

Lecture Notes on Bayesian Nonparametrics

Peter Orbanz

Version: May 16, 2014

These are class notes for a PhD level course on Bayesian nonparametrics, taught at Columbia University in Fall 2013.

This text is a draft.

I apologize for the many typos, false statements and poor explanations no doubt still left in the text; I will post corrections as they become available.

Contents

Chapter 1. Terminology	1
1.1. Models	1
1.2. Parameters and patterns	2
1.3. Bayesian and nonparametric Bayesian models	3
Chapter 2. Clustering and the Dirichlet Process	5
2.1. Mixture models	5
2.2. Bayesian mixtures	7
2.3. Dirichlet processes and stick-breaking	8
2.4. The posterior of a Dirichlet process	10
2.5. Gibbs-sampling Bayesian mixtures	11
2.6. Random partitions	15
2.7. The Chinese restaurant process	16
2.8. Power laws and the Pitman-Yor process	17
2.9. The number of components in a mixture	19
2.10. Historical references	20
Chapter 3. Latent features and the Indian buffet process	23
3.1. Latent feature models	24
3.2. The Indian buffet process	25
3.3. Exchangeability in the IBP	26
3.4. Random measure representation of latent feature models	26
Chapter 4. Regression and the Gaussian process	29
4.1. Gaussian processes	29
4.2. Gaussian process priors and posteriors	30
4.3. Is the definition meaningful?	33
Chapter 5. Models as building blocks	35
5.1. Mixture models	35
5.2. Hierarchical models	36
5.3. Covariate-dependent models	40
Chapter 6. Exchangeability	43
6.1. Bayesian models and conditional independence	43
6.2. Prediction and exchangeability	44
6.3. de Finetti's theorem	46
6.4. Exchangeable partitions	48
6.5. Exchangeable arrays	50
6.6. Applications in Bayesian statistics	52

Chapter 7. Posterior distributions	55
7.1. Existence of posteriors	56
7.2. Bayes' theorem	57
7.3. Dominated models	58
7.4. Conjugacy	60
7.5. Gibbs measures and exponential families	63
7.6. Conjugacy in exponential families	65
7.7. Posterior asymptotics, in a cartoon overview	66
Chapter 8. Random measures	71
8.1. Sampling models for random measure priors	72
8.2. Random discrete measures and point processes	73
8.3. Poisson processes	73
8.4. Total mass and random CDFs	76
8.5. Infinite divisibility and subordinators	77
8.6. Poisson random measures	79
8.7. Completely random measures	80
8.8. Normalization	81
8.9. Beyond the discrete case: General random measures	83
8.10. Further references	85
Appendix A. Poisson, gamma and stable distributions	87
A.1. The Poisson	87
A.2. The gamma	87
A.3. The Dirichlet	88
A.4. The stable	89
Appendix B. Nice spaces	91
B.1. Polish spaces	91
B.2. Standard Borel spaces	92
B.3. Locally compact spaces	93
Appendix C. Conditioning	95
C.1. Probability kernels	95
C.2. Conditional probability	95
C.3. Conditional random variables	96
C.4. Conditional densities	97
Appendix. Index of definitions	99
Appendix. Bibliography	101

CHAPTER 1

Terminology

1.1. Models

The term “model” will be thrown around a lot in the following. By a **statistical model** on a sample space \mathbf{X} , we mean a set of probability measures on \mathbf{X} . If we write $\mathbf{PM}(\mathbf{X})$ for the space of all probability measures on \mathbf{X} , a model is a subset $M \subset \mathbf{PM}(\mathbf{X})$. The elements of M are indexed by a **parameter** θ with values in a **parameter space** \mathbf{T} , that is,

$$M = \{P_\theta | \theta \in \mathbf{T}\}, \quad (1.1)$$

where each P_θ is an element of $\mathbf{PM}(\mathbf{X})$. (We require of course that the set M is measurable in $\mathbf{PM}(\mathbf{X})$, and that the assignment $\theta \mapsto P_\theta$ is bijective and measurable.)

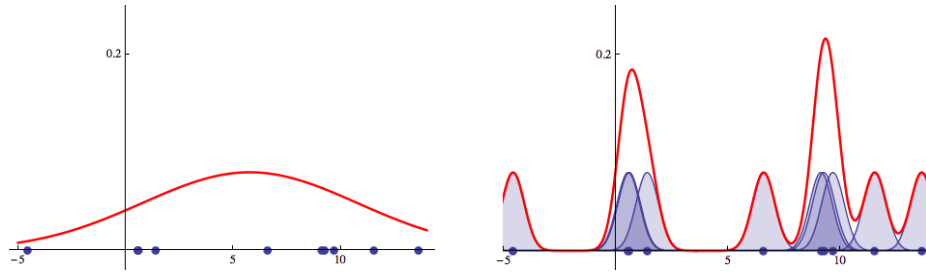
We call a model **parametric** if \mathbf{T} has finite dimension (which usually means $\mathbf{T} \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$). If \mathbf{T} has infinite dimension, M is called a **nonparametric model**. To formulate statistical problems, we assume that n observations x_1, \dots, x_n with values in \mathbf{X} are recorded, which we model as random variables X_1, \dots, X_n . In classical statistics, we assume that these random variables are generated i.i.d. from a measure in the model, i.e.

$$X_1, \dots, X_n \sim_{\text{iid}} P_\theta \quad \text{for some } \theta \in \mathbf{T}. \quad (1.2)$$

The objective of statistical inference is then to draw conclusions about the value of θ (and hence about the distribution P_θ of the data) from the observations.

Example 1.1 (Parametric and Nonparametric density estimation). There is nothing Bayesian to this example: We merely try to illustrate the difference between parametric and nonparametric methods. Suppose we observe data X_1, X_2, \dots in \mathbb{R} and would like to get an estimate of the underlying density. Consider the following two estimators:

FIGURE 1.1. Density estimation with Gaussians: Maximum likelihood estimation (*left*) and kernel density estimation (*right*).



- (1) **Gaussian fit.** We fit a Gaussian density to the data by maximum likelihood estimation.
- (2) **Kernel density estimator.** We again use a Gaussian density function g , in this case as a kernel: For each observation $X_i = x_i$, we add one Gaussian density with mean x_i to our model. The density estimate is then the density $p_n(x) := \frac{1}{n} \sum_{i=1}^n g(x|x_i, \sigma)$. Intuitively, we are “smoothing” the data by convolution with a Gaussian. (Kernel estimates usually also decrease the variance with increasing n , but we skip details.)

Figure 1.1 illustrates the two estimators. Now compare the number of parameters used by each of the two estimators:

- The Gaussian maximum likelihood estimate has 2 degrees of freedom (mean and standard deviation), regardless of the sample size n . This model is parametric.
- The kernel estimate requires an additional mean parameter for each additional data point. Thus, the number of degrees of freedom grows linearly with the sample size n . Asymptotically, the number of scalar parameters required is infinite, and to summarize them as a vector in a parameter space \mathbf{T} , we need an infinite-dimensional space.

◁

1.2. Parameters and patterns

A helpful intuition, especially for Bayesian nonparametrics, is to think of θ as a *pattern* that explains the data. Figure 1.2 (left) shows a simple example, a linear regression problem. The dots are the observed data, which shows a clear linear trend. The line is the pattern we use to explain the data; in this case, simply a linear function. Think of θ as this function. The parameter space \mathbf{T} is hence the set of linear functions on \mathbb{R} . Given θ , the distribution P_θ explains how the dots scatter around the line. Since a linear function on \mathbb{R} can be specified using two scalars, an offset and a slope, \mathbf{T} can equivalently be expressed as \mathbb{R}^2 . Comparing to our definitions above, we see that this linear regression model is parametric.

Now suppose the trend in the data is clearly nonlinear. We could then use the set of *all* functions on \mathbb{R} as our parameter space, rather than just linear ones. Of course, we would usually want a regression function to be continuous and reasonably smooth, so we could choose \mathbf{T} as, say, the set of all twice continuously differentiable functions on \mathbb{R} . An example function θ with data generated from it could then look

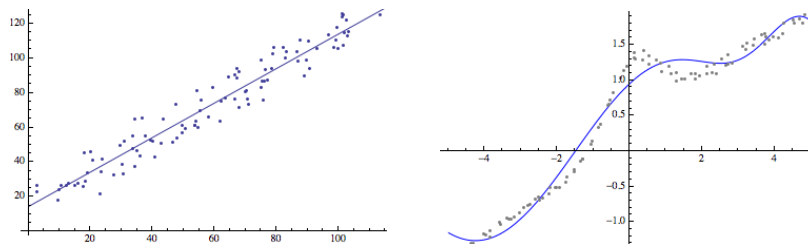


FIGURE 1.2. Regression problems: Linear (*left*) and nonlinear (*right*). In either case, we regard the regression function (plotted in blue) as the model parameter.

like the function in Figure 1.2 (right). The space \mathbf{T} is now infinite-dimensional, which means the model is nonparametric.

1.3. Bayesian and nonparametric Bayesian models

In Bayesian statistics, we model the parameter as a random variable: The value of the parameter is unknown, and a basic principle of Bayesian statistics is that all forms of uncertainty should be expressed as randomness. We therefore have to consider a random variable Θ with values in \mathbf{T} . We make a modeling assumption on how Θ is distributed, by choosing a specific distribution Q and assuming $Q = \mathcal{L}(\Theta)$. The distribution Q is called the **prior distribution** (or **prior** for short) of the model. A **Bayesian model** therefore consists of a model M as above, called the **observation model**, and a prior Q . Under a Bayesian model, data is generated in two stages, as

$$\begin{aligned} \Theta &\sim Q \\ X_1, X_2, \dots | \Theta &\sim_{\text{iid}} P_\Theta. \end{aligned} \tag{1.3}$$

This means the data is **conditionally i.i.d. rather than i.i.d.** Our objective is then to determine the **posterior distribution**, the conditional distribution of Θ given the data,

$$Q[\Theta \in \bullet | X_1 = x_1, \dots, X_n = x_n]. \tag{1.4}$$

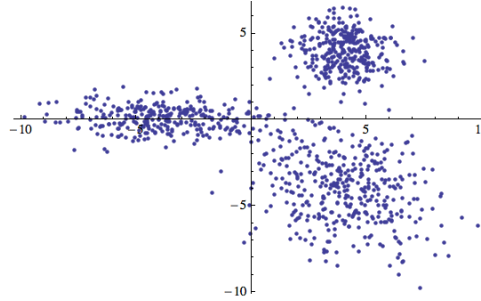
This is the counterpart to parameter estimation in the classical approach. The value of the parameter remains uncertain given a finite number of observations, and Bayesian statistics uses the posterior distribution to express this uncertainty.

A **nonparametric Bayesian model** is a Bayesian model whose parameter space has infinite dimension. To define a nonparametric Bayesian model, we have to define a probability distribution (the prior) on an infinite-dimensional space. A distribution on an infinite-dimensional space \mathbf{T} is a stochastic process with paths in \mathbf{T} . Such distributions are typically harder to define than distributions on \mathbb{R}^d , but we can draw on a large arsenal of tools from stochastic process theory and applied probability.

CHAPTER 2

Clustering and the Dirichlet Process

The first of the basic models we consider is the Dirichlet process, which is used in particular in data clustering. In a **clustering problem**, we are given observations x_1, \dots, x_n , and the objective is to subdivide the sample into subsets, the **clusters**. The observations within each cluster should be mutually similar, in some sense we have to specify. For example, here is a sample containing $n = 1000$ observations in \mathbb{R}^2 :



It is not hard to believe that this data may consist of three groups, and the objective of a clustering method would be to assign to each observation a cluster label 1, 2 or 3. Such an assignment defines a partition of the index set $\{1, \dots, 1000\}$ into three disjoint sets.

2.1. Mixture models

The basic assumption of clustering is that each observation X_i belongs to a single cluster k . We can express the cluster assignment as a random variable L_i , that is, $L_i = k$ means X_i belongs to cluster k . Since the cluster assignments are not known, this variable is unobserved. We can then obtain the distribution characterizing a single cluster k by conditioning on L ,

$$P_k(\bullet) := \mathbb{P}[X \in \bullet | L = k] . \quad (2.1)$$

Additionally, we can define the probability for a newly generated observation to be in cluster k ,

$$c_k := \mathbb{P}\{L = k\} . \quad (2.2)$$

Clearly, $\sum_k c_k = 1$, since the c_k are probabilities of mutually exclusive events. The distribution of X is then necessarily of the form

$$P(\bullet) = \sum_{k \in \mathbb{N}} c_k P_k(\bullet) . \quad (2.3)$$

A model of this form is called a **mixture distribution**. If the number of clusters is finite, i.e. if there is only a finite number K of non-zero probabilities c_k , the

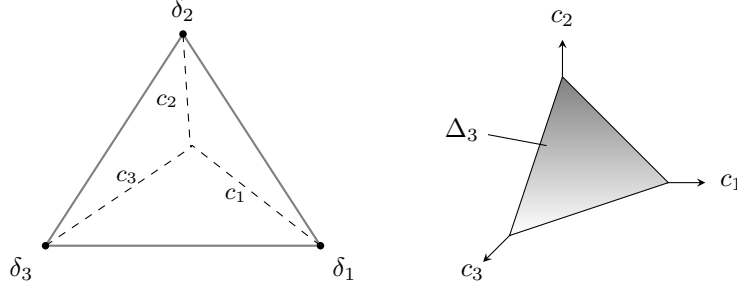


FIGURE 2.1. The simplex Δ_3 . **Each point in the set can be interpreted as a probability measure on three disjoint events.** For any finite K , the simplex Δ_K can be regarded as a subset of the Euclidean space \mathbb{R}^K .

mixture is called a **finite mixture**. Sequences of the form (c_k) are so important in the following that they warrant their own notation: The set of all such sequences is called the **simplex**, and we denote it as

$$\Delta := \left\{ (c_k)_{k \in \mathbb{N}} \mid c_k \geq 0 \text{ and } \sum_k c_k = 1 \right\}. \quad (2.4)$$

Additionally, we write Δ_K for the subset of sequences in which at most the first K entries are non-zero.

We now make a second assumption, namely that all P_k are distributions in a parametric model $\{P_\phi \mid \phi \in \Omega_\phi\}$ whose elements have a conditional density $p(x|\phi)$. If so, we can represent P_k by the density $p(x|\phi_k)$, and P in (2.3) has density

$$p(x) = \sum_{k \in \mathbb{N}} c_k p(x|\phi_k). \quad (2.5)$$

A very useful way to represent this distribution is as follows: Let θ be a **discrete probability measure** on Ω_ϕ . Such a measure is always of the form

$$\theta(\bullet) = \sum_{k \in \mathbb{N}} c_k \delta_{\phi_k}(\bullet), \quad (2.6)$$

for some $(c_k) \in \Delta$ and a sequence of points $\phi_k \in \Omega_\phi$. The Dirac measures¹ δ_{ϕ_k} are also called the **atoms** of θ , and the values ϕ_k the **atom locations**. Now if (c_k) and (ϕ_k) are in particular the same sequences as in (2.5), we can write the mixture density p as

$$p(x) = \sum_{k \in \mathbb{N}} c_k p(x|\phi_k) = \int p(x|\phi) \theta(d\phi). \quad (2.7)$$

The measure θ is called the **mixing measure**. This representation accounts for the name mixture model; see Section 5.1 for more on mixtures.

Equation (2.7) shows that all model parameters—the sequences (c_k) and (ϕ_k) —are summarized in the mixing measure. In the sense of our definition of a model

¹ Recall that the **Dirac measure** or **point mass** δ_ϕ is the probability measure which assigns mass 1 to the singleton (the one-point set) $\{\phi\}$. Its most important properties are

$$\delta_\phi = \begin{cases} 1 & \phi \in A \\ 0 & \phi \notin A \end{cases} \quad \text{and} \quad \int h(\tau) \delta_\phi(d\tau) = h(\phi) \quad (2.8)$$

for any measurable set A and any measurable function h .

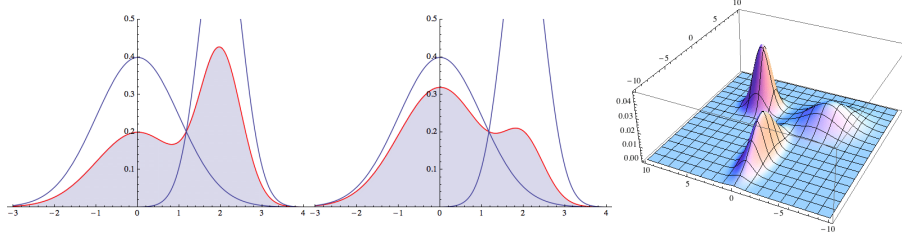


FIGURE 2.2. Gaussian mixture models. *Left:* The two-component model with density $f(x) = \frac{1}{2}g(x|0,1) + \frac{1}{2}g(x|2,0.5)$. The red, filled curve is the mixture density, the individual components are plotted for comparison. *Middle:* Mixture with components identical to the left, but weights changed to $c_1 = 0.8$ and $c_2 = 0.2$. *Right:* Gaussian mixture with $K = 3$ components on \mathbf{R}^2 . A sample of size $n = 1000$ from this model is shown in the introduction of Chapter 2.

in Section 1.1, we can regard (2.7) as the density of a measure P_θ . If \mathbf{T} is a set of discrete probability measures on Ω_ϕ , then $M = \{P_\theta | \theta \in \mathbf{T}\}$ is a model in the sense of Equation (1.1), and we call M a **mixture model**. To be very clear:

All mixture models used in clustering can be parametrized by discrete probability measures.

Without further qualification, the term mixture model is often meant to imply that \mathbf{T} is the set of all discrete probabilities on the parameter space Ω_ϕ defined by $p(x|\phi)$. A **finite mixture model** of order K is a mixture model with \mathbf{T} restricted no more than K non-zero coefficients.

2.2. Bayesian mixtures

We have already identified the parameter space \mathbf{T} for a mixture model: The set of discrete probability measures on Ω_ϕ , or a suitable subspace thereof. A **Bayesian mixture model** is therefore a mixture model with a *random* mixing measure

$$\Theta = \sum_{k \in \mathbb{N}} C_k \delta_{\Phi_k}, \quad (2.9)$$

where I have capitalized the variables C_k and Φ_k to emphasize that they are now random. The prior Q of a Bayesian mixture is the law Q of Θ , which is again worth emphasizing:

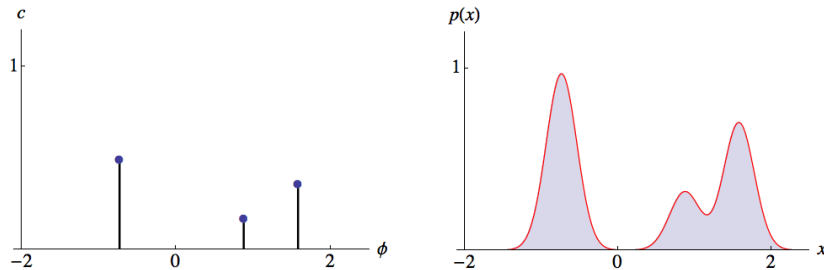
The prior of a Bayesian mixture model is the distribution of a random mixing measure Θ .

To define a Bayesian mixture model, we have to choose the component densities $p(x|\phi)$ (which also defines Ω_ϕ), and we have to find a way to generate a random probability measure on Ω_ϕ as in (2.9).

To do so, we note that to generate Θ , we only have to generate two suitable random sequences (C_k) and (Φ_k) . The easiest way to generate random sequences is to sample their elements i.i.d. from a given distribution, so we begin by choosing a probability measure G on Ω_ϕ and sample

$$\Phi_1, \Phi_2, \dots \sim_{\text{iid}} G. \quad (2.10)$$

A random measure with this property—i.e. (Φ_k) is i.i.d. and independent of (C_k) —is called **homogeneous**.



The weights (C_k) cannot be i.i.d.: We can of course sample i.i.d. from a distribution on $[0, 1]$, but the resulting variables will not add up to 1. In terms of simplicity, the next-best thing to i.i.d. sampling is to normalize an i.i.d. sequence. For a *finite* mixture model with K components, we can sample K i.i.d. random variables V_1, \dots, V_K in $[0, \infty)$ and define

This clearly defines a distribution on Δ_K . The simplest example of such a distribution is the Dirichlet distribution, which we obtain if the variables V_k have gamma distribution (cf. Appendix A.3).

If the number K of mixture components is infinite, normalizing i.i.d. variables as above fails: An infinite sum of strictly positive i.i.d. variables has to diverge, so we would have $T = \infty$ almost surely. Nonetheless, there is again a simple solution: We can certainly sample C_1 from a probability distribution H on $[0, 1]$. Once we have observed C_1 , though, C_2 is no longer distributed on $[0, 1]$ —it can only take values in $[0, 1 - C_1]$. Recall that the C_k represent probabilities; we can think of $I_k := [0, 1 - (C_1 + \dots + C_k)]$ as the remaining probability mass after the first k probabilities C_k have been determined, e.g. for $k = 2$:

Clearly, the distribution of C_{k+1} , conditionally on the first k values, must be a distribution on the interval I_k . Although this means we cannot use H as the distribution of C_{k+1} , we see that *all we have to do is to scale H to I_k* . To generate samples from this scaled distribution, we can first sample V_k from the original H , and then scale V_k as

Since I_k itself scales from step to step as $|I_k| = (1 - V_k)|I_{k-1}|$, we can generate the sequence $C_{1:\infty}$ as

$$V_1, V_2, \dots \sim_{\text{iid}} H \quad \text{and} \quad C_k := V_k \prod_{j=1}^{k-1} (1 - V_j). \quad (2.14)$$

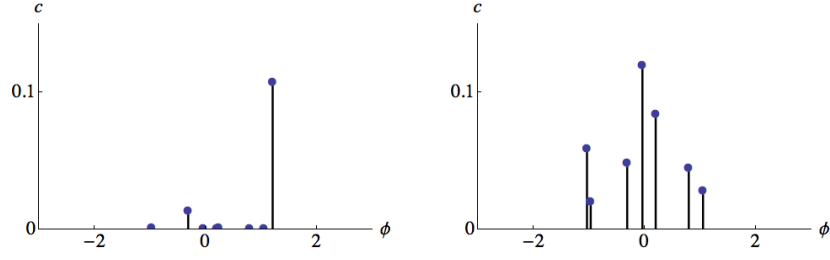


FIGURE 2.4. Two random measures with $K = 10$. In both cases, the atom locations Φ_k are sampled from a $\mathcal{N}(0, 1)$ distribution. The weights C_k are drawn from Dirichlet distributions on Δ_{10} with uniform expected distribution. If the Dirichlet concentration parameter is small ($\alpha = 1$, *left*), the variance of the weights is large. A larger concentration parameter ($\alpha = 10$, *right*) yields more evenly distributed weights.

More generally, we can sample the variables V_k each from their own distribution H_k on $[0, 1]$, as long as we keep them independent,

$$V_1 \sim H_1, V_2 \sim H_2, \dots \quad (\text{independently}) \quad \text{and} \quad C_k := V_k \prod_{j=1}^{k-1} (1 - V_j). \quad (2.15)$$

The sampling procedure (2.15) is called **stick-breaking** (think of the interval as a stick from which pieces $(1 - V_k)$ are repeatedly broken off). Provided $\mathbb{E}[V_k] > 0$, it is not hard to see that (C_k) generated by (2.14) is indeed in Δ .

We can now generate a homogeneous random measure with $K = \infty$ by choosing a specific distribution G in (2.10) and a specific sequence of distributions H_k on $[0, 1]$ in (2.15), and defining

$$\Theta := \sum C_k \delta_{\Phi_k}. \quad (2.16)$$

The basic parametric distribution on $[0, 1]$ is the beta distribution. The homogeneous random probability measure defined by choosing $H_1 = H_2 = \dots$ as a beta distribution is the Dirichlet process.

Definition 2.1. If $\alpha > 0$ and if G is a probability measure on Ω_ϕ , the random discrete probability measure Θ in (2.16) generated by

$$V_1, V_2, \dots \sim_{\text{iid}} \text{Beta}(1, \alpha) \quad \text{and} \quad C_k := V_k \prod_{j=1}^{k-1} (1 - V_j) \quad (2.17)$$

$$\Phi_1, \Phi_2, \dots \sim_{\text{iid}} G$$

is called a **Dirichlet process** (DP) with **base measure** G and **concentration** α , and we denote its law by $\text{DP}(\alpha, G)$. \triangleleft

If we integrate a parametric density $p(x|\phi)$ against a random measure Θ generated by a Dirichlet process, we obtain a mixture model

$$p(x) = \sum_{k \in \mathbb{N}} C_k p(x|\Phi_k), \quad (2.18)$$

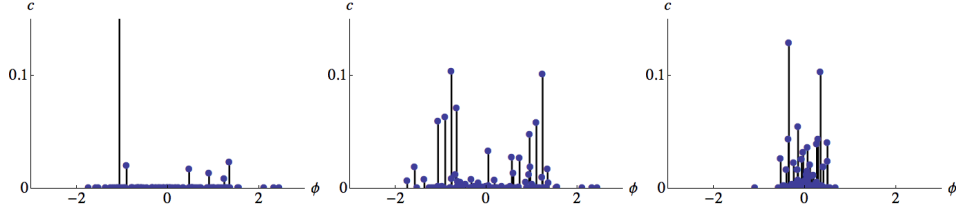


FIGURE 2.5. Random measures sampled from a Dirichlet process with normal base measure. *Left:* For concentration $\alpha = 1$, the atom sizes exhibit high variance. *Middle:* For larger values of the concentration (here $\alpha = 10$), the atom sizes become more even. Compare this to the behavior of the Dirichlet distribution. *Right:* Decreasing the variance of the normal base measure changes the distribution of the atoms; the DP concentration is again $\alpha = 10$.

called a **Dirichlet process mixture**. Observations X_1, X_2, \dots are generated from a DP mixture according to

$$\begin{aligned} \Theta &\sim \text{DP}(\alpha, G_0) \\ \Phi_1, \Phi_2, \dots | \Theta &\sim_{\text{iid}} \Theta \\ X_i &\sim p(x | \Phi_i) \end{aligned} \quad (2.19)$$

The number of non-zero coefficients C_k is now infinite, and the model therefore represents a population subdivided into an *infinite* number of clusters, although, for a finite sample $X_1 = x_1, \dots, X_n = x_n$, we can of course observe at most n of these clusters.

Remark 2.2. You will have noticed that I have motivated several definitions in this section by choosing them to be as simple and “close to i.i.d.” as possible. For Bayesian models, this is important for two reasons:

- (1) Dependencies in the prior (such as coupling between the variables C_k and Φ_k) make it *much* harder to compute posterior distributions—both computationally (in terms of mathematical complexity and computer time) and statistically (in terms of the amount of data required).
- (2) If we choose to use dependent variables, we cannot simply make them “not independent”; rather, we have to choose one specific form of dependency. Any specific form of dependence we choose is a modeling assumption, which we should only impose for good reason.

◁

2.4. The posterior of a Dirichlet process

So far, we have considered how to generate an instance of a random measure Θ . To use Θ as the parameter variable in a Bayesian model, we have to define how observations are generated in this model, and we then have to determine the posterior distribution. Before we discuss posteriors of mixtures, we first consider a simpler model where observations are generated directly from Θ . That is, we sample:

$$\begin{aligned} \Theta &\sim \text{DP}(\alpha, G_0) \\ \Phi_1, \Phi_2, \dots | \Theta &\sim_{\text{iid}} \Theta \end{aligned} \quad (2.20)$$

Each sample Φ_i almost surely coincides with an atom of Θ .

Under the model (2.20), we never actually observe Θ , only the variables Φ_i . What can we say about their distribution? From the definition of the DP in (2.17), we can see that the first observation Φ_1 is simply distributed according to G . That is not the case for Φ_2 , given Φ_1 : If we have observed $\Phi_1 = \phi$, we know Θ must have an atom at ϕ , and we now could observe either $\Phi_2 = \phi$ again, or another atom.

Theorem 2.3 (Ferguson [12, Theorem 1]). *Suppose Θ has a $DP(\alpha, G_0)$ distribution and that observations $\Phi_1 = \phi_1, \dots, \Phi_n = \phi_n$ are generated as in (2.20). Then the posterior distribution of Θ is*

$$\mathbb{P}[\Theta \in \bullet | \Phi_1, \dots, \Phi_n] = DP\left(\alpha G_0 + \sum_{k=1}^n \delta_{\phi_k}\right), \quad (2.21)$$

and the next observation Φ_{n+1} has conditional distribution

$$\mathbb{P}[\Phi_{n+1} \in \bullet | \Phi_1, \dots, \Phi_n] = \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{k=1}^n \delta_{\phi_k}. \quad (2.22)$$

◁

The result shows in particular that the Dirichlet process prior has a *conjugate* posterior, that is, the posterior is again a Dirichlet process, and its parameters can be computed from the data by a simple formula; we will discuss conjugate posteriors in more detail in Section 7.4.

I should stress again that (2.21) is the posterior distribution of a DP under the sampling model (2.20) in which we observe the variables Φ_i , *not* under the DP mixture, in which the variables Φ_i are unobserved. To work with DP mixtures, we address this problem using latent variable algorithms.

2.5. Gibbs-sampling Bayesian mixtures

Random variables like the Φ_i , which form an “intermediate” unobserved layer of the model, are called **latent variables**. If a model contains latent variables, we can usually not compute the posterior analytically, since that would involve conditioning on the latent information. There are inference algorithms for latent variable models, however, and they are usually based on either of two different strategies:

- (1) **Variational algorithms**, which upper- or lower-bound the effect of the additional uncertainty introduced by the latent variables, and optimize these bounds instead of optimizing the actual, unknown solution. The EM algorithm for finite mixtures is a (non-obvious) example of a variational method, although a finite mixture is usually not interpreted as a Bayesian model.
- (2) **Imputation methods**, which sample the latent variables and condition on the sampled values. This is typically done using MCMC.

Inference in DP mixtures and other Bayesian mixtures is based on sampling algorithms that use imputation. I would like to stress that in most Bayesian models,

we use sampling algorithms because the model contains latent variables.

(Most introductory texts on MCMC will motivate sampling algorithms by pointing out that it is often only possible to evaluate a probability density up to scaling, and that such an unnormalized distribution can still be sampled—which is perfectly true, but really beside the point when it comes to latent variable models.)

Gibbs sampling. Suppose we want to simulate samples from a multivariate distribution Q on Ω_ϕ , where we assume $\Omega_\phi = \mathbb{R}^D$, so random draws are of the form $\Phi = (\Phi_1, \dots, \Phi_D)$. A Gibbs sampler loops over the dimensions $d = 1, \dots, D$ and samples Φ_d conditionally on the remaining dimensions. The conditional probability

$$Q[\Phi_d \in \bullet | \Phi_1 = \phi_1, \dots, \Phi_{d-1} = \phi_{d-1}, \Phi_{d+1} = \phi_{d+1}, \dots, \Phi_D = \phi_D] \quad (2.23)$$

is called the **full conditional** distribution of Φ_d . Recall that the **Gibbs sampler** for P is the algorithm which, in its $(j+1)$ st iteration, samples

$$\begin{aligned} \Phi_1^{(j+1)} &\sim Q[\Phi_d \in \bullet | \Phi_2 = \phi_2^{(j)}, \dots, \Phi_D = \phi_D^{(j)}] \\ &\vdots \\ \Phi_d^{(j+1)} &\sim Q[\Phi_d \in \bullet | \Phi_1 = \phi_1^{(j+1)}, \dots, \Phi_{d-1} = \phi_{d-1}^{(j+1)}, \Phi_{d+1} = \phi_{d+1}^{(j)}, \dots, \Phi_D = \phi_D^{(j)}] \\ &\vdots \\ \Phi_D^{(j+1)} &\sim Q[\Phi_d \in \bullet | \Phi_1 = \phi_1^{(j+1)}, \dots, \Phi_{D-1} = \phi_{D-1}^{(j+1)}] \end{aligned}$$

Note that, at each dimension d , the values $\phi_1^{(j+1)}, \dots, \phi_{d-1}^{(j+1)}$ generated so far in the current iteration are already used in the conditional, whereas the remaining dimensions $\phi_{d+1}^{(j)}, \dots, \phi_D^{(j)}$ are filled in from the previous iteration. Since this removal of a single dimension makes notation cumbersome, it is common to write

$$\phi_{-d} := \{\phi_1, \dots, \phi_{d-1}, \phi_{d+1}, \dots, \phi_D\}, \quad (2.24)$$

so the full conditional of Φ_d is $Q[\Phi_d \in \bullet | \Phi_{-d} = \phi_{-d}]$ et cetera.

A naive Gibbs sampler for DP mixtures. In the Dirichlet process mixture model, we generate n observations X_1, \dots, X_n by generating a latent random measure $\Theta = \sum_{k \in \mathbb{N}} C_k \delta_{\Phi_k}$ and sampling from Θ :

$$\begin{aligned} \Theta &\sim \text{DP}(\alpha G_0) \\ \Phi_1, \dots, \Phi_n &\sim \Theta \\ X_i &\sim p(\bullet | \Phi_i). \end{aligned} \quad (2.25)$$

Recall that reoccurrences of atom locations are possible: X_i and X_j belong to the same cluster if $\Phi_i = \Phi_j$. We observe that, under this model:

- The variables Φ_i are conditionally independent of each other given Θ .
- Φ_i is conditionally independent of X_j given Φ_j if $j \neq i$. In other words, the conditional variable $\Phi_i | \Phi_j$ is independent of X_j .

To derive a Gibbs sampler, we regard the variables Φ_i as n coordinate variables.

The joint conditional distribution of (Φ_1, \dots, Φ_n) given the data is complicated, but we can indeed derive the full conditionals

$$\mathcal{L}(\Phi_i | \Phi_{-i}, X_{1:n}) \quad (2.26)$$

with relative ease:

- We choose one Φ_i and condition on the remaining variables Φ_{-i} . Since the Φ_i are conditionally independent and hence exchangeable, we can choose variables in any arbitrary order.
- If we know Φ_{-i} , we can compute the DP posterior $\mathcal{L}(\Theta | \Phi_{-i})$.

We know from Theorem 2.3 that

$$\mathbb{P}[\Theta \in \bullet | \Phi_{-i}] = \text{DP}\left(\alpha G_0 + \sum_{j \neq i} \delta_{\Phi_j}\right). \quad (2.27)$$

We also know that, if we sample $\Theta \sim \text{DP}(\alpha G)$ and $\Phi \sim \Theta$, then Φ has marginal distribution $\mathcal{L}(\Phi) = G$. Combined, that yields

$$\mathbb{P}[\Phi_i \in \bullet | \Phi_{-i}] = \frac{\alpha}{\alpha + (n-1)} G_0 + \frac{1}{\alpha + (n-1)} \sum_{j \neq i} \delta_{\Phi_j}. \quad (2.28)$$

To account for the observed data, we additionally have to condition on $X_{1:n}$. Since $\Phi_i | \Phi_j$ is independent of X_j ,

$$\mathcal{L}(\Phi_i | \Phi_{-i}, X_{1:n}) = \mathcal{L}(\Phi_i | \Phi_{-i}, X_i). \quad (2.29)$$

To obtain the full conditional of Φ_i , we therefore only have to condition (2.28) on X_i . To do so, we can think of $\mathbb{P}[\Phi_i \in \bullet | \Phi_{-i}]$ as a prior for Φ_i , and compute its posterior under a single observation $X_i = x_i$, with likelihood $p(x|\phi)$ as given by the mixture model. Since $p(x|\phi)$ is typically a parametric model, we can apply Bayes' theorem, and obtain the full conditional

$$\mathbb{P}[\Phi_i \in d\phi | \Phi_{-i} = \phi_{-i}, X_i = x_i] = \frac{\alpha p(x_i|\phi) G_0(d\phi) + \sum_{j \neq i} \delta_{\Phi_j}(d\phi)}{\text{normalization}}. \quad (2.30)$$

By substituting the full conditionals into the definition of the Gibbs sampler, we obtain:

Algorithm 2.4.

