{P}$ is a measure μ with:

- $\mathbb{P}$ is non-negative: $\mathbb{P}(A) \geq 0$ for $\forall A \in \Sigma$;
- $\mathbb{P}$ has a null empty set: $\mathbb{P}(\varnothing) = 0$;
- $\mathbb{P}$ is σ-additive: $\mathbb{P}(\bigcup_{i \in I_\Sigma} A_i) = \sum_{i \in I_\Sigma} \mu(A_i)$ for A_i disjoint;
- $\mathbb{P}$ for the whole sample space, X, is unity: $\mathbb{P}(X) = 1$.

The four requirements are called the Kolmogorov axioms (Kolmogorov, 1933). The probability measure is an actual *measure*. It therefore obeys the three requirements: (1) non-negativity for any set, (2) the existence of a null empty set, and (3) σ-additivity. Here we note that a *probability* measure compared to a general measure obeys a fourth requirement, namely the restriction of the measure for the whole space X to 1. This can be seen as some kind of normalization. It influences how two probability measures have to be summed to become again a probability measure.

In Figure A.1 the probability measure is visualized as a mapping from the probability space to the unit interval $[0, 1]$.

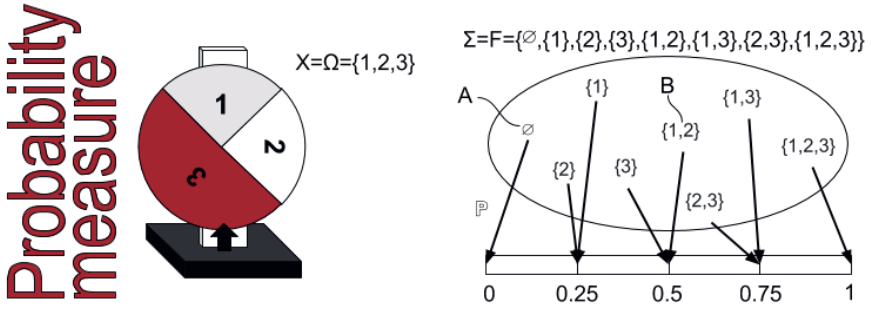

Figure A.1: A probability measure $\mathbb{P}$ mapping the probability space for 3 events to the unit interval. Left: a turning wheel representing three possible outcomes of which the third is twice as likely as the other two outcomes. Right: a probability measure $\mathbb{P}$ assigned to each outcome. The empty set, $A = \emptyset$, has probability measure 0. The set of encountering either 1 or 2, $B = \{1, 2\}$, has probability measure 0.5. Taken from Wikipedia.

A *probability space* $(X, \Sigma, \mathbb{P})$ is a measure space (X, Σ, μ) with the probability measure $\mathbb{P}$ as its measure μ.

▼ **Definition A.8** — ***probability space***

A **probability space** $(X, \Sigma, \mathbb{P})$ is a triple with:

- X a set;
- Σ a σ-algebra over X;
- $\mathbb{P}$ a probability measure from Σ to $[0, 1]$.

We will equivalently use the symbols $(X, \Sigma, \mathbb{P})$ or $(\Omega, \mathbb{F}, \mathbb{P})$ for the probability space, also called probability triple (Rosenthal, 2006). The space X is the event space Ω, the set of *elementary outcomes*. The σ-algebra over subsets of Ω is denoted by $\mathbb{F}$. The probability measure $\mathbb{P}$ assigns a value on the unit interval $[0, 1]$ to every event in $\mathbb{F}$.

A.1.2 Counting Measure

The *counting measure* forms the basis for the definition of discrete probabilities (Schilling, 2005).

▼ **Definition A.9** — ***counting measure***

A **counting measure** ν on a space X is a measure μ with:

- ν is non-negative and integer-valued for $\forall A \in \Sigma$;
- $\nu < \infty$ for $\forall A \in \Sigma$ if A bounded (of finite size);

- $\nu = \infty$ if $\exists A \in \Sigma$ with A unbounded (infinite).

A counting measure is a measure that is integer-valued. Every set A has a measure that is a positive integer or zero. The set A is unbounded if and only if its counting measure is infinite.

A.1.3 Borel Measure

The *Borel σ-algebra* defines a σ-algebra for the real line $\mathbb{R}$.

▼ Definition A.10 — *Borel σ-algebra*

A **Borel σ-algebra** $\mathbb{B}_\sigma$ on $\mathbb{R}$ is the smallest σ-algebra that contains all open subsets of $\mathbb{R}$:

- $\mathbb{B} = \Sigma(U)$ with $U = U \subseteq \mathbb{R}$: U is open.

The Borel σ-algebra contains all open subsets of $\mathbb{R}$. The property of closure under complementation of a σ-algebra means that it also contains the closed subsets of $\mathbb{R}$. If $A = (0,1)$, then $A^c = \{[-\infty, 0], [1, \infty]\}$.

A *Borel measure* assigns values to subsets of $\mathbb{B}_\sigma$.

▼ Definition A.11 — *Borel measure*

A **Borel measure** μ is a function from $\Sigma = \mathbb{B}_\sigma$ to $[-\infty, +\infty]$, with the three measure requirements:

- μ is non-negative: $\mu(A) \geq 0$ for $\forall A \in \Sigma$;
- μ has a null empty set: $\mu(\varnothing) = 0$;
- μ is σ-additive: $\mu(\bigcup_{i \in I_\Sigma} A_i) = \sum_{i \in I_\Sigma} \mu(A_i)$ for A_i disjoint.

The *Borel space* is a measureable space with a Borel σ-algebra rather than a general σ-algebra.

▼ Definition A.12 — *Borel space*

A **Borel space** $(X, \mathbb{B}_\sigma)$ is a pair with:

- X a set;
- $\mathbb{B}_\sigma$ a Borel σ-algebra over X.

A *complete measure space* is a measure space in which every subset of every null set is measurable.

▼ **Definition A.13 — *complete measure space***

A **complete measure space** (X, Σ, μ):

- $S \subseteq N \in \Sigma$ and $\mu(N) = 0 \Rightarrow S \in \Sigma$.

The Borel space is not a complete measure space. There are sets in the Borel σ-algebra that are of measure zero and that contain subsets that are undefined.

A.1.4 Lebesgue Measure

The *Lebesgue measure* defines a size to subsets of $\mathbb{R}^n$ that completes the Borel measure (Lebesgue, 1902). It makes use of the notion of an *outer measure*.

