



Lecture notes on Bayesian machine learning

What is BML, why use it, and how to implement it.

Rémi Bardenet and Julyan Arbel



Copyright © 2021 Rémi Bardenet and Julyan Arbel

This template is adapted from Mathias Legrand's and Vel's *Orange Book* template v2.4, licensed under CC BY-NC-SA 3.0 (<http://creativecommons.org/licenses/by-nc-sa/3.0/>).

Contents

I	What is Bayesian machine learning?	
1	The example of penalized linear regression	11
1.1	Fisher does linear regression	11
1.2	Wald does linear regression	12
1.3	Savage does linear regression	13
1.4	Hoerl and Kennard do linear regression	13
2	Maximizing expected utility	15
2.1	ML problems are decision problems	15
2.2	Bayesians maximize expected utility	16
2.2.1	Posterior subjective expected utility	16
2.3	Specifying a joint model	17
2.4	The Bayes decision rule for common ML problems	18

II**Foundations**

3	Many (incompatible) reasons to be a Bayesian	21
3.1	Because you abide by the likelihood principle	21
3.1.1	The formal LP	21
3.1.2	SEU satisfies the LP	22
3.1.3	The stopping rule principle	22
3.1.4	Pros and cons of the LP	23
3.2	Because you place coherence above all things: subjective Bayes	23
3.2.1	A closer look at the axioms	23
3.2.2	Major criticisms	26
3.3	Because you like coherence and consensus: objective Bayes	27
3.4	Because you are a Waldian frequentist in disguise	27
3.4.1	On the consistency of Bayesian estimators	27
3.4.2	Complete class theorems	27
3.4.3	PAC-Bayes statistical learning	27

III**Implementing Bayesian machine learning**

4	Markov chain Monte Carlo	31
4.1	Basic Monte Carlo	32
4.2	The Metropolis-Hastings algorithm	32
4.3	Gibbs sampling	33
4.4	Combining MCMC kernels	35
4.5	Hamiltonian Monte Carlo	35
4.5.1	An abstract variant of Metropolis-Hastings	36
4.5.2	An augmented target	36
4.5.3	Hamiltonian dynamics	36
4.5.4	Ideal HMC and numerical HMC	37
4.5.5	On the ergodicity of HMC	37
4.5.6	An ubiquitous variant: NUTS	37
4.6	MCMC practice: convergence diagnostics	37

4.7	Alternative methods	38
4.7.1	Other randomized quadratures	38
4.7.2	Deterministic quadrature in large dimensions	38
5	Variational Bayes	39
5.1	The evidence lower bound (ELBO)	40
5.2	Mean-field inference	40
5.2.1	Mean-field VB for LDA	41
5.2.2	Mean-field VB for marginal LDA	43
5.2.3	VB generalizes the EM algorithm	43
5.3	Gradient-based algorithms	43
5.3.1	VB for deep networks	44
5.4	Theoretical guarantees	44
5.5	An alternative derivation of VB, and generalizations	44
5.6	Variants of the ELBO	45

IV

Bayesian nonparametrics

6	Random functions: Gaussian processes	49
6.1	Introduction and definitions	49
6.2	Examples of Gaussian processes	50
6.3	Reproducing kernel Hilbert space	51
7	Random probability measures: Dirichlet processes and the like	53
7.1	Dirichlet process	53
7.2	Mixtures and model-based clustering	62
7.3	Priors beyond the Dirichlet process	66
7.3.1	Pitman–Yor process	69
8	Asymptotic frequentist properties	75
8.1	Introduction	75

8.2	Posterior consistency	76
8.2.1	Doob's Theorem	79
8.2.2	Schwartz approach	79
8.3	Concentration rates	82
8.3.1	iid case	84
8.3.2	Non iid case	85