For iteration $l = 1, \dots, L$:

- For $i = 1, \dots, n$, sample $\Phi_i | \Phi_{-i}, X_i$ according to (2.30).

Although this algorithm is a valid sampler, it has extremely slow mixing behavior. The reason is, roughly speaking:

- The X_i are grouped into $K \leq n$ clusters; hence, there are only K distinct values $\Phi_1^*, \dots, \Phi_K^*$ within the set Φ_1, \dots, Φ_n .
- The algorithm cannot change the values Φ_k^* from one step to the next. To change the parameter of a cluster, it has to (1) create a new cluster and (2) move points from the old to the new cluster one by one. Whenever such a new parameter value is generated, it is sampled from a full conditional (2.30), which means it is based only on a single data point.

In terms of the state space of the sampler, this means that in order to move from the old to the new cluster configuration—even if it differs only in the value of a single cluster parameter Φ_k^* —the sampler has to move through a region of the state space with very low probability.

MacEachern's algorithm. The issues of the naive sampler can be addressed easily by grouping data points by cluster and generating updates of the cluster parameters given the *entire* data in the cluster. The resulting algorithm is the standard sampler for DP mixtures, and is due to MacEachern [40].

Recall that X_i and X_j are considered to be in the same cluster iff $\Phi_i = \Phi_j$. We express the assignments of observations to clusters using additional variables

B_1, \dots, B_n , with

$$B_i = k \quad \Leftrightarrow \quad X_i \text{ in cluster } k . \quad (2.31)$$

We must also be able to express that Φ_i is not contained in any of the current clusters defined by the remaining variables Φ_{-i} , which we do by setting

$$B_i = 0 \quad \Leftrightarrow \quad x_i \text{ not in any current cluster } k \in \{1, \dots, K\} . \quad (2.32)$$

The posterior of a DP mixture given data x_1, \dots, x_n can then be sampled as follows:

Algorithm 2.5. In each iteration l , execute the following steps:

(1) For $i = 1, \dots, n$, sample

$$B_i \sim \text{Multinomial}(a_{i0}, a_{i1}, \dots, a_{iK}) .$$

(2) For $k = 1, \dots, K_l$, sample

$$\Phi_k^* \sim \frac{\left(\prod_{i|B_i=k} p(x_i|\phi) \right) G_0(d\phi)}{\int_{\Omega_\phi} \left(\prod_{i|B_i=k} p(x_i|\phi) \right) G_0(d\phi)} .$$

To convince ourselves that the algorithm is a valid sampler for the posterior, observe that conditioning on the variables B_i permits us to subdivide the data into clusters, and then compute the posterior of the cluster parameter Φ_k^* given the entire cluster:

$$\mathbb{P}[\Phi_k^* \in d\phi | B_{1:n}, X_{1:n}] = \frac{\left(\prod_{i|B_i=k} p(x_i|\phi) \right) G_0(d\phi)}{\text{normalization}} . \quad (2.33)$$

Since Φ_k^* is, conditionally on $B_{1:n}$ and $X_{1:n}$, independent of the other cluster parameters, this is indeed the full conditional distribution of Φ_k^* . Conversely, we can compute the full conditionals of the variables B_i given all remaining variables: Since

$$\mathbb{P}[\Phi_i \in d\phi | B_{-i}, \Phi_{1:K}^*, X_i = x_i] = \frac{\alpha p(x_i|\phi) G_0(d\phi) + \sum_{j \neq i} p(x_j|\phi) \delta_{\Phi_{B_i}^*}(d\phi)}{\text{normalization}} ,$$

we have for any cluster k :

$$\begin{aligned} \mathbb{P}[B_i = \bullet | B_{-i}, \Phi_{1:K}^*, X_i = x_i] &= \mathbb{P}[\Phi_i = \Phi_\bullet^* | B_{-i}, \Phi_{1:K}^*, X_i = x_i] \\ &= \int_{\{\phi_\bullet^*\}} \mathbb{P}[\Phi_i = d\phi | B_{-i}, \Phi_{1:K}^*, X_i = x_i] \\ &= \int_{\{\phi_\bullet^*\}} \frac{p(x_i|\phi)}{N} \delta_{\phi_\bullet^*}(d\phi) \\ &= \frac{p(x_i|\phi_k^*)}{N} =: a_{ik} . \end{aligned} \quad (2.34)$$

The probability that x_i is not in any of the current clusters is the complement of the cluster probabilities:

$$\begin{aligned}
 \mathbb{P}[B_i = 0 | B_{-i}, \Phi_{1:K}^*, X_i = x_i] &= \mathbb{P}[\Phi_i \in \Omega_\phi \setminus \{\phi_1^*, \dots, \phi_K^*\} | B_{-i}, \Phi_{1:K}^*, X_i = x_i] \\
 &= \int_{\Omega_\phi \setminus \{\phi_{1:K}^*\}} \mathbb{P}[\Phi_i = d\phi | B_{-i}, \Phi_{1:K}^*, X_i = x_i] \\
 &= \frac{\alpha}{N} \int_{\{\phi_\bullet^*\}} p(x_i | \phi) G_0(d\phi) =: a_{i0} .
 \end{aligned} \tag{2.35}$$

If these two types of full conditionals are again substituted into the definition of the Gibbs sampler, we obtain precisely Algorithm 2.5.

Remark 2.6. MacEachern’s algorithm is easy to implement if $p(x|\phi)$ is chosen as an exponential family density and the Dirichlet process base measure G_0 as a natural conjugate prior for p . In this case, Φ_k^* in the algorithm is simply drawn from a conjugate posterior. If p and G_0 are not conjugate, computing the distribution of Φ_k^* may require numerical integration (which moreover has to be solved K times in every iteration of the algorithm). There are more sophisticated samplers available for the non-conjugate case which negotiate this problem. The de-facto standard is Neal’s “Algorithm 8”, see [45]. \triangleleft

2.6. Random partitions

A clustering solution can be encoded as a partition of the index set of the sample: Suppose we record observations X_1, \dots, X_{10} and compute a clustering solution that subdivides the data into three clusters,

$$(\{X_1, X_2, X_4, X_7, X_{10}\}, \{X_3, X_5\}, \{X_6, X_8, X_9\}) . \tag{2.36}$$

We can encode this solution as a partition of the index set $[10] = \{1, \dots, 10\}$:

$$(\{1, 2, 4, 7, 10\}, \{3, 5\}, \{6, 8, 9\}) . \tag{2.37}$$

Since we always regard the elements of a finite sample X_1, \dots, X_n as the initial n elements of an infinite sequence X_1, X_2, \dots , we must in general consider partitions of \mathbb{N} rather than $[n]$. To make things more precise, a **partition**

$$\psi = (\psi_1, \psi_2, \dots) \tag{2.38}$$

of \mathbb{N} is a subdivision of \mathbb{N} into a (possibly infinite) number of subsets $\psi_i \subset \mathbb{N}$, such that each $i \in \mathbb{N}$ is contained in exactly one set ψ_k . The sets ψ_k are called the **blocks** of the partition. A partition ψ^n of $[n]$ is defined analogously. The blocks can be ordered according to their smallest element, as we have done in (2.37). It hence make sense to refer to ψ_k as the k th block of ψ .

Suppose we have some method to compute a clustering solution from a given data set. Even if our clustering algorithm is deterministic, the observations X_1, X_2, \dots are random, and the result is hence a **random partition**

$$\Psi = (\Psi_1, \Psi_2, \dots) , \tag{2.39}$$

that is, a partition-valued random variable. Recall how we used the variables L_i to encode cluster assignments (by setting $L_i = k$ if X_i is in cluster k). In terms of partitions, this means

$$L_i = k \quad \Leftrightarrow \quad i \in \Psi_k , \tag{2.40}$$

and a random sequence (L_1, L_2, \dots) is hence precisely equivalent to a random partition (Ψ_1, Ψ_2, \dots) .

Given a discrete probability measure $\theta = \sum c_k \delta_{\phi_k}$, we can generate a random partition Ψ of \mathbb{N} by sampling the variables L_1, L_2, \dots in (2.40) with probabilities

$$\mathbb{P}(L_i = k) = c_k . \quad (2.41)$$

Any discrete probability measure θ hence parametrizes a distribution $P_\theta(\Psi \in \bullet)$ on random partitions. If Θ is a *random* discrete probability measure with distribution Q , we can define a distribution on partitions by integrating out Θ ,

$$P(\Psi \in \bullet) := \int_{\mathbf{T}} P_\theta(\Psi \in \bullet) Q(d\theta) . \quad (2.42)$$

We can sample from this distribution in two steps, as $\Theta \sim Q$ and $\Psi|\Theta \sim P_\Theta$.

2.7. The Chinese restaurant process

Recall our discussion in Section 1.2, where we interpreted the parameter of a Bayesian model as a representation of the solution that we are trying to extract from the data. A clustering solution is a partition of the sample index set, whereas the parameter in Bayesian mixture is a discrete probability measure—the partition represents the actual subdivision of the sample, the random discrete measure the mixture model we are fitting to the sample. We could hence argue that a more appropriate choice for the model parameter would be a partition, in which case the prior should be a distribution on partitions.

As we have seen above, a discrete random measure induces a distribution on partitions. The **Chinese restaurant process** with concentration α is the distribution $P(\Psi \in \bullet)$ on partitions that we obtain if we choose Q as a Dirichlet process with parameters (α, G_0) in (2.42). The choice base measure G_0 has no effect on Ψ (provided G_0 is non-atomic), and the CRP hence has only a single parameter.

According to (2.42), we can sample a CRP partition by sampling a random measure Θ from a Dirichlet process, throwing away its atom locations, and then sampling the block assignment variables L_i from the distribution given by the weights C_k . That is of course the case for all partitions induced by random measures, but in the specific case of the CRP, we can greatly simplify this procedure and sample Ψ using the following procedure:

Sampling scheme 2.7. For $n = 1, 2, \dots$,

- (1) insert n into an existing block Ψ_k with probability $\frac{|\Psi_k|}{\alpha + (n-1)}$.
- (2) create a new block containing only n with probability $\frac{\alpha}{\alpha + (n-1)}$.

The algorithm can be rewritten as a distribution,

$$L_{n+1} \sim \sum_{k=1}^{K_n} \frac{\sum_{i=1}^n \mathbb{I}\{L_i = k\}}{\alpha + n} \delta_k + \frac{\alpha}{\alpha + n} \delta_{K+1} . \quad (2.43)$$

Asymptotically (i.e. for $n \rightarrow \infty$), a partition generated by the formula is distributed according to $\text{CRP}(\alpha)$. Compare this to the predictive sampling distribution of the Dirichlet process.²

More generally, we can substitute a random partition prior for a random measure prior whenever the random measure is homogeneous. Homogeneity is required because it makes the weights of Θ , and hence the induced partition Ψ , independent of the atom locations—if Θ is inhomogeneous, we have to sample the atoms as well in order to determine Ψ . In the homogeneous case, we can sample Ψ , and then fill in independent atoms later if needed. For clustering applications, we do of course need the atom locations as parameters for the parametric components $p(x|\phi)$, but if Θ is homogeneous, we can sampling process as follows. First, generate a partition:

- (1) Sample (C_i) .
- (2) Sample Ψ given (C_i) .

Then, given the partition, generate the actual observations:

- (3) For each block Ψ_k , sample a parameter value $\Phi_k \sim G_0$.
- (4) For all $i \in \Psi_k$, sample $X_i | \Phi_k \sim p(\bullet | \Phi_k)$.

This representation neatly separates the information generated by the model into the clustering solution (the partition Ψ) and information pertaining to the generation of observations (the parameters Φ_k and the observations X_i themselves).

Random partitions have been thoroughly studied in applied probability; Pitman’s monograph [51] is the authoritative reference.

2.8. Power laws and the Pitman-Yor process

If we generate n observations X_1, \dots, X_n from a discrete random measure, we can record the number K_n of distinct clusters in this data (by comparing the latent variables Φ_1, \dots, Φ_n). If the random measure is a Dirichlet process, then K_n grows logarithmically in n . For a large range of real-world problems, this kind of logarithmic growth is not a realistic assumption, since many important statistics are known to follow so-called power laws.

Definition 2.8. We say that a probability distribution P on positive numbers is a **power law distribution** if its density with respect to Lebesgue or counting measure is of the form

$$p(x) = c \cdot x^{-a} \tag{2.44}$$

for constants $c, a \in \mathbb{R}_+$. ◁

Examples of statistics with a power law distribution include the frequencies of words in the English language, the number of followers per user on Twitter, the sizes of cities, the sizes of bodies of water (from puddles to lakes and oceans) in nature, the distribution of wealth, etc. Power laws typically arise as the outcome of aggregation processes.

It is possible to obtain a clustering model which generates power-law distributed clusters by tweaking the Dirichlet process. The result is one of the most fascinating objects used in Bayesian nonparametrics, the Pitman-Yor process. Recall that

² The name “Chinese restaurant process” derives from the idea that the blocks of the partition are tables in a restaurant and the numbers are customers who join the tables. Some colleagues find such “culinary analogies” very useful; I do not.

the predictive distribution of observation X_{n+1} given X_1, \dots, X_n under a Dirichlet process mixture is

$$p(x_{n+1}|x_1, \dots, x_n) = \sum_{k=1}^{K_n} \frac{n_k}{n + \alpha} p(x_{n+1}|\phi_k^*) + \frac{\alpha}{n + \alpha} \int p(x_{n+1}|\phi) G_0(\phi) d\phi.$$

Informally, to obtain a power law, we should have more very small and very large clusters, and fewer clusters of medium size. A possible way to obtain such a law would be to modify the DP in a way that makes it harder for a very small cluster to transition to medium size. A very simple solution to apply a “discount” $d \in [0, 1]$ as follows:

$$p(x_{n+1}|x_1, \dots, x_n) = \sum_{k=1}^{K_n} \frac{n_k - d}{n + \alpha} p(x_{n+1}|\phi_k^*) + \frac{\alpha + K_n \cdot d}{n + \alpha} \int p(x_{n+1}|\phi) G_0(\phi) d\phi$$

To understand the influence of d , think of sampling observations from the model consecutively. Whenever a new cluster k is generated, it initially contains $n_k = 1$ points. Under the DP (where $d = 0$), the probability of observing a second point in the same cluster is proportional to n_k . If d is close to 1, however, it becomes much less likely for the cluster to grow, so many clusters stay small. Some will grow, though, and once they contain a few points, d has little effect any more, so new observations tend to accumulate in these clusters which have outgrown the effect of the discount. For $d > 0$, the model hence tends to produce a few large and many very small clusters.

We can define a random measure prior with this predictive distribution by modifying the stick-breaking construction of the Dirichlet process:

Definition 2.9. The homogeneous random measure $\xi = \sum_{k \in \mathbb{N}} C_k \delta_{\Phi_k}$ defined by weights

$$V_k \sim \text{Beta}(1 - d, \alpha + kd) \quad \text{and} \quad C_k := V_k \prod_{j=1}^{k-1} (1 - V_j) \quad (2.45)$$

and atom locations

$$\Phi_1, \Phi_2, \dots \sim_{\text{iid}} G_0 \quad (2.46)$$

is called a **Pitman-Yor process** with concentration α , diversity parameter d , and base measure G_0 . \triangleleft

For a non-technical introduction to the Pitman-Yor process, have a look at Yee Whye Teh’s article on Kneser-Ney smoothing, which applies the Pitman-Yor process to an illustrative problem in language processing [61].

In the applied probability literature, it is common to denote the concentration parameter (our α) as θ , and the second parameter (our d) as α . In the Bayesian literature, using α for the concentration is well-established. The two conventions can be traced back to the classic papers by Kingman [31] and Ferguson [12], respectively.

Remark 2.10. Several different names float around the literature referring to a number of very closely related objects: The name Dirichlet process refers to the random measure $\Theta = \sum_k C_k \delta_{\Phi_k}$ in Definition 2.1. We have already noted above that, since the DP is homogeneous, its only non-trivial aspect is really the distribution of the weight sequence (C_k) , which is completely controlled by the concentration

α . The weights can be represented in two ways: Either, we can keep them in the order they are generated by the stick-breaking construction (2.14). In this case, $\mathcal{L}((C_k)_{k \in \mathbb{N}})$ is usually called the **GEM(α) distribution**, named rather obscurely for the authors of [20], [10] and [42]. Alternatively, we can rank the C_k by size to obtain an ordered sequence (C'_k) , whose distribution was dubbed the **Poisson-Dirichlet distribution** by Kingman [31], usually abbreviated $\text{PD}(\alpha)$. Since the induced random partition does not depend on G_0 either, its law, the Chinese restaurant process, also has just one parameter, and is commonly denoted $\text{CRP}(\alpha)$, as we have done above.

We can generalize all of these distributions to the case described by the Pitman-Yor process above, but now have to include the additional parameter d . Thus, we obtain the $\text{GEM}(\alpha, d)$ distribution of the weights sampled in (2.45), the **two-parameter Poisson-Dirichlet** distribution $\text{PD}(\alpha, d)$ for the ranked weights, the **two-parameter Chinese restaurant process** $\text{CRP}(\alpha, d)$ for the induced random partition of \mathbb{N} , and of course the Pitman-Yor process with parameters (α, d, G_0) for the random measure Θ . Historically, the two-parameter case was first introduced as the two-parameter Poisson-Dirichlet in [50] and [52]. Ishwaran and James [24] turned it into a random measure and coined the term Pitman-Yor process. \triangleleft

2.9. The number of components in a mixture

An old and very important problem in mixture modeling is how to select the number of mixture components, i.e. the order K of the mixture. Whether and how DP mixtures and similar models provide a solution to this problem is one of the most commonly misunderstood aspects of Bayesian nonparametrics:

*Bayesian nonparametric mixtures are **not** a tool for automatically selecting the number of components in a finite mixture. If we assume that the number of clusters exhibited in the underlying population is finite, Bayesian nonparametric mixtures are misspecified models.*

I will try to argue this point in more detail: If we use a DP mixture on a sample of size n , then in any clustering solution supported by the posterior, a random, finite number $K_n \leq n$ of clusters is present. Hence, we obtain a posterior distribution on the number of clusters. Although this is not technically model selection—since there is just one model, under which the different possible values of K_n are mutually exclusive events—we indeed have obtained a solution for the number of clusters. However, a DP random measure has an infinite number of atoms almost surely. Hence, the modeling assumption implicit in a DP mixture is that

as $n \rightarrow \infty$, we will inevitably (with probability 1) observe an infinite number of clusters.

To choose an adequate strategy for determining the number of components, we have to distinguish three types of problems:

- (1) **K is finite and known**, which is the assumption expressed by a finite mixture model of order K .
- (2) **K is finite but unknown**. In this case, the appropriate model would be a finite mixture model of unknown order. *This problem should **not** be modeled by a DP or other infinite mixture.*
- (3) **K is infinite**, as assumed by the Dirichlet process/CRP or the Pitman-Yor process.

In many clustering problems, (3) is really the appropriate assumption: In topic modeling problems, for example, we assume that a given text corpus contains a finite number of topics, but that is only because the corpus itself is finite. If the corpus size increases, we would expect new topics to emerge, and it would usually be very difficult to argue that the number of topics eventually runs up against some finite bound and remains fixed.

What if we really have reason to assume that (2) is adequate? In a classical mixture-model setup—estimate a finite mixture by approximate maximum likelihood using an EM algorithm, say—choosing K in (2) is a model selection problem, since different K result in different models whose likelihoods are not comparable. This problem is also called the **model order selection** problem, and popular solutions include penalty methods (AIC or BIC), stability, etc. [see e.g. 43].

We can also take a Bayesian mixture approach and assume K to be unknown, but if we believe K is finite, we should choose a model that generates a random *finite* number K . For example, we could define a prior on K as

$$K := K' + 1 \quad \text{where} \quad K' \sim \text{Poisson}(\lambda), \quad (2.47)$$

(where we add 1 since a Poisson variable may be 0) and then sample from a finite Bayesian mixture with K components.

There are of course situations in which it can be beneficial to deliberately mis-specify a model, and we can ask whether, even though an infinite mixture model is misspecified for case (2), it may perhaps still yield the right answer for K . Miller and Harrison [44] have recently clarified that this is not the case for DP or Pitman-Yor process mixtures using a wide range of mixture components (such as Gaussians, multinomials, etc.): If we generate a sample from a *finite* mixture, and then compute the posterior under an *infinite* DP or PYP mixture model, then asymptotically, the posterior concentrates on solutions with an infinite number of clusters. That is exactly what the model assumes: An infinitely large sample exhibits an infinite number of clusters almost surely.

Remark 2.11. To obtain an infinite (nonparametric) mixture, it is by no means necessary to take a Bayesian approach—the mixing distribution can be estimated by nonparametric maximum likelihood, similar to a Kaplan-Meier estimator [43, §1.2] Bayesian finite mixtures are less common; see [54] for an overview. \triangleleft

2.10. Historical references

The Dirichlet process (and the corresponding approach to priors) was introduced by Ferguson [12], who attributes the idea to David Blackwell. Kingman [31] introduced the closely related Poisson-Dirichlet distribution; his paper is full of insights and still a very worthwhile read. Blackwell [4] showed that a DP random measure is discrete; see [32, Chapter 8.3] for an accessible and more general derivation of his result. Almost simultaneously with Ferguson’s paper, [5] proposed an interpretation of the Dirichlet process as a generalized Pólya urn. Urn models offer a third perspective on partition models, in addition to random discrete measures and random partitions—from an urn, we sample balls (= observations) of different colors (= cluster labels). This is the perspective taken by **species sampling models** (where each color is called a *species*). See [8] for more on species sampling, and [19, 49] for more on urn models.

Antoniak [1] proposed a model called a *mixture of Dirichlet processes* (MDP), which is sometimes mistaken as a Dirichlet Process mixture. The MDP puts a prior on the parameters of the DP base measure. A draw from a MDP is discrete almost surely, just as for the DP. Steven MacEachern has pointed out to me that Antoniak’s paper also contains a Dirichlet process mixture: Antoniak introduces the idea of using a parametric likelihood with a DP or MDP, which he refers to as “random noise” (cf his Theorem 3) and as a sampling distribution (cf Example 4). If this is used with a DP, the resulting distribution is identical to a Dirichlet process mixture model. However, Lo [36] was the first author to study models of this form from a mixture perspective.

Initially, interest focused primarily on overcoming the discreteness property of the DP, with models such as the DP mixture model, tailfree processes [9] and Pólya trees [13]. NTR processes were introduced by [9], and the idea was taken up by [14] in the context of survival analysis. context by [60], who apply the Dirichlet process to right-censored data and obtain a Kaplan-Meier estimator in the limit $\alpha \rightarrow 0$.

The name “Chinese restaurant process”, is due to L. E. Dubins and J. Pitman [see 51], who introduced it as a distribution on infinite permutations, in which case each “table” is a cycle of a permutation, rather than a block in a partition, and the order in which elements are inserted at a table matters. By projecting out the order within cycles, one obtains a partition of \mathbb{N} , and the induced distribution on partitions is the CRP as commonly used in Bayesian nonparametrics. The two-parameter Poisson-Dirichlet is due Pitman and Yor [52], partly in joint work with Perman [50]. Its random measure form, the Pitman-Yor process, was introduced by Ishwaran and James [24].

CHAPTER 3

Latent features and the Indian buffet process

I am including the Indian buffet process, or IBP, in these notes as one of the three basic models, along with the Dirichlet and Gaussian process. That drastically overstates its importance, at least in terms of how widely these three models are used. The IBP is of conceptual interest, though, and neatly illustrates some fundamental ideas.

In clustering, we have looked at partitions of data, where each data point or object belongs to one (and only one) group. There is a range of problems in which we are more interested in solutions in which clusters can overlap, i.e. each data point can belong to multiple groups. Instead of a partition, say,

$$(\{1, 2, 4, 7\}, \{3\}, \{5\}, \{6, 9\}, \{8\}) , \quad (3.1)$$

we would hence consider something like

$$(\{1, 2, \textcolor{red}{3}, 4, 7\}, \{3\}, \{\textcolor{red}{4}, 5\}, \{\textcolor{red}{3}, 6, 9\}, \{8, \textcolor{red}{9}\}) . \quad (3.2)$$

Such a “relaxed partition” is, technically speaking, a **family of sets**. The term family implies that elements can reoccur—there may be two or more groups containing the same objects (an equivalent term is *multiset*). Just as for a partition, we will refer to the elements of the family as **blocks**. A convenient way to encode both partitions and families of sets is as a binary matrix \mathbf{z} , where $z_{ik} = 1$ iff i is in block k . To illustrate:

	block #		block #
object #	$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$		$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \textcolor{red}{1} & 1 & 0 & \textcolor{red}{1} & 0 \\ 1 & 0 & \textcolor{red}{1} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & \textcolor{red}{1} \end{pmatrix}$
	partition (3.1)		family of sets (3.2)

Clearly, a partition is obtained as a special case of a family of sets—namely, if the sum of every row of \mathbf{z} is 1. For families of sets, we allow multiple occurrences, but for the problems discussed below, each number should occur in at most a finite number of blocks, i.e. we require the sum of each row in \mathbf{z} to be finite, even if \mathbf{z} has an infinite number of columns.

3.1. Latent feature models

The use of overlapping blocks was motivated in machine learning by a type of models that resemble linear factor analyzers: We observe data X_1, X_2, \dots in, say, \mathbb{R}^d , and think of each vector X_i as a list of measurements describing an object i . To explain the data, we assume that there is an underlying (unobserved) set of properties, called **features**, which each object may or may not possess. The measurement X_i depends on which features object i exhibits.

Definition 3.1. Let X_1, \dots, X_n be observations represented as d -dimensional vectors. A **latent feature model** assumes that there is a fixed $d \times K$ -matrix ϕ and a binary matrix $\mathbf{z} \in \{0, 1\}^{n \times K}$ which parametrize the distribution of X as

$$X_{ij} \sim P_{\mathbf{b}_{ij}} \quad \text{where} \quad \mathbf{b}_{ij} = (\mathbf{z}_{i1}\phi_{j1}, \dots, \mathbf{z}_{iK}\phi_{jK}) . \quad (3.3)$$

Each dimension k is called a **feature**. The number $K \in \mathbb{N} \cup \{\infty\}$ of features is called the **order** of the model. \triangleleft

If we interpret the matrix \mathbf{z} as the model parameter, the model is clearly non-parametric (regardless of whether or not K is finite). Equation (3.3) says that the distribution of X_{ij} is completely determined by the effects ϕ_{ij} ; the entries of \mathbf{z} act as switches which turn effects on or off. To obtain a tractable model, we impose two further assumptions:

- (1) The effects ϕ_{jk} are scalars which combine *linearly*, i.e. the law of X_{ij} is parameterized by the sum $\sum_k \mathbf{z}_{ik}\phi_{jk}$.
- (2) All additional randomness is an additive noise contribution.

The model (3.3) then becomes

$$X_{ij} = \sum_k \mathbf{z}_{ik}\phi_{jk} + \varepsilon_{ij} = (\mathbf{z}\phi)_{ij} + \varepsilon_{ij} . \quad (3.4)$$

Example 3.2 (Collaborative filtering). A particular application is a prediction problem that arises in marketing: Each vector X_i represents an individual (a “user”), and each dimension j a product—movies are the prototypical example. We read X_{ij} as a rating, i.e. a value that expresses how much user i likes movie j . Observed data in this setting is usually sparse: Each user has on average only seen and rated a small subset of movies.

The basic approach is to identify other users with similar tastes as user i , i.e. who have rated a large fraction of movies similarly. We then use the ratings these users have assigned to movie j as a predictor for X_{ij} . This approach is called **collaborative filtering**. The perhaps simplest implementation would be to identify users with similar preferences and then simply average their ratings for movie j . Since the data is sparse, this estimate may well be volatile (or not even well-defined, if no similar user has rated the movie). A more robust approach is to group movies into types—for illustration, think of dramas, documentaries, etc—and use these types as summary variables.

Collaborative filtering can be implemented as the latent feature model (3.4) by interpreting the components as follows:

feature k	movie type k
\mathbf{z}_{ik}	user i likes movies of type k (iff $\mathbf{z}_{ik} = 1$)
ϕ_{kj}	contribution of feature k to rating of movie j
ε_{ij}	remaining randomness in X_{ij}

The illustration of movie types as dramas, documentaries etc has to be taken with a grain of salt, since the types are not predefined, but rather latent classes that are effectively estimated from data. \triangleleft

3.2. The Indian buffet process

If we regard the binary matrix \mathbf{z} as the model parameter, a Bayesian approach has to define a prior law for a random binary matrix \mathbf{Z} . The basic model of this type is a generalization of the Chinese restaurant process from partitions to overlapping blocks. The model samples a matrix \mathbf{Z} as follows:

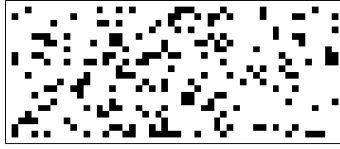
Sampling scheme 3.3. For $n = 1, 2, \dots$,

- (1) insert n into *each* block Ψ_k separately with probability $\frac{|\Psi_k|}{n}$.
- (2) create $\text{Poisson}(\frac{\alpha}{n})$ new blocks, each containing only n .

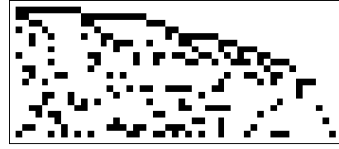
If we interpret the i th row of the matrix as a “preference profile” for user i , the distribution of this profile should obviously not depend on i . In other words, the law of \mathbf{Z} should be invariant under permutations of the rows (or *row-exchangeable*). That requirement is clearly not satisfied by the matrix generated by Algorithm 3.3: The expected number of 1s per row increases with i . This problem is addressed as follows:

- We generate a matrix using Algorithm 3.3.
- We order the matrix in a unique way that eliminates the order: Any two matrices which differ only by the order in which the rows occur map to the same ordered matrix.
- The ordered matrix is regarded as a representative of an equivalence class.

Thus, we do not eliminate the dependence of the distribution on the order of rows by defining a different distribution on matrices, but rather by defining a distribution on *equivalence classes* of matrices. The ordering method we use is called **left-ordering**:



unordered



left-ordered

It sorts all columns (=features) exhibited by object 1 to the very left of the matrix; the next block of columns are all features not exhibited by object 1, but by object 2; etc. Each of the resulting blocks is again left ordered. A more concise way to express this is: Read each column k of \mathbf{Z} as a binary number, with z_{1k} the most significant bit, and order these numbers decreasingly from left to right. We hence augment the sampling scheme 3.3 by an additional step:

Sampling scheme 3.4.

- For $n = 1, 2, \dots$,
 - (1) insert n into *each* block Ψ_k separately with probability $\frac{|\Psi_k|}{n}$.
 - (2) create $\text{Poisson}(\frac{\alpha}{n})$ new blocks, each containing only n .
- Left-order the matrix and output the resulting matrix \mathbf{Z} .

The distribution on (equivalence classes of) binary matrices defined by Algorithm 3.4 is called the **Indian buffet process** (IBP), due to Griffiths and Ghahramani [21]. Compare this to the analogous CRP sampling procedure in Algorithm 2.7.

3.3. Exchangeability in the IBP

In a random matrix sampled from an IBP, both the rows and the columns are exchangeable, which means that the distribution of \mathbf{Z} is invariant under rearrangement of the rows or the columns. More precisely, if we sample a matrix \mathbf{Z} of size $n \times K$, then for any permutations π of $\{1, \dots, n\}$ and π' of $\{1, \dots, K\}$, we have

$$(Z_{ij}) \stackrel{d}{=} (Z_{\pi(i)\pi'(j)}) . \quad (3.5)$$

It is easy to see that, due to the left-ordering step, the columns are exchangeable. However, left-ordering also makes the rows exchangeable, which is a little less obvious—the left-ordering illustration above seems to suggest that rows further down the matrix tend to contain more 1s. That is not actually the case: The rows sums all have the same distribution $\text{Poisson}(\alpha)$.

To understand how that happens, recall the additivity and thinning properties of the Poisson distribution (see (A.3) and (A.4) in Appendix A). Now consider the IBP sampling scheme in Algorithm 3.4:

- In the first step ($n=1$), we generate $\text{Poisson}(\alpha)$ blocks by definition.
- For the second row ($n=2$), we have already generated $\text{Poisson}(\alpha)$ blocks in the previous step. Each of these blocks has size 1, so the second row contains a 1 in each of these columns with independent probability $\frac{1}{2}$. By (A.4), the number of 1s generated in this way is $\text{Poisson}(\frac{\alpha}{2})$. Additionally, $\text{Poisson}(\frac{\alpha}{2})$ new blocks are generated by step 2 of the algorithm. The total number of blocks is hence $\text{Poisson}(\frac{\alpha}{2} + \frac{\alpha}{2})$ by (A.3).
- For $n = 3, 4, \dots$, the argument requires more book-keeping, because blocks can now have sizes larger than one, but if we work out the details, we again find that the row sum has distribution $\text{Poisson}(\alpha)$.

Thus, each row sum *marginally* has a $\text{Poisson}(\alpha)$ distribution, and that is true regardless of left-ordering. However, before left-ordering, the rows are *not* exchangeable, because of the order in which 1s occur in each row: The new blocks always appear on the right, so in the n th row, we see $\text{Poisson}(\frac{(n-1)\alpha}{n})$ blocks with gaps in between, and then $\text{Poisson}(\frac{\alpha}{n})$ blocks without gaps. Left-ordering removes this pattern.