▼ **Definition A.14 — *outer measure***

An **outer measure** ϕ on a space $\mathbb{R}$ is a measure μ with:

- ϕ is non-negative and real-valued for $\forall A \in \Sigma$;
- ϕ has a null empty set: $\phi(\varnothing) = 0$;
- ϕ is σ-subadditive: $\phi(\bigcup_{i \in I_\Sigma} A_i) < \sum_{i \in I_\Sigma} \mu(A_i)$ for $\forall A_i$;
- ϕ is monotone: $A \subseteq B$ implies $\phi(A) \leq \phi(B)$;
- ϕ is translation-invariant: $\phi(A + x) = \phi(A)$ for $\forall A \in \Sigma$ and $\forall x \in \mathbb{R}$.

An outer measure relaxes σ-additivity of disjoint sets of X to σ-subadditivity for any sequence of sets. Intuitively, the outer measure of a set is an upper bound on the size of a set.

▼ **Definition A.15 — *Lebesgue outer measure***

A **Lebesgue outer measure** λ on a space $\mathbb{R}^n$ is an outer measure ϕ with:

- $\lambda(A) = \inf\left\{\sum_{k=1}^{\infty} l(I_k) : (I_k)_{k \in \mathbb{N}} \text{ is a sequence of open intervals with } A \subseteq \bigcup_{k=1}^{\infty} I_k\right\}$.

Here $A \subseteq \mathbb{R}$ is a subset of the real line. The Lebesgue outer measure λ is the infimum (greatest lower bound) of the sum of the lengths $l(I) = b - a$ of the intervals $I = [a, b]$.

The *Lebesgue measure* is defined through the Lebesgue outer measure.

▼ **Definition A.16 — *Lebesgue measure***

A **Lebesgue measure** m on a space $\mathbb{R}^n$ is a Lebesgue outer measure λ with:

- $m(B) = \lambda(B \cup A) + \lambda(B \cup A^c)$.

A.1.5 Random Measure and Random Process

A measurable function is defined between two measurable spaces.

▼ **Definition A.17** — ***measurable function***

A **measurable function** $f : X \to Y$ fulfills:

- $f^{-1}(E) \in \Sigma \qquad \text{for} \qquad \forall E \in T,$

with both (X, Σ) and (Y, T) measurable spaces.

A measurable function *preserves the structure* of the corresponding measurable spaces (captured through the σ-algebras).

A random element or (X, Σ)-valued random variable is a measurable function between two measurable spaces, with as domain a measurable space that is a probability space.

▼ **Definition A.18** — ***random element***

A **random element** or **(X, Σ)-valued random variable** X is a measurable function from probability space $(\Omega, \mathbb{F}, \mathbb{P})$ to measurable space (X, Σ).

An (X, Σ)-valued random variable is visualized in Figure A.2.

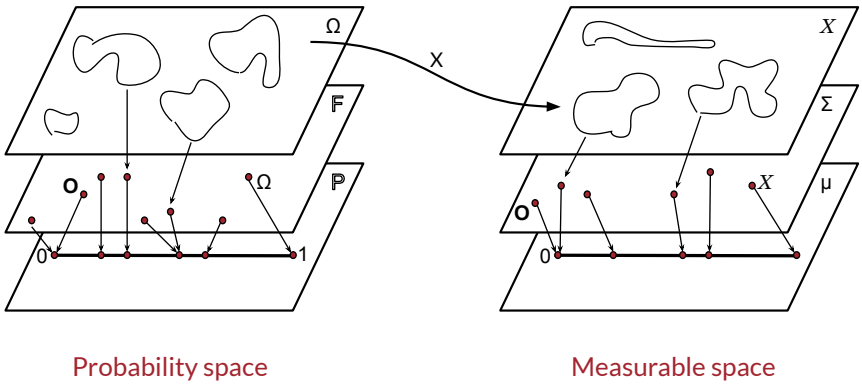

Figure A.2: An (X, Σ)-valued random variable X is a measurable function from $(\Omega, \mathbb{F}, \mathbb{P})$ (at the left) to (X, Σ) (at the right). The planes at the top depict the samples spaces Ω and X. The planes in the middle depict the σ-algebras $\mathbb{F}$ and Σ. The planes at the bottom depict measures: at the left $\mathbb{P}$, and at the right an induced measure μ. The null set is of measure 0. The set Ω is a of measure 1.

For random variables for which we do not specify the codomain explicitly, the choice for the codomain is the real line $\mathbb{R}$ and the corresponding Borel σ-algebra on the reals.

▼ **Definition A.19 — *random variable***

A **random variable** X is a measurable function from probability space $(\Omega, \mathbb{F}, \mathbb{P})$ to the real line with the Borel σ-algebra $(\mathbb{R}, \mathbb{B}_{\mathbb{R}})$.

A $(\mathbb{R}, \mathbb{B}_{\mathbb{R}})$-valued random variable is also called a real-valued random variable assuming a natural choice for the σ-algebra, or called a random variable assuming the reals.

> Random elements are a generalization of random variables. A *complex*-valued random variable or *complex random variable* is a measurable function from Ω to $\mathbb{C}$. An *elephant*-valued random variable or *random elephant* is a measurable function from Ω to a suitable space of elephants (Kingman, 1993).

This allows us to define a *measure*-valued random variable, a random measure.

▼ **Definition A.20 — *random measure***

A **random measure** is a function $\xi : \Omega \times X \to [0, +\infty]$ from probability space $(\Omega, \mathbb{F}, \mathbb{P})$ to measurable space (X, Σ) such that $\xi(\cdot, X)$ is a random variable on $(\Omega, \mathbb{F}, \mathbb{P})$ and $\xi(\omega, \cdot)$ is a measure on Σ.

We are now in the position to define a random process (the Dirichlet process in this thesis is an example of such a process). A *random process* is an ordered set of random variables. The set can be a sequence of random variables in a time series. It can be a series of steps in the spatial domain, called a random field.

▼ **Definition A.21 — *random process***

A **random process** X is a collection $\{X_t : t \in T\}$ with X_t an (S, Σ)-valued random variable on Ω and $(\Omega, \mathbb{F}, \mathbb{P})$ a probability space, (S, Σ) a measurable space, and T a totally *ordered* set.

A random process is a probability distribution with a domain that is a set of probability distributions. A random process is a distribution over distributions, a hierarchy over distribution.

Before we close this section, we will introduce two more concepts. The *distribution* of a random variable and the *probability density function* of a random variable.