V

Bayesian deep learning

9	Deep Gaussian processes and variational autoencoders	91
9.1	Introduction	91
9.2	Deep Gaussian processes	91
9.3	Variational autoencoders	91
10	Bayesian neural networks	93
10.1	Introduction	93
10.2	Prior distributions	93
10.2.1	Connection prior/initialization	93
10.2.2	Neural-network Gaussian process (NN-GP)	93
10.2.3	Neural tangent kernel (NTK)	93
10.2.4	Edge of Chaos	93
10.2.5	Unit priors get heavier with depth	93
10.2.6	Other priors	93
10.3	Posterior sampling algorithms	93
10.3.1	MCMC	93
10.3.2	Laplace approximation	93
10.3.3	Monte Carlo dropout	93
10.3.4	Variational inference	94
10.3.5	Expectation propagation	94
10.3.6	SGD-based methods	94
10.3.7	Last-layer methods	94
10.3.8	Deep ensembles	94

10.3.9	Cold posteriors	94
10.4	Generalization	94
10.4.1	PAC-Bayes	94
	Bibliography	95



What is Bayesian machine learning?

1	The example of penalized linear regression	11
1.1	Fisher does linear regression	
1.2	Wald does linear regression	
1.3	Savage does linear regression	
1.4	Hoerl and Kennard do linear regression	
2	Maximizing expected utility	15
2.1	ML problems are decision problems	
2.2	Bayesians maximize expected utility	
2.3	Specifying a joint model	
2.4	The Bayes decision rule for common ML problems	

1. The example of penalized linear regression

In this chapter, we introduce different historical approaches to learning and inference on a running example, penalized linear regression. We shall exaggerate the stances of some famous scientists, for the sake of illustration.

1.1 Fisher does linear regression

Say we have a data set (x_i, y_i) , $1 \leq i \leq N$, where $x_i \in \mathbb{R}^d$ are the *features* and $y_i \in \mathbb{R}$ the *response*. We want to study the influence of features on the response, and we ask British statistician Ronald Fisher (1890–1962) for help. He recommends positing a simple statistical model, i.e. a parametrized collection of PDFs for $y|x$. Calling the parameter θ and abusively denoting the parametrized distributions by $y|x, \theta$, we posit

$$p(y_i|x_i, \theta) = \mathcal{N}(y_i|x_i^T \theta, \sigma^2), 1 \leq i \leq N,$$

i.i.d., with known σ for simplicity. To characterize the influence of features on the response, Fisher recommends that we estimate θ , along with a confidence interval to communicate our uncertainty. Fisher was the one to introduce the maximum likelihood estimator

$$\hat{\theta}_{\text{MLE}} \triangleq \arg \max_{\theta} p(\mathbf{y}|X, \theta) = \arg \max_{\theta} \prod_{i=1}^N \mathcal{N}(y_i|x_i^T \theta, \sigma^2).$$

Now

$$\log \prod_{i=1}^N \mathcal{N}(y_i|x_i^T \theta, \sigma^2) = \log \mathcal{N}(\mathbf{y}|X\theta, \sigma^2 I) \propto -\|\mathbf{y} - X\theta\|^2,$$

where the proportionality sign allows us to drop additive terms that do not depend on θ . This entails that $\hat{\theta}_{\text{MLE}}$ is nothing but the least squares estimator

$$\hat{\theta}_{\text{MLE}} = (X^T X)^{-1} X^T \mathbf{y},$$

where we assumed that X has full rank. There is nothing inherently good in using the MLE rather than another estimator, but it often has good *frequentist* properties, i.e., properties established by integrating over the posited data generation model. For instance, it is easy to prove the following.