3.4. Random measure representation of latent feature models

The latent feature model (3.4) is parametrized by a pair (\mathbf{z}, ϕ) , consisting of a binary matrix \mathbf{z} and a vector ϕ . A Bayesian formulation of the model hence

	Partition	Family of sets
random matrix \mathbf{Z}	exchangeable rows	exchangeable rows
constraint on \mathbf{Z}	each row sums 1 a.s.	each row sum is finite a.s.
random measure	weight sum to 1 a.s.	weight sum finite a.s.
probabilities C_i	mutually exclusive events	disjoint events
basic model	CRP	IBP

TABLE 3.1. Partitions vs families of sets.

has to define a prior on random pairs (\mathbf{Z}, Φ) . A very elegant way to generate this information is using a discrete random measure

$$\Theta = \sum_{k \in \mathbb{N}} C_k \delta_{\Phi_k} , \quad (3.6)$$

where the weights take values in $[0, 1]$. We then generate the matrix \mathbf{Z} by sampling

$$Z_{ik} | \Theta \sim \text{Bernoulli}(C_k) , \quad (3.7)$$

and model the observations as

$$X_{ij} | \Theta = (Z\Phi)_j + \varepsilon_{ij} . \quad (3.8)$$

Thus, the i.i.d. noise ε aside, all information in the model is summarized in Θ . This representation is due to [66], who also showed that for a specific random measure called the *beta process*, this representation yields the IBP. More precisely, if Θ is generated by a beta process, the random matrix \mathbf{Z} in (3.7) is, after left-ordering, distributed according to the IBP. This definition of the IBP via the beta process is the precise analogue of the way the CRP can be defined via the Dirichlet process. We will discuss the beta process in Chapter 8.

I should emphasize that the weights C_k of the random measure do *not* have to sum to 1, since they represent mutually independent probabilities. Thus, Θ is a random measure, but not a random probability measure. Recall, however, that we have constrained the matrix \mathbf{Z} to finite row sums (each object only exhibits a finite number of features). Since $\mathbb{E}[Z_{ik}] = C_k$, we see immediately (by Borel-Cantelli) that this requires

$$\Theta(\Omega_\phi) = \sum_{k \in \mathbb{N}} C_k < \infty \quad \text{a.s.} \quad (3.9)$$

Table 3.1 compares models assuming disjoint and overlapping blocks.

CHAPTER 4

Regression and the Gaussian process

Consider a simple regression problem, where observations consist of covariates $x_i \in \mathbb{R}_+$ and responses $y_i \in \mathbb{R}$. The solution of such a problem is a regression function $\theta : \mathbb{R} \rightarrow \mathbb{R}_+$. Given the regression function, we predict the response at an unobserved location x to be $y = \theta(x)$. In terms of Bayesian nonparametrics, this means that the parameter variable Θ is a random regression function (recall the discussion in Section 1.2). To define a prior distribution, we have to define a probability distribution on a suitable space \mathbf{T} of functions which we consider viable solutions. The Gaussian process is such a distribution on functions; in many ways, it is the simplest distribution we can hope to define on continuous functions.

The space of *all* functions between two spaces $\mathbf{X} \rightarrow \mathbf{Y}$ is the product space $\mathbf{Y}^{\mathbf{X}}$. (Think of each element of $\mathbf{Y}^{\mathbf{X}}$ as an infinitely long list, with one entry for each element of $x \in \mathbf{X}$; the entry specifies the function value at x . The finite-dimensional analogue would be to think of a vector (x_1, \dots, x_d) in the Euclidean space $\mathbb{R}^d = \mathbb{R}^{\{1, \dots, d\}}$ as a function $i \mapsto x_i$.) In the case of our regression problem, this would be the uncountable-dimensional space $\mathbb{R}^{\mathbb{R}_+}$. This space is not a good choice for \mathbf{T} —almost all functions in this space jump almost everywhere, and we would like a regression solution to be reasonably smooth.¹ As more suitable function spaces, we will consider the Hilbert space $\mathbf{T} := L_2(\mathbb{R}_+, \mathbb{R})$ of Lebesgue square-integrable functions and the space $\mathbf{T} := C(\mathbb{R}_+, \mathbb{R})$ of continuous functions $\mathbb{R}_+ \rightarrow \mathbb{R}$.

4.1. Gaussian processes

Let \mathbf{T} be a space of functions from a set $S \subset \mathbb{R}^d$ to \mathbb{R} . If Θ is a random element of \mathbf{T} , and we fix a point $s \in S$, then $\Theta(s)$ is a random variable in \mathbb{R} . More generally, if we fix n points $s_1, \dots, s_n \in S$, then $(\Theta(s_1), \dots, \Theta(s_n))$ is a random vector in \mathbb{R}^n .

Definition 4.1. Let μ be a probability measure on the function space \mathbf{T} . The distributions

$$\mu_{s_1, \dots, s_n} := \mathcal{L}(\Theta(s_1), \dots, \Theta(s_n)), \quad (4.1)$$

defined by μ are called the **finite-dimensional marginals** or **finite-dimensional distributions** of μ . If μ_{s_1, \dots, s_n} is an n -dimensional Gaussian for each finite set $s_1, \dots, s_n \in S$ of points, μ is called a **Gaussian process** (GP) on \mathbf{T} . \triangleleft

This definition is standard in the literature, but it can cause a lot of confusion, and I would like to issue a dire warning:

¹The idea of almost all and almost everywhere, which seems vague without having defined a probability measure first, can be made precise topologically: The functions which are at least piece-wise continuous, meaning well-behaved, form a topologically meager subset.

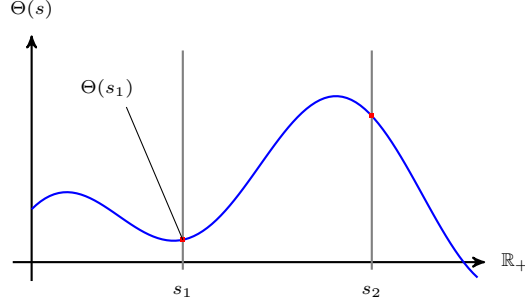


FIGURE 4.1. A random function $\Theta : \mathbb{R} \rightarrow \mathbb{R}$ defines a random scalar $\Theta(s)$ for every point $s \in \mathbb{R}$.

Warning 4.2. The definition of the GP implicitly assumes that (a) the measure μ on \mathbf{T} exists, and (b) that it is uniquely defined by the finite-dimensional marginals, which is neither obvious nor universally true. \triangleleft

Roughly speaking, uniqueness is usually guaranteed, but existence is a much more subtle question; it depends on the space \mathbf{T} and on which finite-dimensional marginals we would like μ to have. For now, we will assume that μ exists and is uniquely specified by Definition 4.1.

We define functions $\mathbf{m} : S \rightarrow \mathbb{R}$ and $\mathbf{k} : S \times S \rightarrow \mathbb{R}$ as

$$\mathbf{m}(s) := \mathbb{E}[\Theta(s)] \quad \mathbf{k}(s_1, s_2) := \text{Cov}[\Theta(s_1), \Theta(s_2)] \quad (4.2)$$

and call them the **mean function** and **covariance function** of μ . If μ is a Gaussian process, its definition says that each finite-dimensional marginal μ_{s_1, \dots, s_n} is the Gaussian distribution with mean vector and covariance matrix

$$\begin{pmatrix} \mathbf{m}(s_1) \\ \vdots \\ \mathbf{m}(s_n) \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \mathbf{k}(s_1, s_1) & \dots & \mathbf{k}(s_1, s_n) \\ \vdots & & \vdots \\ \mathbf{k}(s_n, s_1) & \dots & \mathbf{k}(s_n, s_n) \end{pmatrix}. \quad (4.3)$$

Provided a GP μ is uniquely defined by Definition 4.1, it is hence completely defined by its mean and covariance functions, and we can parametrize Gaussian processes on a given space \mathbf{T} as $\text{GP}(\mathbf{m}, \mathbf{k})$.

4.2. Gaussian process priors and posteriors

In a regression problem, the solution explaining the data is a function, so we can in principle use a Gaussian process as a prior in a Bayesian regression model. To make this feasible, however, we have to be able to compute a posterior distribution.

We first have to define an observation model: Suppose our data is of the form $(s_1, x_1), \dots, (s_n, x_n)$, where $s_i \in S$ are observation points (covariates) and $x_i \in \mathbb{R}$ is the observed value at s_i (the response). We assume that there is a function $\theta : S \rightarrow \mathbb{R}$, the regression function, from which the x_i are generated as noisy observations:

$$X_i = \Theta(s_i) + \varepsilon_i \quad \text{where } \varepsilon_i \sim \mathcal{N}(0, \sigma^2). \quad (4.4)$$

We assume, of course, that the noise contributions $\varepsilon_1, \varepsilon_2, \dots$ are i.i.d. The posterior distribution we are looking for is the distribution

$$\mathbb{P}[\Theta \in \bullet | X_1 = x_1, \dots, X_n = x_n], \quad (4.5)$$

and hence a measure on the function space \mathbf{T} . Since we know (or for now pretend to know) that a distribution on \mathbf{T} is uniquely determined by the finite-dimensional marginal distributions, it is sufficient to determine the distributions

$$\mathcal{L}(\Theta(s_{n+1}), \dots, \Theta(s_{n+m}) | X_1 = x_1, \dots, X_n = x_n), \quad (4.6)$$

for any finite set of *new* observation locations $\{s_{n+1}, \dots, s_{n+m}\}$. To keep notation sane, we abbreviate

$$A := \{n+1, \dots, n+m\} \quad \text{and} \quad B := \{1, \dots, n\} \quad (4.7)$$

and write

$$\Theta(s_A) := (\Theta(s_{n+1}), \dots, \Theta(s_{n+m})) \quad \text{and} \quad X_B := (X_1, \dots, X_n). \quad (4.8)$$

To condition on the variables X_i , we first have to take a closer look at their distribution: In (4.4), $\Theta(s_i)$ is the sum of two independent Gaussian variables variance $\mathbf{k}(s_i, s_i)$ and σ^2 , and hence again Gaussian with variance $\mathbf{k}(s_i, s_i) + \sigma^2$. For $i \neq j$, only the contributions $\Theta(s_i)$ and $\Theta(s_j)$ couple (since the noise is independent). Hence, X_B has covariance matrix

$$\Sigma_{BB} := \begin{pmatrix} \mathbf{k}(s_1, s_1) + \sigma^2 & \dots & \mathbf{k}(s_1, s_n) \\ \vdots & \ddots & \vdots \\ \mathbf{k}(s_n, s_1) & \dots & \mathbf{k}(s_n, s_n) + \sigma^2 \end{pmatrix}. \quad (4.9)$$

The joint covariance of the $(n+m)$ -dimensional vector $(\Theta(s_A), X_B)$ is then

$$\text{Cov}[(\Theta(s_A), X_B)] = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{AB}^t & \Sigma_{BB} \end{pmatrix}. \quad (4.10)$$

Again, since the noise is independent, the only contributions to the covariance come from the GP, and hence

$$\Sigma_{AB} = \left(\mathbf{k}(s_i, s_j) \right)_{i \in A, j \in B} \quad \text{and} \quad \Sigma_{AA} = \left(\mathbf{k}(s_i, s_j) \right)_{i, j \in A}. \quad (4.11)$$

Determining the GP posterior hence comes down to conditioning in a multidimensional Gaussian. How that works is explained by the following simple lemma:

Lemma 4.3 (Conditioning in Gaussian distributions). *Let (A, B) be a partition of the set $\{1, \dots, d\}$ and let $\mathbf{X} = (\mathbf{X}_A, \mathbf{X}_B)$ be a Gaussian random vector in $\mathbb{R}^d = \mathbb{R}^A \times \mathbb{R}^B$, with*

$$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \quad \text{and} \quad \text{Cov}[\mathbf{X}] = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{AB}^t & \Sigma_{BB} \end{pmatrix}. \quad (4.12)$$

Then the conditional distribution of $\mathbf{X}_A | (\mathbf{X}_B = \mathbf{x}_B)$ is again Gaussian, with mean

$$\mathbb{E}[\mathbf{X}_A | \mathbf{X}_B = \mathbf{x}_B] = \mu_A - \Sigma_{AB} \Sigma_{BB}^{-1} (\mathbf{x}_B - \mu_B) \quad (4.13)$$

and covariance

$$\text{Cov}[\mathbf{X}_A | \mathbf{X}_B = \mathbf{x}_B] = \Sigma_{AA} - \Sigma_{AB}^t \Sigma_{BB}^{-1} \Sigma_{AB}. \quad (4.14)$$

◁

We can now read off the posterior of the Gaussian process, simply by substituting into the lemma. Since the finite-dimensional marginal distributions are all Gaussian, the posterior is a GP.

Theorem 4.4. *The posterior of a $GP(\mathbf{0}, \mathbf{k})$ prior under the observation (4.4) is again a Gaussian process. Its finite-dimensional marginal distributions at any finite set $\{s_{n+1}, \dots, s_{n+m}\}$ of locations is the Gaussian with mean vector*

$$\mathbb{E}[\Theta(s_A)|X_B = x_B] = \Sigma_{AB}(\Sigma_{BB} + \sigma^2 \mathbf{I})^{-1} x_B \quad (4.15)$$

and covariance matrix

$$\text{Cov}[\Theta(s_A)|X_B = x_B] = \Sigma_{AA} - \Sigma_{AB}^t(\Sigma_{BB} + \sigma^2 \mathbf{I})^{-1} \Sigma_{AB}. \quad (4.16)$$

◁

What we have left to do is to give a proof of Lemma 4.3, which is a “disclaimer proof”: No deep insights, it simply clarifies that there is no black magic involved.

PROOF OF LEMMA 4.3. The conditional density $g(\mathbf{x}_A|\mathbf{x}_B)$ is given by

$$g(\mathbf{x}_A, \mathbf{x}_B) = g(\mathbf{x}_A|\mathbf{x}_B)g(\mathbf{x}_B). \quad (4.17)$$

It is useful to think of $\tilde{\mathbf{X}}_A := (\mathbf{X}_A|\mathbf{X}_B = \mathbf{x}_B)$ as a separate, “conditional” random variable. This variable is independent of \mathbf{X}_B (which is exactly what the product in (4.17) says), and the joint covariance of $\tilde{\mathbf{X}}_A$ and \mathbf{X}_B is hence block-diagonal,

$$\text{Cov}[\tilde{\mathbf{X}}_A, \mathbf{X}_B] = \begin{pmatrix} \tilde{\Sigma}_{AA} & \mathbf{0} \\ \mathbf{0} & \Sigma_{BB} \end{pmatrix}. \quad (4.18)$$

We have to determine $\tilde{\Sigma}_{AA}$. We can do so by decomposing the quadratic form

$$f(\mathbf{x}) := (\mathbf{x} - \mathbb{E}[\mathbf{X}])^t \text{Cov}[\mathbf{X}]^{-1} (\mathbf{x} - \mathbb{E}[\mathbf{X}]). \quad (4.19)$$

Since $g(\mathbf{x}) \propto e^{-\frac{1}{2}f(\mathbf{x})}$, the multiplicative decomposition (4.17) corresponds to an additive decomposition of f into components corresponding to $\tilde{\mathbf{X}}_A$ and \mathbf{X}_B . We know from linear algebra² that any symmetric matrix with invertible blocks A, B, Z can be inverted as

$$\begin{pmatrix} A & Z \\ Z^t & B \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ -B^{-1}Z^t & \mathbf{1} \end{pmatrix} \begin{pmatrix} (A - ZB^{-1}Z^t)^{-1} & \mathbf{0} \\ \mathbf{0} & B^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{1} & -ZB^{-1} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}. \quad (4.20)$$

Substituting in the components of $\text{Cov}[\mathbf{X}]$, we have

$$\text{Cov}[\mathbf{X}]^{-1} = \underbrace{\begin{pmatrix} \mathbf{1} & \mathbf{0} \\ -\Sigma_{BB}^{-1}\Sigma_{BA} & \mathbf{1} \end{pmatrix}}_{\text{modifies } \mu_A^t} \underbrace{\begin{pmatrix} (\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}}_{\substack{\text{inverse of} \\ \text{block-diagonal covariance (4.18)}}} \underbrace{\begin{pmatrix} \mathbf{1} & -\Sigma_{AB}\Sigma_{BB}^{-1} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}}_{\text{modifies } \mu_A} \quad (4.21)$$

We substitute into the quadratic form (4.19), multiply out the terms, and obtain

$$\mathbb{E}[\tilde{\mathbf{X}}_A] = \mu_A - \Sigma_{AB}\Sigma_{BB}^{-1}(\mu_B - \mathbf{x}_B) \quad \text{and} \quad \text{Cov}[\tilde{\mathbf{X}}_A] = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{AB} \quad (4.22)$$

as claimed. \square

²If you would like to read up further on this, look for the term *Schur complement* in the linear algebra literature. A concise reference is Appendix A.5.5 in Boyd/Vandenberghe, “Convex Optimization”.

4.3. Is the definition meaningful?

To ensure that the definition of the Gaussian process makes sense, we have to answer two separate questions:

- (1) Does a measure satisfying the definition of GP (\mathbf{m}, \mathbf{k}) exist on \mathbf{T} ?
- (2) Given \mathbf{T} and functions \mathbf{m} and \mathbf{k} , is GP (\mathbf{m}, \mathbf{k}) unique?

We first answer question (2), which is easier, and the answer very generally is “yes”. I will not state this result rigorously, since the precise meaning of a probability measure on \mathbf{T} depends on which topology we choose on the function space, but for all practical purposes, we can always rest assured:

Theorem sketch 4.5. *Let \mathbf{T} be defined as above. Any distribution μ on \mathbf{T} is uniquely determined by its finite-dimensional marginals.* \triangleleft

This approach to the construction of a stochastic process, which uses an *infinite* number of *finite*-dimensional distributions to define a single *infinite*-dimensional distribution, is called a **projective limit** construction. It is the most general (and most powerful) technique available for the construction of stochastic processes; I will not elaborate further here since the details can get fairly technical.

The question whether μ exists has many different answers, depending on the choice of \mathbf{T} . The answer is short and crisp if \mathbf{T} is a Hilbert space:

Theorem 4.6 (Prokhorov). *Let $\mathbf{T} = \mathbf{L}_2(S, \mathbb{R})$. Then the Gaussian process GP (\mathbf{m}, \mathbf{k}) on \mathbf{T} exists if and only if $\mathbf{m} \in \mathbf{T}$ and*

$$\int_S \mathbf{k}(s, s) ds < \infty. \quad (4.23)$$

\triangleleft

The integral in (4.23) is called the **trace** of the covariance function.³ The simplicity of the criterion (4.23) is not so surprising if we note that \mathbf{k} is by definition a positive definite function, and hence a Mercer kernel. Since we know that Mercer kernels are inherently related to Hilbert spaces, we would expect a simple answer.

If we want \mathbf{T} to be a space of continuous functions, existence results become more cumbersome. Here is an example: Recall that a function θ is called **locally Lipschitz-continuous** (of order r) if every $s \in S$ has an open neighborhood $U_\varepsilon(s)$ on which

$$|\theta(s) - \theta(t)| \leq C|s - t|^r \quad \text{for all } t \in U_\varepsilon(s). \quad (4.24)$$

The Lipschitz constant C is independent of s , but the definition weakens Lipschitz continuity, which requires the equation to hold for all t in S (rather than just all t in some neighborhood). The following criterion is sufficient (though not necessary) for the existence of a GP:

Theorem 4.7 (Kolmogorov, Chentsov). *Let \mathbf{T} be the set of functions $\theta : \mathbb{R}^d \rightarrow \mathbb{R}$ which are locally Lipschitz of order r . Then the Gaussian process μ on \mathbf{T} exists if*

³A few keywords, in case you would like to read up further details in the literature: The function k defines a linear operator \mathbf{K} (a linear mapping from \mathbf{T} to itself) by $(\mathbf{K}f)(s) := \int k(s, t)f(t)dt$. In this context, k is called an **integral kernel**. Operators defined in this form are called **Hilbert-Schmidt integral operators**, and are a specific example of Hilbert-Schmidt operators (which means they are bounded and have finite Hilbert-Schmidt norm). An operator satisfying (4.23) is called a **trace class** operator.

there are any constants $\alpha > 0$ and $C > 0$ such that

$$\mathbb{E}[|\Theta(s) - \Theta(t)|^\alpha] \leq C|s - t|^{1+r\alpha} . \quad (4.25)$$

◁

Models as building blocks

The basic Bayesian nonparametric models we have seen so far can be combined to obtain more complicated models. Two popular combination strategies are (1) hierarchical models, which define “layers” of hidden variables, where each layer is generated by a basic model conditionally on the previous one, and (2) covariate-dependent models, which are families of basic models indexed by a covariate such as time or space (e.g. a time series of Dirichlet process mixtures).

These types of models arguably account for the lion’s share of research in Bayesian nonparametrics, particularly in machine learning applications, where such model combinations have yielded a number of natural and compelling solutions.

5.1. Mixture models

In the context of clustering, we have specifically considered mixture models with a finite or countable number of “components”. In general, a **mixture model** on \mathbf{X} is a probability distribution

$$\mu(\bullet) = \int_{\Omega_\phi} \mathbf{p}(\bullet, \phi) m(d\phi) , \quad (5.1)$$

where \mathbf{p} is a probability kernel $\Omega_\phi \rightarrow \mathbf{PM}(\mathbf{X})$. The probability measure m on Ω_ϕ is called the **mixing measure**. The mixtures introduced in Chapter 2 are the special case where the mixing measure m is discrete (and possibly generated at random as $m = \Theta$). If the probability \mathbf{p} in (5.1) has a conditional density $p(x|\phi)$ with respect to some σ -finite measure ν on \mathbf{X} , then μ has a density f with respect to the same ν , given by

$$f(x) = \int p(x|\phi) m(d\phi) , \quad (5.2)$$

Integrals of the form (5.1) correspond to a two-stage sampling procedure: If we generate X_1, X_2, \dots as

- (1) sample $\Phi_i \sim m$
- (2) sample $X_i | \Phi_i \sim \mathbf{p}(\bullet | \Phi_i)$,

then each sample is distributed as $X_i \sim \mu$.

Example 5.1. An illustrative example of a mixture with continuous mixing measure is Student’s t -distribution: If we choose $p(x|\phi)$ in (5.2) as a Gaussian density with mean μ and variance σ^2 , where we fix μ and set $\phi := \sigma^2$, and choose the mixing measure m as an inverse-gamma distribution on σ^2 , then f is a Student t -density. By mixing over all possible widths of the Gaussian, we obtain a distribution with heavy tails; see Figure 5.1. \triangleleft

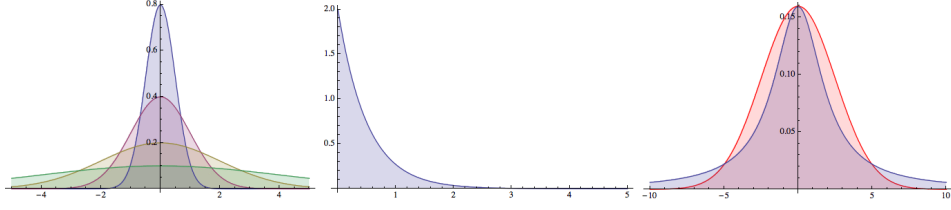


FIGURE 5.1. Continuous mixtures: If the probability kernel $\mathbf{p}(\bullet, \sigma^2)$ is a zero-mean Gaussian distribution with variance σ^2 (left), and the mixing measure m is chosen as a gamma distribution (middle), the resulting mixture distribution is Student's t -distribution (right). Note how mixing over all scales of the Gaussian generates heavy tails in the mixture.

Remark 5.2 (Bayesian models as mixtures). Compare the mixture sampling scheme above to a Bayesian model for an exchangeable sequence $X_{1:\infty}$:

- (1) $\Theta \sim Q$
- (2) $X_1, X_2, \dots | \Theta \sim P_\Theta$

We see that any such Bayesian model is a mixture, with the prior Q as its mixing measure, but we have to be careful: In a Bayesian model, we sample *one* realization of the parameter, and then generate the *entire* sample given this realization, whereas in a mixture, each X_i is generated by a separate value Φ_i . Thus, if each X_i in the Bayesian model takes values in \mathbf{X} , the model is a mixture on the sample space \mathbf{X}^∞ , with mixture components of the form P_θ^∞ . \triangleleft

5.2. Hierarchical models

Suppose we specify a Bayesian model for a data source, which we represent as the joint distribution $\mathcal{L}(X_{1:\infty})$ of an infinite random sequence. The standard Bayesian modeling assumption (1.3) decomposes the joint distribution of the sequence as

$$\mathbb{P}(X_{1:\infty} \in \bullet) = \int_{\mathbf{T}} P_\theta^\infty(\bullet) Q(d\theta) . \quad (5.3)$$

We observe that we can apply the same idea recursively, and split up Q as, say,

$$Q(\Theta \in d\theta) = \int_{\mathbf{T}'} Q[\Theta|\Theta'] Q'(\Theta' \in d\theta') , \quad (5.4)$$

for some additional random variable Θ' with law Q' and values in a space \mathbf{T}' .

Why should we? If the random object Θ is very simple (e.g. a random scalar), there is indeed no good reason to do so, but if Θ is more complicated, then the recursive decomposition above can be used to *simplify* Θ by *imposing hierarchical structure*. By imposing a hierarchical structure, we introduce layers of random variables that are not observed, i.e. latent variables.¹ A useful rule of thumb to keep in mind is:

Hierarchical models are latent variable models which impose conditional independence structure between separate layers of latent variables.

In fact, we have already seen such a structure in Bayesian mixture models, which becomes more apparent if we define the model backwards:

Example 5.3 (DP mixture). Start with the joint distribution $\mathbb{P}(X_{1:\infty} \in \bullet)$ of the data, and assume that it is a mixture as in (5.1), with a smooth component density p in (5.2). Then each observation X_i is explained by a separate value Φ_i sampled from the mixing measure, so the model parameter would be $\Theta = (\Phi_{1:\infty})$, and we have no hope to recover it from data, since we would have to estimate each Φ_i from a single data point. Since the Φ_i are exchangeable, we can model them by a *random* mixing measure Θ' with distribution Q' . If we choose Q' in (5.4) as a Dirichlet process, $Q[\Theta|\Theta']$ as the joint distribution of $\Theta = (\Phi_i)$ given the DP random measure Θ' , and P_θ as a measure with density $p(\bullet|\theta)$, we obtain precisely the Dirichlet process mixture model. (Compared to our notation in Chapter 2, the random measure has now moved up one step in the hierarchy and is denoted Θ' rather than Θ .) \triangleleft

Hierarchical nonparametric models are particularly popular in machine learning, for a number of reasons:

- In many machine learning models (such as HMMs), the layers of the hierarchy have a natural interpretation.
- We can use both the basic models of Bayesian nonparametrics, such as DPs and GPs, and standard models from machine learning, as building blocks. They can be combined into hierarchies to represent more complex models.
- If we can Gibbs-sample each layer in a hierarchy, we can Gibbs-sample the hierarchy.

Perhaps the most prominent examples are two closely related models known as the infinite hidden Markov model and the hierarchical Dirichlet process.

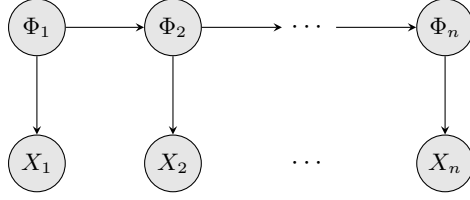
Example 5.4 (Bayesian HMM). Recall that a **hidden Markov model** is a model for time-series data X_1, X_2, \dots , where each observation X_i is indexed by a discrete time index i . The model generates observations by first generating a sequence (Φ_1, Φ_2, \dots) of parameters with values in a space Ω_ϕ , which in this context is called the **state space**. An HMM specifically imposes the assumption that the sequence $\Phi_{1:\infty}$ is a Markov chain. We additionally define a parametric model with density $p(x|\phi)$ and parameter space Ω_ϕ , and explain observations X_i as

$$X_i|\Phi_i \sim p(x|\Phi_i). \quad (5.5)$$

The parametric model defined by p is also called the **emission model**. Depicted as a graphical model, an HMM looks like this:

¹ Most authors motivate hierarchical models differently: If we define a prior controlled by a hyperparameter, and a suitable setting for that hyperparameter is not known, we can “be Bayesian about it” and define a hyperprior on the hyperparameter. If that hyperprior is controlled by yet another hyperparameter, we can add a further hyperprior, etc.

That may well be how many hierarchical models in the literature have come about, but conceptually, I have personally never found this explanation very satisfactory: By consecutively defining hyper-priors, hyper-hyper-priors etc., we cannot escape the fact that the entire hierarchy ultimately still represents a single measure Q on \mathbf{T} , so with each additional layer, we are only putting off the inevitable. The utility of hierarchies is that they impose a hierarchical structure on the random object Θ .



Recall further that a Markov chain on a finite state space Ω_ϕ is completely characterized by two parameters, (1) the distribution ν_0 of the first state Φ_1 , and (2) the **transition matrix** \mathbf{t} , which specifies the probabilities to move from state i to state j :

$$\mathbf{t} = (t_{ij}) \quad \text{where} \quad t_{ij} := \mathbb{P}[\Phi_{n+1} = \phi^j | \Phi_n = \phi^i] \quad \text{for all } \phi^i, \phi^j \in \Omega_\phi,$$

where I am using superscripts to enumerate the elements of the state space as $\Omega_\phi = \{\phi^1, \dots, \phi^K\}$.

Now observe that, if we fix a specific time index i , the observation X_i is distributed according to a finite mixture model, with component distribution $p(x|\phi)$ and mixing measure $m(\bullet) = \sum_{k=1}^{|\Omega_\phi|} c_k \delta_{\phi_k}(\bullet)$. What are the mixture weights c_k ? If we consider the distribution of X_i conditionally on the previous state $\Phi_{n-1} = \phi^i$, then clearly

$$c_k := \mathbb{P}[\Phi_n = \phi_k | \Phi_{n-1} = \phi^i] = t_{ik}. \quad (5.6)$$

If we instead marginalize out the first $(n-1)$ states, then c_k is the probability for the Markov chain (ν, \mathbf{t}) to end up in state k after n steps. Hence, we can regard a HMM with finite state space of size K as a sequence of mixture models with K components. The mixtures are tied together by the Markov chain.

We obtain a Bayesian HMM by defining a prior distribution on the parameters (ν, \mathbf{t}) of the Markov chain. The initial distribution ν is a probability on K events, so we could use a Dirichlet distribution on the simplex \triangle_K as a prior. Similarly, each row of the transition matrix \mathbf{t} is a probability distribution (again on K events), so we could e.g. sample ν and each row of \mathbf{t} independently from one and the same Dirichlet distribution. \triangleleft

Beal, Ghahramani, and Rasmussen [2] noticed that a HMM with (countably) infinite state space $\Omega_\phi = \{\phi^1, \phi^2, \dots\}$ can be obtained by making the Bayesian HMM nonparametric:

Example 5.5 (Infinite HMM [2]). Suppose we sample the distribution defining the i th row of \mathbf{t} from a Dirichlet process. More precisely, we sample a random measure $\Theta_i \sim \text{DP}(\alpha, G)$ and define

$$T_{ij} := C_j^i \quad \text{for } \Theta_i = \sum_{k \in \mathbb{N}} C_k^i \delta_{\Phi_k^i}. \quad (5.7)$$

Then each atom Φ_k^i describes a separate state and there is a countably infinite number of atoms, so the random matrix \mathbf{T} is infinite, and the state space is indeed countably infinite.

There is, however, one complication: If we sample each Θ_i independently from a Dirichlet process with continuous base measure, the sets of atoms defined by any two such random measures Θ_i and Θ_j are almost surely disjoint. (Defining the base measure on a countably infinite space does not solve this problem, it just introduces additional counting problems.) The solution is to instead make the Θ_i

conditionally independent by tying them together in a hierarchy: First, we sample a random measure Θ' from a Dirichlet process. Then, conditionally on Θ' , each Θ_i is sampled from another Dirichlet process, with Θ' as its base measure:

$$\begin{aligned}\Theta' &\sim Q' = \text{DP}(\alpha, G) \\ \Theta_1, \Theta_2, \dots | \Theta' &\sim_{\text{iid}} Q[\bullet | \Theta'] = \text{DP}(\alpha, \Theta') .\end{aligned}\tag{5.8}$$

Beal *et al.* [2] called this hierarchy a **hierachical Dirichlet process** or **HDP**.

Suppose a given atom Φ_k of Θ' has a large weight C'_k . A look at the DP sampling formula (2.22) shows that such atoms tend to occur early in the sequential sampling of Θ_i . Hence, atoms with large weight C'_k also tend to have large weights C_k^i in the measures Θ_i . The HMM defined by the resulting infinite transition matrix \mathbf{T} therefore concentrates much of its probability mass on a small subset of states (which is precisely what makes this model interesting). Beal *et al.* [2] called this model an **infinite hidden Markov model**. From a machine learning perspective, the infinite state space means the model can “add new states” as more data becomes available. See [2] for more details and for inference, and [15] for an interesting theoretical perspective. \triangleleft

In [2], the HDP was considered specifically as a device to generate the random transition matrix \mathbf{T} . Teh, Jordan, Beal, and Blei [65] observed that this model is much more widely applicable, by regarding it as a generic hierarchical representation of a family of discrete random measures $\{\Theta_1, \Theta_2, \dots\}$ which all share the same atoms. The name *hierarchical Dirichlet process* is generally used to refer to their version of the model.

Example 5.6 (Hierarchical Dirichlet process [65]). Consider the clustering setup again, where we model an exchangeable sequence $X_{1:\infty}$ e.g. by a DP mixture. The parameter is a random measure Θ . Now suppose that the observed data is naturally subdivided into (known) subsets, so in fact we observe multiple sample sequences $X_{1:n_1}^1, X_{1:n_2}^2, \dots$. We split up the prior and hierarchically decompose Θ into $\{\Theta_1, \Theta_2, \dots\}$, generated conditionally independently from a single Θ' as in (5.8). Each Θ_k is interpreted as the parameter explaining one set $X_{1:n_k}^k$ of observations.

A popular example are text document models (topic models), where each $X_{1:n_k}^k$ represents a single text document, and the individual observations X_i^k individual words. The basic assumption in such models is that a **topic** model is a distribution over the terms in a given vocabulary, represented as the parameter vector of a multinomial distribution with one category per term, and that text documents are mixtures of topics. Since each document may mix topics according to its own proportions, we need to estimate one mixture for each document k (represented by Θ_k), but all of these mixtures share the same topics (the atoms of the measure Θ'), and overall, some topics are more common in text documents than others (expressed by the weights C'_i of Θ'). See [65, 64] for more details. \triangleleft

Remark 5.7 (Gibbs-sampling hierarchies). I have already mentioned above that one of the appealing aspects of hierarchical models is that they can be constructed by using basic models as components, and that we can Gibbs-sample the hierarchy if we can Gibbs-sample each component.

I would like to complement this with a word of caution: Each time we add an additional layer in our hierarchy, the size of the state space which has to be explored

by the Gibbs sampler is multiplied by the number of states added in the additional layer. Thus, as an admittedly imprecise rule of thumb, the effective dimension of the state space grows roughly exponentially in the depth of the hierarchy. A hierarchical structure of course somewhat constrains complexity. However, for many of the more complex models we have seen in the machine learning literature in recent years, it seems unlikely that any sampler for the model posterior can actually be run long enough to have mixed. I believe we should ask ourselves seriously how much we really know about these models. \triangleleft

5.3. Covariate-dependent models

Observational data may involve **covariates**, i.e. observed variables that we do not bother to model as random—for example, because we are not trying to predict their values—but rather condition upon. In Bayesian nonparametrics, this problem was first addressed systematically by MacEachern [41], and although his work focussed on the Dirichlet process, I hope the (brief) discussion below clarifies that his ideas are much more generally applicable. For more on covariate-dependent models, see [16].