We have encountered a random variable, and a probability measure $\mathbb{P}$ on the original probability space. Now, one might wonder whether there is a measure that is logically assigned to elements on the measurable space that is the codomain of this random variable. Is there a natural measure μ that can transform this measurable space into a measure space? It turns

out there is. There is a measure *induced* on this space by the random variable.[1] This measure μ is known as the *distribution* or *law* of a random variable (Rosenthal, 2006):

▼ **Definition A.22 — *distribution of a random variable***

Given a random variable X from $(\Omega, \mathbb{F}, \mathbb{P})$ to $(\mathbb{R}, \mathbb{B}_\sigma)$, the **distribution** μ of X is the induced probability measure: $\mu(B) = \mathbb{P}(X^{-1}(B))$ for all Borel sets $B \in \mathbb{B}_\sigma$.

The distribution of X is the probability measure μ induced on $(\mathbb{R}, \mathbb{B}_\sigma)$. This makes this space a measurable space $(\mathbb{R}, \mathbb{B}_\sigma, \mu)$. We will write X as being *distributed as* μ in the following shorthand notation:

$$X \sim \mu. \tag{A.1}$$

A measure ν is *absolutely continuous* with respect to a measure λ if, for every set E, $\lambda(E) = 0$ implies $\nu(E) = 0$. We also write this as $\nu \ll \lambda$. The measure ν is *dominated* by λ. The Radon-Nikodym theorem states that for two σ-finite measures one measure can be expressed as an integral of the other.

▼ **Definition A.23 — *Radon-Nikodym theorem***

The **Radon-Nikodym theorem** states that given a measurable space (X, Σ) and two σ-finite measures, ν, λ with $\nu \ll \lambda$, that there exists a Σ-measurable function $f : X \to [0, \infty)$, such that for any measurable set $A \subseteq X$,

$$\nu(A) = \int_A f \, d\lambda. \tag{A.2}$$

This allows us to define the Radon-Nikodym derivative: $f = \frac{d\nu}{d\lambda}$. The probability density function f of a random variable X is the Radon-Nikodym derivative of the induced measure (with respect to some base measure, normally the Lebesgue measure).

▼ **Definition A.24 — *probability density function***

The **probability density function** f of a random variable X is the Radon-Nikodym derivative of the induced measure μ on $(\mathbb{R}, \mathbb{B}_\sigma)$ with respect to a base measure λ,

$$f = \frac{d\mu}{d\lambda}. \tag{A.3}$$

For a discrete random variable the counting measure can be used as a base measure. For continuous random variables the Lebesgue measure is usually chosen as base measure.

[1] The measure induced on a measurable space by another measurable space by means of a measurable function is also known as a push-forward measure.

A.2 Bayesian Inference

Let x be a (S, Σ_S, μ_S)-valued random variable, y a (T, Σ_T, μ_T)-valued random variable, then we can construct z, a (C, Σ_C, μ_C)-valued random variable with the latter being a subset of the product set of x and y: $C \in S \otimes T$.

▼ **Definition A.25 — *product space***

A **product space** $(S \otimes T, \Sigma_{S\otimes T})$ has σ-algebra $\Sigma_{S\otimes T} = \sigma(F \otimes G : F \in \Sigma_S, G \in \Sigma_T)$ with (S, Σ_S, μ_S) and (T, Σ_T, μ_T) two σ-finite measure spaces.

▼ **Definition A.26 — *product measure***

A **product measure** $\mu_{S\otimes T}$ is a measure $\mu_{S\otimes T}(F \otimes G) = \mu_S(F) \otimes \mu_T(G)$ with (S, Σ_S, μ_S) and (T, Σ_T, μ_T) two σ-finite measure spaces.

The **joint probability distribution** P_C is a probability measure on the product σ-algebra Σ_C with $C \in S \otimes T$. As function of the random variables x and y the joint probability distribution is written as $_{X,Y}(x, y)$, $f(x, y)$, or $p(x, y)$.

A σ-algebra is *independent* in the following sense.

▼ **Definition A.27 — *independent σ-algebra***

Let $(\Omega, \mathbb{F}, P)$ be a probability space and $\mathbb{A}$ and $\mathbb{B}$ be a sub-σ-algebras of $\mathbb{F}$. $\mathbb{A}$ and $\mathbb{B}$ are **independent σ-algebras** if:

- $P(A \cap B) = P(A)P(B)$ $\forall A \in \mathbb{A}$ and $B \in \mathbb{B}$.

Two random variables x and y are independent if and only if the σ-algebras that they generate are independent.

▼ **Definition A.28 — *conditional probability distribution***

Let $(\Omega, \mathbb{F}, P)$ be a probability space, $\mathbb{G} \subseteq \mathbb{F}$ a sub-σ-algebra of $\mathbb{F}$, and $X : \Omega \to \mathbb{R}$ a real-valued random variable ($\mathbb{F}$-measurable with respect to the Borel σ-algebra $\mathbb{B}_\sigma$ on $\mathbb{R}$). There exists a function $\mu : \mathbb{B}_\sigma \times \Omega \to \mathbb{R}$ such that $\mu(\cdot, \omega)$ is a probability measure on $\mathbb{B}_\sigma$ for each $\omega \in \Omega$ and $\mu(H, \cdot) = P(X \in H|\mathbb{G})$ (almost surely) for every $H \in \mathbb{B}_\sigma$. For any $\omega \in \Omega$, the function $\mu(\cdot, \omega) : \mathbb{B}_\sigma \to \mathbb{R}$ is called a **conditional probability distribution** of X given $\mathbb{G}$.

Informally[2], a conditional probability is described with a sub-σ-algebra which only presents part of the structure of the full σ-algebra. As function of the random variables x and y the conditional probability distribution of y given x is written as $f_{Y|X}(y|x)$, $f(y|x)$, or $p(y|x)$.

A typical conditional probability distribution is that of the data given parameters. Another often used conditional probability distribution is that of the data given a statistic (summary) of that data. This statistic can be a so-called sufficient statistic.

▼ Definition A.29 — *sufficient statistic*

A conditional probability distribution of the data X given a *sufficient statistic* $t = T(X)$ does not depend on parameter θ:

- $P(x|t,\theta) = P(x|t)$

Random variables, or more generally, random elements x and θ define a Bayesian[3] model with observations x and parameters θ.

▼ Definition A.30 — *Bayesian model*

A **Bayesian model** $f(x,\theta)$ defines a function, a joint probability distribution, over observations x and parameters θ with both x and θ random elements.