Proposition 1.1.1 Under $\mathbf{y} \sim \mathcal{N}(X\theta, \sigma^2 I)$, and assuming X has full rank,

$$\hat{\theta}_{\text{MLE}} \sim \mathcal{N}(\theta, \sigma^2(X^T X)^{-1}). \quad (1.1)$$

Proof. Left as an exercise. ■

Fisher then argues that (1.1) gives you a wealth of properties that make $\hat{\theta}_{\text{MLE}}$ an interesting estimator. For starters, $\hat{\theta}_{\text{MLE}}$ is unbiased ($\mathbb{E}\hat{\theta}_{\text{MLE}} = \theta$) and $\hat{\theta}_{\text{MLE}} \rightarrow \theta$ in probability. Moreover, the (random) ellipsoid

$$\mathcal{E}_\alpha = \{\theta \in \mathbb{R}^d \text{ such that } \sigma^{-2}(\theta - \hat{\theta}_{\text{MLE}})^T X^T X (\theta - \hat{\theta}_{\text{MLE}}) \leq \alpha\}$$

contains θ with known probability $1 - \delta(\alpha)$. It thus makes sense, to Fisher, to find the smallest value $\alpha_{0.99}$ such that $1 - \delta(\alpha) \geq 0.99$, and report the *confidence region* $\mathcal{E}_{\alpha_{0.99}}$ to quantify his uncertainty about θ . The property that, under repetition of the data generation, the (random) ellipsoid $\mathcal{E}_{\alpha_{0.99}}$ contains θ about 99% of the time, is called *coverage*.

1.2 Wald does linear regression

Imagine that you had asked Hungarian-American statistician Abraham Wald (1902–1950) for help instead of Fisher. He would have argued that estimating θ , or outputting a region of \mathbb{R}^d that encodes your uncertainty, are two different actions to be taken under uncertainty.

Consider estimation. Wald would ask you to specify the loss $L(\theta, \hat{\theta})$ that you incur by estimating θ by $\hat{\theta} \in \mathbb{R}^d$. You might answer $L_\alpha(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^\alpha$ for some $\alpha > 0$. Wald would then say that the accuracy of an estimator $\hat{\theta} = \hat{\theta}(\mathbf{y})$ is characterized by its risk function

$$R(\cdot, \hat{\theta}) : \theta \mapsto \mathbb{E}_{\mathbf{y}|X, \theta} L(\theta, \hat{\theta}(\mathbf{y})). \quad (1.2)$$

If the risk function of an estimator is smaller than that of another estimator, we say that the former dominates the latter. Wald's minimum requirement for an estimator is that the estimator is *admissible*, i.e., that it is not dominated. Maybe surprisingly, the MLE for regression with the squared loss L_2 is *not* admissible! This is an extension by **TBC** of an important theorem known as the James-Stein theorem (**TBC**). We leave the original James-Stein theorem as Exercise ???. We shall see an even simpler estimator that dominates the MLE in Section ??.

So what should we do, according to Wald, if the MLE is no option? It would be natural to find an estimator with as small as possible a risk function. Unfortunately, the risk function being a function, many pairs of estimators are incomparable. Wald might recommend to sum up a risk function by a single number, and look, for instance, for an estimator minimizing the worst-case risk, a so-called *minimax* estimator

$$\hat{\theta}_{\text{minimax}} \in \arg \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta}). \quad (1.3)$$

Another solution to sum up the risk function is to integrate it against a measure for which the risk is integrable. This leads to an estimator that we call the *Bayes* estimator, anticipating over Chapter 2,

$$\hat{\theta}_{\text{Bayes}} \in \arg \min_{\hat{\theta}} \mathbb{E}_{\theta} R(\theta, \hat{\theta}) = \mathbb{E}_{\theta} \mathbb{E}_{\mathbf{y}|X, \theta} R(\theta, \hat{\theta}). \quad (1.4)$$

We now have a joint distribution over θ, \mathbf{y} . As long as the loss $L(\theta, \hat{\theta}(\mathbf{y}))$ is integrable w.r.t. this joint, the tower property of the expectation yields

$$\mathbb{E}_{\theta} \mathbb{E}_{\mathbf{y}|X, \theta} L(\theta, \hat{\theta}) = \mathbb{E}_{\mathbf{y}} \mathbb{E}_{\theta|X, \mathbf{y}} L(\theta, \hat{\theta}). \quad (1.5)$$