I will generically denote the covariate information as a (non-random) variable z with values in a space \mathbf{Z} . An intuitive example is time-dependence, where \mathbf{Z} is a set of time points. Say we are modeling some effect X over time. If we try to predict both the effect X and the time at which it occurs, we would include a time-valued random variable Z in the model, and attempt to predict (X, Z) . If we are instead interested in predicting X at given times z , we would regard time as a covariate, and try to predict $X(z)$.

Now suppose we have a model $M = \{P_\theta | \theta \in \mathbf{T}\}$ that we consider adequate for $X(z)$ at a fixed covariate value z . Since $X(z)$ is effectively a function of z , we have to substitute M by a z -indexed family of models $M(\mathbf{Z})$. That means the parameter θ becomes a function $\theta(z)$ of z . As a parameter for the complete covariate-dependent model, we hence have to consider functions

$$\theta(\bullet) : \mathbf{Z} \rightarrow \mathbf{T} . \quad (5.9)$$

At the very least, we will want this function to be measurable, so in its most general form, the covariate dependent model defined by M would be

$$M(\mathbf{Z}) = \{P_{\theta(\bullet)} | \theta \in \mathbf{B}(\mathbf{Z}, \mathbf{T}), P \in M\} , \quad (5.10)$$

where $\mathbf{B}(\mathbf{Z}, \mathbf{T})$ is the space of Borel functions $\mathbf{Z} \rightarrow \mathbf{T}$. We can always think of this as a \mathbf{T} -valued regression problem, and as in any regression problem, we will have to impose additional smoothness properties on the function $\theta(\bullet)$.

Definition 5.8. Let \mathbf{Z} be a standard Borel space and $\mathbf{B}(\mathbf{Z}, \mathbf{T})$ the set of Borel-measurable functions $\mathbf{Z} \rightarrow \mathbf{T}$. Let $M = \{P_t | t \in \mathbf{T}\}$ be a model on \mathbf{X} . Let Q be a prior distribution on $\mathbf{B}(\mathbf{Z}, \mathbf{T})$. We call a Bayesian model with prior Q and sample space $\mathbf{B}(\mathbf{Z}, \mathbf{T})$ a **covariate-dependent model** with covariate space \mathbf{Z} if

$$\mathcal{L}(X(z) | \Theta(\bullet)) = P_{\Theta(z)} \quad \text{for all } z \in \mathbf{Z} . \quad (5.11)$$

\triangleleft

This definition is by no means carved in stone—I have just made it up here in the hope of making ideas precise. In most applications, the definition would have

to be extended to include some form of censoring, since we rarely observe samples $X_i(z)$ for every points z .

MacEachern [41] introduced the specific case in which $\Theta(z)$ is a Dirichlet process for every $z \in \mathbf{Z}$.

Definition 5.9. Let $\mathbf{T} = \mathbf{PM}(V)$ and let $\Phi : \Omega \rightarrow \mathbf{B}(\mathbf{Z}, \mathbf{PM}(V))$ be a covariate-dependent parameter. If the law of Φ^z is a Dirichlet process for all $z \in \mathbf{Z}$, i.e. if there are measurable mappings $\alpha : \mathbf{Z} \rightarrow \mathbb{R}_{>0}$ and $G_0 : \mathbf{Z} \rightarrow \mathbf{PM}(V)$ such that

$$\mathcal{L}(\Phi^z) = \text{DP}(\alpha(z), G_0(z)) \quad \text{for all } z \in \mathbf{Z}, \quad (5.12)$$

then $\mathcal{L}(\Phi)$ is called a **dependent Dirichlet process**. \triangleleft

In general, it is far from trivial to specify a prior distribution on random measurable functions $\mathbf{Z} \rightarrow \mathbf{T}$. (Recall how difficult it is to turn a Gaussian distribution into the simplest distribution on continuous functions, the GP.) For a specific problem, we can start with a prior on \mathbf{T} and try to find a representation that can be reformulated as a function of the covariate. MacEachern [41] noticed that this is possible for the Dirichlet process using the stick-breaking construction: A DP random measure is of the form $\Theta = \sum_k C_k \delta_{\Phi_k}$. We can hence turn it into a covariate-dependent parameter $\Theta(\bullet)$ by making its components covariate-dependent,

$$\Theta(z) = \sum_{k \in \mathbb{N}} C_k(z) \delta_{\Phi_k(z)}. \quad (5.13)$$

For the component parameters Φ_k , that “simply” means we have to define random functions $\mathbf{Z} \rightarrow \Omega_\phi$; if Ω_ϕ is Euclidean, we can do so using a Gaussian process. The C_k require a bit more thought: The stick-breaking construction (2.17) shows that we can generate the C_k from i.i.d. variables V_k that are marginally beta-distributed, so we need to generate random functions $V_k : \mathbf{Z} \rightarrow [0, 1]$ such that marginally, $\mathcal{L}(V_k(z)) = \text{Beta}(1, \alpha(z))$, where α may now also be a function of \mathbf{Z} .

The arguably most widely used solution is to transform a Gaussian process using cumulative distribution functions: Suppose Y is a real-valued random variable and F its CDF. Then the random variable $F(Y)$ is uniformly distributed on $[0, 1]$. If F^{-1} is the right-continuous inverse

$$F^{-1}(w) := \{y \in \mathbb{R} | F(y) > w\} \quad (5.14)$$

of F and $U \sim \text{Uniform}[0, 1]$, then $F^{-1}(U) \stackrel{d}{=} Y$. Hence, if \tilde{V} is a random function $\mathbf{Z} \rightarrow \mathbb{R}$ sampled from a Gaussian process, we can define F_z as the CDF of the (Gaussian) marginal distribution of $\tilde{V}(z)$ and G_z as the CDF of $\text{Beta}(1, \alpha(z))$. Then

$$V(z) := G_z^{-1} \circ F_z(\tilde{V}(z)) \quad (5.15)$$

is a random function $\mathbf{Z} \rightarrow [0, 1]$ with marginals $\mathcal{L}(V(z)) = \text{Beta}(1, \alpha(z))$. We can then obtain a dependent Dirichlet process as

$$\Theta(\bullet, z) := \sum_{n=1}^{\infty} \left(V_n(z) \prod_{j=1}^{n-1} (1 - V_j(z)) \right) \delta_{\Phi_n(z)}(\bullet). \quad (5.16)$$

Intriguingly, Θ is a random measurable function $\mathbf{Z} \rightarrow \mathbf{PM}(\mathbf{T})$, and hence a random probability kernel (or random conditional probability).

If modeling dependence of the weights C_k on z is not considered important, we can greatly simplify this model by setting

$$\Theta(\bullet, z) := \sum_{n=1}^{\infty} \left(V_n \prod_{j=1}^{n-1} (1 - V_j) \right) \delta_{\Phi_n(z)}(\bullet), \quad (5.17)$$

called the **single- p model** in [41]. Note that (5.17) is simply a Dirichlet process on a function space, where each atom location $\Phi_k(\bullet)$ is a function, with a Gaussian process as its base measure.

CHAPTER 6

Exchangeability

Recall how we informally described statistical inference in Section 1.2 as the process of extracting an underlying pattern (represented by the model parameter) from observational data. A Bayesian approach models the unknown pattern as a random variable. Again (very) informally, the idea is that we decompose the randomness in the data source into two parts, as

$$data = underlying\ pattern + sample\ randomness \quad (6.1)$$

(where I would ask you to read the “+” symbolically, not as an arithmetic operator). In a specific Bayesian model, these two parts correspond to the prior and the observation model. We can only hope to extract an underlying pattern from observations, if (1) a common pattern exists and (2) it is not completely obfuscated by the sample randomness. Exchangeability properties provide criteria for when this is possible. The best-known result of this type is of course de Finetti’s theorem, but it is actually just the basic (and historically first) example of a larger class of theorems, which explain the consequences of exchangeability for a wide range of random structures. In this chapter, I will briefly discuss three important results: The theorems of de Finetti, of Kingman, and of Aldous and Hoover. For more details, see [48].

6.1. Bayesian models and conditional independence

In applications of Bayesian modeling, we typically see Bayes’ theorem used in a way that looks more or less like this:

$$Q[d\theta | X_{1:n} = x_{1:n}] =_{a.s.} \frac{\prod_{i=1}^n p_{\theta}(x_i)}{\int_{\mathbf{T}} \prod_{i=1}^n p_{\theta'}(x_i) Q(d\theta')} Q(d\theta) , \quad (6.2)$$

where Q is a prior distribution and p a likelihood density. A closer look shows that this setup implies a substantial modeling assumption beyond the choice of Q and p : We have assumed that, for a fixed instance θ of Θ , the joint likelihood of the sample $X_{1:n}$ factorizes, i.e. that

$$p_{n,\theta}(x_1, \dots, x_n) = \prod_{i=1}^n p_{\theta}(x_i) . \quad (6.3)$$

Since Θ is a random variable, this means we assume that observations X_1, X_2, \dots are conditionally independent (and identically distributed) given Θ . More formally, we call random variables X_1, \dots, X_n **conditionally independent** given a random variable Θ if

$$\mathbb{P}[X_{1:n} \in dx_1 \times \dots \times dx_n | \Theta] =_{a.s.} \prod_{i=1}^n \mathbb{P}[X_i \in dx_i | \Theta] . \quad (6.4)$$

Conditional independence of X and X' given Θ is often also denoted $X \perp\!\!\!\perp_{\Theta} X'$. If the conditional distribution $\mathbb{P}[X_i \in dx_i | \Theta]$ on the right-hand side above is identical for all X_i , we say that the X_i are **conditionally i.i.d.** given Θ . In this case, we can obviously write

$$\mathbb{P}[X_{1:n} \in dx_1 \times \dots \times dx_n | \Theta] =_{\text{a.s.}} \prod_{i=1}^n P_{\Theta}(dx_i) \quad (6.5)$$

for some family of distributions P_{θ} on \mathbf{X} .

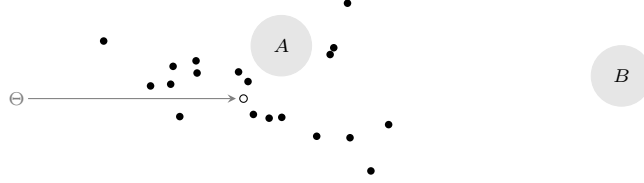
It is really important to understand that conditional independence is not simply an arbitrary modeling assumption—arguably, it is the heart and soul of Bayesian modeling. In terms of our “data=pattern + sample randomness” idea above, conditional independence means that, given the pattern Θ , all remaining randomness in the sample completely decouples (is stochastically independent between samples). In other words:

If observations are conditionally independent given Θ , all joint information in data sampled from the source is contained in Θ .

From a Bayesian statistics perspective, this means Θ (and only Θ) is the information we want to extract from data. The fundamental question we ask in this chapter is:

Under which conditions can we assume conditional independence of observations given some random quantity?

Example 6.1. To illustrate the difference between known and unknown Θ , suppose we sample data from a Gaussian with known, fixed covariance, and Θ is simply the Gaussian’s mean:



What does the observed data tell us about the probability of whether the next observation will occur in the shaded area A or B , respectively? That depends on whether or not we know where Θ is:

- (1) If we do not know the location of Θ , the data indicates the next sample is more likely to be located in A than in B . That means the observed sample $X_{1:n}$ carries information about the next sample point X_{n+1} , and $X_{1:n}$ and X_{n+1} are hence stochastically dependent—they couple through the unknown location Θ .
- (2) If we do know Θ , we can precisely compute the probability of A and B , and the observed sample provides no further information. All information $X_{1:n}$ can possibly provide about X_{n+1} is contained in Θ , so knowing Θ decouples $X_{1:n}$ and X_{n+1} .

Similarly, in the regression example in Figure 1.2, the observations are conditionally independent given the regression function. \triangleleft

6.2. Prediction and exchangeability

We can approach the problem from a different angle by taking a predictive perspective, as we have done in Example 6.1: Under what conditions can we predict

observation X_{n+1} from a recorded sample $X_{1:n}$? For simplicity, we first consider two samples of equal size n :

$$\underbrace{X_1, \dots, X_n}_{\text{already observed}}, \underbrace{X_{n+1}, \dots, X_{2n}}_{\text{future observations}} \quad (6.6)$$

Suppose we use some statistical tool to extract information from $X_{1:n}$. If we hope to use this information to make predictions about $X_{n+1:2n}$, then whatever it is that we have extracted must still be valid for $X_{n+1:2n}$. The most general form of “information” extractable from $X_{1:n}$ is of course the joint distribution $\mathcal{L}(X_{1:n})$. We see that there are two possible cases in which we may be able to predict $X_{n+1:2n}$:

- $X_{1:n}$ and $X_{n+1:2n}$ contain—up to finite-sample effects—the same information. In terms of distributions, this would mean

$$\mathcal{L}(X_{1:n}) = \mathcal{L}(X_{n+1:2n}) . \quad (6.7)$$

- $\mathcal{L}(X_{1:n})$ and $\mathcal{L}(X_{n+1:2n})$ differ, but the difference can be estimated from $X_{1:n}$ —for example, if $X_{1:\infty}$ represents a time series with a drift, the two laws would differ, but we may be able to estimate the drift and correct for it.

Since we are looking for a general and reasonably simple result, we only consider the first case; the second one would be a mess of assumptions and special cases.

We note that the first case implies we could swap the two blocks of samples in (6.6), and predict $X_{1:n}$ from $X_{n+1:2n}$ just as well as the other way around: Since any two random variables satisfy $\mathcal{L}(Y|Z)\mathcal{L}(Z) = \mathcal{L}(Z|Y)\mathcal{L}(Y)$, and since conditional probabilities are a.s. unique, (6.7) implies

$$\mathcal{L}(X_{n+1:2n}|X_{1:n}) =_{\text{a.s.}} \mathcal{L}(X_{1:n}|X_{n+1:2n}) . \quad (6.8)$$

Again combined with (6.7), this additionally means that the joint distribution of $X_{1:2n}$ does not change if we swap the blocks:

$$\mathcal{L}(X_{n+1:2n}, X_{1:n}) = \mathcal{L}(X_{1:n}, X_{n+1:2n}) \quad (6.9)$$

For general prediction problems, the block structure assumed in (6.6) is rather arbitrary. Instead, we would rather ask under which conditions we can predict any one observation in a sample from the remaining ones. In terms of swapping variables around, that means: If we consider prediction of X_{n+1} given $X_{1:n}$, we can swap X_{n+1} with any element of $X_{1:n}$:

$$\underbrace{X_1, \dots, X_m, \dots, X_n}_{\text{predict}}, \underbrace{X_{n+1}}_{\text{observed}} \longrightarrow \underbrace{X_1, \dots, X_{n+1}, \dots, X_n}_{\text{observed}}, \underbrace{X_m}_{\text{predict}}$$

As observations come in one by one, each element of the sequence is at some point the most recent one. The assumption hence implies we can swap any variable with any other one to its left. By making such swaps repeatedly, we can generate any possible rearrangement of the sample (since the set of transpositions—of permutations which only swap two elements—forms a generator of the symmetric group). The counterpart of (6.9) is then

$$\mathcal{L}(X_1, \dots, X_{n+1}) = \mathcal{L}(X_{\pi(1)}, \dots, X_{\pi(n+1)}) \quad (6.10)$$

for any permutation π of $\{1, \dots, n+1\}$.

This should of course hold for any n —unless we have reason to assume that out-of-sample prediction is only possible up to a certain sample size (e.g. if a drift kicks in after some fixed number of samples, or in similarly exotic settings). We hence consider an infinite sequence $X_{1:\infty}$, and demand that (6.10) holds for any permutation that affects at most the first $n+1$ elements, for any n . In other words, the joint distribution has to be invariant under permutations of \mathbb{N} that exchange an arbitrary but finite number of elements. The set of all such permutations is called the **infinite symmetric group** and denoted \mathbb{S}_∞ .

Definition 6.2. The random sequence $X_{1:\infty} = (X_1, X_2, \dots)$ is called **exchangeable** if its joint distribution does not depend on the order in which the values X_i are observed. More formally, if

$$(X_1, X_2, \dots) \stackrel{d}{=} (X_{\pi(1)}, X_{\pi(2)}, \dots) \quad \text{for all } \pi \in \mathbb{S}_\infty, \quad (6.11)$$

or, expressed in terms of the joint distribution, $\pi(\mathcal{L}(X)) = \mathcal{L}(X)$ for every π . \triangleleft

From a statistical modeling perspective, we can paraphrase the definition as:

exchangeability = the order of observations does not carry relevant information

Remark 6.3. It can be shown that we can alternatively define exchangeability in terms of all bijections π of \mathbb{N} (i.e. we additionally include those permutations which change an infinite number of elements); the two definitions are equivalent. \triangleleft

We set out originally to find a criterion for whether a sequence is conditionally i.i.d. Which sequences are exchangeable? An i.i.d. sequence clearly is, simply because its distribution is a product and products commute. It is easy to see that the same is true for a sequence which is conditionally i.i.d. given some Θ , since its conditional distribution given Θ is a product as in (6.5), and there is just a single Θ for all X_i . Thus,

$$\{ \text{conditionally i.i.d. sequences} \} \subset \{ \text{exchangeable sequences} \}.$$

To obtain a criterion, we hence have to ask which exchangeable sequences are *not* conditionally i.i.d. (and which additional conditions we may have to impose to exclude such troublemakers). The rather amazing answer, given by de Finetti's theorem, is that there are no troublemakers: A sequence $X_{1:\infty}$ is exchangeable if and only if it is conditionally i.i.d. given *some* Θ . Exchangeability is hence precisely the criterion we are looking for. Zabell [68] gives a very insightful account of prediction and exchangeability.

6.3. de Finetti's theorem

If an infinite sequence $X_{1:\infty}$ is conditionally i.i.d. given Θ , then by definition,

$$\mathbb{P}[X_{1:\infty} \in dx_1 \times dx_2 \times \dots | \Theta] = \prod_{i \in \mathbb{N}} \mathbb{P}[X_i \in dx_i | \Theta]. \quad (6.12)$$

If we define $P_\theta(\bullet) := \mathbb{P}[X_i \in \bullet | \Theta = \theta]$, we obtain a family of measures

$$M := \{P_\theta | \theta \in \mathbf{T}\}. \quad (6.13)$$

Since Θ is random, P_Θ is a random variable with values in $\mathbf{PM}(\mathbf{X})$, and hence a random probability measure. We abbreviate the factorial distribution as

$$P_\theta^\infty(dx_1 \times dx_2 \times \dots) = \prod_{i \in \mathbb{N}} P_\theta(dx_i). \quad (6.14)$$

Now, if we choose *any* family M , and any random variable Θ with values in \mathbf{T} , then P_Θ is the joint distribution of a conditionally i.i.d. sequence. Since we already know that all conditionally i.i.d. sequences are exchangeable, so in general, we have to permit any measure in $\mathbf{PM}(\mathbf{X})$, and hence $M = \mathbf{PM}(\mathbf{X})$. Therefore, we simply choose $\mathbf{T} = \mathbf{PM}(\mathbf{X})$, so any parameter value θ is a probability measure, and $P_\theta = \theta$. We can now interpret Θ as a random probability measure.

Theorem 6.4 (de Finetti). *An infinite random sequence $X_{1:\infty}$ is exchangeable if and only if there is a random probability measure Θ on \mathbf{X} such that*

$$\mathbb{P}[X_{1:\infty} \in \bullet | \Theta] =_{a.s.} \Theta^\infty(\bullet) \quad (6.15)$$

<

If you have seen de Finetti's theorem before, you are probably more used to the form (6.16) below, which is a direct consequence of (6.15): The quantities on both sides of (6.15) are random variables, and the randomness in both is given by the randomness in Θ . If we integrate both sides of the equation against $\mathcal{L}(\Theta)$ —that is, if we compute expectations with respect to Θ —we obtain:

Corollary 6.5. *A random sequence X is exchangeable if and only if*

$$\mathbb{P}(X \in \bullet) = \int_{\mathbf{PM}(\mathbf{X})} \theta^\infty(\bullet) \nu(d\theta) \quad (6.16)$$

for some distribution ν on $\mathbf{PM}(\mathbf{X})$.

<

We have to be a bit careful here: (6.16) states that two random variables are equal in expectation, whereas (6.15) says that the same two variables are equal almost surely, which is a much stronger statement. The argument above, that we integrate both sides of (6.15), therefore only establishes that exchangeability implies (6.16), but not the converse. However, the right-hand side of (6.16) is a mixture, and we know that mixtures can be sampled in two stages, which here means sampling $\Theta \sim \nu$ and then sampling $X_{1:n} | \Theta$ from Θ^∞ , so $X_{1:\infty}$ is conditionally i.i.d. We already know that conditionally i.i.d. sequences are exchangeable.

The theorem has another important implication: Since $X_{1:\infty}$ is conditionally i.i.d., we can apply the law of large numbers [e.g. 28, Theorem 4.23] *conditionally* on Θ . For any measurable set A , it tells us that, almost surely given Θ , the probability of A under the empirical measure defined by $X_{1:n}$ converges to the probability $\Theta(A)$. We therefore have, as the second direct consequence of Theorem 6.4:

Corollary 6.6. *If the random sequence X is exchangeable, then*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \delta_{X_i(\omega)}(\bullet) \xrightarrow{\text{weakly}} \Theta(\omega) \quad \nu\text{-a.s.}, \quad (6.17)$$

where Θ is the random measure in (6.15).

<

To make the step from the argument above (almost sure convergence for every A) to the actual corollary (weak convergence a.s.), we have tacitly used the Polish topology of \mathbf{X} .

Finally, I would like to point out that we can alternatively state de Finetti's theorem directly in terms of random variables, rather than in terms of their distributions: Suppose $\theta \in \mathbf{PM}(\mathbf{X})$ is a probability measure on \mathbf{X} . We denote the i.i.d.

random sequence sampled from this measure as

$$X_\theta^\circ := (X_1, X_2, \dots) \quad \text{where} \quad X_1, X_2, \dots \sim_{\text{iid}} \theta, \quad (6.18)$$

that is, $\mathcal{L}(X_\theta^\circ) = \theta^\infty$. We then get an *exchangeable* sequence by randomizing θ , i.e. if Θ is a random probability measure on \mathbf{X} , then X_Θ° is an exchangeable sequence. de Finetti's theorem can then be restated as

$$X_{1:\infty} \text{ exchangeable} \quad \Leftrightarrow \quad X_{1:\infty} =_{\text{a.s.}} X_\Theta^\circ \quad \text{for some } \Theta \in \mathbf{RV}(\mathbf{PM}(\mathbf{X})). \quad (6.19)$$

This perspective will be very useful for the more advanced representation theorems discussed below, which are much more elegantly stated in terms of random variables than in terms of distributions.

6.4. Exchangeable partitions

Not all problems are naturally represented by random sequences—for some problems, a random graph or random matrix, for example, may be a better fit for the data. For such problems, we can still draw on exchangeability properties for Bayesian modeling, we just have to substitute the representation theorem for exchangeable sequences (de Finetti) by a suitable representation for another exchangeable structure. In Chapter 2, we have considered random partitions of \mathbb{N} as solutions of clustering problems, and exchangeable partitions will be the first example of a more intricate exchangeable structure we consider.

We already noted in Chapter 2 that, given any method that maps a sample to a clustering solution, any random sample X_1, \dots, X_n induces a random partition of $[n]$. If the X_i are exchangeable, the induced random partition also has an exchangeability property: Suppose we permute the sequence (X_i) and obtain $(X_{\pi(i)})$. If the clustering solution for, say, X_1, \dots, X_5 is

$$(\{1, 2, 4\}, \{3, 5\}), \quad (6.20)$$

the solution for $X_{\pi(1)}, \dots, X_{\pi(5)}$ would be

$$(\{\pi(1), \pi(2), \pi(4)\}, \{\pi(3), \pi(5)\}). \quad (6.21)$$

Since (X_i) and $(X_{\pi(i)})$ are equally distributed, the two partitions above have the same probability of occurrence, and are hence equivalent for statistical purposes.

Remark 6.7. In particular, the permutation may change the order of the blocks (e.g. if π would swap 1 and 3 in the example above). The enumeration of the blocks by their index k is hence completely arbitrary. This fact is known in the clustering literature as the *label switching problem*, and can be rather inconvenient, since it implies that any statistical method using the clustering solution as input has to be invariant to permutation of the cluster labels. As our discussion above shows, it is a direct consequence of exchangeability of the observations. \triangleleft

To formalize what we mean by exchangeability of random partitions, recall from Section 2.6 how we encoded a random partition $\Psi = (\Psi_1, \Psi_2, \dots)$ of \mathbb{N} by a random sequence (L_i) , with $L_i = k$ if $i \in \Psi_k$.

Definition 6.8. An **exchangeable random partition** Ψ is a random partition of \mathbb{N} whose law is invariant under the action of any permutation π on \mathbb{N} . That is,

$$(L_1, L_2, \dots) \stackrel{\text{d}}{=} (L_{\pi(1)}, L_{\pi(2)}, \dots) \quad (6.22)$$

for all $\pi \in \mathbb{S}_\infty$. \triangleleft

The counterpart to de Finetti's theorem for partitions is Kingman's representation theorem. We will state this theorem in a form similar to (6.19). To do so, we have to define a specific type of random partitions Ψ_θ° which play a role analogous to that of i.i.d. sequences in de Finetti's theorem. These random partitions were introduced by Kingman, who called them “paint-box partitions”.

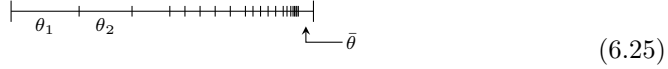
The natural parameter space for these partitions turns out to be the set of a specific type of sequences which are called mass partitions in some parts of the literature [see e.g. 3]. By a **mass partition** θ , we mean a partition of the unit interval of the form

$$\theta = (\theta_1, \theta_2, \dots, \bar{\theta}) \text{ with } \bar{\theta} := 1 - \sum_{i \in \mathbb{N}} \theta_i \quad (6.23)$$

which satisfies

$$\theta_1 \geq \theta_2 \geq \dots \geq 0 \quad \text{and} \quad \sum_i \theta_i \leq 1. \quad (6.24)$$

A mass partition might look like this:



This is just the kind of partition we generated using the stick-breaking construction in (2.15), with the difference that the interval lengths generated by stick-breaking need not be monotonically decreasing (although they decrease in expectation). The Poisson-Dirichlet distribution (cf. Remark 2.10) is a distribution on mass partitions.

Given a mass partition θ , we can sample a partition of \mathbb{N} by throwing uniform random variables U_1, U_2, \dots on the unit interval (think of the indices of these variables as the elements of \mathbb{N}). Since the mass partition subdivides $[0, 1]$ into subintervals, we just have to record which U_i end up in the same subinterval, and regard their indices as elements of the same block:

Definition 6.9. Let θ be a mass partition, and define a random partition Ψ_θ° of \mathbb{N} as follows: Let $U_{1:\infty}$ be a sequence of i.i.d. uniform random variables in $[0, 1]$ and let

$$L_i := k \quad \Leftrightarrow \quad U_i \in \left[\sum_{j=1}^{k-1} \theta_j, \sum_{j=1}^k \theta_j \right). \quad (6.26)$$

Then Ψ_θ° is called the **paint-box partition** with parameter θ . \triangleleft

Note that U_i may be larger than $\sum_{j \in \mathbb{N}} \theta_j$, i.e. in the example in (6.25), it would end up in the rightmost interval of length $\bar{\theta}$. If so, the definition implies that i forms a block of its own in Ψ_θ° , since the probability of inserting a second number into the same block is zero. These singleton blocks are called **dust**. Clearly, the partition Ψ_θ° is exchangeable.

If the paint-box partitions play a role analogous to i.i.d. sequences, then the set of all mass partitions plays a role analogous to that of the parameter space $\mathbf{PM}(\mathbf{X})$ in the de Finetti representation. We denote this set as

$$\Delta^\downarrow := \{ \theta \in [0, 1]^\mathbb{N} \mid \theta \text{ mass partition} \}. \quad (6.27)$$

So far, this is just a set of points; to turn it into a space, we need to define a topology, and we have already noted several times that we like our spaces to be Polish. We can metrize Δ^\downarrow by defining the metric $d(\theta, \theta') := \max_{k \in \mathbb{N}} |\theta_k - \theta'_k|$. The

metric space (Δ^d, d) is compact [3, Proposition 2.1], and hence Polish (since all compact metric spaces are Polish).

Theorem 6.10 (Representation theorem, Kingman). *A random partition Ψ is exchangeable if and only if*

$$\Psi =_{a.s.} \Psi_{\Theta}^{\circ} \quad (6.28)$$

for some random mass partition $\Theta \in \mathbf{RV}(\Delta^d)$. \triangleleft

As in the sequence case, this result immediately implies an integral decomposition:

Corollary 6.11 (Integral decomposition). *A random partition Ψ is exchangeable if and only if*

$$\mathbb{P}(\Psi \in \bullet) = \int_{\Delta^d} \mathbf{p}(\bullet, \theta) \nu(d\theta), \quad (6.29)$$

where \mathbf{p} denotes the paintbox distribution $\mathbf{p}(\bullet, \theta) := \mathcal{L}(\Psi_{\theta}^{\circ})(\bullet)$. \triangleleft

Perhaps more important is the fact that the mass partition θ —the model parameter, from a statistical perspective—can asymptotically be recovered from observations:

Corollary 6.12 (Law of large numbers). *If Ψ is an exchangeable random partition, then*

$$\frac{\sum_{i=1}^n \mathbb{I}\{L_i = k\}}{n} \xrightarrow{n \rightarrow \infty} \theta_k \quad (6.30)$$

\triangleleft

For this reason, the elements θ_k of θ are also called the **asymptotic frequencies** of Ψ .

6.5. Exchangeable arrays

The next type of structure we consider are (infinite) collections of variables indexed by d indices,

$$x := (x_{i_1, \dots, i_d})_{i_1, \dots, i_d \in \mathbb{N}} \quad \text{where } x_{i_1, \dots, i_d} \in \mathbf{X}_0. \quad (6.31)$$

Such a structure x is called a **d -array**. Clearly, sequences are 1-arrays, but d -arrays are much more versatile: A matrix, for example, is a 2-array where \mathbf{X}_0 is an algebraic field (so that adding and multiplying entries of x , and hence matrix multiplication, is well-defined). A simple graph—a graph without multiple edge—is represented by a 2-array with $\mathbf{X}_0 = \{0, 1\}$, the adjacency matrix of the graph. If the matrix is symmetric, the graph is undirected. To keep things simple, I will only discuss 2-arrays in this section. Although results for general d -arrays are a straightforward generalization, they are notationally cumbersome. See [48, Section 6] for more on d -arrays.

Suppose X is a random 2-array. To define exchangeability for arrays, we have to decide which components of X we are going to permute. In sequences, we simply permuted individual elements. If data is represented as a 2-array X , however, this typically implies that the row-column structure carries some form of meaning—otherwise we could just write the entries into a sequence—and an adequate notion of exchangeability should therefore preserve rows and columns. That is, if two entries are in the same row, they should still be in the same row after the permutation has been applied, even if the order of elements within the row may have changed (and

similarly for columns). Hence, rather than permuting entries of X , we permute only its rows and columns. There are two ways of doing so: We could either apply the same permutation π to the rows and to the columns, or we could use one permutation π_1 on the rows and another π_2 on the columns.

Definition 6.13. A random 2-array $X = (X_{ij})_{i,j \in \mathbb{N}}$ is **jointly exchangeable** if

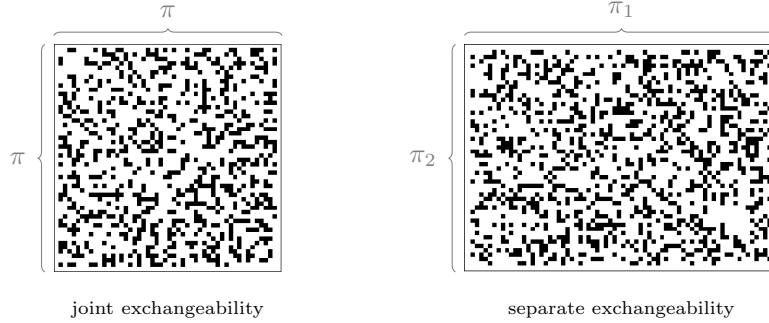
$$(X_{ij}) \stackrel{d}{=} (X_{\pi(i)\pi(j)}) \quad \text{for every } \pi \in \mathbb{S}_\infty . \quad (6.32)$$

X is **separately exchangeable** if

$$(X_{ij}) \stackrel{d}{=} (X_{\pi_1(i)\pi_2(j)}) \quad \text{for every pair } \pi_1, \pi_2 \in \mathbb{S}_\infty . \quad (6.33)$$

◁

For a binary matrix (with 1s encoded as black dots), this looks like this:



A simple way to generate a random matrix would be to define a function f with values in \mathbf{X}_0 , say with two arguments. Now sample two sequences of random variables (U_1, U_2, \dots) and (U'_1, U'_2, \dots) . If we set $X_{ij} := f(U_i, U'_j)$ we get a random 2-array. If the sequences (U_i) and (U'_j) have independent elements and are independent of each other, X_{ij} is clearly separately exchangeable. If we set $X_{ij} = f(U_i, U_j)$ instead, it is jointly exchangeable. It is not hard to see that this cannot be all jointly or separately exchangeable arrays, though: If we start with either of the arrays above and randomize each entry independently—that is, if we include a third argument $f(\bullet, \bullet, U_{ij})$, where the random variables U_{ij} are independent—we do not break exchangeability. Changing the distribution of the random variables U_i etc. does not give us any additional expressive power, since we can always equivalently change the function f . Hence, we can simply choose all variables as i.i.d. uniform (or some other convenient, simple distribution).

Definition 6.14. Let $\mathbf{F}(\mathbf{X}_0)$ be the space of measurable functions $\theta : [0, 1]^3 \rightarrow \mathbf{X}_0$. Let (U_i) and (V_i) be two i.i.d. sequences and (U_{ij}) an i.i.d. 2-array, all consisting of Uniform $[0, 1]$ random variables. For any $\theta \in \mathbf{F}$, define two random arrays J_θ° and S_θ° as

$$J_\theta^\circ := (J_{ij}) \quad \text{with} \quad J_{ij} := \theta(U_i, U_j, U_{ij}) \quad (6.34)$$

and

$$S_\theta^\circ := (S_{ij}) \quad \text{with} \quad S_{ij} := \theta(U_i, V_j, U_{ij}) . \quad (6.35)$$

◁

Rather amazingly, it turns out that these arrays J_θ° and S_θ° play a role analogous to that of i.i.d. sequences and paintbox partitions—that is, *any* exchangeable array can be obtained by making the function θ random.