In a **supervised learning** task both x and θ are known. In an **unsupervised learning** task x is known, but θ is unknown. The random variable θ is called a hidden or latent variable. The random variable θ can be any random element: a random vector, a random matrix, a random process.

Let the observations x be a sequence $x_0, x_1, \ldots$, then the observations x_i can be *independent and identically* distributed.

▼ Definition A.31 — *independent and identically distributed*

A collection of random variables $x = \{x_0, x_1, \ldots\}$ is **independent and identically distributed (i.i.d.)** if:

- the probability distribution $p(x_i)$ is the same for $\forall x_i \in x$
- each x_i is independent with respect to x_j with $i \neq j$.

In other words, random variables having the same distribution are said to be identically distributed.

[2] Even more informally, in "254A, Notes 0: A review of probability theory" Tao describes how conditioning can be seen as removing a *partial* amount of randomness consistent with the probabilistic way of thinking. By conditioning a random variable to be fixed, one can turn that random variable into a deterministic one, while preserving the random nature of other variables.

[3] A historic perspective on the term Bayesian can be found in (Fienberg et al., 2006).

The observations x_i can be distributed in an *exchangeable* sequence in which any order is equally likely.

▼ **Definition A.32 — *exchangeable***

A sequence of random variables $x = \{x_0, x_1, \ldots\}$ is **exchangeable** if for any finite permutation ρ of the indices $0, 1, \ldots$:

- the joint probability distribution of the permuted sequence $p(x_{\rho(0)}, x_{\rho(1)}, \ldots)$ equals that of the original sequence $p(x_0, x_1, \ldots)$.

The joint probability distribution of i.i.d. observations given parameters can be written as a product:

$$p(x_0, \ldots, x_{k-1}|\theta) = \prod_{i=0}^{k-1} p(x_i|\theta). \tag{A.4}$$

▼ **Definition A.33 — *likelihood function***

The **likelihood function** is defined as:

$$L(\theta; x) = p(x = X|\theta). \tag{A.5}$$

The likelihood indicates the probability that a particular value $x = X$ is observed when the parameter is considered to be θ.

The likelihood function allows us to find an optimal set of parameter values given the observations. We can find those parameters that maximize the likelihood, $L(\theta)$, given the observations, X, see Aldrich (1997). This maximization method is called maximum likelihood estimation (MLE).

▼ **Definition A.34 — *maximum likelihood estimation***

Maximum likelihood estimation is defined as the method optimizing:

$$\theta^* \in \underset{\theta}{\operatorname{argmax}}\, L(\theta; x). \tag{A.6}$$

The maximum likelihood method finds the maximum of $p(x|\theta)$ for all possible parameter values θ. The maximum in maximum likelihood estimation does not need to be unique (Steel, 1994). The notation makes this explicit by writing θ^* as a member (denoted by the $\in$ symbol) of the outcomes of the argmax operation (and does not use the equal sign).

In the case we have information about the parameters θ we can model this with a probability distribution.

▼ **Definition A.35 — *prior probability distribution***

A **prior probability distribution** defines a probability distribution $p(\theta)$ over parameters θ without a dependency on the observations x.

Given the definition of a prior probability distribution, we can define *maximum a posteriori* estimation.

▼ **Definition A.36 — *maximum a posteriori***

Maximum a posteriori estimation is defined as:

$$\theta^* \in \underset{\theta}{\operatorname{argmax}} \sum_{i=0}^{k-1} \log p(x_i|\theta) + \log p(\theta). \tag{A.7}$$

If we are not only interested in the parameter θ^* that maximizes $p(x|\theta)$ and $p(\theta)$, but in the complete distribution for $p(\theta)$ we need Bayes' theorem described by Laplace (1820).

▼ **Definition A.37 — *Bayesian inference***

Bayesian inference using Bayes' theorem is defined as:

$$p(\theta|x) = \frac{\overbrace{p(x|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(x)}_{\text{normalization constant}}} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}. \tag{A.8}$$

Bayes' theorem describes the posterior probability $p(\theta|x)$ as the likelihood times the prior probability distribution divided by a normalization constant, also called the evidence. The normalization constant is not a function of the parameters θ. If a function is known except for the normalization constant, it is indicated by the "proportional to" symbol $\propto$.

$$f(\theta|x) \propto p(x|\theta)p(\theta) \tag{A.9}$$

In Bayesian inference $p(\theta|x)$ is calculated. In contrast, in maximum likelihood and maximum a posteriori only parts of Eq. A.8 are calculated, respectively $p(x|\theta)$ and $p(x|\theta)p(\theta)$. In Section 2.2 inference methods will be described that approximate Bayesian inference. Approximation is required in the case closed-form expressions are not available. If the inference task only requires maximum a posteriori, approximation methods are also available (Daume, 2007), but this is outside of the scope of the current thesis.

It is important to note that Bayes' rule does not always apply. Recall the definition of the probability density function (Definition A.24) in Appendix A.1.5 for which we needed the notion of absolutely continuity. The posterior is not always absolutely continuous with respect to the prior. In particular for nonparametric Bayesian models this is not necessarily

the case. For example, the Dirichlet process as a prior has a posterior that is typically orthogonal to the prior. However, using appropriate care it is still the case that the posterior is well-defined and one can perform Bayesian inference without using Bayes' theorem. To read more on the exact conditions under which this is possible, we refer the reader to (Ghosal and Van der Vaart, 2017).

There are two supervised learning models, a generative model and a discriminative model. Below we provide their definitions and in Figure A.3 we give three examples for each model.

▼ **Definition A.38** — ***generative model***

A **generative** model defines the joint probability distribution $p(x, \theta)$.

▼ **Definition A.39** — ***discriminative model***

A **discriminative** model defines the conditional probability distribution $p(\theta|x)$ directly.

Figure A.3 shows three generative and three discriminative models. They are chosen for their structure. From left to right, the structure between the random variables gets enriched. The first column shows no particular structure. The second column shows a sequence structure. The third column shows a graph structure. Figure A.3 visualizes three generative models: (1) the Naive Bayes Model (Russell et al., 1995), (2) the Hidden Markov Model (Baum and Petrie, 1966), and (3) the Directional Model (Koller and Friedman, 2009). It shows also three discriminative models: (1) Logistic Regression, (2) Linear-chain Conditional Random Fields, and (3) general Conditional Random Fields.