Since $\hat{\theta}$ is only a function \mathbf{y} , minimizing (1.5) boils down to setting

$$\hat{\theta}_{\text{Bayes}} = \hat{\theta}_{\text{Bayes}}(\mathbf{y}) = \arg \inf_{\hat{\theta}} \mathbb{E}_{\theta|X, \mathbf{y}} L(\theta, \hat{\theta}).$$

For the squared loss $L = L_2$, the Bayes estimator is thus the mean of the *posterior* distribution, i.e. $\mathbb{E}_{\theta|X, \mathbf{y}} \theta$. For the one-loss $L = L_1$, the Bayes estimator is a generalized median of the same posterior distribution. In particular, note that the Bayes estimator depends on the loss function, and is not necessarily the posterior expectation.

1.3 Savage does linear regression

1.4 Hoerl and Kennard do linear regression



2. Maximizing expected utility

After formally defining decision problems, we show that basic machine learning problems such as classification, regression, model choice, etc. are decision problems. Then we introduce subjective expected utility, the single unique guideline of all Bayesians, and go over its consequences for ML decisions. Justifying subjective expected utility shall wait until Chapter II.

For this chapter, we mostly used two references. (Parmigiani and Inoue, 2009) focuses on ideas and is a formidable entry point to (Bayesian) decision theory, while (Schervish, 2012) is a great textbook-level reference for advanced reading, with mathematical details.

2.1 ML problems are decision problems

Formally, a decision problem is defined as

1. a set \mathcal{S} of *states*. For technical reasons, we also require a σ -algebra $\Sigma_{\mathcal{S}}$ that makes $(\mathcal{S}, \Sigma_{\mathcal{S}})$ a Borel space (Schervish, 2012).
2. a set of *rewards* \mathcal{R} . We also require a σ -algebra $\Sigma_{\mathcal{R}}$, which should contain all singletons.
3. a set \mathcal{A} of measurable functions from \mathcal{S} to \mathcal{R} , called *actions*.
4. a utility function $u : \mathcal{R} \rightarrow \mathbb{R}$.

We think of states as encoding all information about the situation at hand. Actions are what we are tasked to choose, and picking action a while the situation is described by a given state s leads to reward $a(s)$. Note that we assume here that the same set of actions \mathcal{A} remains available in every state s .¹

Most basic ML problems are of this kind; see Figure 2.1 for a few classical formalizations. Note that most choices made in this table are arbitrary, and correspond to the simplest variant of each problem. For instance, in classification, one might penalize false negatives and false positives

¹In future versions of the course, we might make the framework more general, to include, e.g. Markov decision processes.

	\mathcal{S}	\mathcal{R}	\mathcal{A}	$u(r)$
Regression	$(\mathcal{X} \times \mathcal{Y})^{n+1}$	$\mathcal{Y} = \mathbb{R}$	$\{a_g : s \mapsto y - g(x; x_{1:n}, y_{1:n})\}$	$-\ r\ ^2$
Classification	$(\mathcal{X} \times \mathcal{Y})^{n+1}$	$\mathcal{Y} = \{0, 1\}$	$\{a_g : s \mapsto y - g(x; x_{1:n}, y_{1:n})\}$	$\mathbb{1}_{\{r=0\}}$
Point estimation	$\mathcal{Y}^n \times \Theta$	Θ	$\{a_g : s \mapsto \theta - g(y_{1:n})\}$	$-\ r\ ^2$
Interval estimation	$\mathcal{Y}^n \times \Theta$	$\{0, 1\} \times \mathbb{R}_+$	$\{a_g : s \mapsto (\mathbb{1}_{\{\theta \in g(y_{1:n})\}}, g(y_{1:n}))\}$	$r_1 + \gamma r_2$
Model choice	$\mathcal{Y}^n \times (\cup_{m=1}^M \{m\} \times \Theta_m)$	$\{0, 1\}$	$\{a_g : s \mapsto \mathbb{1}_{\{m=g(y_{1:n})\}}\}$	r

Figure 2.1: Some classical formalizations of ML problems as decision problems. Actions are labeled by functions g (“predictors”), the domain and codomain of which should be obvious from the definition; for instance g outputs a $\{0, 1\}$ label in classification, and a Borel subset of Θ in “interval” estimation.

differently; see Exercises. Note also that, since Wald, 1950 and as done in Section ??, it is also customary to define loss functions instead of utilities, as $L(a, s) = -u(a(s))$. At this point of the document, both notations are as expressive, and we shall use them interchangeably. The distinction will come later in Chapter II, when we discuss state-dependent utilities.