FIGURE 6.1. Functions w (“graphons”) representing different types of random graph models. *Left to right:* Undirected graph with linear edge density (the standard example in [37]), nonparametric block model for separately exchangeable data [29], Mondrian process model for separately exchangeable data [55], graphon with Gaussian process distribution for undirected graph [35]. Figure from [48].

Theorem 6.15 (Aldous, Hoover). *A random 2-array $X = (X_{ij})$ with entries in a Polish space \mathbf{X}_0 is jointly exchangeable if and only if*

$$X \stackrel{d}{=} J_{\Theta}^{\circ} \quad \text{for some } \Theta \in \mathbf{RV}(\mathbf{F}(\mathbf{X}_0)) . \quad (6.36)$$

It is separately exchangeable if and only if

$$X \stackrel{d}{=} S_{\Theta}^{\circ} \quad \text{for some } \Theta \in \mathbf{RV}(\mathbf{F}(\mathbf{X}_0)) . \quad (6.37)$$

◁

Remark 6.16 (Exchangeable random graphs and graph limits). Suppose X is in particular the adjacency matrix of a random simple graph (where simple means there is at most one edge between two vertices if the graph is undirected, or at most one edge in each direction in the directed case). Then X is a random binary matrix, which is symmetric iff the graph is undirected.

Because X is binary, we can simplify the random function Θ from three to two arguments: For a fixed function $\theta \in \Theta(\{0,1\})$ and a uniform random variable U , $\theta(x,y,U)$ is a random element of $\{0,1\}$. If we define

$$w(x,y) := \mathbb{P}[\theta(x,y,U) = 1] , \quad (6.38)$$

then w is an element of the set \mathbf{W} of measurable functions $[0,1]^2 \rightarrow [0,1]$. For a fixed function $w \in \mathbf{W}$, we can sample the adjacency matrix X of an exchangeable random graph G_w° as:

$$\begin{aligned} U_1, U_2, \dots &\sim_{\text{iid}} \text{Uniform}[0,1] \\ X_{ij} &\sim \text{Bernoulli}(w(U_i, U_j)) \end{aligned} \quad (6.39)$$

Theorem 6.15 then implies that *any* exchangeable random graph G can be obtained by mixing over w : If and only if G is exchangeable,

$$G \stackrel{d}{=} G_W^{\circ} \quad \text{for some } W \in \mathbf{RV}(\mathbf{W}) . \quad (6.40)$$

The functions w are also called **graphons** or **graph limits** in random graph theory, see [37]. I will not go into further details, but rather refer to [48]. ◁

6.6. Applications in Bayesian statistics

You will have noticed that all exchangeability theorems discussed above—de Finetti, Kingman, Aldous-Hoover, the special case of Aldous-Hoover for graphs—had a common structure: We consider a random structure X (sequence, graph, partitions,...). We assume that this structure has an exchangeability property.

exchangeable structure	ergodic structures	representation	\mathbf{T}
sequences in \mathbf{X}	i.i.d. sequences	de Finetti	$\mathbf{PM}(\mathbf{X})$
partitions of \mathbb{N}	paintbox partitions J_θ°	Kingman	Δ^\downarrow
graphs	graphs G_θ° in (6.39)	Aldous-Hoover	\mathbf{W}
arrays (jointly)	arrays J_θ° in (6.34)	Aldous-Hoover	\mathbf{F}
arrays (separately)	arrays S_θ° in (6.35)	Aldous-Hoover	\mathbf{F}

TABLE 6.1.

If so, the relevant representation theorem specifies some space \mathbf{T} and a family of random variables X_θ° , parametrized by elements $\theta \in \mathbf{T}$. The representation result then states that X is exchangeable if and only if it is of the form

$$X \stackrel{d}{=} X_\Theta \quad \text{for some random element } \Theta \in \mathbf{RV}(\mathbf{T}) . \quad (6.41)$$

The special structures X_θ are also called the **ergodic structures**.

Here is one of my favorite examples of how to apply representation results in Bayesian modeling:

Example 6.17 (Priors for graph-valued data [35]). Suppose we consider data represented by a single, large graph (a network, say). As more data is observed, the graph grows. Can we define a Bayesian model for such data? We can interpret the observed graph G_n (with n vertices) as a small snapshot from an infinite graph G (just as we would interpret n sequential observations as the first n elements in an infinite sequence). If we assume G to be exchangeable, by Theorem 6.15, there is a random function W with $G \stackrel{d}{=} G_W^\circ$. In other words, we explain our observed graph as the first n vertices sampled from W according to (6.39).

Thus, in order to define a prior distribution for data modeled as an exchangeable graph, we *have to define a prior distribution on the function space \mathbf{W}* . We could define a parametric model (by choosing a finite-dimensional subspace of \mathbf{W}), or a nonparametric one; in terms of the nonparametric models we have already discussed, we could generate \mathbf{W} using a Gaussian process, as in [35]. We can also consider models for graph data defined in the literature; if these models implicitly assume exchangeability, we can categorize them according to what type of function W characterizes the particular model. Figure 6.1 shows examples. \triangleleft

Let me try to sketch the bigger picture: For a given type of exchangeable structure X , the ergodic random structures X_θ° define a family of distributions

$$P_\theta := \mathcal{L}(X_\theta^\circ) . \quad (6.42)$$

These distributions are generic—each representation theorem tells us how to sample from P_θ for a given θ (e.g. by the paint-box sampling scheme in Kingman’s theorem if X is an exchangeable partition). Thus, *any* statistical model of exchangeable X is of the form

$$M = \{P_\theta | \mathbf{T}_0 \subset \mathbf{T}\} , \quad (6.43)$$

where the parameter space of the model is some subset \mathbf{T}_0 of the space \mathbf{T} characterized by the relevant representation theorem. Defining a Bayesian model then means defining a prior Q on \mathbf{T}_0 . Table 6.1 summarizes the examples we have seen in this chapter.

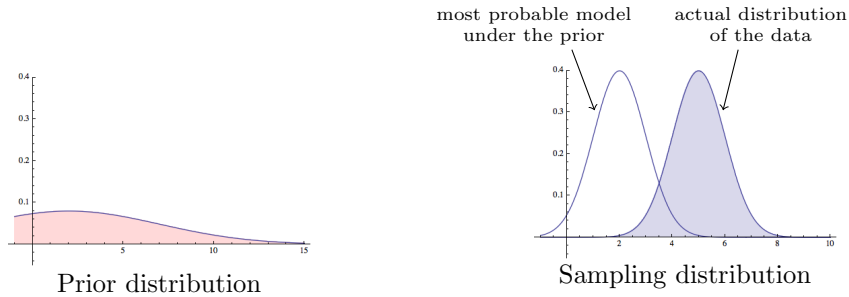
In particular, we can consider the Dirichlet process again: To define a prior distribution for an exchangeable partition, we can invoke Kingman's theorem, and hence have to define prior distribution on the space $\mathbf{T} = \Delta^{\downarrow}$ of mass partitions. Clearly, a stick-breaking distribution as in (2.14) does just that, if we subsequently order the weights generated by stick-breaking by decreasing size, as assumed in the definition of Δ^{\downarrow} . (From a sampling perspective, ranking by size makes no actual difference, since it clearly leaves the paint-box distribution invariant; it is simply a device to enforce uniqueness of the parametrization.) If we specifically choose the stick-breaking construction of the DP as our prior Q , the resulting exchangeable random partition is the CRP.

CHAPTER 7

Posterior distributions

This chapter discusses theoretical properties of posterior distributions. To motivate the questions we will try to address, let me run through a drastically oversimplified example of Bayesian inference:

Example 7.1 (Unknown Gaussian mean). We assume that the data is generated from a Gaussian on \mathbb{R} with fixed variance σ^2 ; the mean θ is unknown. Hence, the observation model is $p(x|\theta, \sigma) = g(x|\theta, \sigma)$ (where g is the Gaussian density on the line). We assume that the mean Θ is random, but since we do not know its distribution, we have to make a modeling assumption for Q . Suppose we have reason to believe that Θ is roughly distributed according to a Gaussian with density $g(\theta|\mu_0 = 2, \sigma_0 = 5)$:

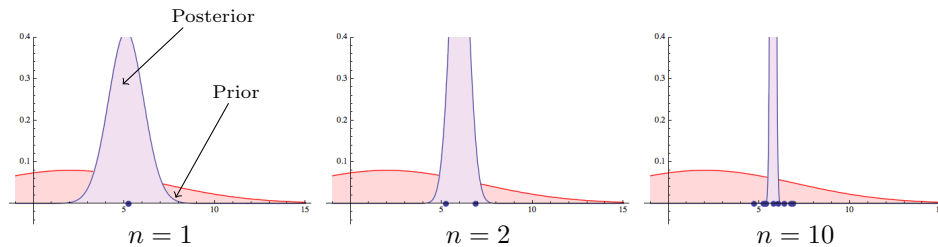


Since circa 68% of the mass of a Gaussian is located within one standard deviation of the mean, this prior distribution expresses the assumption that $\mu_0 = 2$ is the most probable value of Θ and that $\Theta \in [-3, 7]$ with probability ≈ 0.68 .

Our objective is now to compute the posterior distribution. In a simple model like this the posterior can be computed using Bayes' theorem—which we have not discussed in detail yet, we will do so in Section 7.2 below. Using the theorem, we find that the posterior under n observations with values x_1, \dots, x_n is again a Gaussian with density $g(\theta|\mu_n, \sigma_n)$ and parameters

$$\mu_n := \frac{\sigma^2 \mu_0 + \sigma_0^2 \sum_{i=1}^n x_i}{\sigma^2 + n \sigma_0^2} \quad \text{and} \quad \sigma_n := \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n \sigma_0^2}. \quad (7.1)$$

To plot these posteriors and compare them to the prior density, I have sampled $n = 10$ data points from the observation model with parameter $\theta = 6$, i.e. from $\mathbf{p}(\bullet, 6)$. The posteriors after the first few observations look like this:



The observations are plotted as blue dots on the line. Although we know of course that we should never try to perform statistical estimation from 10 observations, we see that the posterior concentrates very rapidly at the actual parameter value. \triangleleft

We see that Example 7.1 raises a list of implicit questions, which all happen to have an easy answer in the case above, but require more effort in general:

- (I) **How should we choose the prior?** (*We will see that this question cannot be answered in isolation from choosing the observation model; this is hence the modeling problem of Bayesian statistics, analogous to the modeling problem in the classical case.*)
- (II) **Does the posterior always exist?** (*As long as we work on a parameter space with reasonable topological properties: Yes.*)
- (III) **If the posterior exists, how do we determine it?** (*There are several tools, including Bayes' theorem, conjugacy, and sampling, and each is only applicable in certain settings. There is no universally applicable answer.*)
- (IV) **Asymptotically, will we find the right answer? How much data do we need to get a reasonable approximation?** (*The first step here is to carefully define what we mean by "right answer". Once we have done so, this leads to mathematical statistics for Bayesian models, which is every bit as rich and faceted as for the classical case.*)

The existence problem (II) is discussed in Section 7.1 below, Bayes' theorem in Section 7.2 and conjugacy in Section 7.4. Section 7.7 provides a very brief overview and further references on the asymptotics problem.

7.1. Existence of posteriors

The question which can be answered in most general terms is that for existence of a posterior: Suppose we have defined a parameter random variable Θ with law Q and values in \mathbf{T} , and an observation model $M = \{P_\theta | \theta \in \mathbf{T}\}$ consisting of measures on some space \mathbf{X} . We generate an observation as

$$\begin{aligned} \Theta &\sim Q \\ X|\Theta &\sim P_\Theta. \end{aligned} \tag{7.2}$$

The observation variable X may be pretty much anything; it may describe a single point, a sequence of length n , a random structure, or an infinitely large sample.

The posterior is the probability kernel

$$\mathbf{q}(\bullet, x) := \mathbb{P}[\Theta \in \bullet | X = x], \tag{7.3}$$

and our question is simply under which condition the object \mathbf{q} exists and has the properties of a probability kernel (a regular conditional probability, see Appendix C.1). The existence result then follows directly from the standard result on the existence of regular conditional probabilities (Theorem C.2). This depends only

on the topological properties of the parameter space; the choice of the model, the prior, and even of the sample space \mathbf{X} are irrelevant.

Proposition 7.2. *If \mathbf{T} is a standard Borel space, \mathbf{X} a measurable space, and a Bayesian model is specified as in (7.2), the posterior (7.3) exists.* \triangleleft

See Appendix B.2 for more on standard Borel spaces.

7.2. Bayes' theorem

The posterior guaranteed by the existence result above is an abstract mathematical object; the proof of existence is not constructive in any practically feasible sense. To actually use a Bayesian model, we have to find a way to compute the posterior from the prior, the observation model and the observed sample. The first such way we discuss is Bayes' theorem, which is rather generally applicable for parametric Bayesian models, though often not for nonparametric ones.

Suppose we observe a sequence $X_{1:n}$ of observations. If we assume that the sequence is exchangeable, de Finetti's theorem tells us that the X_i are conditionally i.i.d. given some random probability measure Θ on \mathbf{X} . Suppose we know that Θ takes its values in some subset $\mathbf{T} \subset \mathbf{PM}(\mathbf{X})$. To use a more familiar modeling notation, we can define $Q := \mathcal{L}(\Theta)$ and

$$P_\theta(\bullet) := \theta(\bullet) \quad \text{for all } \theta \in \mathbf{T} . \quad (7.4)$$

The data is then explained as

$$\begin{aligned} \Theta &\sim Q \\ X_1, \dots, X_n | \Theta &\sim_{\text{iid}} P_\Theta . \end{aligned} \quad (7.5)$$

If we know Q and P_θ , how do we determine the posterior $\mathbb{P}[\Theta \in \bullet | X_{1:n}]$?

Bayes' theorem is the density-based approach to this problem. Recall the basic textbook result on existence of conditional densities (included in the appendix as Lemma C.4): Suppose suitable densities $p(x_{1:n}, \theta)$ of the joint distribution and $p(x_{1:n})$ of the marginal distribution exist. By Lemma C.4, the conditional distribution $\mathbb{P}[\Theta \in \bullet | X_{1:n}]$ is determined by the density

$$p(\theta | x_{1:n}) = \frac{p(x_{1:n}, \theta)}{p(x_{1:n})} . \quad (7.6)$$

The trick is to specify the density $p(x_{1:n}, \theta)$ in its second argument θ with respect to the prior measure Q . Then $p(\theta | x_{1:n})$ is also a density with respect to the prior, and we have

$$\mathbb{P}[\Theta \in d\theta | X_{1:n}] = p(\theta | x_{1:n}) Q(d\theta) . \quad (7.7)$$

Thus, we have obtained a density that allows us to start with the one measure on \mathbf{T} that we know—the prior Q —and transform it into the posterior, using a transformation parametrized by the data.

Bayes' theorem, formally stated as Theorem 7.3 below, provides:

- (1) A sufficient condition for the relevant densities to exist.
- (2) A specific expression for the density $p(\theta | x_{1:n})$ in terms of the density of the model P_θ .

More precisely, it shows that a sufficient condition for the existence of the density (7.7) is the existence of a conditional density $p(x|\theta)$ of the observation model: There must be *some* σ -finite measure μ on \mathbf{X} such that

$$P_\theta(X \in dx) = p(x|\theta)\mu(dx) \quad \text{for all } \theta \in \mathbf{T}. \quad (7.8)$$

(This is really rather remarkable, since it provides a condition for the absolute continuity of the posterior with respect to the prior, both measures on \mathbf{T} , purely in terms of the observation model on \mathbf{X} .) If so, the theorem also shows that

$$p(\theta|x_{1:n}) = \frac{\prod_{i=1}^n p(x_i|\theta)}{p(x_1, \dots, x_n)}, \quad (7.9)$$

and our transformation rule for turning the prior into the posterior is hence

$$Q[d\theta|X_1 = x_1, \dots, X_n = x_n] = \frac{\prod_{i=1}^n p(x_i|\theta)}{p(x_1, \dots, x_n)} Q(d\theta). \quad (7.10)$$

Identity (7.10) is known as the **Bayes equation**. Formally stated, Bayes theorem looks like this:

Theorem 7.3 (Bayes' Theorem). *Let $M = \mathbf{p}(\bullet, \mathbf{T})$ be an observation model and $Q \in \mathbf{PM}(\mathbf{T})$ a prior. Require that there is a σ -finite measure μ on \mathbf{X} such that $\mathbf{p}(\bullet, \theta) \ll \mu$ for every $\theta \in \mathbf{T}$. Then the posterior under conditionally i.i.d. observations X_1, \dots, X_n as in (7.5) is given by (7.10), and $\mathbb{P}\{p(X_1, \dots, X_n) \in \{0, \infty\}\} = 0$.*

◁

The proof is straightforward: We know from Lemma C.4 that the conditional density is of the form $p(x|\theta)/p(x)$. Two things might go wrong:

- We have to verify that the prior dominates the posterior.
- We have to make sure that the quotient $p(x|\theta)/p(x)$ exists. Since $p(x|\theta)$ is given, we only have to verify that $p(x) \notin \{0, 1\}$ with probability 1.

PROOF. It is sufficient to proof the result for $n = 1$. The probability that $p(x|\theta)/p(x)$ does *not* exist is

$$\mathbb{P}\{p(X) \in \{0, \infty\}\} = \int_{p^{-1}\{0\}} p(x)\mu(dx) + \int_{p^{-1}\{\infty\}} p(x)\mu(dx). \quad (7.11)$$

The first term is simply $\int 0 d\mu = 0$. Since $p(x|\theta)$ is a conditional density, $p(x)$ is a μ -density of $\mathcal{L}(X)$. As the μ -density of a finite measure, it can take infinite values at most on a μ -null set, which means the second term also vanishes.

$$\mathbb{P}(X \in dx, \Theta \in d\theta) = \mathbf{p}(dx, \theta)Q(d\theta) = p(x|\theta)\mu(dx)Q(d\theta) \quad (7.12)$$

This means that $p(x|\theta)\mathbf{1}(\theta) = p(x|\theta)$ is a *joint* density of $\mathcal{L}(X, \Theta)$ with respect to $\mu \otimes Q$, which implies $\mathcal{L}(X, \Theta) \ll \mu \otimes Q$. By Lemma C.4, this is sufficient for the existence of a conditional density, which according to (C.10) is given by (7.10). \square

7.3. Dominated models

If you have read papers on Bayesian nonparametrics before, you may have noticed that the Bayes equation is not used very frequently in the nonparametric context. The problem is that many Bayesian nonparametric models do not satisfy the conditions of Bayes' theorem: If $\mathbb{P}[d\theta|X_{1:n}]$ is the posterior of a Dirichlet process, for example, then there is *no* σ -finite measure ν which satisfies $\mathbb{P}[d\theta|X_{1:n} = x_{1:n}] \ll \nu$ for all $x_{1:n}$. In particular, the prior does not, and so there is no density $p(\theta|x_{1:n})$.

Since this type of problem is fairly fundamental to Bayesian nonparametrics, I will discuss it in some more detail in this section, even though it is rather technical by nature.

A set M of probability measures is called **dominated** if there is a σ -finite measure μ on \mathbf{X} such that $P \ll \mu$ for every $P \in M$. We then also say that M is dominated by μ . Since the posterior is a conditional, it defines a family of distributions

$$\mathcal{Q}_n := \{\mathbb{P}[\Theta \in \bullet | X_{1:n} = x_{1:n}] | x_{1:n} \in \mathbf{X}^n\} \quad (7.13)$$

on \mathbf{T} . The condition for the Bayes equation to exist is hence that \mathcal{Q}_n is dominated by the prior Q for all sample sizes n , and we can paraphrase Bayes' theorem as:

If the observation model M is dominated, then \mathcal{Q}_n is dominated by the prior for all n .

To understand the concept of a dominated model better, it is useful to define

$$\mathbf{N}(\mu) := \{A \in \mathcal{B}(\mathbf{X}) | \mu(A) = 0\} \quad \text{and} \quad \mathbf{N}(M) := \bigcap_{\mu \in M} \mathbf{N}(\mu). \quad (7.14)$$

$\mathbf{N}(\mu)$ is the set of all null sets of μ . Informally, think of $\mathbf{N}(\mu)$ as a pattern in the set of Borel sets; in general, this pattern differs for different measures. Absolute continuity can now be stated as

$$\nu \ll \mu \quad \Leftrightarrow \quad \mathbf{N}(\nu) \supset \mathbf{N}(\mu). \quad (7.15)$$

Hence, M is dominated iff there is a σ -finite μ such that $\mathbf{N}(M) \supset \mathbf{N}(\mu)$. A dominating measure μ for M has to assign positive mass to every measurable set not contained in $\mathbf{N}(M)$. Since $\mathbf{N}(M)$ becomes smaller the more different null set patterns M contains, the number of distinct null set patterns in a dominated model must be limited—simply because a σ -finite measure does not have that much mass to spread around. At most, the number of distinct patterns can be countable:

Lemma 7.4 (Halmos and Savage [22]). *Every dominated set $M \subset \mathbf{M}(\mathbf{X})$ of measures has a countable subset M' such that every $\mu \in M$ satisfies $\mathbf{N}(\mu) = \mathbf{N}(\mu')$ for some $\mu' \in M'$.* \triangleleft

The lemma does not imply any restrictions on the size of a dominated set, as the next example illustrates.

Example 7.5. Let G_x be the unit-variance Gaussian measure on \mathbb{R} with mean x , and δ_x the Dirac measure at x . The two models

$$M_G := \{G_x | x \in \mathbb{R}\} \quad \text{and} \quad M_D := \{\delta_x | x \in \mathbb{R}\} \quad (7.16)$$

contain exactly the same number of measures. Lemma 7.4 shows, however, that M_D is not dominated. The set is an extreme example, since $\mathbf{N}(M_D) = \emptyset$. A dominating measure thus would have to assign positive mass to every non-empty set and hence could not possibly be σ -finite. In contrast, all measures in M_G have the same null sets, and $\mathbf{N}(M_G)$ are precisely the null sets of Lebesgue measure. \triangleleft

As we have already mentioned above, the Dirichlet process posterior is not dominated by the prior, nor in fact by any σ -finite measure. More generally, that is the case for all priors based on random discrete probability measures. I will explain the argument in two stages, first for the prior, then for any σ -finite measure.

Proposition 7.6. *Let $\Theta = \sum_k C_k \delta_{\Phi_k}$ be any random discrete probability measure on Ω_ϕ and require that the joint distribution $\mathcal{L}(\Phi_{1:\infty})$ is non-atomic. Then the posterior of Θ is not dominated by the prior, that is,*

$$\mathcal{Q}_n \not\ll \mathcal{Q} \quad \text{for all } n \in \mathbb{N}. \quad (7.17)$$

◁

PROOF. Suppose the first observation takes value $\Phi_1 = \phi$. Then we know that Θ takes its value in the (measurable) subset

$$M_\phi := \{\theta \in \mathbf{PM}(\mathbf{X}) \mid \theta \text{ has atom at } \phi\}. \quad (7.18)$$

Since $\mathcal{L}(\Phi_{1:n})$ is non-atomic, the prior assigns probability zero to M_ϕ . We hence have

$$\mathcal{L}(\Theta)(M_\phi) = 0 \quad \text{and} \quad \mathcal{L}(\Theta \mid \Phi_1 = \phi)(M_\phi) = 1 \quad (7.19)$$

and hence $\mathcal{L}(\Theta \mid \Phi_1 = \phi) \not\ll \mathcal{L}(\Theta)$. For $n > 1$, we can simply replace ϕ by $\phi_{1:n}$ in the argument. \square

We can generalize this result from the prior to any σ -finite measure on \mathbf{T} with only marginally more effort:

Proposition 7.7. *Let $\Theta = \sum_k C_k \delta_{\Phi_k}$ be any random discrete probability measure on Ω_ϕ and require that (1) the joint distribution $\mathcal{L}(\Phi_{1:\infty})$ is non-atomic and (2) Θ has an infinite number of non-zero weights C_k . Then the family*

$$\mathcal{Q}_\infty := \bigcup_{n \in \mathbb{N}} \mathcal{Q}_n \quad (7.20)$$

of posteriors of Θ for all sample sizes is not dominated by any σ -finite measure, i.e.

$$\mathcal{Q}_\infty \not\ll \nu \quad \text{for all } \nu \in \mathbf{PM}(\mathbf{T}). \quad (7.21)$$

◁

PROOF. Let S be a countable subset of Ω_ϕ . Similar as above, let

$$M_S := \{\theta = \sum_k C_k \delta_{\Phi_k} \mid \{\phi_1, \phi_2, \dots\} = S\}. \quad (7.22)$$

For any two distinct countable sets $S_1 \neq S_2$ of points in Ω_ϕ , the sets M_{S_1} and M_{S_2} of measures are disjoint (unlike the sets M_ϕ above, which is why we have changed the definition). For any two sets $S_1 \neq S_2$, there exists a finite sequence $\phi_{1:n}$ which can be extended to a sequence in S_1 , but not to one in S_2 . Hence, $\mathcal{L}(\Theta \mid \Phi_{1:n} = \phi_{1:n})(M_{S_2}) = 0$. Since Ω_ϕ is uncountable, there is an uncountable number of distinct sets S , and by Lemma 7.4, \mathcal{Q}_∞ cannot be dominated. \square

7.4. Conjugacy

The most important alternative to Bayes theorem for computing posterior distributions is conjugacy. Suppose M is an observation model, and we now consider a *family* $\mathcal{Q} \subset \mathbf{PM}(\mathbf{T})$ of prior distributions, rather than an individual prior. We assume that the family \mathcal{Q} is indexed by a parameter space \mathbf{Y} , that is, $M = \{Q_y \mid y \in \mathbf{Y}\}$. Many important Bayesian models have the following two properties:

- (i) The posterior under any prior in \mathcal{Q} is again an element of \mathcal{Q} ; hence, for any specific set of observations, there is an $y' \in \mathbf{Y}$ such that the posterior is $Q_{y'}$.

- (ii) The posterior parameter y' can be computed from the data by a simple, tractable formula.

This is basically what we mean by conjugacy, although—for historical reasons—the terminology is a bit clunky, and conjugacy is usually defined as property (i), even though property (ii) is the one that really matters for inference. For Bayesian non-parametrics, conjugacy is almost all-important: We have already seen in Theorem 2.3 that the DP is conjugate, but most Bayesian nonparametric models at some level rely on a conjugate posterior.

To make things more precise, we require as always that \mathcal{Q} is measurable as a subset of $\mathbf{PM}(\mathbf{T})$, and that the parametrization $y \mapsto Q_y$ is bijective and measurable. The prior parameter y is often called a **hyperparameter**, and although it can be randomized, we may simply consider it a non-random control parameter.

Definition 7.8. An observation model $M \subset \mathbf{PM}(\mathbf{X})$ and the family of priors \mathcal{Q} are called **conjugate** if, for any sample size n and any observation sequence $x_{1:n} \in \mathbf{X}^n$, the posterior under any prior $Q \in \mathcal{Q}$ is again an element of \mathcal{Q} . \triangleleft

The definition captures precisely property (i) above. Since $y \mapsto Q_y$ is measurable, we can represent the family \mathcal{Q} as a probability kernel

$$\mathbf{q} : \mathbf{Y} \rightarrow \mathbf{PM}(\mathbf{T}) \quad \text{where} \quad \mathbf{q}(\bullet, y) = Q_y. \quad (7.23)$$

We can now think of the model parameter—our usual Θ —as a parameterized random variable Θ^y , with law $\mathcal{L}(\Theta^y) = \mathbf{q}(\bullet, y)$ (see Theorem C.3 in the appendix, which also shows that Θ^y depends measurably on y). If the model is conjugate in the sense of Definition 7.8, then for every prior Q_y in \mathcal{Q} and every observation sequence $x_{1:n}$, there is a $y' \in \mathbf{Y}$ such that

$$\mathbb{P}[\Theta^y \in d\theta | X_{1:n} = x_{1:n}] = \mathbf{q}(d\theta, y'). \quad (7.24)$$

If we define a mapping as $T_n(y, x_{1:n}) := y'$, for the value y' in (7.24), the posterior is given by

$$\mathbb{P}[\Theta^y \in d\theta | X_{1:n} = x_{1:n}] = \mathbf{q}(d\theta, T_n(y, x_{1:n})). \quad (7.25)$$

Since the posterior depends measurably on y and $x_{1:n}$, and

$$T_n(y, x_{1:n}) = \mathbf{q}^{-1}(\mathbb{P}[\Theta^y \in \bullet | X_{1:n} = x_{1:n}]),$$

the mapping T_n is always measurable. I will refer to the sequence (T_n) of mappings as a **posterior index** for the model; there is no standard terminology.

All these definitions leave plenty of room for triviality: For any observation model M , the family $\mathcal{Q} := \mathbf{PM}(\mathbf{T})$ is trivially conjugate, and the identity map on $\mathbf{Y} \times \mathbf{X}^n$ is always a posterior index. The definitions are only meaningful if property (ii) above is satisfied, which means that T_n should be of known, easily tractable form for all n .

In parametric models, the only class of models with interesting conjugacy properties, i.e. for which (i) holds nontrivially and (ii) also holds, are exponential family models and their natural conjugate priors (up to some borderline exceptions). I will define exponential families in detail in the next section; for now, it suffices to say that if M is an exponential family model with sufficient statistics S , and \mathcal{Q} the natural conjugate family of priors for M with parameters (λ, y) , then the posterior index is well-known to be

$$T_n(y, x_{1:n}) = (\lambda + n, y + \sum_{i \leq n} S(x_i)) \quad \text{for any } n. \quad (7.26)$$

For lack of a better name, I will refer to models with posterior index of the general form (7.26) as **linearly conjugate**. Linear conjugacy is not restricted to parametric models: If the prior is $\text{DP}(\alpha, G)$, we choose y in (7.26) as $y := \alpha \cdot G$ and $\lambda := y(\Omega_\phi) = \alpha$. By Theorem 2.3, the posterior is computed by updating the prior parameters as

$$(\alpha, G, \phi_1, \dots, \phi_n) \mapsto \frac{1}{n + \alpha} \left(\alpha G + \sum_{i=1}^n \delta_{\phi_i} \right), \quad (7.27)$$

and the model is hence linearly conjugate with posterior index

$$T_n(\alpha, \alpha G, \phi_1, \dots, \phi_n) := \left(\alpha + n, \alpha G + \sum_{i=1}^n \delta_{\phi_i} \right). \quad (7.28)$$

Similarly, it can be shown (although it requires a bit of thought) that the posterior of the GP in Theorem 4.4 is linearly conjugate.

The somewhat confusing nomenclature used for conjugacy in the literature—where conjugacy is defined as property (i), but the desired property is (ii)—is due to the fact that, in the parametric case, (i) and (ii) coincide: If M and \mathcal{Q} are parametric families and satisfy (i), then they are (more or less) exponential families, in which case they are linearly conjugate and hence satisfy (ii). This is no longer true in nonparametric models:

Example 7.9. In Chapter 2, we had seen how a homogeneous random probability measure Θ can be generated by generating weights C_k from the general stick-breaking construction (2.15), and then attaching atom locations Φ_k sampled i.i.d. from a measure G on Ω_ϕ . If in particular $\Omega_\phi = \mathbb{R}_+$ or $\Omega_\phi = [a, b]$ (so that it is totally ordered with a smallest element), Θ is called a **neutral-to-the-right (NTTR) process**¹. It can be shown that the posterior of a NTTR process Θ under observations $\Phi_1, \Phi_2, \dots | \Theta \sim_{\text{iid}} \Theta$ is again a NTTR process [9, Theorem 4.2]. The model is hence conjugate in the sense of Definition 7.8, that is, it is closed under sampling. There is, however, in general no known explicit form for the posterior index, except in specific special cases (in particular the Dirichlet process). \triangleleft

There are many things we still do not know about conjugacy in the nonparametric case, but basically all important models used in the literature—Dirichlet and Pitman-Yor process models, Gaussian process regression models, beta process used with IBPs, etc.—are linearly conjugate. It can be shown that linearly conjugate nonparametric models are closely related to exponential family models [47]: Roughly speaking, a nonparametric (and hence infinite-dimensional) prior can be represented by an infinite family of finite-dimensional priors—these are the finite-dimensional marginals that we discussed in Chapter 4 for the Gaussian process (in which case they are multivariate Gaussians). For the Dirichlet process, the finite-dimensional marginals are, unsurprisingly, Dirichlet distributions (see Section 8.9).

A nonparametric model is linearly conjugate if and only if the finite-dimensional marginals are linearly conjugate, so linearly conjugate Bayesian nonparametric models can be constructed by assembling suitable families of exponential family distributions into an infinite-dimensional model. I will omit the details here, which are rather technical, and refer to [47].

Remark 7.10 (“Non-conjugate” DP mixtures). You will find references in the literature on DP mixtures and clustering referring to the “non-conjugate” case.

Recall from Section 2.5 that, in order to derive the update equations in a simple Gibbs sampler for DP mixtures, it was convenient to choose the parametric mixture components $p(x|\phi)$ as exponential family models and the base measure G of the DP as a conjugate prior. There are hence two levels of conjugacy in this model: Between the DP random measure and its posterior, and between p and G . References to non-conjugate DP mixtures always refer to the case where p and G are not conjugate; samplers for this case exist, but they still rely on conjugacy of the DP posterior. \triangleleft

7.5. Gibbs measures and exponential families

Arguably the most important class of models in parametric Bayesian statistics are exponential families. These models are special in a variety of ways: They are the only parametric models that admit finite-dimensional sufficient statistics, they are maximum entropy models, and they admit conjugate priors.

Exponential family models are specific classes of Gibbs measures, which are defined using the concept of entropy. We start with a σ -finite measure μ on \mathbf{X} , and write $\mathcal{P}(\mu)$ for the set of all probability measures which have a density with respect to μ ,

$$\mathcal{P}(\mu) := \{P \in \mathbf{PM}(\mathbf{X}) \mid P \ll \mu\}. \quad (7.29)$$

The elements of $\mathcal{P}(\mu)$ are often called the **probability measures generated by** μ . Choose a measure $P \in \mathcal{P}(\mu)$ with density f under μ . The **entropy** of P is

$$H(P) := \mathbb{E}_P[-\log f(X)] = - \int_{\mathbf{X}} f(x) \log f(x) \mu(dx). \quad (7.30)$$

Regarded as a functional $H : \mathcal{P}(\mu) \rightarrow \mathbb{R}_+$, the entropy is concave.