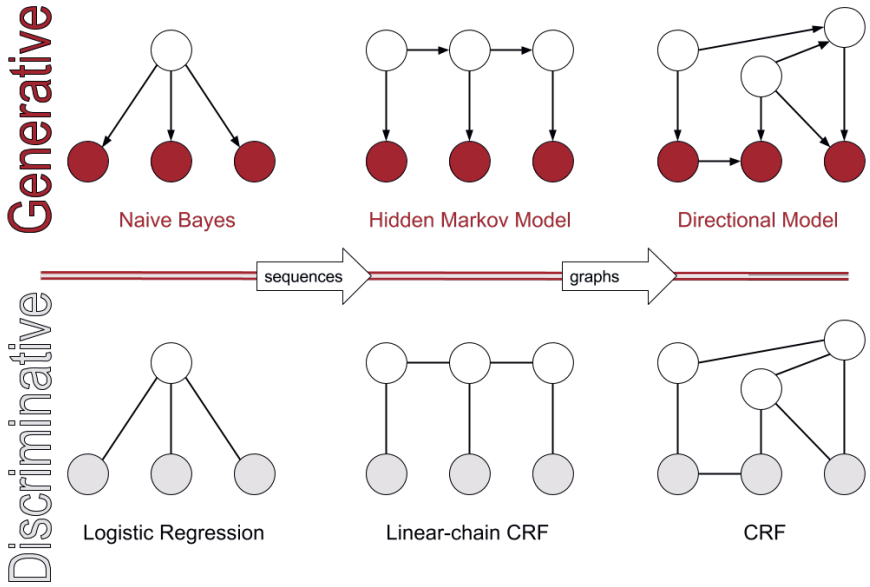

Figure A.3: Generative models: Naive Bayes Model, Hidden Markov Model, and Directional Model. Discriminative models: Logistic Regression, Linear-chain Conditional Random Fields, and general Conditional Random Fields. Figure adapted from Sutton and McCallum (2011).

There is no definitive reason to use a generative model rather than a discriminative model or vice-versa. Here we confine ourselves to two remarks. First, a discriminative model seems to have a lower asymptotic error, but a generative model seems to approaches its (higher) asymptotic error faster. This has been studied using a Naive Bayes classifier versus Logistic Regression (Jordan, 2002). This would mean that a discriminative model would be better for large datasets, while a generative model would be better for small datasets. However, Xue and Titterington (2008) doubt the existence of such precisely defined regimes depending on dataset size. According to them the behaviour seems to stem from the correctness of the conditional or the joint model specification. Second, the prior $p(\theta)$ in the generative model provides a principled way to handle missing information, while the direct modeling of decision boundaries in a discriminative model often leads to better performance in a classification task (Jaakkola et al., 1999). Apart from generative models and discriminative models, there are also hybrid models (Bouchard and Triggs, 2004; Raina et al., 2003; Bosch et al., 2008). In the thesis we will restrict ourselves to generative models.

A.3 Model Composition

A model can be composed out of a set of probability distibutions. We list three of such possible compositions. The Naive Bayes model is a *product* of probability distributions with a prior distribution (Definition A.40). The finite mixture model is a *sum* over a finite number

of probability distributions where each one is weighted (Definition A.41). The infinite mixture model is a *sum* over an infinite number of probability distributions where each one is weighted (Definition A.42).

▼ **Definition A.40 — *naive Bayes model***

The **naive Bayes model** is a product over a finite number $k \neq \infty$ of probability distributions $p(x_i|\theta)$ multiplied by the prior distribution $p(\theta)$:

$$p(\theta|x) \propto p(\theta)\prod_{i=0}^{k-1} p(x_i|\theta). \tag{A.10}$$

A finite mixture model is a sum over a finite number of probability distributions.

▼ **Definition A.41 — *finite mixture model***

A **finite mixture model** is a sum over a finite number $k \neq \infty$ of probability distributions $p(x_i)$, with each distribution weighted by a factor w_i with $\sum_i w_i = 1$.

$$p(x) = \sum_{i=0}^{k-1} w_i p(x_i). \tag{A.11}$$

The mixture model is finite in the sense that there are only $k \neq \infty$ distributions summed up. The weights of the individual distributions $p(x_i)$ are normalized (sum up to one) such that the weighted sum over the probability distributions is itself a probability distribution.

An infinite mixture model is a sum over an infinite number of probability distributions.

▼ **Definition A.42 — *infinite mixture model***

A **infinite mixture model** is a sum over an infinite number of probability distributions $p(x_i)$, with each distribution weighted by a factor w_i with $\sum_i w_i = 1$.

$$p(x) = \sum_{i=0}^{\infty} w_i p(x_i). \tag{A.12}$$

The infinite mixture model is a sum over an infinite number of probability distributions with weights that sum up to one. In this way it assigns a finite value to a countably infinite set of functions.

If the number of probability distributions is uncountable infinite, we speak about a compound distribution.

▼ **Definition A.43 — *compound probability distribution***

A **compound probability distribution** for a probability density function $p(\theta)$ (nonnegative and integrating to 1) is given by

$$p(x) = \int_{\Omega} p(\theta)p(x|\theta)d\theta. \tag{A.13}$$

Informally, $p(\theta)$ has the same function as the weight in a mixture model. From this presentation it is also clear that a compound distribution is a special case of a marginal distribution. The joint distribution $p(x,\theta) = p(\theta)p(x|\theta)$. The compound distribution is obtained through its marginal distribution: $\int p(x,\theta)d\theta$. In the thesis we will encounter infinite mixture models or compound probability distributions in Chapters 3 and 4.

A.4 General Random Elements

In section A.1 random elements were described in general. Random elements can vary from random vectors, random distributions, random clusters (partitions), to random trees. Table A.1 describes the random elements and the corresponding examples of random processes in the literature. Below we mention them with the appropriate references.

Table A.1: A list of seven mathematical structures and for each of these structures one or more random processes that can generate the structure. For example, a distribution on distributions can be generated by a Beta Process, Gamma Process, Dirichlet Process, or a Polya Tree.