2.2 Bayesians maximize expected utility

By definition, a Bayesian is someone who, facing a decision problem, picks a joint distribution p over $(\mathcal{S}, \Sigma_{\mathcal{S}})$ and commits to choosing actions according to the following principle. After observing $T \in \Sigma_{\mathcal{S}}$, they will pick action

$$a^* \in \arg \max_{a \in \mathcal{A}} \mathbb{E}[u(a(s)) | T]. \quad (2.1)$$

Usually, one states (2.1) in the particular case where $T = \mathcal{S}$, i.e.

$$a^* \in \arg \max_{a \in \mathcal{A}} \mathbb{E}_{s \sim p} u(a(s)). \quad (2.2)$$

Prescription (2.2) is called the *subjective* expected utility principle (SEU), to insist on the fact that p is an arbitrary choice from the decision maker. We choose to rather define a Bayesian as one following the conditional SEU (2.1), because for subtle reasons to be discussed in Section ??, abiding by (2.2) does not imply that we should use conditional expectations to select actions after we have observed an event. At this stage, we also have not discussed how Bayesians choose p , and there are many ways to do so; see Section ??. Finally, to give an example of Bayesian decision, the ridge regression estimator is the Bayes action for a particular decision problem and joint distribution over states; see Section ??.

2.2.1 Posterior subjective expected utility

In ML, it is customary to split the state variable into (s_O, s_U) , where $O, U \subset \{1, \dots, \dim \mathcal{S}\}$ are disjoint subsets that respectively index observed states (“data”) and unknown states. In particular, we can rewrite (2.2) as

$$a^* \in \arg \max_{a \in \mathcal{A}} \mathbb{E}_{s_O} \mathbb{E}_{s_U | s_O} u(a(s)). \quad (2.3)$$

Now assume that actions are labeled by a “predictor”, which maps s_O to one or several of the unknown variables of interest, say s_I for some $I \subset U$. This is the case for all rows in Figure 2.1. In classification or regression, for instance, the variable of interest is the new label y , and actions are

labeled by measurable predictors of this variable of interest: evaluating $g(x; x_{1:n}, y_{1:n})$ is thought of as training a given algorithm (say, an SVM) over $\{(x_i, y_i), 1 \leq i \leq n\}$ and evaluating the corresponding predictor at the new feature vector x . Now, to maximize (2.3) over \mathcal{A} , it is enough maximize it over g . And since g is a function of s_O only, the optimal g is

$$g^* : s_O \mapsto \arg \max_g \mathbb{E}_{s_U | s_O} u(a_g(s)). \quad (2.4)$$

Indeed, by maximizing the innermost expectation in (2.2) for each fixed value of s_O , we maximize the whole expectation. The resulting g^* is called the *Bayes decision rule*. The expectation in (2.4) is called *posterior* expected utility, in reference to the fact that, in practice, we often compute it for a single value of s_O , *after* s_O has been observed.

2.3 Specifying a joint model

We assume here familiarity with probabilistic graphical models, to the point of telling from a directed acyclic graph (DAG) whether two sets of nodes are independent given a third one. The reader needing a recap is referred to (K. Murphy, 2012, Sections 10.1 to 10.5).

To specify a joint distribution over states, we typically give a list of conditional distributions. To make sure we consistently define a joint distribution, we give a directed acyclic graph, and list the conditional of each node given its parents.

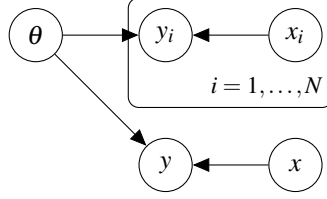


Figure 2.2: A possible DAG for a classification task.