Gibbs measures are measures which maximize the entropy under an expectation constraint. More precisely, let $S : \mathbf{X} \rightarrow \mathbf{S}$ be a measurable function with values in a Banach space \mathbf{S} . We fix a value $s \in \mathbf{S}$, and ask: Which distribution among all P that satisfy $\mathbb{E}_P[S] = s$ has the highest entropy? This is an equality-constraint optimization problem,

$$\begin{aligned} & \max_{P \in \mathcal{P}(\mu)} H(P) \\ \text{s.t.} \quad & \mathbb{E}_P[S] = s. \end{aligned} \quad (7.31)$$

Since H is concave, the problem has a unique solution in $\mathcal{P}(\mu)$ for every value of s , provided the constraint is satisfiable. We can reformulate the optimization as a Lagrange problem:

$$\max_{P \in \mathcal{P}(\mu)} H(P) - \langle \theta, \mathbb{E}_P[S] \rangle. \quad (7.32)$$

If $\mathbf{S} = \mathbb{R}^d$, we can simply read this as $H(P) - (\theta_1 \mathbb{E}_P[S_1] + \dots + \theta_d \mathbb{E}_P[S_d])$, i.e. we have d equality constraints and d Lagrange multipliers θ_i . There is nothing particularly special about the finite-dimensional case, though, and θ may be an element

¹The stick-breaking construction (2.15) was by in the early 1970s by Doksum [9], who also showed that the DP is a special case of a NTTR prior. The fact that the DP is obtained by choosing H in (2.14) as a beta distribution, however, was only pointed out several years later by Sethuraman and Tiwari [58], and not published in detail until much later [57]. Doksum [9, Theorem 3.1] also gives a precise characterization of random measures for which stick-breaking constructions of the form (2.15) exist: Θ is NTTR if and only if there is a subordinator (an positive Lévy process) Y_t such that $\Theta([0, t]) \stackrel{d}{=} 1 - \exp(-Y_t)$ for all $t \in \Omega_\phi$.

of an infinite-dimensional Hilbert space with inner product $\langle \bullet, \bullet \rangle$. The Banach space \mathbf{S} need not even have a scalar product, however, and in general, θ is an element of the norm dual \mathbf{S}^* of \mathbf{S} , and $\langle \bullet, \bullet \rangle$ in (7.32) is the evaluation functional $\langle x, x^* \rangle = x^*(x)$.²

Denote by $\mathbf{T} \subset \mathbf{S}^*$ the set of θ for which (7.32) has a solution. For every $\theta \in \mathbf{T}$, the solution is a unique measure $\mu^\theta \in \mathcal{P}(\mu)$. These measures μ^θ are called **Gibbs measures**, and we write

$$\mathcal{G}(S, \mu) := \{\mu^\theta \mid \theta \in \mathbf{T}\}. \quad (7.33)$$

Solving the problem explicitly is not completely straightforward, since the entropy is, concavity aside, not a particularly nice functional.³ If $\mathbf{S} = \mathbb{R}^d$ (and hence $\mathbf{S}^* = \mathbb{R}^d$ as well), it can be shown under mild technical conditions that Gibbs measures are given by the densities

$$\mu^\theta(dx) = Z_\theta^{-1} e^{\langle S(x), \theta \rangle} \mu(dx) \quad \text{with} \quad Z_\theta := \mu(e^{\langle S(\bullet), \theta \rangle}), \quad (7.34)$$

where I am using the “French” notation $\mu(f) = \int f(x) \mu(dx)$. The normalization term Z_θ is called the **partition function** in physics.

Recall that a statistic S is called **sufficient** for a set M of probability measures if all measures in M have identical conditional distribution given S .

Proposition 7.11. *If $\mathcal{G}(S, \mu)$ has the density representation (7.34), then S is a sufficient statistic for $\mathcal{G}(S, \mu)$.* \triangleleft

In this case, the model $\mathcal{G}(S, \mu)$ is also called an **exponential family** with respect to μ with sufficient statistic S . (In the general Banach space case, terminology is less well established.) The sufficiency result above is of course beaten to death in every introductory textbook, but usually proven in an unnecessarily technical fashion (e.g. invoking Neyman-Pearson), and I spell out a proof here only to clarify that the argument is really elementary.

PROOF. Suppose we condition μ^θ on $S(X) = s$ (i.e. someone observes $X = x$ but only reports the value s to us). The only remaining uncertainty in X is then which specific point $x \in S^{-1}\{s\}$ has been observed. A look at (7.34) shows that the distribution of X within $S^{-1}\{s\}$ depends on the choice of μ , but not at all on the value of θ . Since μ is the same for all μ^θ , this means that $\mu^\theta[\bullet \mid S(X) = s]$ does not depend on θ , so S is sufficient for $\mathcal{G}(S, \mu)$. \square

²Recall that a **Banach space** is a vector space \mathbf{X} with a norm $\|\cdot\|$, which is **complete** (it contains all limits of sequences when convergence is defined by $\|\cdot\|$). Associated with every Banach space is its **norm dual** \mathbf{X}^* . Formally, \mathbf{X}^* is the space of linear mappings $\mathbf{X} \rightarrow \mathbb{R}$, and equipped with the generic norm $\|x^*\| := \sup_{\|x\|=1} x^*(x)$, it is again a Banach space.

The intuition is roughly this: The simplest Banach spaces are Hilbert spaces, such as Euclidean space. We know that, on \mathbb{R}^d , every linear functional $\mathbb{R}^d \rightarrow \mathbb{R}$ can be represented as a scalar product $\langle x, x^* \rangle$ for some fixed element x^* of \mathbb{R}^d itself. This defines a duality between linear functionals *on* \mathbf{X} and points *in* \mathbf{X} , and we can thus identify \mathbf{X}^* and \mathbf{X} . For non-Hilbert \mathbf{X} , there is no notion of a scalar product, but we can verify that the duality between elements of \mathbf{X} and \mathbf{X}^* still holds. Even though we can no longer regard \mathbf{X} and \mathbf{X}^* as one and the same space, they are always twins with regard to their analytic properties (dimensionality, separability, etc). It is therefore useful (and customary) to write the mapping x^* as $x^*(x) =: \langle x, x^* \rangle$, and to think of the operation $\langle \bullet, \bullet \rangle$ as something like a scalar product between elements of two different spaces. (If \mathbf{X} is in particular a Hilbert space, it is its own dual, and $\langle \bullet, \bullet \rangle$ is precisely the scalar product). In this sense, the pair consisting of a Banach space and its dual is *almost* a Hilbert space, albeit with a slightly split personality.

7.6. Conjugacy in exponential families

Suppose we choose any exponential family model $M = \mathcal{G}(S, \mu)$ on \mathbf{X} . If we sample n observations $x_{1:n}$ i.i.d. from μ^θ , the joint density is the product of the densities (7.34). It is useful to define

$$S(x_{1:n}) := S(x_1) + \dots + S(x_n) \quad \text{and} \quad f_n(s, \theta) := Z_\theta^{-n} e^{\langle s, \theta \rangle}. \quad (7.35)$$

We can then write the joint density of $x_{1:n}$ concisely as

$$\prod_{i=1}^n Z_\theta^{-1} e^{\langle S(x_i), \theta \rangle} = f_n(S(x_{1:n}), \theta). \quad (7.36)$$

If we choose any prior Q on \mathbf{T} , Bayes' theorem 7.3 is applicable, since $\mathcal{G}(S, \mu)$ is by definition dominated by μ . Substituting into the Bayes equation (7.10), we see that the posterior is

$$Q[d\theta | X_{1:n} = x_{1:n}] = \frac{f_n(S(x_{1:n}), \theta)}{Q(f_n(S(x_{1:n}), \bullet))} Q(d\theta). \quad (7.37)$$

The definition of $f_\bullet(\bullet)$ is not just a shorthand, though, but rather emphasizes a key property of Gibbs measures:

$$f_n(s, \theta) \cdot f_{n'}(s', \theta) = f_{n+n'}(s + s', \theta). \quad (7.38)$$

That means in particular that we can define a conjugate prior by using a density with shape $f_{n'}(s', \theta)$ (i.e. which is equal to f up to scaling, since we still have to normalize f to make it a density with respect to θ).

In more detail: If

$$Q(d\theta) \propto f_{n'}(s', \theta) \nu(d\theta), \quad (7.39)$$

the posterior is, up to normalization,

$$\mathbb{P}[d\theta | X_{1:n} = x_{1:n}] \propto f_n(S(x_{1:n}), \theta) f_{n'}(s', \theta) \nu(d\theta) = f_{n'+n}(s' + S(x_{1:n}), \theta) \nu(d\theta).$$

Note that this still works if we replace the integer n' by any positive scalar λ . If we substitute in the definition of f and summarize, we obtain:

Proposition 7.12. *Let \mathbf{T} be the parameter space of an exponential family model $\mathcal{G}(S, \mu)$, and ν any σ -finite measure on \mathbf{T} . The measures*

$$Q^{\lambda, \gamma}(d\theta) := \frac{e^{\langle \gamma, \theta \rangle - \lambda \log Z_\theta}}{\nu(e^{\langle \gamma, \bullet \rangle - \lambda \log Z_\bullet})} \nu(d\theta) \quad \text{for } \lambda > 0, \gamma \in \mathbf{S} \quad (7.40)$$

form a conjugate family of priors for $\mathcal{G}(S, \mu)$ with posterior index

$$T_n((\lambda, \gamma), x_{1:n}) = (\lambda + n, \gamma + S(x_{1:n})). \quad (7.41)$$

◁

³ The straightforward way to solve the Lagrange problem would be to maximize (7.32) with respect to P . Again, there is nothing special about the finite-dimensional case, since we can define directional derivatives in a general Banach space just as in \mathbb{R}^d , so analytic minimization is in principle possible. (The direct generalization of the vector space derivative from \mathbb{R}^d to a Banach space is called the **Fréchet derivative**, see [38, Chapter 7 & 8] if you want to learn more.) That does not work for the entropy, though: Even on a simple domain (say, distributions on $[0, 1]$), the entropy is concave and upper semi-continuous [e.g. 59, Chapter I.9], but nowhere continuous in the weak topology. In particular, it is not differentiable. There is to the best of my knowledge also no general known form for the Fenchel-Legendre conjugate, and variational calculus has to be invoked as a crutch to solve the optimization problem where possible.

The measures $Q^{\lambda, \gamma}$ are called the **natural conjugate priors** for the model $\mathcal{G}(S, \mu)$, and we write

$$\mathcal{G}^\circ(S, \mu, \nu) := \{Q^{\lambda, \gamma} \mid \lambda > 0, \gamma \in \mathbf{S}\} . \quad (7.42)$$

Clearly, $\mathcal{G}^\circ(S, \mu, \nu)$ is itself an exponential family. More specifically,

$$\mathcal{G}^\circ(S, \mu, \nu) = \mathcal{G}(S^\circ, \nu) \quad \text{where } S^\circ(\theta) := (\theta, \log Z_\theta) . \quad (7.43)$$

Useful tables of the most important models and their conjugate priors—Gaussian and Gauss-Wishart, Multinomial and Dirichlet, Poisson and gamma, etc.—can be found in many textbooks, or simply on Wikipedia.

Remark 7.13. Another way to describe the derivation above is as follows: We start with the sampling distribution $\mathcal{L}_\theta(X)$ and assume there is a sufficient statistics S . Since a sufficient statistic completely determines the posterior, we can forget all information in X not contained in $S(X)$ and pass to the image measure $\mathcal{L}_\theta(S(X_{1:n}))$. Suppose $p_n(s|\theta)$ is a density of this image measure under some measure $\mu(dx)$. A natural conjugate prior is then obtained by choosing some measure $\nu(d\theta)$ and re-normalizing $p_n(s|\theta)$ such that it becomes a density in θ , viz.

$$q(\theta|n, s) := \frac{p_n(s|\theta)}{\int p_n(s|\theta)\nu(d\theta)} . \quad (7.44)$$

This program only works out nicely in the exponential family case, since exponential family models are basically the only models which admit a finite-dimensional sufficient statistic (“basically” because there are some borderline cases which are almost, but not quite, exponential families). Since the resulting prior family is of the form $q(\theta|n, s)$, it is parametric only if \mathcal{L}_θ is an exponential family. \triangleleft

The argument in Remark 7.13 is in fact how natural conjugate priors were originally defined [53]. This perspective is helpful to understand the relationship between conjugate pairs: For example, the conjugate prior for the variance of a univariate Gaussian is a gamma distribution. The relevant sufficient statistic for the variance is $S(x) := x^2$, and a simple application of the integral transformation theorem shows that the density of $S(X)$ has the shape of a gamma if X is Gaussian. By regarding this function as a density for θ (i.e. by normalizing with respect to θ), we obtain the actual gamma density. Remarkably, the standard sufficient statistic of the Poisson also yields a gamma distribution, see Section A.2.

7.7. Posterior asymptotics, in a cartoon overview

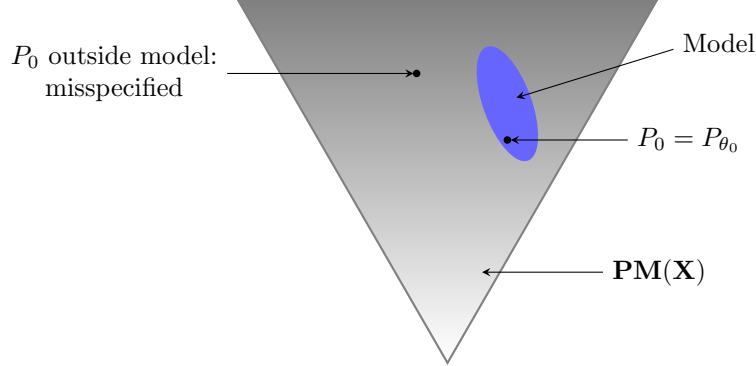
The theory of posterior asymptotics and posterior concentration is one of the few parts of Bayesian nonparametrics on which there is a fairly coherent literature, and I will not attempt to cover this topic in any detail, but rather refer to better and more competent descriptions than I could possibly produce. The forthcoming book [18], once published, will no doubt be the authoritative reference for years to come. In the meantime, [17] and the lecture notes [33] may be good places to start.

In a nutshell, the two main questions addressed by the theory of posterior asymptotics are:

- (1) **Consistency:** In the limit of infinite sample size, does the posterior concentrate at the correct value of Θ ?
- (2) **Convergence rates:** How rapidly does the posterior concentrate, i.e. how much data do we need in order to obtain a reliable answer?

The meaning of consistency depends crucially on how we define what the “correct value” of Θ is.

Recall that in frequentist statistics, we assume there exists a true distribution P_0 on the sample space that accounts for the data. If our model includes this distribution, i.e. if $P_0 \in M$, we say that the model is correctly specified. If not, M is **misspecified**. Here is my stick figure depiction of a model $M = \{P_\theta | \theta \in \mathbf{T}\}$ as a subset of the space of probability measures $\mathbf{PM}(\mathbf{X})$:



Still in the frequentist context, an estimator for the model parameter θ is a function $\hat{\theta}_n(x_{1:n})$ of the sample (or, more precisely, a family of functions indexed by n , since the number of arguments changes with sample size). If the model is correctly specified, there is some true parameter value θ_0 satisfying $P_{\theta_0} = P_0$. We say that $\hat{\theta}_n$ is a **consistent** estimator for the model M if, for every $\theta \in \mathbf{T}$,

$$\lim_n \hat{\theta}_n(X_{1:n}) \rightarrow \theta_0 \quad \text{almost surely} \quad (7.45)$$

if $X_1, X_2, \dots \sim_{\text{iid}} P_\theta$.

Consistency of Bayesian models. In the Bayesian setup, we model Θ as a random variable, which complicates the definition of consistency in two ways:

- The obvious complication is that, instead of an estimator with values in \mathbf{T} , we are now dealing with a posterior distribution *on* \mathbf{T} , and we hence have to define convergence in terms of where on \mathbf{T} the posterior concentrates as the sample size grows.
- A somewhat more subtle point is that, in order to ask whether the posterior concentrates at the correct value of Θ , we have to define what we mean by “correct value”. It turns out that the precise choice of this definition has huge impact on the resulting consistency properties.

Suppose we assume the observations X_1, X_2, \dots , from which the posterior is computed, are actually sampled from a given Bayesian model. A natural consistency requirement then seems to be that the posterior should concentrate at that value of the parameter which generated the data:

Definition 7.14. A Bayesian model with parameter space \mathbf{T} , observation model $M = \{P_\theta | \theta \in \mathbf{T}\}$ and prior distribution Q is **consistent in the Bayesian sense** if, for observations generated as

$$\begin{aligned} \Theta &\sim Q \\ X_1, X_2, \dots | \Theta &\sim_{\text{iid}} P_\Theta, \end{aligned} \quad (7.46)$$

the posterior satisfies

$$Q[\bullet | X_{1:n}] \xrightarrow[n \rightarrow \infty]{\text{weakly}} \delta_{\Theta(\omega)} \quad Q\text{-almost surely .} \quad (7.47)$$

◁

It is tempting to call this “weak consistency”, but unfortunately, some authors call it “strong consistency”. The terminology “consistent in the Bayesian sense” is a crutch I invented here—there does not seem to be a universally accepted nomenclature yet.

There is an almost universal consistency result, known as **Doob’s theorem**: If a Bayesian model on a standard Borel space is identifiable, i.e. if the parametrization mapping $\theta \rightarrow P_\theta$ is bimeasurable, then the model is consistent in the sense of Definition 7.14. (See e.g. [67, Theorem 10.10] for a precise statement.)

The problem with this notion of consistency is that its statement holds only *up to a null set under the prior distribution*. At second glance, this assumption seems almost brutal: Suppose we choose our prior as a point mass, i.e. we pick a single distribution P_0 on \mathbf{X} and set $Q := \delta_{P_0}$. The model is then $M = \{P_0\}$. The posterior is again always δ_{P_0} ; it does not even take into account what data we observe. Yet, the model is identifiable and hence consistent in the sense of Definition 7.14. Thus, we can obtain a universal consistency result for *any* model—if the given form of the model is not identifiable, we only have to reparametrize it in a sensible manner—but the price to pay is to explain away the rest of the universe as a null set.

Another way of defining consistency is by disentangling the data source and the prior: We assume the data is generated by a data source described by a unknown distribution P_0 . The data analyst specifies a prior Q , and we ask whether the corresponding posterior will asymptotically converge to the correct distribution P_0 *almost surely under the distribution of the data*.

Definition 7.15. A Bayesian model with parameter space \mathbf{T} , observation model $M = \{P_\theta | \theta \in \mathbf{T}\}$ and prior distribution Q is **consistent in the frequentist sense** if, for every $\theta_0 \in \mathbf{T}$ and observations generated as

$$X_1, X_2, \dots \sim_{\text{iid}} P_{\theta_0} , \quad (7.48)$$

the posterior satisfies

$$Q[\bullet | X_{1:n}] \xrightarrow[n \rightarrow \infty]{\text{weakly}} \delta_{\theta_0} \quad P_{\theta_0}\text{-almost surely .} \quad (7.49)$$

◁

With this notion of consistency, Bayesian models can be inconsistent and converge to completely wrong solutions. Since there is no longer a one-size-fits all result, the asymptotic theory of Bayesian models becomes much richer, and actually establishing that a model is consistent for a given problem is a much stronger statement than consistency in the sense of Doob’s theorem. See [33] for example results.

Convergence rates. Consistency is a purely asymptotic property, and instead of asking only whether we *would* find the right solution in the infinite limit of an asymptotically large sample, we can additionally ask how much data is required to get reasonably close to that solution—in other words, how rapidly the posterior distribution concentrates with increasing sample size.

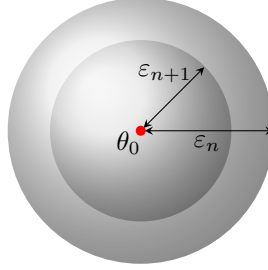
Quantifying concentration is a somewhat technical problem, but the basic idea is very simple: To measure how tightly the posterior $Q[\bullet | X_1, \dots, X_n]$ concentrates around θ_0 , we place a ball $B_{\varepsilon_n}(\theta_0)$ of radius ε_n around θ_0 . Basically, we want the posterior mass to concentrate inside this ball, but of course even a posterior that concentrates more and more tightly around θ_0 may still spread some small fraction on its mass over the entire space. We therefore permit a small error $\tau > 0$, and require only that

$$Q[B_{\varepsilon_n}(\theta_0) | X_1, \dots, X_n] > 1 - \tau \quad P_{\theta_0}\text{-almost surely} . \quad (7.50)$$

If we do not want the choice of τ to be a liability, we have to require that this holds for *any* $\tau \in (0, 1)$, although we can of course choose ε_n according to τ . For a given τ , we can then write $\varepsilon_n(\tau)$ for the smallest radius satisfying (7.50). If the posterior indeed concentrates as sample size grows, we obtain a sequence of shrinking radii

$$\varepsilon_n(\tau) > \varepsilon_{n+1}(\tau) > \dots \quad (7.51)$$

describing a sequence of shrinking, concentric balls around θ_0 :



We have thus reduced the problem from quantifying the convergence of a sequence of probability measures (the posteriors for different values of n) to the simpler problem of quantifying convergence of a sequence of numbers (the radii ε_n).

A typical convergence rate result expresses the rate of convergence of the sequence (ε_n) by means of an upper bound, formulated as a function of the sample size:

$$\forall \tau > 0 : \quad \varepsilon_n(\tau) < c(\tau) f(n) . \quad (7.52)$$

The function f is called a **rate**. It would of course be possible to derive results in this form for one specific Bayesian model (i.e. a specific prior distribution Q). It is usually more interesting, however, to instead obtain results for an entire class of models—both because it makes the result more generally applicable, and because it explicitly shows how the convergence rate depends on the complexity of the model. As we know from other statistical models, we have to expect that fitting complicated models requires more data than in simple ones. The function f hence has to take into account the model complexity, and convergence rate results are of the form

$$\forall \tau > 0 : \quad \varepsilon_n(\tau) < c(\tau) f(n, \text{model complexity}) . \quad (7.53)$$

Quantifying model complexity is another non-trivial question: In parametric models, we can simply count the number of degrees of freedom of the model (the number of effective dimensions of the parameter space), but for infinite-dimensional parameter spaces, more sophisticated tools are required. Empirical process theory and statistical learning theory provide an arsenal of such complexity measures (such as covering numbers, metric entropies, VC dimensions, etc.), and many of these

tools are also applied to measure model complexity in Bayesian nonparametrics. Once again, I refer to [\[33\]](#) for more details.

CHAPTER 8

Random measures

In this chapter, we will discuss random measures in more detail—both random probability measures and general random measures. Random measures play a fundamental role in Bayesian statistics: Whenever a data source is modeled as an exchangeable sequence, de Finetti’s theorem tells us that there is some random measure Θ such that the observations are explained as

$$\begin{aligned} \Theta &\sim Q \\ X_1, \dots, X_n | \Theta &\sim_{\text{iid}} \Theta . \end{aligned} \tag{8.1}$$

Although we can assume a parametric model, Θ is in general an infinite-dimensional quantity, and historically, Bayesian nonparametric priors were originally conceived to model Θ directly—the Dirichlet process was proposed in [12] as a prior distribution for Θ .

That is not quite how Bayesian nonparametric models are used today. Instead, one usually uses the more familiar approach of splitting Θ into a “likelihood” component (a sampling distribution) P_t and a random variable T with law Q^T , viz.

$$\Theta = P_T , \tag{8.2}$$

so that the generative model is

$$\begin{aligned} T &\sim Q^T \\ X_1, \dots, X_n | T &\sim_{\text{iid}} P_T . \end{aligned} \tag{8.3}$$

There are arguably two main reasons why this approach is preferred:

- (1) Tractable random measure priors (like the Dirichlet process) turn out to generate discrete measures, which are not useful as models of Θ (Theorem 2.3 says that if the DP is used directly in this way, the posterior simply interpolates the empirical distribution of the data with a fixed distribution G).
- (2) Splitting Θ into a likelihood P_t and a random pattern T —and possibly splitting T further into a hierarchical model—has proven much more useful than modeling Θ “monolithically”. One reason is tractability; another is that a suitably chosen T often provides a more useful summary of the data source than Θ .

We will briefly discuss some ways in which random measures are actually deployed in Section 8.1. The lion’s share of this chapter is then devoted to random discrete measures; the theory of such measures revolves around the Poisson process. Constructing priors on smooth random measures is technically much more challenging: Smoothness requires long-range dependencies, which make the mathematical structure of such models much more intricate, and introduce coupling in the posterior that is usually defies our mathematical toolkit. Non-discrete random measures are briefly discussed in Section 8.9.

8.1. Sampling models for random measure priors

When random discrete measures are used in Bayesian models, then usually in a model of the form (8.3). Important examples are:

- (i) In clustering models, T itself is a random probability measure; the weights parametrize a random partition, the atom locations serve as parameters for the (parametric) distributions of individual clusters.
- (ii) In latent feature models, a (non-normalized) random measure can be used to generate the latent binary matrix describing assignments to overlapping clusters (as described in Section 3.4).
- (iii) Dirichlet process mixtures (and other nonparametric mixtures) can be used in density estimation problems. In this case, the models (8.3) and (8.1) coincide: T is the random mixture density $p(x)$ in (2.18), which is interpreted as the density of Θ in (8.1).
- (iv) In survival analysis, T is a hazard rate. A random measure on \mathbb{R}_+ can be integrated to obtain random cumulative distribution functions, which in turn can be used to model hazard rates. I will not discuss this approach in detail, see e.g. [23, 34].

For most of the models we consider, including case (i)–(iii) above, we need to consider two ways of sampling from a random measure:

- The **multinomial sampling model**: $\hat{\xi}$ is a random *probability* measure and we sample $\Phi_1, \Phi_2, \dots | \hat{\xi} \sim_{\text{iid}} \hat{\xi}$.
- The **Bernoulli sampling model**: $\xi = \sum_{k \in \mathbb{N}} C_k \delta_{\Phi_k}$ is a (usually non-normalized) random measure with $C_k \in [0, 1]$ and we sample a random matrix \mathbf{Z} , for each column k , as $Z_{1k}, Z_{2k}, \dots | \Theta \sim \text{Bernoulli}(C_k)$.

Again, I am making up terminology here—the multinomial model, by far the most important case, is usually not discussed explicitly. The Bernoulli model was named and explicitly described by Thibeaux and Jordan [66], who called it a **Bernoulli process**.

Multinomial sampling. The sampling model used with basically with all random *probability* measure priors—the Dirichlet process, normalized completely random measures, Pitman-Yor process, etc.—in one way or another is of the form

$$\begin{aligned} \hat{\xi} &\sim Q^\xi \\ \Phi_1, \Phi_2, \dots | \hat{\xi} &\sim_{\text{iid}} \hat{\xi}. \end{aligned} \tag{8.4}$$

Different applications of this sampling scheme may look very different, depending on which level of the hierarchy the random measure occurs at—if the Φ_i are observed directly, we have a model of the form (8.1); in a DP mixture clustering model, the Φ_i are unobserved cluster parameters, etc.

Notation 8.1. If ν is a measure on some space Ω_ϕ , and $I := (A_1, \dots, A_n)$ a partition of Ω_ϕ into measurable sets, I will use the notation

$$\nu(I) := (\nu(A_1), \dots, \nu(A_n)) \tag{8.5}$$

throughout the remainder of this chapter. \triangleleft

I call the sampling scheme (8.4) “multinomial” since, if $I := (A_1, \dots, A_n)$ is a finite partition of Ω_ϕ , then

$$\mathbb{P}[\Phi \in A_j | \hat{\xi}] = \hat{\xi}(A_j), \tag{8.6}$$

and the random index $J \in [n]$ defined by $\Phi \in A_J$ is multinomially distributed with parameter $\hat{\xi}(I)$. It can in fact be shown, with a bit more effort, that sampling $\Phi \sim \nu$ for any probability distribution ν on Ω_ϕ , can be regarded as a limit of the sampling procedure $J \sim \text{Multinomial}(\nu(I))$ —roughly speaking, for $n \rightarrow \infty$ and $|A_i| \searrow 0$ (see [47] for details).

8.2. Random discrete measures and point processes

We now focus on random discrete measures, i.e. ξ is of the form

$$\xi(\bullet) = \sum_{k \in \mathbb{N}} C_k \delta_{\Phi_k}(\bullet), \quad (8.7)$$

where the weights $C_k \in \mathbb{R}_+$ may or may not sum to 1. We will use three different representations of such random measures: In terms of two sequences (C_k) and (Φ_k) , as we have done so far; in terms of a random cumulative distribution function representing the weights (C_k) ; and in terms of a point process.

Recall that a **point process** on a space \mathbf{X} is a random, countable collection of points on \mathbf{X} . In general, points can occur multiple times in a sample, and a general point process is hence a random countable *multiset* of points in \mathbf{X} . If the multiset is a set, i.e. if any two points in a sample are distinct almost surely, the point process is called **simple**. In other words, a simple point process is a random variable

$$\Pi : \Omega \rightarrow 2^{\mathbf{X}} \quad \text{where} \quad \Pi(\omega) \text{ countable a.s.} \quad (8.8)$$

We are only interested in the simple case in the following.

If Π is a point process, we can define a random measure by simply counting the number of points that Π places in a given set A :

$$\xi(A) := |\Pi \cap A| = \sum_{X \in \Pi} \delta_X(A). \quad (8.9)$$

If we enumerate the points in the random set Π as X_1, X_2, \dots , we can read this as a random measure $\xi = \sum_k C_k \delta_{X_k}$ with weights $C_k = 1$ (since the point process is simple). Such a measure is called a **random counting measure**.

The random discrete measures we have used as priors have scalar weights C_k . We can generate such scalar weights in \mathbb{R}_+ using a point process by throwing points onto \mathbb{R}_+ . A discrete random measure on Ω_ϕ with non-trivial weights can then be defined using a simple point process Π on $\mathbf{X} := \mathbb{R}_{\geq 0} \times \Omega_\phi$ as

$$\xi(\bullet) := \sum_{(C, \Phi) \in \Pi} C \cdot \delta_\Phi(\bullet). \quad (8.10)$$

8.3. Poisson processes

By far the most important point process is the Poisson process, for which the number of points in any fixed set is Poisson-distributed. (Recall the Poisson distribution (A.2) from Appendix A.)

Definition 8.2. Let μ be a measure on a Polish space \mathbf{X} . A point process Π^μ on \mathbf{X} is called a **Poisson process** with parameter μ if

$$|\Pi^\mu \cap A| \perp\!\!\!\perp |\Pi^\mu \cap (\mathbf{X} \setminus A)| \quad (8.11)$$

and

$$|\Pi^\mu \cap A| = \begin{cases} \sim \text{Poisson}(\mu(A)) & \text{if } \mu(A) < \infty \\ \infty \text{ a.s.} & \text{if } \mu(A) = \infty \end{cases} \quad (8.12)$$

for every measurable set A in \mathbf{X} . \triangleleft

The Poisson process is explained for sets A of infinite measure $\mu(A)$ by “slicing up” μ into a countable number of finite components: We require

$$\mu = \sum_{n \in \mathbb{N}} \mu_n \quad \text{for some measures } \mu_n \text{ with } \mu_n(\mathbf{X}) < \infty. \quad (8.13)$$

Perhaps the best way to illustrate the definition of the Poisson process is to provide an explicit sampling scheme:

Theorem 8.3 (Sampling a Poisson process). *Let μ be a measure on a standard Borel space \mathbf{X} . If μ is non-atomic and satisfies (8.13), the Poisson process Π^μ on \mathbf{X} exists, and can be sampled as follows:*

(1) *If $\mu(\mathbf{X}) < \infty$, then $\Pi^\mu \stackrel{d}{=} \{X_1, \dots, X_N\}$, where*

$$N \sim \text{Poisson}(\mu(\mathbf{X})) \quad \text{and} \quad X_1, \dots, X_N \sim_{iid} \frac{\mu}{\mu(\mathbf{X})}. \quad (8.14)$$

(2) *If $\mu(\mathbf{X}) = \infty$, then*

$$\Pi^\mu = \bigcup_{n \in \mathbb{N}} \Pi^{\mu_n}. \quad (8.15)$$

\triangleleft

The theorem immediately implies several useful properties:

Corollary 8.4. *Let μ be a measure and (ν_n) a sequence of measures, all of which are non-atomic and satisfy (8.13). Let $\phi : \mathbf{X} \rightarrow \mathbf{X}$ be a measurable mapping. Then the following holds:*

$$\phi(\Pi^\mu) = \Pi^{\phi(\mu)} \quad \text{if } \mu \text{ is } \sigma\text{-finite}. \quad (8.16)$$

$$\Pi^\mu \cap A = \Pi^{\mu(\cdot \cap A)} \quad \text{for any set } A \in \mathcal{B}(\mathbf{X}). \quad (8.17)$$

$$\bigcup_{n \in \mathbb{N}} \Pi^{\mu_n} = \Pi^{\sum_n \mu_n} \quad (8.18)$$

\triangleleft

Informally speaking, Theorem 8.3 shows that sampling from a Poisson process with parameter μ is “almost” i.i.d. sampling from μ , but sampling is well-explained even if μ is infinite and cannot be normalized to a probability measure. To obtain a coherent generalization of sampling to the case $\mu(\mathbf{X}) = \infty$, we consider the *entire* point set as a single draw, rather than drawing points individually. To substitute for the independence property between separate i.i.d. draws, we now need an independence property that holds *within* the sample; this is given by the independence (8.11) of the process between disjoint sets. Not that (8.11) implies the total number of points must be random: If we were to posit a fixed number n of samples, then observing k samples in $\mathbf{X} \setminus A$ would imply $|\Pi \cap A| < n - k$, so the numbers of points on A and $\mathbf{X} \setminus A$ would not be independent.

Complete randomness. Since (8.11) holds for any measurable set A , it implies the numbers of points in any two disjoint sets are independent. In the point process literature, this property is known as *complete randomness* or *pure randomness*.

Definition 8.5. A point process Π is called **completely random** if

$$(\Pi \cap A) \perp\!\!\!\perp (\Pi \cap A') \quad (8.19)$$

for every pair of disjoint measurable sets $A, A' \in \mathcal{B}(\mathbf{X})$. \triangleleft

The Poisson process is completely random by definition. A rather baffling fact, however, is that *only* the Poisson process is completely random.

Proposition 8.6. *If a simple point process Π on an uncountable standard Borel space \mathbf{X} is completely random, the set function defined by the expected number of points per set,*

$$\mu(A) := \mathbb{E}[\Pi \cap A] \quad \text{for all Borel sets } A, \quad (8.20)$$

is a non-atomic measure on \mathbf{X} . If μ is σ -finite, Π is a Poisson process with parameter μ . \triangleleft

This illustrates why the Poisson process is of such fundamental importance: As we discussed above, we can think of complete randomness as the point process analogue of i.i.d. sampling. In this sense, it is perhaps more accurate to think of the Poisson process as a sampling paradigm (similar to i.i.d. sampling from μ), rather than as a model with parameter μ . I will give a proof of Proposition (8.6) here, but the main purpose of doing so is to clarify that it is absolutely elementary.

PROOF. For the proof, I will abbreviate the (random) number of points in a set A as $N(A) := |\Pi \cap A|$. The first step is to show that μ in (8.20) is a measure. Let A_1, A_2, \dots be a (possibly infinite) sequence of mutually disjoint sets. Disjointness implies

$$N(\cup_i A_i) = \sum_i N(A_i). \quad (8.21)$$

Since the sets are disjoint and Π is completely random, the random variables $N(A_i)$ are mutually independent and their expectations hence additive, so

$$\mu(\cup_i A_i) = \mathbb{E}[N(\cup_i A_i)] = \sum_i \mathbb{E}[N(A_i)] = \sum_i \mu(A_i). \quad (8.22)$$

Thus, μ is countably additive. Since also $\mu(\emptyset) = 0$ and $\mu(A) \geq 0$, it is indeed a measure on \mathbf{X} . Since points are distinct almost surely, $\mu(\{x\}) = 0$ for any $x \in \mathbf{X}$, i.e. μ is non-atomic.