Structure	Example
Distribution on functions	Gaussian Process
Distribution on distributions	Beta Process Gamma Process Dirichlet Process Polya Tree
Distribution on partition assignments	Chinese Restaurant Process Pitman-Yor Process
Distribution on partition sizes	Stick-breaking Process
Distribution on hierarchical partitions	Dirichlet Diffusion Tree Kingman's coalescence
Distribution on sparse binary matrices	Indian Buffet Process
Distribution on integer-valued matrices	Gamma-Poisson Process
Distribution on kd-trees	Mondrian Process

The Gaussian Process (Rasmussen and Williams, 2006) describes a distribution on functions. The Beta Process (Hjort, 1990), the Gamma Process (Ferguson, 1974), the Dirichlet Process and the Polya Tree (Ferguson, 1973) describe a distribution on distributions. The Chinese Restaurant Process (Aldous, 1985) and Pitman-Yor Process (Pitman and Yor, 1997) describe

a distribution on partitions (in the form of cluster assignments). The Stick-breaking Process describes a distribution on partition sizes (with no information on assignments themselves). The Dirichlet Diffusion Tree (Neal, 2001) and Kingman's coalescence (Kingman, 1965) describe a distribution on hierarchical partitions. The Indian Buffet Process (Ghahramani and Griffiths, 2005) describes a distribution over sparse binary matrices. The Gamma-Poisson Process (Titsias, 2008) describes a distribution over integer-valued matrices. The Mondrian Process (Roy and Teh, 2009) describes a distribution over kd-trees.

A.5 Plate Notation

Random processes and mixture models are visually represented by a method called *plate notation* (cf. Buntine, 1994; Koller and Friedman, 2009). Sets of variables are represented in a plate, a rectangular region (see Figure A.4).

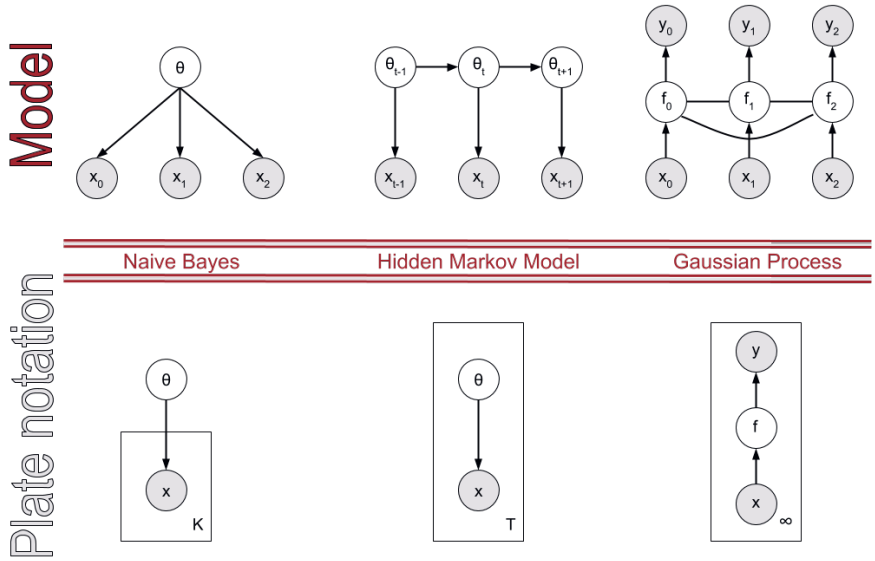

Figure A.4: Top: graphical model of a Naive Bayes, hidden Markov model, and Gaussian process. Bottom: corresponding plate notation of the Naive Bayes, hidden Markov model, and Gaussian process. Observed variables are denoted by a circle that is shaded.

Plate notation is a representation that does not preserve all dependencies between variables. For example, the dependencies between the states in the Hidden Markov Model (e.g., between θ_0 and θ_1) are not represented. The Gaussian process has a potentially infinite number of parameters. The use of plate notation for nonparametric models can be found in (Fox et al., 2007).

A.6 Completely Random Measure and Lévy Measure

Some random process are mathematically represented by a completely random measure (Kingman, 1967), which is defined as follows.

▼ **Definition A.44** — ***completely random measure***

A **completely random measure** is a random measure $\mu : \Omega \times X \to [0, +\infty]$ from probability space $(\Omega, \mathbb{F}, \mathbb{P})$ to measurable space (X, Σ) with

- for any collection of disjoint sets $A_1, \ldots, A_k \in \Sigma$ and $A_i \cap A_j = \emptyset$ for $i \neq j$ a mutual independency between $\mu(A_1), \ldots, \mu(A_k)$.

Kingman (1967) shows that a completely random measure can be decomposed into three components:

1. a deterministic function;
2. a countable set of non-negative random masses at deterministic locations;
3. a countable set of non-negative random masses at random locations.

The first component is a deterministic function. The second component has non-negative random masses, also called atoms, on deterministic locations. The third component is the one of interest. It has a set of random masses (atoms) that can be represented as a Poisson random measure on $\mathbb{R}^+ \otimes X$ with mean measure ν which is known as the Lévy intensity measure (Favaro et al., 2013).

Table A.2: Lévy measure of the Beta Process (Wang and Carin, 2012), Gamma Process (Knowles et al., 2014), the Dirichlet Process (Lijoi and Prünster, 2010) (indirectly through $F = 1 - e^{-\nu}$) .

Random Process	Lévy measure
Beta Process	$\nu(da, dw) = H(da)\alpha w^{-1}(1-w)^{\alpha-1}dw$
Gamma Process	$\nu(da, dw) = H(da)w^{-1}e^{-\alpha w}dw$
Dirichlet Process	$\nu(da, dw) = H(da)e^{-w\alpha(x,\infty)}(1-e^{-w})^{-1}dw$

For Lévy measure decompositions of other processes such as the Indian buffet process, we refer to Wang and Carin (2012).

A.7 Exchangeability

Here we recall Definition A.32 for exchangeable sequences. De Finetti's theorem states that there is parameter θ such that the data x_i is conditionally independent given this parameter for exchangeable sequences (cf. De Finetti, 1937).

▼ **Definition A.45 — *De Finetti's theorem***

A sequence $\{x_0, x_1, \ldots\}$ of (X, Σ_X)-valued random variables is an infinitely exchangeable sequence if and only if there exist a measure $\mu(d\theta)$ on θ such that

$$p(x_0, \ldots, x_{k-1}) = \int_\Omega \prod_{i=0}^{k-1} p(x_i|\theta)\mu(d\theta) \qquad \forall k \geq 1. \tag{A.14}$$

In words, de Finetti's theorem states that if we have *exchangeable* data, we have a parameter θ, a likelihood $p(x|\theta)$, and some measure μ on θ, such that the data $(x_0, \ldots, x_{k-1})$ is *conditionally independent*. Hence, although the data is not i.i.d., there are underlying, unobservable, quantities that are i.i.d. and exchangeable sequences are mixtures of these quantities. The theorem proofs that if the observations are exchangeable, they must be a random sample from some model and there must exist a prior probability distribution over the parameters of that model, hence requiring a Bayesian approach.