To show that μ is a Poisson process with parameter μ , we have to show that $N(A) \sim \text{Poisson}(\mu(A))$ holds for any Borel set A . What is the probability of observing precisely k points in A ? Since μ is non-atomic, we can subdivide A into a partition (A_1, \dots, A_n) of measurable sets with

$$\mu(A_1) = \dots = \mu(A_n) = \frac{\mu(A)}{n}, \quad (8.23)$$

for any $n \in \mathbb{N}$. Rather than counting points in each set A_i , we can greatly simplify matters by only distinguishing whether a set contains points or not, by defining

$$I_{A_i} := \begin{cases} 1 & N(A_i) > 0 \\ 0 & N(A_i) = 0 \end{cases}. \quad (8.24)$$

Since each I_{A_i} is binary, it is a Bernoulli variable with some success probability p_{in} . As the points of Π are distinct, each point is contained in a separate set A_i for n sufficiently large, and so

$$\lim_n \sum_{i=1}^n I_{A_i} = N(A) . \quad (8.25)$$

For a Bernoulli variable, the success probability p_{in} is just the expectation $\mathbb{E}[I_{A_i}]$. If n is large enough that points are in separate intervals, we have $\mathbb{E}[I_{A_i}] \rightarrow \mathbb{E}[N(A_i)]$, or more formally,

$$\lim_n \frac{p_{in}}{\mathbb{E}[N(A_i)]} = \lim_n \frac{p_{in}}{\mu(A)/n} = 1 , \quad (8.26)$$

so for sufficiently large n , we can assume that I_{A_i} has success probability $\mu(A)/n$. By complete randomness, the variables I_{A_i} are independent, so the probability of observing precisely k successes is binomially distributed,

$$\mathbb{P}\{I_{A_1} + \dots + I_{A_n} = k\} = \binom{n}{k} \left(\frac{\mu(A)}{n}\right)^k \left(1 - \frac{\mu(A)}{n}\right)^{n-k} . \quad (8.27)$$

In the limit $n \rightarrow \infty$, as we have argued above, this converges to the probability of observing k points of Π in A . A few lines of arithmetic show that

$$\lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\mu(A)}{n}\right)^k \left(1 - \frac{\mu(A)}{n}\right)^{n-k} = \frac{\mu(A)^k}{k!} \lim_n \left(1 - \frac{\mu(A)}{n}\right)^n . \quad (8.28)$$

We know from basic calculus that $(1 - \frac{a}{n})^n \rightarrow e^{-a}$, so we obtain

$$\mathbb{P}\{N(A) = k\} = e^{-\mu(A)} \frac{\mu(A)^k}{k!} , \quad (8.29)$$

and Π is indeed a Poisson process. \square

8.4. Total mass and random CDFs

If the random measure ξ is not normalized, the mass $\xi(\Omega_\phi)$ it assigns to the entire space is in general a non-negative random variable. We denote this variable

$$T_\xi := \xi(\Omega_\phi) = \sum_k C_k \quad (8.30)$$

and call it the **total mass** of ξ . This variable may carry a lot of information about ξ ; indeed, for the most important class of non-normalized random discrete measures—homogeneous Poisson random measures, which we define below—the distribution of T_ξ completely determines the distribution of the weights C_k .

For the purposes of Bayesian nonparametrics, we are only interested in the case where T_ξ is almost surely finite. Recall the applications of random measures we have seen so far; they suggest three possible uses for ξ :

- (1) ξ is a random probability measure (so $T_\xi = 1$ almost surely).
- (2) ξ is not normalized, but we use it to *define* a random probability measure, by dividing by its total mass. That requires $T_\xi < \infty$ a.s.
- (3) ξ is used in a latent feature model as in Sec. 3.4. We already argued in Section 3.4 that this also requires $T_\xi < \infty$ a.s.

We are also generally only interested in homogeneous random measures (recall: the atoms Φ_i are i.i.d. and independent of the weights). For all that follows, we hence make the following

General assumption 8.7. *All random measures we discuss are assumed to be homogeneous and have finite total mass almost surely.*

If ξ is homogeneous, we can assume without loss of generality that

$$\Omega_\phi = [0, 1] \quad \text{and} \quad \Phi_1, \Phi_2, \dots \sim_{\text{iid}} \text{Uniform}[0, 1] \quad (8.31)$$

That is possible because the weights do not depend on the atom locations—to turn ξ into a random measure on an arbitrary Polish space Ω_ϕ instead of $[0, 1]$, we can simply replace the uniform scalar atoms and replace them by other i.i.d. random variables. All non-trivial structure in ξ is encoded in the weight sequence.

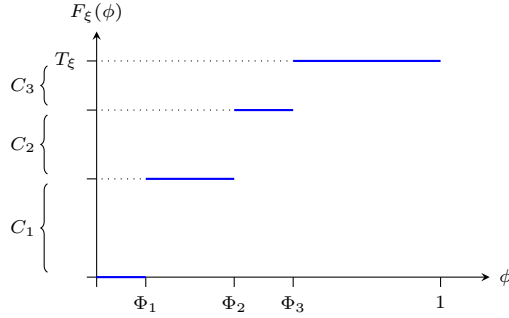
One great advantage of this representation is that measures on $[0, 1]$ can be represented by their cumulative distribution functions. A homogeneous random measure ξ on $[0, 1]$, can hence be represented by the *random CDF*

$$F_\xi(\phi) := \xi([0, \phi]) . \quad (8.32)$$

The total mass of ξ is then

$$T_\xi =_{\text{a.s.}} F_\xi(1) . \quad (8.33)$$

Since ξ is discrete, the random function F_ξ is piece-wise constant and, as a CDF, non-decreasing, for example:



We can alternatively regard this as the path of a real-valued, non-decreasing, piece-wise constant stochastic process on $[0, 1]$. Those are a lot of adjectives, but processes of this type are particularly easy to handle mathematically—roughly speaking, because all action happens at a countable number of points (the jumps), and since non-decreasingness can be enforced in a completely local manner by requiring the jumps to be non-negative, without introducing complicated long-range dependencies between different points of the path.

8.5. Infinite divisibility and subordinators

Suppose we are want to define a homogeneous random measure ξ ; as we have noted in the previous section, we are interested in measures with finite total mass T_ξ . Recall our basic design principle that components of our random objects should be as independent from each other as possible. Since ξ is homogeneous, the atom locations are i.i.d., but we *cannot* sample the weights C_k i.i.d.; if we do, $T_\xi = \infty$ holds almost surely (infinite sums of positive i.i.d. variables have to diverge, by Borel-Cantelli).

How much independence between the weights C_k can we get away with? Arguably the next best thing to i.i.d. weights C_k would be to make T_ξ infinitely

divisible: Recall that a scalar random variable T is called **infinitely divisible** if, for every $n \in \mathbb{N}$, there is a probability measure μ_n such that

$$\mathcal{L}(T) = \underbrace{\mu_n * \dots * \mu_n}_{n \text{ times}}.$$

In other words, for every n , there exists a random variable $T^{(n)}$ with $\mathcal{L}(T^{(n)}) = \mu_n$ such that T can be represented as the sum of n i.i.d. copies of $T^{(n)}$,

$$T \stackrel{d}{=} \sum_{i=1}^n T_i^{(n)}. \quad (8.34)$$

A random variable can be infinitely divisible and almost surely finite—gamma and Gaussian variables are both infinitely divisible. Thus, although we cannot represent T as an *infinite* sum of i.i.d. variables, we can subdivide into *any finite* number of i.i.d. components.

It turns out that, if we assume the total mass T_ξ of a homogeneous random measure is infinitely divisible, then its CDF F_ξ is necessarily a special type of Lévy process called a *subordinator*. Recall that we can think of a real-valued stochastic process F on $[0, 1]$ as a random function $F : [0, 1] \rightarrow \mathbb{R}$. We have already discussed random functions in Chapter 4, where we were particularly interested in distributions on continuous functions. The CDFs of discrete measures jump at a countable number of points, so we now require instead that F is almost surely a so called right-continuous function with left-hand limits (rcll function). That means simply that F is piece-wise continuous with an at most countable number of jumps, and if it jumps at a point ϕ , the function value $F(\phi)$ at the jump location already belongs to the *right* branch of the graph.

If we choose a sub-interval $I = [a, b]$ in $[0, 1]$, then $F(b) - F(a)$, i.e. the (random) amount by which F increases between a and b , is called the **increment** of F on I . Recall that F is called a **Lévy process** if:

- (1) It is almost surely an rcll function.
- (2) It has independent increments: If I and J are two disjoint intervals, the increments of F on I and on J are independent random variables.

There is a direct correspondence between Lévy processes and (scalar) infinitely divisible variables: If F is a Lévy process, the scalar variable $F(1)$ (or indeed $F(\phi)$ for any $\phi \in (0, 1]$) is infinitely divisible, and the converse is also true:

Proposition 8.8 (e.g. [56, Theorem 7.10]). *A scalar random variable T is infinitely divisible iff there is a real-valued Lévy process $F(\phi)$ on $[0, 1]$ such that $F(1) \stackrel{d}{=} T$. \triangleleft*

Informally, we can “smear out” the scalar variable T into a stochastic process on $[0, 1]$, and this process can always be chosen as a Lévy process.

We can then use the special structure of Lévy processes to simplify further: The Lévy-Khinchine theorem [28, Theorem 15.4] tells us that any Lévy process can be decomposed into three, mutually independent components as

$$\text{Lévy process path} = \text{non-random linear function} + \text{centered Brownian motion} + \text{Poisson process jumps}.$$

Since we know our random function F is non-decreasing, it cannot have a Brownian motion component: The increment of a centered Brownian motion on any interval $I \subset [0, 1]$ is positive or negative with equal probability. Without Brownian motion,

the effect of the non-random linear function would be to turn the piece-wise constant path in the figure above into a piece-wise linear one, where each segment has the same slope. For a discrete random measure, the CDF should not change between jumps, so the linear component has to vanish. What we are left with is a Poisson process on $[0, 1]$. This process is given by a Poisson process Π^μ with on $\mathbb{R}_+ \times [0, 1]$ as

$$F(\phi) = \sum \{ C \mid (C, \Phi) \in \Pi^\mu \text{ and } \Phi < \phi \} . \quad (8.35)$$

If the mean measure of the Poisson process has product structure $\mu = \mu_c \times \mu_\phi$ on $\mathbb{R}_+ \times [0, 1]$, a stochastic process F of the form (8.35) called a **subordinator**. The product structure of μ means that F has independent increments. In summary:

homogeneous random discrete measures with infinitely divisible total mass

\updownarrow

CDFs generated by a subordinator

We also see that there is a direct correspondence between the random CDF representation and the point process representation of such measures through (8.35).

8.6. Poisson random measures

Suppose we define a random measure ξ using a point process in (8.36). Our discussion in the previous section shows that, in order for the total mass T_ξ to be infinitely divisible, we need to choose the point process specifically as a Poisson process. If Π^μ is a Poisson process on $\mathbb{R}_+ \times \Omega_\phi$, then

$$\xi(\bullet) := \sum_{(C, \Phi) \in \Pi^\mu} C \cdot \delta_\Phi(\bullet) . \quad (8.36)$$

is called a **Poisson random measure**. If ξ is also homogeneous, the parameter measure μ of the Poisson factorizes as

$$\mu = \mu_C \otimes \mu_\phi \quad (8.37)$$

into non-atomic measures μ_C on \mathbb{R}_+ and μ_ϕ on Ω_ϕ . This is precisely the case in which the CDF of ξ is a subordinator, if we choose $\Omega_\phi = [0, 1]$.

Although the class of Poisson random measures is huge, there are only a few which play an actual role in Bayesian nonparametrics. The most important examples are arguably gamma and stable random measures, which can be used to derive Dirichlet and Pitman-Yor processes. Both are standard models in applied probability. A model which was hand-tailored for Bayesian nonparametrics is the beta process or beta random measure [23], which has applications in survival analysis and can be used to represent the IBP.

A family of homogeneous Poisson random measures is defined by specifying the measure μ_C in (8.37), which controls the distribution of the weights. If we sample a weight sequence (C_k) from a Poisson process with parameter μ_C , we can then choose a probability measure on some space Ω_ϕ , sample a sequence of atoms, and attach them to (C_k) to obtain a random measure on Ω_ϕ .

Definition 8.9. A homogeneous Poisson random measure ξ as in (8.36) is called a gamma, stable, or beta process, respectively, if μ_C in (8.37) is of the form:

Random measure	Parameter measure μ_C
gamma process	$\mu_C(dc) = \alpha c^{-1} e^{-\beta c} dc$
stable process	$\mu_C(dc) = \gamma c^{-\alpha-1} dc$
beta process	$\mu_C(dc) = \gamma c^{-1} (1-c)^{\alpha-1} dc$

<

Example 8.10 (beta process). The beta process was originally introduced by Hjort [23] as a prior for survival models. Thibeaux and Jordan [66] pointed out that the IBP can be generated by first sampling a random measure ξ from a beta process, and then sampling a binary matrix \mathbf{Z} from ξ according to the Bernoulli sampling scheme in Section 8.1.

Teh and Görür [62] showed how this perspective can be used to generalize the IBP to distributions where the row sums of \mathbf{Z} —i.e. the sizes of groups or clusters—follow a power law distribution, in analogy to the cluster sizes generated by a Pitman-Yor process. If ξ is a homogeneous Poisson random measure with

$$\mu(dc) = \gamma \frac{\Gamma(1+\alpha)}{\Gamma(1-\alpha)\Gamma(\alpha+d)} c^{-1-d} (1-c)^{\alpha-1+d}, \quad (8.38)$$

and if \mathbf{Z} is a random binary matrix generated from ξ according to the Bernoulli sampling model, then:

- For $d = 0$, \mathbf{Z} is distributed according to an Indian buffet process.
- For $d \in (0, 1]$, the column sums of \mathbf{Z} follow a power law distribution.

The random measure (8.38) is obtained by modifying a beta process according to the intuition

Poisson random measure + stable random measure \rightarrow power law ,

and Teh and Görür [62] refer to it as a **beta-stable random measure**. For more background on beta processes, I recommend the survey [63]. <

8.7. Completely random measures

Suppose we use a Poisson random measure ξ to derive a prior for a Bayesian nonparametric model. That may mean that we normalize ξ to obtain a random probability measure and then sample from it, that ξ is a beta process sampled with a Bernoulli sampling model, etc. Regardless of the specific sampling model we use, we usually observe atom location of ξ . Suppose Φ_1, \dots, Φ_n are observed atom location. If we compute the posterior of ξ given Φ , we *know* that a random measure ξ_n sampled from this posterior has atoms at Φ_1, \dots, Φ_n . Hence, ξ_n is no longer a Poisson random measure (since the parameter measure μ of a Poisson process must be atomless).

Poisson random measures can be generalized to a very natural class of measures, called completely random measures, which includes both Poisson random measures and fixed atom locations: Recall the complete randomness property (8.19) for point processes. Now consider the analogous property for random measures: If ξ is a random measure on Ω_ϕ , we require that for any two measurable sets A and A'

$$\xi(A) \perp\!\!\!\perp \xi(A') \quad \text{whenever } A, A' \text{ disjoint.} \quad (8.39)$$

In analogy to point processes, we call a random measure satisfying (8.39) a **completely random measure**, or **CRM** for short.

If ξ has finite total mass almost surely—and in fact under much more general conditions not relevant for us, see [32]—a CRM can always be represented as follows: If ξ is completely random, then

$$\xi =_{\text{a.s.}} \xi_n + \xi_f + \xi_r \quad (8.40)$$

where ξ_n is a non-random measure on Ω_ϕ , ξ_f is a random measure

$$\xi_f \stackrel{\text{d}}{=} \sum_{\phi_i \in A} C_i \delta_{\Phi_i} \quad (8.41)$$

with a fixed, countable set $A \subset \Omega_\phi$ of atoms, and ξ_r is a Poisson random measure [30, Theorem 1].

In particular, a CRM does not have to be discrete, but only the non-random component ξ_n can be smooth: Informally, suppose B is a set and we know what the smooth component looks like on B . Then, by smoothness, that provides information on how it behaves on $\Omega_\phi \setminus B$, at least close to the boundary, so if the smooth component is random, its restrictions to B and $\Omega_\phi \setminus B$ are stochastically dependent, and (8.39) is violated.

For Bayesian nonparametrics, the non-random component is not of interest, so we always assume $\xi_n = 0$. In the prior, we usually have no reason to assume atoms at specific locations. Hence, the distribution of ξ_r in (8.40) is sampled from the prior, and ξ_f appears in the posterior. A very readable derivation of completely random measures is given by Kingman [32, Chapter 8], but it is useful to keep in mind that the only CRMs of interest to Bayesian nonparametrics are usually those which satisfy assumption 8.7 and are of the form

$$\begin{array}{ccc} \xi & =_{\text{a.s.}} & \xi_f + \xi_r \\ \text{appears in posterior} & \xrightarrow{\quad \uparrow \quad} & \xleftarrow{\quad \quad} \text{sampled from prior} \end{array} \quad (8.42)$$

Even if the prior is a completely random measure, though, the posterior need *not* be a CRM. One example of a CRM prior whose posterior is indeed of the form (8.42) is the beta process or, more generally, the beta-stable, combined with a Bernoulli sampling model as in Example 8.10.

8.8. Normalization

So far in this chapter, we have only discussed unnormalized random measures. We will now turn to the arguably more important case of random probability measures. Suppose ξ is a random measure with a.s. finite total mass T_ξ . We can define a random probability measure $\hat{\xi}$ from ξ by normalization, as

$$\hat{\xi}(\bullet) := \frac{\xi(\bullet)}{T_\xi} . \quad (8.43)$$

If ξ is a completely random measure as above, $\hat{\xi}$ is also known as a **normalized completely random measure**.

Before we can define $\hat{\xi}$, we have to verify that its total mass is finite. That may not be a trivial matter in general; if ξ is in particular a Poisson random measure with parameter measure μ , a sufficient condition for $T_\xi < \infty$ is

$$\int_{\mathbb{R}_+} \int_{\Omega_\phi} \min\{1, c\} \mu(d\phi, dc) < \infty . \quad (8.44)$$

Example 8.11. The most important example is the gamma random measure (cf. Definition 8.9). In this case, it can be shown that the random mass $\xi(A)$ assigned by ξ to a Borel set A is a gamma variable with parameters $(\alpha G_0(A), 1)$. In particular, T_ξ is $\text{Gamma}(\alpha, 1)$, and hence finite a.s., so we do not have to invoke condition (8.44). The random measure obtained by normalizing ξ is the Dirichlet process,

$$\hat{\xi} \sim \text{DP}(\alpha G_0) . \quad (8.45)$$

We can easily see that this is the case by subdividing the space Ω_ϕ into a finite partition $I := (A_1, \dots, A_n)$ of Borel sets. Then each entry $\xi(A_i)$ of the random vector $\xi(I)$ (using notation (8.5)) is a $\text{Gamma}(\alpha G_0(A_i), 1)$ random variable. A normalized vector of such gamma variables has the Dirichlet distribution with concentration α and expectation $G_0(I)$ (see Appendix A.3). Thus, for any partition of Ω_ϕ , the vector $\hat{\xi}(I)$ is Dirichlet-distributed, and $\hat{\xi}$ is indeed a DP. \triangleleft

Although it is perhaps not completely obvious at first glance, a normalized discrete random measure $\hat{\xi}$ in (8.43) is in general *not* independent of T_ξ . For illustration, suppose we sample a finite random measure $\xi = \sum_{k=1}^3 C_k \delta_{\Phi_k}$, where the weights C_k are generated i.i.d. from a degenerate distribution consisting of two point masses at, say, 1 and 10:



We sample three weights, compute their sum T_ξ , and normalize to obtain $\hat{\xi}$. Even after we have normalized, if we are told that $T_\xi = 12$, we can precisely read off the values of the C_k , and hence of the weights \hat{C}_k of $\hat{\xi}$, except for the order in which they occur. Thus, T_ξ and $\hat{\xi}$ are clearly dependent. The distribution of ξ does not have to be degenerate—we could replace the mixture of point masses with a mixture of, say, two narrow Gaussians, and we could still tell if $T_\xi \approx 12$ that one of the weights \hat{C}_k is about ten times as large as all others. The same can happen for an infinite number of atoms in a Poisson random measure, if e.g. the Poisson mean measure peaks sharply at two points.

In fact, ξ and T_ξ are *always* dependent, except in one special case:

Proposition 8.12. *If ξ is a completely random measure, then $\hat{\xi} \perp\!\!\!\perp T_\xi$ holds if and only if ξ is a gamma random measure, that is, if $\hat{\xi}$ is a Dirichlet process.* \triangleleft

This characterization of the gamma and Dirichlet processes is a direct consequence of the special properties of the gamma distribution, see Theorem A.1 in Appendix A.

In Remark 2.2, we argued that we want our random measures to be as simple as possible, in the sense that we try to limit coupling between their components. In this sense, Proposition 8.12 says that the Dirichlet process is the simplest object we can possibly hope to obtain—unless we keep the number of atoms finite, in which case it would be the Dirichlet distribution, again due to Theorem A.1. We also argued that restricting coupling between components of the random measure in the prior keeps the posterior tractable, which suggests that the Dirichlet process posterior should be particularly simple. Indeed:

Theorem 8.13 (James, Lijoi, and Prünster [25]). *Suppose Θ is a normalized completely random measure and homogeneous. If observations are generated as $X_1, X_2, \dots \sim_{\text{iid}} \Theta$, the posterior distribution $\mathcal{L}(\Theta|X_1, \dots, X_n)$ is the law of a homogeneous normalized completely random measure if and only if Θ is a Dirichlet process.* \triangleleft

The Pitman-Yor process is not a normalized completely random measure: Note that the variables V_k are not i.i.d., since $\mathcal{L}(V_k)$ depends on k .

8.9. Beyond the discrete case: General random measures

Random discrete measures are comparatively easy to define, since it is sufficient to define the two random sequences (C_k) and (Φ_k) . Defining general, possibly smooth random measures, is a much harder problem: Informally speaking, in order to enforce smoothness, we have to introduce sufficiently strong stochastic dependencies between points that are close in Ω_ϕ .

There is a generally applicable way to define arbitrary random measures on a Polish space Ω_ϕ ; I will only discuss the case of random probability measures, since the general case is a bit more technical. Here is the basic idea: Suppose ξ is any random probability measure on Ω_ϕ , i.e. a random variable with values in $\mathbf{PM}(\Omega_\phi)$, and let $P := \mathcal{L}(\xi)$ be its distribution. Now choose a partition $I = (A_1, \dots, A_d)$ of Ω_ϕ into a finite number d of measurable sets. If we evaluate ξ on each set A_i , we obtain a vector

$$\xi(I) = (\xi(A_1), \dots, \xi(A_d)) . \quad (8.46)$$

with d non-negative entries and sum 1, and hence a random element of the simplex $\Delta_d \subset \mathbb{R}^d$. We can define a different random vector $\xi(I)$ for each possible partition of Ω_ϕ into finitely many sets. Let \mathbf{I} be the set of all such partitions. If we denote the distribution of $\xi(I)$ by $P_I := \mathcal{L}(\xi(I))$, we obtain a family of distributions

$$\mathcal{P} := \{P_I | I \in \mathbf{I}\} . \quad (8.47)$$

Theorem 8.14 below implies that \mathcal{P} completely determines P . This means we can construct P (and hence ξ) by positing a suitable family of distributions \mathcal{P} . These distributions P_I are much easier to specify than P , since they live on the finite-dimensional sets Δ_d , whereas P lives on the infinite-dimensional space $\mathbf{PM}(\Omega_\phi)$.

What I have described so far is a uniqueness statement, not a construction result: I have started with the object ξ we want to construct, and told you that we can derive distributions from ξ which then define ξ . What we really need is a set of criteria for whether a given family \mathcal{P} of distributions defines a random probability measure ξ .

One property is clearly necessary: If all P_I are supposed to be derived from the same P , they must cohere over different partitions. For example, suppose I is obtained from another partition $J = (A_1, \dots, A_d, A_{d+1})$ by merging two sets in J , say $I = (A_1, \dots, A_d \cup A_{d+1})$. If ξ is a random measure on Ω_ϕ , it must satisfy

$$\xi(I) \stackrel{\text{d}}{=} (\xi(A_1), \dots, \xi(A_d) + \xi(A_{d+1})) . \quad (8.48)$$

Hence, if $\xi_I \sim P_I$ and $\xi_J \sim P_J$, they must accordingly satisfy

$$(\xi_{I,1}, \dots, \xi_{I,d}) \stackrel{\text{d}}{=} (\xi_{J,1}, \dots, \xi_{J,d} + \xi_{J,d+1}) . \quad (8.49)$$

More generally, we have to consider the case where we obtain I from J by merging an arbitrary number of sets in J , which is notationally a bit more cumbersome: Suppose $J = (A_1, \dots, A_d)$, and I is a coarsening of J , i.e. there is a partition $\psi^D = (\psi_1, \dots, \psi_k)$ of $[d]$ such that

$$I = (\cup_{i \in \psi_1} A_i, \dots, \cup_{i \in \psi_k} A_i) . \quad (8.50)$$

We then define a mapping $\text{pr}_{\text{JI}} : \Delta_J \rightarrow \Delta_I$ as

$$\text{pr}_{\text{JI}}(p_1, \dots, p_d) := \left(\sum_{i \in \psi_1} p_i, \dots, \sum_{i \in \psi_k} p_i \right) . \quad (8.51)$$

The general form of requirement (8.49) is then

$$\text{pr}_{\text{JI}}(\xi_J) \stackrel{\text{d}}{=} \xi_I \quad \text{or equivalently} \quad \text{pr}_{\text{JI}}(P_J) = P_I . \quad (8.52)$$

The family (8.47) of distributions is called **projective** if it satisfies (8.52) whenever I is a coarsening (8.50) of J .

A second obvious requirement is the following: If the expected measure (i.e. the expected value) of a random measure ξ on Ω_ϕ is $\mathbb{E}[\xi] = G$, then $\mathbb{E}[\xi(I)] = G(I)$. Hence, if the family \mathcal{P} is supposed to define a random measure, there must be a probability measure G on Ω_ϕ such that

$$\mathbb{E}[\xi_I] = G(I) \quad \text{for all } I \in \mathbf{I} . \quad (8.53)$$

Remarkably, it turns out that conditions (8.52) and (8.53) are all we need to construct random measures:

Theorem 8.14 (Orbanz [46]). *Let Ω_ϕ be a Polish space. A family $\{P_I | I \in \mathbf{I}\}$ of distributions $P_I \in \mathbf{PM}(\Delta_I)$ uniquely defines the distribution P of a random measure ξ on Ω_ϕ if and only if (i) it is projective and (ii) there is some probability measure G on Ω_ϕ such that*

$$\mathbb{E}_{P_I}[\xi_I] = G(I) . \quad (8.54)$$

If so, P satisfies $\mathbb{E}_P[\xi] = G$. \triangleleft

Example 8.15 (Dirichlet process). We can in particular choose each distribution P_I as a Dirichlet distribution on Δ_I (see Section A.3). We choose the same concentration parameter α for all P_I , fix a probability measure G on Ω_ϕ , and define the expectation parameter of each P_I as $g_I := G(I)$. It is then not hard to show that the resulting family \mathcal{P} is projective, and by definition, it satisfies (A.3). The resulting random measure ξ on Ω_ϕ is a Dirichlet process with concentration α and base measure G . The Dirichlet process was originally constructed (roughly) in this way by Ferguson [12], and then shown to be almost surely discrete by Blackwell [4]. Our definition of the Dirichlet process, via the stick-breaking construction—where discreteness is obvious—is informed by hindsight. See [46] for details. \triangleleft

Although Theorem 8.14 in principle allows us to construct arbitrary random measures, there are very few examples of continuous random measures used in Bayesian nonparametrics. One is the Dirichlet process mixture: The DP mixture model (2.18) with a smooth parametric component density p defines a smooth random measure. We have only used such mixtures for clustering, but they are also used for density estimation problems, where the random density (2.18) is used to fit a target distribution. Another example is the *Pólya tree prior* [13], which contains the DP as a special case, but for certain parameter settings generates continuous distributions almost surely—with the caveat that continuous in this case only means

absolutely continuous with respect to Lebesgue measure, and a Polya tree measure actually is piece-wise smooth rather than smooth.

I would name to main reasons why there is little work on priors on smooth measures:

- Tractability: Forcing a random measure to be smooth requires stochastic dependence. Distributions on such measures are hard to construct to begin with, but to use them as priors, we have to condition on data, and dependencies typically become much more complicated in the posterior. Comparing to those random measures which are actually used—in particular the CRMs and normalized CRMs discussed above—shows that they are defined precisely to keep dependencies between different subsets of Ω_ϕ at an absolute minimum.
- More bespoke models have proven more useful: As discussed already in the introduction to this chapter, the main motivation to construct general, smooth random measure priors was originally to model the unknown random measure in de Finetti’s theorem directly. Three decades of research in Bayesian nonparametrics show rather compellingly that that approach seems to be too brute-force for most problems; modelling a more specific pattern turns out to be more useful.

8.10. Further references

The book to read on Poisson processes is [32]; if you have any interest in random discrete measures, I would recommend to read at least Chapters 2, 5.1, 8 and 9. A very accessible exposition of point processes, Lévy processes and related topics is given by Cinlar [6]. For more general point process theory, I have also found the two volumes by Daley and Vere-Jones [7] useful. For posterior properties of normalized CRMs, see [26] and [34]. Mixture models based on normalized CRMs can be sampled with an algorithm similar to Neal’s Algorithm 8 [11].

APPENDIX A

Poisson, gamma and stable distributions

The distributions we encounter most commonly when working with random discrete measures are the Poisson, gamma and Dirichlet distributions. The next few pages collect their most important properties for purposes of Bayesian non-parametrics.

A.1. The Poisson

Recall that the Poisson distribution is the distribution we obtain from the series expansion

$$e^\lambda = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \quad (\text{A.1})$$

of the exponential function: If we normalize by multiplication with $e^{-\lambda}$ and multiply in a point mass δ_k at each k , we obtain a probability measure

$$P_\lambda(\bullet) := \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \delta_k(\bullet) \quad (\text{A.2})$$

on $\mathbb{N} \cup \{0\}$, called the **Poisson distribution** with parameter λ . The Poisson distribution is usually defined for $\lambda > 0$, but our definition includes the case $\lambda = 0$, for which $P_\lambda = \delta_0$.

The Poisson distribution has two very useful properties that we use at various points in these notes:

- (1) **Additivity:** If $K_1 \sim \text{Poisson}(\alpha_1)$ and $K_2 \sim \text{Poisson}(\alpha_2)$ then

$$(K_1 + K_2) \sim \text{Poisson}(\alpha_1 + \alpha_2) . \quad (\text{A.3})$$

- (2) **Thinning:** The number of successes in a Poisson number of coin flips is Poisson, namely if $K \sim \text{Poisson}(\alpha)$ and $X_1, \dots, X_K \sim_{\text{iid}} \text{Bernoulli}(p)$, then

$$\sum_{i=1}^K X_i \sim \text{Poisson}(p\alpha) . \quad (\text{A.4})$$

A.2. The gamma

The gamma distribution is the distribution on \mathbb{R}_+ with Lebesgue density

$$f_{\alpha,\beta}(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \quad \alpha, \beta > 0 . \quad (\text{A.5})$$

The gamma has a “magic” property which makes it unique among all distributions on positive scalars, and which directly accounts for many special properties of the Dirichlet process:

Theorem A.1 (Lukacs [39]). *Let X and Y be two non-degenerate and positive random variables, and suppose that they are independently distributed. The random variables $U := X + Y$ and $V := X/Y$ are independently distributed if and only if both X and Y have gamma distribution with the same scale parameter β .* \triangleleft

Remark A.2 (Gamma and Poisson). The gamma can be obtained as the natural conjugate prior of the Poisson: If the variables K_1, \dots, K_n are i.i.d. $\text{Poisson}(\lambda)$, the sum $S_n(K_{1:n}) = \sum_i K_i$ is a sufficient statistic for λ . By additivity (A.3) of the Poisson, the sum is distributed as $S_n \sim \text{Poisson}(n\lambda)$. Following Remark 7.13, we can obtain the natural conjugate prior by passing to the image measure under S —which is again $\text{Poisson}(n\lambda)$, since S is the identity—and renormalizing it as a density with respect to λ . Since

$$P_{n\lambda}(k) = e^{-n\lambda} \frac{(n\lambda)^k}{k!} = e^{-n\lambda} \frac{(n\lambda)^k}{\Gamma(k+1)}, \quad (\text{A.6})$$

the normalization constant is given by

$$\int_{\mathbb{R}_+} e^{-n\lambda} (n\lambda)^k d\lambda = \frac{1}{n} \int_{\mathbb{R}_+} e^{-n\lambda} (n\lambda)^k d(n\lambda) = \frac{1}{n} \Gamma(k+1). \quad (\text{A.7})$$

(The first equality is a change of variables, the second the definition of the gamma function.) The conjugate prior density is hence

$$e^{-n\lambda} \frac{n^k \lambda^k}{\frac{1}{n} \Gamma(k+1)} = e^{-n\lambda} \frac{n^{k+1} \lambda^k}{\Gamma(k+1)} = f_{k+1,n}(\lambda), \quad (\text{A.8})$$

that is, a gamma density with parameter $\alpha = k+1$ and $\beta = n$. \triangleleft

The additivity of the Poisson is inherited by the gamma. Since the first parameter α corresponds to the value of the Poisson draw and the gamma variable λ to the Poisson parameter, (A.3) translates into the following property:

(1) **Additivity:** If $\lambda_1 \sim \text{Gamma}(\alpha_1, \beta)$ and $\lambda_2 \sim \text{Gamma}(\alpha_2, \beta)$ then

$$(\lambda_1 + \lambda_2) \sim \text{Gamma}(\alpha_1 + \alpha_2, \beta). \quad (\text{A.9})$$

We have already seen above that the sum of n i.i.d. $\text{Poisson}(\lambda)$ variables has distribution $\text{Poisson}(n\lambda)$. If we double the number of samples, the Poisson parameter doubles to $2n\lambda$. Since n corresponds to the second parameter of the gamma, Poisson additivity induces another useful property:

(2) **Scaling:** If $\lambda \sim \text{Gamma}(\alpha, \beta)$, then

$$c \cdot \lambda \sim \text{Gamma}(\alpha, c\beta) \quad \text{for any constant } c > 0. \quad (\text{A.10})$$

A.3. The Dirichlet

Suppose we sample K positive random variables λ_k . Since the λ_k take positive scalar values, the random vector

$$C_{1:K} := \left(\frac{\lambda_1}{\sum_k \lambda_k}, \dots, \frac{\lambda_K}{\sum_k \lambda_k} \right) \quad (\text{A.11})$$

is a random element of the simplex \triangle_K , that is, a random probability distribution on K events. If the λ_k are independent gamma variables $\lambda_k \sim \text{Gamma}(\alpha_k, \beta)$, the

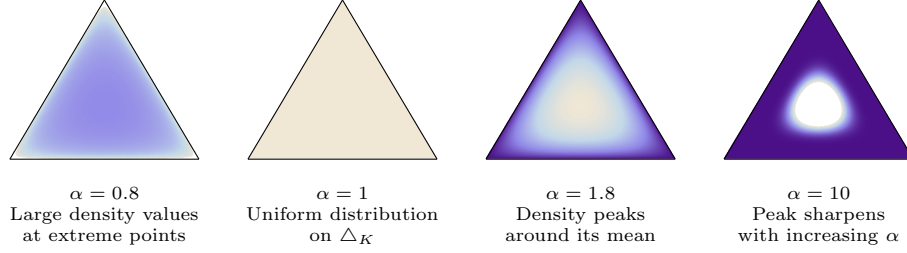


FIGURE A.1. Dirichlet distribution on Δ_3 , with uniform expectation, for various concentration parameters (dark colors = small density values).

distribution of $C_{1:K}$ is the **Dirichlet distribution**, given by the density

$$f(c_{1:K} | \alpha, g_{1:K}) := \frac{1}{K(\alpha, g_{1:K})} \exp\left(\sum_{k=1}^K (\alpha g_k - 1) \log(c_k)\right)$$

with respect to Lebesgue measure (restricted to the subset $\Delta_K \subset \mathbb{R}^K$). The Dirichlet is parametrized by its mean $g_{1:K} = \mathbb{E}[C_{1:K}] \in \Delta_K$ and a concentration parameter $\alpha > 0$. These parameters are derived from the parameters of the gamma variables λ_k as

$$\alpha \text{ (in the Dirichlet)} = \beta \text{ (in the gamma)} \quad (\text{A.12})$$

and

$$g_{1:K} := \left(\frac{\alpha_1}{\sum_k \alpha_k}, \dots, \frac{\alpha_K}{\sum_k \alpha_k} \right). \quad (\text{A.13})$$

The two different uses of α are a bit unfortunate, but are so common in the literature that I will not meddle. Figure A.1 illustrates the effect of the concentration parameter in the case of uniform expectation $g_{1:3} = (1/3, 1/3, 1/3)$ on Δ_3 .