The theorem is not limited to exchangeable *sequences*. In contrast, there are similar theorems for other exchangeable objects (Orbanz and Roy, 2015). Five examples (see Table A.3) of exchangeable structures have a theorem describing an underlying measure that can be sampled i.i.d. are: (1) exchangeable sequences (De Finetti, 1930), (2) increments (Bühlmann, 1960), (3) partitions (Kingman, 1978), (4) arrays (Aldous, 1981), and (5) Markov chains (Diaconis and Freedman, 1980).

Table A.3: Five exchangeable structures and their theorems.

Mathematical Object	Theorem
Exchangeable Sequence	de Finetti
Exchangeable Increment	Bühlmann
Exchangeable Partition	Kingman
Exchangeable Array	Aldous-Hoover
Exchangeable Markov Chain	Diaconis-Freedman

A.8 Stick-breaking Representation

Below we introduce the *stick-breaking representation* by Freedman and Diaconis (1983), also known as the residual allocation model (Sawyer and Hartl, 1985; Hoppe, 1986).

▼ **Definition A.46 — *stick-breaking***

An infinite sequence of random variables $\phi = \{\phi_0, \phi_1, \ldots\}$ has a **stick-breaking representation** with parameters α and β denoted by $\phi \sim GEM(\alpha, \beta)$.

$$w_k \overset{iid}{\sim} Beta(1-\beta, \alpha + k\beta) \qquad k = 1, \ldots, K \tag{A.15}$$

$$\phi_k = w_k \prod_{i=1}^{k-1} (1 - w_i) \tag{A.16}$$

The stick-breaking process samples repeatedly from a $Beta(1-\beta, \alpha+k\beta)$ distribution. The result of the process is a vector of k weights ϕ_k. The abbrevation *GEM* stands for Griffiths, Engen, and McCloskey (Ewens, 1990; Ethier, 1990). There is also a variant of GEM with a single parameter α which can be obtained by setting $\beta = 0$. In that case w_k are drawn from a $Beta(1, \alpha)$ distribution. Note that although w_k are sampled i.i.d., the resulting stick sizes ϕ_k are not independent. Stick size ϕ_k depends not only on w_k, but also on the weights $w_1, \ldots w_{k-1}$ drawn previously.

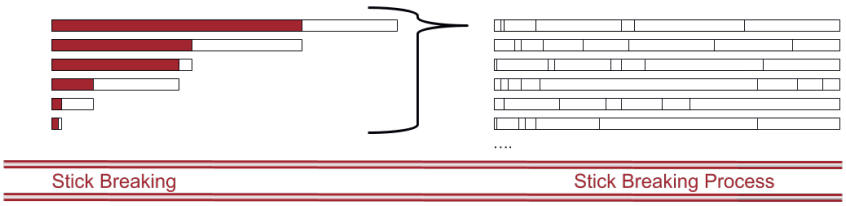

Figure A.5: The stick-breaking representation. Left: at the first row, the stick is broken at x_0, at the next rows the remaining part of the stick is broken x_i with $i > 0$. Only six iterations are shown. Right: samples of a stick-breaking process. The first row shows the stick ratios from the stick-breaking representation at the left. The next rows show other samples from the same process.

Figure A.5 visualizes the stick-breaking process. A stick of fixed length 1 gets broken at a position w_0 sampled from a Beta distribution. The remainder of the stick is broken again at position $w_1(1 - w_0)$. This process continues for an infinite number of times. A stick-breaking process generates in this manner a sequence of non-negative values that sum up to one. The stick-breaking representation can on itself give rise to more sophisticated stochastic processes (Dunson et al., 2012). Computationally it can also fulfill a useful role. Namely, it is possible to approximate a distribution over partitions by truncating a stick-breaking process. The stick-breaking procedure is then only performed a limited number of times (Kurihara et al., 2007).

In Section 2.1 the relevance of the stick-breaking process for the Dirichlet process will be shown. In that case the values generated by the stick-breaking process represent the weights of the partitions induced by the Dirichlet Process.

APPENDIX B

IMPLEMENTATION

We describe two initialization algorithms. The first algorithm initializes Gibbs sampling over parameters. The second algorithm initializes Gibbs sampling over clusters.

B.1 Initialization of Gibbs Sampling over Parameters

Algorithm 8 as shown in Section 3.3 does not describe how the parameters are initialized. The algorithm to initialize the parameters θ_i is given in Algorithm 15.

Algorithm 15 Gibbs sampling over parameters. The initialization of θ_i.

procedure GIBBS ALGORITHM 1 INITIALIZATION(w, λ_0, α) ▷ Accepts points w, hyperparameters λ_0, α and returns k initial line coordinates
  $\lambda_1 = U_{up}(w_1, \lambda_0)$ ▷ Update hyperparameter with w_1 (Eq. 3.21)
  $\theta_1 \sim NIG(\lambda_1)$ ▷ Sample θ_1 from NIG (Eq. 3.24)
  for all $i = 2 : N$ **do**
    $M = i - 1$ ▷ Let M define the number of parameters assigned up to now
    $r_i = \alpha \int F(w_i; \theta) dH$ ▷ Weighted posterior predictive of w_i (Eq. 3.29)
    for all $j = 1 : M$ **do**
      $L_{i,j} = F(w_i; \theta_j)$ ▷ Likelihood of a line given an observation (Eq. 3.9)
    end for
    $p(\theta_{new}) = \frac{r_i}{r_i + \sum_i L_{i,j}}$ ▷ Probability of sampling a new parameter (Eq. 3.31)
    $u \sim U(0, 1)$
    if $p(\theta_{new}) > u$ **then** ▷ Sample with probability $p(\theta_{new})$
      $\lambda_n = U_{up}(w_i, \lambda_0)$ ▷ Update hyperparameters with w_i (Eq. 3.21)
      $\theta_i \sim NIG(\lambda_n)$ ▷ Sample θ_i from NIG (Eq. 3.24)
    else
      $i \sim Mult(M, p(\theta_{old}))$ ▷ Sample i from existing parameters, θ_{old}
      $\theta_i = \theta_{old=i}$ ▷ Pick θ_i given index i
    end if
  end for
  return initialized θ_k for k lines
end procedure

Let us recall the posterior predictive Eq. 3.28:

$$\theta_i \mid \theta_{-i}, w_i \sim r_i H_i + \sum_{j \neq i} F(w_i; \theta_j)\delta_{\theta_j}. \tag{B.1}$$

We initialize through:

$$\begin{aligned} \theta_1 \mid w_1 &\sim H_1 \\ \theta_i \mid \theta_1, \dots \theta_{i-1}, w_i &\sim r_i H_i + \sum_{j=1}^{i-1} F(w_i; \theta_j)\delta_{\theta_j}. \end{aligned} \tag{B.2}$$

Given that j runs up to $i-1$, we do not have to specify $i \neq j$ in different lines of the algorithm (compare with Algorithm 8). The initialization algorithm is so similar from the Gibbs sampling algorithm itself, that it is recommended to write the implementation in such a way that the same function can be used.