For any random element of $C_{1:K}$ of Δ_K , the individual entries C_k are dependent random variables, since they couple through the normalization constraint $\sum_k C_k = 1$. If $C_{1:K}$ is defined by normalizing a vector of positive variables as in (A.11), the variables C_k additionally couple through the total mass: If $T := \sum \lambda_k$ in (A.11), and hence $C_k = \lambda_k/T$, then C_k and T are in general stochastically dependent, which introduces additional stochastic dependence between any two C_i and C_j through T . Theorem A.1 above shows that the *only* exception to this rule is Dirichlet distribution: In the Dirichlet, the C_k couple only through normalization. In this sense, Dirichlet random variables have the simplest structure among all random variables with values in Δ_K .

A.4. The stable

The additivity property (A.3) shows that sums of Poisson random variables are again Poisson; since summation changes the Poisson parameter, the sum and the individual summands differ in their means and variances (both of which are controlled by λ). Can we obtain a similar property for \mathbb{R} -valued random variables? Instead of positing a specific parametric model like the Poisson and demanding that we remain within the model when we take sums, though, we now demand simply that sum and summands differ only in terms of mean and variance. Two random variables X and Y differ only in their mean and variance iff they satisfy $X \stackrel{d}{=} aY + b$

for some constants a and b . Hence, we ask how i.i.d. variables X_0, X_1, \dots, X_n have to be distributed to satisfy

$$\sum_{i=1}^n X_i \stackrel{d}{=} a_n X_0 + b_n, \quad (\text{A.14})$$

for two suitable sequences (a_n) and (b_n) of constants. The additive constant is less interesting: Clearly, if we center the X_i , then the sum is also centered, so we can always eliminate b_n . It can be shown that the constants a_n must always be of the form $a_n = n^{1/\alpha}$ for some $\alpha \in (0, 2]$. We hence define our class of random variables as follows: A distribution $P_\alpha \in \mathbf{PM}(\mathbb{R})$ is called a **stable distribution** or **α -stable** if

$$\sum_{i=1}^n X_i \stackrel{d}{=} n^{1/\alpha} X_0 + b_n \quad \text{whenever } X_0, X_1, \dots \sim_{\text{iid}} P_{\alpha, (b_n)} \quad (\text{A.15})$$

for some $\alpha \in (0, 2]$ and $b_1, b_2, \dots \in \mathbb{R}$. The constant α is called the **index**. If $b_n = 0$ for all n , $P_\alpha = P_{\alpha, (b_n)}$ is called **strictly stable**.

The stable distribution does not in general have a Lebesgue-density, except in some special case, notably for $\alpha = 2$, in which case P_α is a Gaussian. However, (A.15) clearly implies that a strictly stable distribution P_α is infinitely divisible, and it hence has a Lévy -Khinchine representation: A random variable is α -stable if and only if it is (i) normal (if $\alpha = 2$) or (2) has Lévy measure

$$\rho(x) = \begin{cases} c_\oplus x^{-\alpha-1} & \text{if } x > 0 \\ c_\ominus x^{-\alpha-1} & \text{if } x < 0 \end{cases}, \quad (\text{A.16})$$

where at least one of the constants $c_\oplus, c_\ominus \geq 0$ is non-zero [28, Proposition 15.9].

Reading up on the stable is a bit of a mess, since definitions, parametrization and naming are not uniform in the literature; almost every author treats the stable slightly differently. Perhaps the closest thing to a standard reference is [69]. As a concise reference available online, I recommend [27].

APPENDIX B

Nice spaces

Non-trivial probability models require a modicum of topological assumptions: The most relevant spaces for probability and statistics are Polish spaces and standard Borel spaces, which are essentially two sides of the same coin. Also relevant are locally compact spaces, which are spaces on which we can properly work with density representations of probability measures.

B.1. Polish spaces

A topological space \mathbf{X} is called a **Polish space** if it is complete, separable and metrizable. **Complete** means the limit of every convergent sequence of points in \mathbf{X} is again in \mathbf{X} . **Separable** means \mathbf{X} has a dense subset that is countable. \mathbf{X} is metrizable if there exists a metric that generates the topology (it is possible to define topologies that cannot be generated by any metric).¹ Roughly speaking, a Polish topology is the minimum structural requirement necessary to ensure that real and functional analysis are applicable on a space. A Polish structure ensures, for example, that:

- Conditional probabilities are well-behaved.
- All probability measures are Radon measures, i.e. their value on any set can be approximated to arbitrary precision by their values on compact sets—which at first glance may not seem to be a big deal, but for almost all purposes of statistics and most purposes of probability theory, non-Radon measures are basically useless.

As we trade off generality against nice properties, Polish spaces emerge as the golden mean for most applications of probability and statistics. They have most of the pleasant analytical properties of Euclidean spaces (though not necessarily the geometric ones, such as a vector space structure and a scalar product). The class of Polish spaces is much larger, though, and practically all spaces of interest for the purposes of statistics; examples include:

- (1) All finite spaces (in the discrete topology).
- (2) Euclidean space.

¹If you are not used to working with topologies, think of \mathbf{X} as a set of points. We choose some metric d on \mathbf{X} . A metric defines a notion of convergence, so now we can ask whether \mathbf{X} is complete in this metric. It also defines whether a subset A is dense—it is if we can get arbitrarily close to any point in \mathbf{X} by choosing an appropriate point in A , where closeness is measured by d . If the space is separable and complete, we have a complete, separable *metric* space. The open sets in this space are called the topology of \mathbf{X} , and the σ -algebra they generate are the Borel sets. There may be many different metrics d , though, which all generate the same topology; if so, many analytical properties (such as continuity) and all measure-theoretic properties of \mathbf{X} are independent of the specific choice of d . We hence do not fix a specific d , and say that \mathbf{X} , with the topology we have chosen, is a *metrizable* space.

- (3) Any separable Banach space, in particular all separable Hilbert spaces and L_p spaces.
- (4) The space $C(\mathbb{R}_+, \mathbb{R})$ of continuous functions (in the topology of compact convergence).
- (5) The space $\mathbb{D}(\mathbb{R}_+, \mathbb{R})$ of càdlàg functions (in the Skorohod topology).
- (6) The set $\mathbf{PM}(\mathbf{X})$ of probability measures is Polish in the topology of weak convergence if and only if \mathbf{X} is Polish.
- (7) Cantor space, i.e. the set $\{0, 1\}^\infty$ (in the discrete topology).
- (8) Any countable product of Polish spaces, such as $\mathbb{R}^\mathbb{N}$, in the product topology. (Uncountable products of Polish spaces, such as $\mathbb{R}^\mathbb{R}$, are *not* Polish. They are not even Hausdorff spaces, unless in the trivial case where each factor is a singleton.)

B.2. Standard Borel spaces

To reap the benefits of a Polish topology for most measure-theoretic purposes, the space we use does not actually have to be Polish—rather, it is sufficient if the measurable sets we use are generated by *some* space which is Polish. This topology need not be the same we use for analytic purposes—to define convergence of sequences, continuity of functions, etc—the two topologies only have to generate the same measurable sets.

Definition B.1. A measurable space $(\mathbf{X}, \mathcal{A})$ is called a **standard Borel space**² if there is a Polish topology \mathcal{T} on \mathbf{X} that generates the σ -algebra \mathcal{A} . \triangleleft

Clearly, if \mathbf{X} is a Polish space and $\mathcal{B}(\mathbf{X})$ are the Borel sets on \mathbf{X} , then $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$ is standard Borel. But the definition is considerably more general: The system of measurable sets generated by a topology is *much* larger than the topology itself, and we can often make a topology considerably finer or coarser (i.e. considerably increase or decrease the number of open sets) without changing the Borel sets. The finer the topology, the fewer sequences converge, and the fewer sets are hence dense. If the topology on a separable (uncountable) space is made finer and finer, the space will cease to be separable at some point.

We have already seen that standard Borel spaces guarantee conditionals. Another important property is that they are, roughly speaking, spaces of countable complexity:

Lemma B.2. *Let \mathbf{X} be a standard Borel space. Then there exists a countable system of measurable sets A_1, A_2, \dots which **separates** points in \mathbf{X} , that is, for any two distinct points $x_1, x_2 \in \mathbf{X}$, there is a set A_i in the system such that $x_1 \in A_i$ and $x_2 \notin A_i$.* \triangleleft

If A_1, A_2, \dots is a separating sequence, we can uniquely characterize each element $x \in \mathbf{X}$ by the sequence

$$\mathbf{I}(x) := (\mathbb{I}_{A_1}(x), \mathbb{I}_{A_2}(x), \dots) \quad (\text{B.1})$$

of indicator functions. Thus, each element of a standard Borel space is determined by a countable sequence of scalars, a property which we can informally think of as

² Some authors call standard Borel spaces simply “Borel spaces”, a very sensible terminology which I would use here if not for its ambiguity: Other authors use the term Borel space for any measurable space generated by a topology, or simply for any measurable space, or specifically only for uncountable spaces.

a countable dimension, even though the entries of the sequence do not correspond to axes in a vector space.

B.3. Locally compact spaces

Locally compact spaces are relevant for statistical modeling purposes since they are basically the spaces on which we can use densities to represent distributions. The reason why we do not use densities representations for the Gaussian process or Dirichlet process, for example, is precisely because these distributions live on infinite-dimensional spaces that are not locally compact.

The Radon-Nikodym theorem does of course tell us that densities of probability measures exist on any measurable space. The problem is that a density p is always the representation of one measure P with respect to another measure, say μ :

$$P(dx) = p(x)\mu(dx) \quad (\text{B.2})$$

The measure μ is often called the **carrier measure**. For modeling purposes, this representation is only useful if the measure μ is “flat”: If μ is, say, a Gaussian, then a large value of p in a given region could indicate either that P puts a lot of mass in the region, or that μ is small in the region. That means p by itself is not informative; to make it informative, we have to integrate against μ , and have not simplified the representation of P at all.

More formally, for μ to be a useful carrier measure, we need it to be translation invariant. To formulate translation invariance, we first need a translation operation, which we denote $+$ (since on most standard spaces it coincides with addition). A translation should be reversible, so we require $(\mathbf{X}, +)$ to be group. Since we regard \mathbf{X} as a space with a topological structure, rather than just a set of points, $+$ has to be compatible with that structure, i.e. the $+$ -operation must be continuous in the topology of \mathbf{X} . A group whose operation is continuous is called a **topological group**.

We call a measure μ on $(\mathbf{X}, +)$ **translation invariant** if $\mu(A + x) = \mu(A)$ for every Borel set A and point x in \mathbf{X} . Informally, this means that, if we shift a set A around in \mathbf{X} by means of the operation $A + x$, the mass of A under μ does not change. Thus, $\mu(A)$ depends on the shape and size of A , but not on where in the space A is located. The general class of spaces on which such measures exist are locally compact spaces.

Definition B.3. A space \mathbf{X} is called **locally compact** if every point has a compact neighborhood, i.e. if for every $x \in \mathbf{X}$, there is a compact subset of \mathbf{X} that contains x . \triangleleft

An attempt at explanation: Very informally, there are different ways to create a non-compact space from compact ones. We could, say, glue together a countable number of unit intervals (which are compact) to produce a space like the real line (which is not compact). That means we are changing the global structure of the space, but the local structure around an individual point remains the same—the neighborhood of any given point is still a line. Obviously each point is enclosed in a compact set, so the resulting non-compact space is still locally compact. Alternatively, we could multiply the intervals into a Cartesian product. In this case, the local structure around each point changes—it is now something infinite-dimensional—and the resulting space is not locally compact.

Theorem B.4. *Let $(\mathbf{X}, +)$ be topological group. If \mathbf{X} is locally compact and separable, there is a measure λ on \mathbf{X} , unique up to scaling by a positive constant, which satisfies $\lambda(A + x) = \lambda(A)$ for all Borel sets A and points $x \in \mathbf{X}$.* \triangleleft

The measure λ is called **Haar measure** on the group $(\mathbf{X}, +)$. Lebesgue measure is Haar measure on the group (\mathbb{R}^d, d) , scaled to satisfy $\lambda([0, 1]^d) = 1$.

Most nonparametric priors have no useful density representations, because the parameter spaces we encounter in Bayesian nonparametrics are infinite-dimensional, and infinite-dimensional spaces are not locally compact. We have to be careful with this statement, because only vector spaces have a straightforward definition of dimension (by counting basis elements), and spaces such as $\mathbf{PM}(\mathbf{X})$ have no natural vector space structure. In the case of vector spaces, however, we can make the statement precise:

Lemma B.5. *A topological vector space is locally compact if and only if it is finite-dimensional.* \triangleleft

APPENDIX C

Conditioning

Bayesian statistics involves a lot of conditional probabilities, and in Bayesian nonparametrics, we cannot just get away with using conditional densities. Measure-theoretic conditioning is, unfortunately, the problem child in most probability textbooks, and I am frankly at a loss to provide a single concise reference. Below, I have summarized some basic properties for reference.

C.1. Probability kernels

A **probability kernel** is a measurable, measure-valued mapping. That is, if \mathbf{X} and \mathbf{Y} are Borel spaces, a measurable mapping $\mathbf{p} : \mathbf{Y} \rightarrow \mathbf{PM}(\mathbf{X})$ is a probability kernel. For each $y \in \mathbf{Y}$, $\mathbf{p}(y)$ is a probability measure on \mathbf{X} , and it is hence useful to write \mathbf{p} as a function of two arguments:

$$\begin{array}{ccc} & \mathbf{p}(A, y) & \\ & \uparrow \quad \uparrow & \\ \text{measurable set in } \mathbf{X} & \xrightarrow{\quad} & \text{point in } \mathbf{Y} \end{array}$$

Our two most important uses for probability kernels are conditional probabilities and statistical models. A conditional probability of $X \in \mathbf{RV}(\mathbf{X})$ given $Y \in \mathbf{RV}(\mathbf{Y})$ is a probability kernel $\mathbf{Y} \rightarrow \mathbf{PM}(\mathbf{X})$, namely

$$\mathbf{p}(\bullet, y) := \mathbb{P}[X \in \bullet | Y = y] . \quad (\text{C.1})$$

Similarly, a model $M = \{P_\theta | \theta \in \mathbf{T}\}$ can be regarded as a probability kernel by defining

$$\mathbf{p}(\bullet, \theta) := P_\theta(\bullet) . \quad (\text{C.2})$$

We recover the set M as $M = \mathbf{p}(\bullet, \mathbf{T})$.

C.2. Conditional probability

A conditional probability of X given Y is formally a probability kernel \mathbf{p} with the interpretation

$$\mathbb{P}[\bullet | Y = y] := \mathbf{p}(\bullet, y) . \quad (\text{C.3})$$

The intuitive notion of a conditional distribution implies some technical requirements which the mapping \mathbf{p} must satisfy to be of any use:

Definition C.1. Let X and Y be two random variables with values in \mathbf{X} and \mathbf{Y} respectively. A measurable mapping $\mathbf{p} : \mathbf{Y} \rightarrow \mathbf{PM}(\mathbf{X})$ (i.e. a probability kernel) is called a **conditional probability** of X given Y if it satisfies

$$\forall A \in \mathcal{B}(\mathbf{X}) : \int_B \mathbf{p}(A, y) Y[\mathbb{P}](dy) = \mathbb{P}\{X \in A, Y \in B\} \quad (\text{C.4})$$

for all $A \in \mathcal{B}(\mathbf{X})$ and $B \in \mathcal{B}(\mathbf{Y})$ and

$$\mathbb{P}[\{Y = y\} | Y = y] = 1 \quad Y[\mathbb{P}]\text{-a.s.} \quad (\text{C.5})$$

<

The definition makes three requirements on the mapping \mathbf{p} , namely (C.4), measurability, and (C.5). Each of has an intuitive meaning:

- (1) Equation (C.4) simply says that $\mathbf{p}(A, y)$ is the probability of $X \in A$ given that $Y = y$, although the statement is disguised as an integral: If the event $\{Y = y\}$ has non-zero measure, we would state this in the form

$$\mathbb{P}(X \in A | Y = y) = \frac{\mathbb{P}(\{X \in A\} \cap \{Y = y\})}{\mathbb{P}(\{Y = y\})} =: \mathbf{p}(A, y).$$

Since in general $\{Y = y\}$ may very well be a null set—say, if Y is a Gaussian variable and y a point on the line—(C.4) instead requires that \mathbf{p} integrates *as if* it was an expression of the form above. That such an implicit definition is meaningful is not at all obvious: We have to prove that (C.4) actually determines \mathbf{p} for almost all y ; see Theorem C.2 below.

- (2) If \mathbf{p} is not measurable, elementary statements about conditional probabilities become meaningless. For example, the probability (under the distribution of Y) that the conditional probability $\mathbf{p}(A, y)$ has value $t \in [0, 1]$ is $\mathbb{P}(Y^{-1}(\mathbf{p}(A, \bullet)^{-1}\{t\} \cap B))$, which is only defined if \mathbf{p} is measurable.
- (3) Equation (C.4) simply states that, if we already know that $Y = y$, any event in \mathbf{X} that would imply otherwise has zero probability. Although this requirement is semantically simple and clearly necessary, it is a bit harder to formulate and proof in detail.¹

We have to make sure that \mathbf{p} exists when we need it. The next theorem shows that this is the case *whenever the sample space of X is a standard Borel space*, which is pretty much everything we ever need to know about the existence of conditionals.

Theorem C.2 (conditional probability). *Let X and Y be random variables with values in measurable spaces \mathbf{X} and \mathbf{Y} . If \mathbf{X} is a standard Borel space, there is a probability kernel $\mathbf{p} : \mathbf{Y} \rightarrow \mathbf{PM}(\mathbf{X})$ which is a probability kernel of X given Y in the sense of Definition C.1. holds for all $A \in \mathcal{B}(\mathbf{X})$ and $B \in \mathcal{B}(\mathbf{Y})$. The kernel \mathbf{p} is uniquely determined up to modification on a $Y[\mathbb{P}]$ -null set on \mathbf{Y} .* <

C.3. Conditional random variables

In modeling problems, we frequently encounter “conditionally” distributed random variables of the form $X|Y$. A fact very useful for technical purposes is that $X|Y = y$ can indeed be regarded as a random variable “parameterized” by y .

Theorem C.3 ([28, Lemma 3.22]). *Let \mathbf{X} be a Polish and \mathbf{Y} a measurable space. Let $X : \Omega \rightarrow \mathbf{X}$ be a random variable and $\mathbf{p} : \mathbf{Y} \rightarrow \mathbf{PM}(\mathbf{X})$ a probability kernel. Then there is a measurable mapping $X' : \Omega \times \mathbf{Y} \rightarrow \mathbf{X}$ such that*

$$\mathbb{P}(X'(\bullet, y) \in A) = \mathbf{p}(A, y) \tag{C.6}$$

for $Y[\mathbb{P}]$ -almost all y . <

¹ Technically, (C.5) means that (a) in the abstract probability space Ω , the fibres $Y^{-1}(y)$ of the mapping Y are $X^{-1}\mathcal{B}(\mathbf{X})$ -measurable and (b) for almost all $y \in \mathbf{Y}$, the pullback measure $X^\# \mathbf{p}(\bullet, y)$ concentrates on the fibre $Y^{-1}(y)$.

Depending on the context, it can be useful to denote X' as $X^y(\omega) := X'(\omega, y)$. Note that X^y depends measurably on y . What the theorem above says is, in other words, that given a conditional probability of the form $\mathbb{P}[X \in \bullet | Y = y]$ on a Polish space \mathbf{X} , there are random variables X^y such that

$$\mathcal{L}(X^y) = \mathbb{P}[X \in \bullet | Y = y] . \quad (\text{C.7})$$

We can hence interpret X^y as the “conditional random variable” $X|Y = y$.

C.4. Conditional densities

Let X and Y be random variables with values in standard Borel spaces \mathbf{X} and \mathbf{Y} . Now choose a σ -finite measure μ on \mathbf{X} . Since the conditional probability of X given Y is a probability measure on \mathbf{X} for every $y \in \mathbf{Y}$, we can ask whether it has a density with respect to μ , i.e. if there is a measurable function p such that

$$\mathbb{P}[X \in dx | Y = y] = p(x|y)\mu(dx) \quad \mathcal{L}(Y)\text{-a.s.} \quad (\text{C.8})$$

If so, p is called a **conditional density** of X given Y . As a probability measure, each distribution $\mathbb{P}[X \in dx | Y = y]$ is of course absolutely continuous with respect to *some* σ -finite measure, but the question is whether a single μ can be found for all values of y . (It hence comes down to the question whether the family $\{\mathbb{P}[\bullet | Y = y] | y \in \mathbf{Y}\}$ is dominated, cf. Sec. 7.2.) It will therefore not come as a surprise that a sufficient condition can be formulated based on absolute continuity of the *joint* distribution:

Lemma C.4. *Require that the joint distribution $P := \mathcal{L}(X, Y)$ satisfies*

$$P \ll \mu \otimes \nu \quad \text{and define} \quad p(x, y) := \frac{P(dx \times dy)}{\mu(dx)\nu(dy)} . \quad (\text{C.9})$$

Then $\mathbb{P}[X \in dx | Y = y] \ll \mu(dx)$ holds $\mathcal{L}(Y)$ -a.s., i.e. the conditional density $p(x|y)$ exists. It is given by

$$p(x|y) = \frac{p(x, y)}{p(y)} \quad \text{where} \quad p(y) := \int_{\mathbf{X}} p(x, y) \mathbb{P}(X \in dx) . \quad (\text{C.10})$$

Additionally, the function p is a density of $\mathcal{L}(X)$ with respect to μ . \triangleleft

Index of definitions

- α -stable, 90
- d -array, 50
- additivity of gamma variables, 88
- additivity of Poisson variables, 87
- asymptotic frequencies, 50
- atom locations, 6
- atoms, 6
- Banach space, 64
- base measure, 9
- Bayes equation, 58
- Bayesian mixture model, 7
- Bayesian model, 3
- Bernoulli process, 72
- Bernoulli sampling model, 72
- beta process, 80
- beta-stable random measure, 80
- blocks of a family of sets, 23
- blocks of a partition, 15
- carrier measure, 93
- Chinese restaurant process, 16
- clustering problem, 5
- clusters, 5
- collaborative filtering, 24
- complete, 64
- complete space, 91
- completely random, 75
- completely random measure, 80
- concentration, 9
- conditional density, 97
- conditionally i.i.d., 44
- conditionally independent, 43
- conjugate, 61
- consistent, 67
- consistent in the Bayesian sense, 67
- consistent in the frequentist sense, 68
- covariance function, 30
- covariate-dependent model, 40
- covariates, 40
- CRM, 80
- dependent Dirichlet process, 41
- Dirac measure, 6
- Dirichlet process, 9
- Dirichlet process mixture, 10
- discrete probability measure, 6
- dominated model, 59
- Doob's theorem, 68
- dust, 49
- emission model, 37
- entropy, 63
- ergodic structures, 53
- exchangeable, 46
- exchangeable random partition, 48
- exponential family, 64
- family of sets, 23
- feature, 24
- features, 24
- finite mixture, 6
- finite-dimensional distributions, 29
- finite-dimensional marginals, 29
- Fréchet derivative, 65
- gamma process, 80
- Gaussian process, 29
- GEM(α) distribution, 19
- Gibbs measures, 64
- Gibbs sampler, 12
- graph limits, 52
- graphons, 52
- Haar measure, 94
- HDP, 39
- hidden Markov model, 37
- hierachical Dirichlet process, 39
- Hilbert-Schmidt integral operators, 33
- homogeneous, 7
- hyperparameter, 61
- imputation, 11
- increment, 78
- index of a stable distribution, 90
- Indian buffet process, 26
- infinite hidden Markov model, 39
- infinite symmetric group, 46
- infinitely divisible, 78
- integral kernel, 33
- jointly exchangeable, 51

- kernel density estimator, 2
- Lévy process, 78
- latent feature model, 24
- latent variables, 11
- left-ordering, 25
- linearly conjugate, 62
- locally compact, 93
- locally Lipschitz-continuous, 33
- mass partition, 49
- mean function, 30
- misspecified, 67
- mixing measure, 6, 35
- mixture distribution, 5
- mixture model, 7, 35
- model order selection, 20
- multinomial sampling model, 72
- natural conjugate priors, 66
- neutral-to-the-right (NTTR) process, 62
- nonparametric Bayesian model, 3
- nonparametric model, 1
- norm dual, 64
- normalized completely random measure, 81
- observation model, 3
- order, 24
- paint-box partition, 49
- parametric, 1
- partition, 15
- partition function, 64
- Pitman-Yor process, 18
- point mass, 6
- point process, 73
- Poisson distribution, 87
- Poisson process, 73
- Poisson random measure, 79
- Poisson-Dirichlet distribution, 19
- Polish space, 91
- posterior distribution, 3
- posterior index, 61
- power law distribution, 17
- prior, 3
- prior distribution, 3
- probability kernel, 95
- probability measures generated by a measure, 63
- projective family, 84
- projective limit, 33
- random counting measure, 73
- random partition, 15
- rate of posterior convergence, 69
- scaling of gamma variables, 88
- separable space, 91
- separately exchangeable, 51
- separating sequence of sets, 92
- simple point process, 73
- simplex, 6
- single- p model, 42
- species sampling models, 20
- stable distribution, 90
- stable process, 80
- standard Borel space, 92
- state space, 37
- statistical model, 1
- stick-breaking, 9
- strictly stable, 90
- subordinator, 79
- sufficient, 64
- thinning of Poisson variables, 87
- topic, 39
- topological group, 93
- total mass, 76
- trace, 33
- trace class, 33
- transition matrix, 38
- translation invariant, 93
- two-parameter Chinese restaurant process, 19
- two-parameter Poisson-Dirichlet, 19
- variational algorithm, 11

Bibliography

- [1] Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric estimation. *Ann. Statist.*, **2**, 1152–1174.
- [2] Beal, M. J., Ghahramani, Z., and Rasmussen, C. (2002). The infinite hidden Markov model. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 577–584.
- [3] Bertoin, J. (2006). *Random Fragmentation and Coagulation Processes*. Cambridge University Press.
- [4] Blackwell, D. (1973). Discreteness of Ferguson selections. *Annals of Statistics*, **1**(2), 356–358.
- [5] Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.*, **1**, 353–355.
- [6] Cinlar, E. (2011). *Probability and Stochastics*. Springer.
- [7] Daley, D. and Vere-Jones, D. (2008). *An introduction to the theory of point processes*, volume I and II. Springer, 2nd edition.
- [8] De Blasi, P., Favaro, S., Lijoi, A., Mena, R., Prünster, I., and Ruggiero, M. (2014). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. on Pattern Analysis and Machine Intelligence*.
- [9] Doksum, K. A. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Annals of Probability*, **2**, 183–201.
- [10] Engen, S. (1978). *Stochastic abundance models*. Chapman and Hall, London; Halsted Press [John Wiley & Sons], New York. With emphasis on biological communities and species diversity, Monographs on Applied Probability and Statistics.
- [11] Favaro, S. and Teh, Y. W. (2013). MCMC for normalized random measure mixture models. *Statist. Sci.*, **28**(3), 335–359.
- [12] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**(2).
- [13] Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.*, **2**(4), 615–629.
- [14] Ferguson, T. S. and Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data. *Annals of Statistics*, **7**, 163–186.
- [15] Fortini, S. and Petrone, S. (2012). Predictive construction of priors in Bayesian nonparametrics. *Braz. J. Probab. Stat.*, **26**(4), 423–449.
- [16] Foti, N. J. and Williamson, S. A. (2014). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*.
- [17] Ghosal, S. (2010). Dirichlet process, related priors and posterior asymptotics. In N. L. Hjort *et al.*, editors, *Bayesian Nonparametrics*, pages 36–83. Cambridge University Press.

- [18] Ghosal, S. and van der Vaart, A. (?). Fundamentals of nonparametric bayesian inference. Forthcoming.
- [19] Gnedin, A. V., Hansen, B., and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws. *Probability Surveys*, **4**, 146–171.
- [20] Griffiths, R. C. (1979). Exact sampling distributions from the infinite neutral alleles model. *Adv. in Appl. Probab.*, **11**(2), 326–354.
- [21] Griffiths, T. L. and Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In *Adv. Neural Inf. Process. Syst.*
- [22] Halmos, P. R. and Savage, L. J. (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Stat.*, **20**, 225–241.
- [23] Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.*, **18**, 1259–1294.
- [24] Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**(453), 161–173.
- [25] James, L. F., Lijoi, A., and Prünster, I. (2005). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Stat.*, **33**, 105–120.
- [26] James, L. F., Lijoi, A., and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.*, **36**, 76–97.
- [27] Janson, S. (2011). Stable distributions.
- [28] Kallenberg, O. (2001). *Foundations of Modern Probability*. Springer, 2nd edition.
- [29] Kemp, C., Tenenbaum, J., Griffiths, T., Yamada, T., and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proc. of the Nat. Conf. on Artificial Intelligence*, volume 21, page 381.
- [30] Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics*, **21**(1), 59–78.
- [31] Kingman, J. F. C. (1975). Random discrete distributions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **37**, 1–22.
- [32] Kingman, J. F. C. (1993). *Poisson Processes*. Oxford University Press.
- [33] Kleijn, B., van der Vaart, A. W., and van Zanten, J. H. (2012). Lectures on nonparametric Bayesian statistics.
- [34] Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press.
- [35] Lloyd, J. R., Orbanz, P., Ghahramani, Z., and Roy, D. M. (2012). Random function priors for exchangeable arrays. In *Adv. in Neural Inform. Processing Syst.* **25**, pages 1007–1015.
- [36] Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.*, **12**(1), 351–357.
- [37] Lovász, L. (2013). *Large Networks and Graph Limits*. American Mathematical Society.
- [38] Luenberger, D. G. (1969). *Optimization by Vector Space Methods*. J. Wiley & Sons.
- [39] Lukacs, E. (1955). A characterization of the gamma distribution. *Annals of Mathematical Statistics*, **26**(2).

- [40] MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation*, **23**, 727–741.
- [41] MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical report, Ohio State University.
- [42] McCloskey, J. W. T. (1965). *A model for the distribution of individuals by species in an environment*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Michigan State University.
- [43] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons.
- [44] Miller, J. W. and Harrison, M. T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. To appear in JMLR.
- [45] Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.
- [46] Orbanz, P. (2011). Projective limit random probabilities on Polish spaces. *Electron. J. Stat.*, **5**, 1354–1373.
- [47] Orbanz, P. (2012). Nonparametric priors on complete separable metric spaces. Preprint.
- [48] Orbanz, P. and Roy, D. M. (2013). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, to appear.
- [49] Pemantle, R. (2007). A survey of random processes with reinforcement. *Probab. Surv.*, **4**, 1–79.
- [50] Perman, M., Pitman, J., and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields*, **92**(1), 21–39.
- [51] Pitman, J. (2006). *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer.
- [52] Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, **25**(2), 855–900.
- [53] Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Harvard University Press.
- [54] Robert, C. P. (1995). Mixtures of distributions: inference and estimation. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*. Chapman & Hall.
- [55] Roy, D. M. and Teh, Y. W. (2009). The Mondrian process. In *Advances in Neural Information Processing Systems*, volume 21, page 27.
- [56] Sato, K.-I. (1999). *L'evy Processes and Infinitely Divisible Distributions*. Cambridge University Press.
- [57] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica*, **4**, 639–650.
- [58] Sethuraman, J. and Tiwari, R. C. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Statistical decision theory and related topics, III, Vol. 2 (West Lafayette, Ind., 1981)*, pages 305–315. Academic Press, New York.
- [59] Shields, P. C. (1996). *The Ergodic Theory of Discrete Sample Paths*. American Mathematical Society.
- [60] Susarla, V. and Ryzin, J. V. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical*

- Association*, **71**(356), 897–902.
- [61] Teh, Y. W. (2006). A Bayesian interpretation of interpolated Kneser-Ney. Technical report, School of Computing, National University of Singapore.
 - [62] Teh, Y.-W. and Görür, D. (2007). Indian buffet processes with power-law behavior. In *Adv. Neural Inf. Process. Syst.*
 - [63] Teh, Y. W. and Jordan, M. (2014). A gentle introduction to the Dirichlet process, the beta process and Bayesian nonparametrics. *To appear*.
 - [64] Teh, Y. W. and Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press.
 - [65] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, (476), 1566–1581.
 - [66] Thibeaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *J. Mach. Learn. Res. Proceedings*, volume 2, pages 564–571.
 - [67] van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.
 - [68] Zabell, S. L. (2005). *Symmetry and its discontents*. Cambridge Studies in Probability, Induction, and Decision Theory. Cambridge University Press, New York. Essays on the history of inductive probability, With a preface by Brian Skyrms.
 - [69] Zolotarev, V. M. (1986). *One-dimensional stable distributions*. American Mathematical Society.