B.2 Initialization of Gibbs Sampling over Clusters

Algorithm 9 as shown in Section 3.4 requires initialization of the hyperparameters λ_k per cluster k. In contrast to Algorithm 15 we need to initialize not just θ_k, but also the hyperparameters per cluster. This can be done by calling Eq. 3.21 successively by each observation w_i assigned to cluster k. We also require θ_k themselves to calculate $F(w_i; \theta_j)$ for $j \neq i$ in Eq. 3.9 and $p(\theta_{-i})$, or more specific, $p(\theta_{old})$.

Algorithm 16 Gibbs sampling over clusters. The initialization of θ_k and λ_k.

procedure GIBBS ALGORITHM 2 INITIALIZATION(w, λ_0, α) ▷ Accepts points w and hyperparameters λ_0 and α, returns k hyperparameters λ_k and initial parameters θ_k
  for all $k = 1 : K$ **do**
    $m_k = 0$ ▷ Set number of data points per cluster to 0
  end for
  for all $i = 1 : N$ **do**
    $k = U(\{1, \dots, K\})$ ▷ Sample k from discrete uniform distribution
    cluster$(w_i) = k$ ▷ Assign cluster index k to observation w_i
    if $m_k = 0$ **then**
      $\lambda_k = U_{up}(w_i, \lambda_0)$ ▷ Set hyperparameter λ_k with prior pred. given w_i
    else
      $\lambda_k = U_{up}(w_i, \lambda_k)$ ▷ Update hyperparameter λ_k with posterior pred. given w_i
    end if
    $m_k = m_k + 1$
  end for
  for all $k = 1 : K$ **do**
    $\theta_k \sim NIG(\lambda_k)$ ▷ Sample θ_k from NIG with up to date λ_k
  end for
  return initialized parameters θ_k and hyperparameters λ_k for k lines
end procedure

List of Figures

1.1 Examples of point clouds 2

2.1 Dirichlet process 8
2.2 Chinese restaurant process 10
2.3 Matrix representation 10
2.4 Rejection sampling 14
2.5 Gibbs sampling 16

3.1 A mixture of lines 24
3.2 Reintroduction of the Dirichlet process mixture 25
3.3 The Dirichlet process mixture with the realizations integrated out 25
3.4 The Dirichlet process mixture over clusters 26
3.5 The Dirichlet process mixture highlighting the posterior predictive for given parameters 26
3.6 The likelihood of the infinite line model 27
3.7 The conjugate priors of the infinite line model. 28
3.8 The Dirichlet process mixture with focus on posterior predictive 30
3.9 Performance of Algorithm 8 38
3.10 Performance of Algorithm 9 38
3.11 The performance of the Hough transform 39
3.12 Examples of the line estimation process 40
3.13 Examples of incorrect assignments in line estimation 40
3.14 Traceplots of the Markov chains 41

4.1 A mixture of segments 44
4.2 The Dirichlet process mixture reiterated 44
4.3 The conjugate priors of the infinite segment model. 46
4.4 Comparison of segment detection with line detection 49
4.5 Bayesian point estimates with varying types of sampling errors 50
4.6 Traceplots of the Markov chains 51

5.1 A split step in the dyadic sampler 57
5.2 Dyadic versus triadic MCMC 59
5.3 Two examples of fitting a mixture of lines to data points 63
5.4 Comparison of inference methods for line estimation 65

6.1 Architecture of autoencoder and a Bayesian classifier 68
6.2 Variational autoencoder 69
6.3 Variational autoencoder reconstruction 70
6.4 Variational autoencoder scatterplot 70
6.5 Variational autoencoder latent sweep 71
6.6 Ordinary autoencoder reconstruction 72
6.7 Ordinary autoencoder scatterplot 72
6.8 Ordinary autoencoder on 2D lines 73
6.9 Sparse autoencoder reconstruction 73
6.10 Sparse autoencoder scatterplot 74
6.11 Convolutional autoencoder on 2D lines 75
6.12 Reconstruction of a point cloud consisting of multiple cubes 76
6.13 Reconstruction of point clouds using EMD as loss 77
6.14 EMD uniformity explained with grains 78
6.15 Two optimal transportation plans, one non-uniform, the other uniform. . . . 79
6.16 PEMD explained with an example. 80
6.17 Shifted earth mover's distance and partitioning earth mover's distance visualizations with multiple objects 80
6.18 Autoencoder reconstruction using partitioning EMD 82
6.19 Performance of the triadic sampler on 3D cubes 84
6.20 A sample of point to 3D cube assignment of the triadic sampler 85

A.1 Probability measure 103
A.2 Random variable 106
A.3 Generative versus discriminative models 114
A.4 Plate notation 117
A.5 Stick-breaking representation 120

List of Tables

3.1 Contingency table 35

5.1 Clustering performance of the dyadic and triadic sampler 65

6.1 Clustering performance of the triadic sampler on the cube dataset 83

A.1 Structures and Processes 116
A.2 Levy measure 118
A.3 Exchangeable structures 119

List of Abbreviations

ABC	approximate Bayesian computation
CD	Chamfer distance
CRP	Chinese restaurant process
DP	Dirichlet process
DPM	Dirichlet process mixture
EMD	earth mover's distance
GEM	Griffiths, Engen, and McCloskey
HDP	hierarchical Dirichlet process
IBP	Indian buffet process
ILM	infinite line model
ISM	infinite segment model
MAP	maximum a posteriori
MCMC	Markov chain Monte Carlo
ML	maximum likelihood
MLE	maximum likelihood estimation
NB	naive Bayes
NIG	Normal-Inverse-Gamma
PEMD	partitioning earth mover's distance
ReLU	rectified linear unit
SAMS	sequentially allocated merge-split
SEMD	shifted earth mover's distance