■ **Example 2.1 — Logistic regression.** We take the set of states for classification from Table ??, and add a set of parameters Θ to describe the joint, so that $\mathcal{S} = (\mathbb{X} \times \mathbb{Y})^n \times (\mathbb{X} \times \mathbb{Y}) \times \theta$. Note that, formally, it is not necessary to include Θ in the state space if the value of your actions does not depend on θ , but we prefer to add them to keep track of all variables to later integrate over. Consider the DAG of fig. 2.3. To specify the factors, we for instance take $y_i | x_i, \theta \sim \text{Ber}(\sigma(x_i^T \theta))$ and $y | x, \theta \sim \text{Ber}(\sigma(x^T \theta))$. Take also $\theta \sim \mathcal{N}(0, \sigma^2 I)$, and any prior on each of the x_i . Then we have fixed

$$p(x_{1:n}, y_{1:n}, x, y, \theta) = p(\theta) p(x) p(y | x, \theta) \prod_{i=1}^N [p(x_i) p(y_i | x_i, \theta)].$$

■

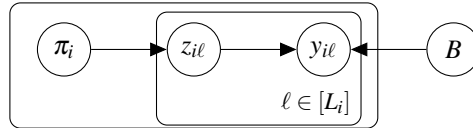


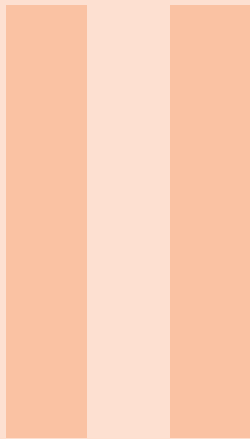
Figure 2.3: The DAG for LDA (<empty citation>).

■ **Example 2.2 — Latent Dirichlet allocation; (<empty citation>).** Consider the problem of topic assignment in a corpus of N documents, i.e. document $i \in [N]$ is represented by a list $(y_{i\ell})_{1 \leq \ell \leq L_i}$ of

words in $[V]$, with $V \geq 1$ the size of the vocabulary. On top of the words of all texts, we want to be able to ask questions on the topics contained in a text, so we will associate each word $y_{i\ell}$ to a topic $z_{i\ell} \in [K]$. To describe the distribution of a word conditionally on its topic, we also need a marginal distribution of words within each topic. Stacking these distributions on top of each other, we keep a matrix B with K rows, each of which is in Δ_V . Finally, we also need a marginal distribution for the topics within a given document. We will denote it by $\pi_i \in \Delta_K$.

We consider the DAG in ???. The factors are given by, for all $i \in [N]$ and $\ell \in [L_i]$, $y_{i\ell} | z_{i\ell} = k, B \sim \text{Cat}(B_k)$, $z_{i\ell} \sim \text{Cat}(\pi_i)$. Moreover, it is customary to take independent Dirichlet priors with the same parameter $\text{Dir}(\gamma)$ on all rows of B , as well as independent Dirichlet priors $\pi_i \sim \text{Dir}(\alpha)$ for each $i \in [N]$. The motivation behind the Dirichlet prior shall be explained in section 4.3. ■

2.4 The Bayes decision rule for common ML problems



Foundations

3	Many (incompatible) reasons to be a Bayesian	21
3.1	Because you abide by the likelihood principle	
3.2	Because you place coherence above all things: subjective Bayes	
3.3	Because you like coherence and consensus: objective Bayes	
3.4	Because you are a Waldian frequentist in disguise	

3. Many (incompatible) reasons to be a Bayesian

So far, the main appeal of subjective expected utility (SEU) has been its conceptual simplicity, and the fact that it answers all decision problems in the same manner. In this chapter, we review some attempts at justifying SEU more formally, i.e. show that SEU logically follows from some simple, consensual principle. When you hear someone say that “Non-Bayesians are incoherent”, that “a prior encapsulates the information available *before* an experiment is made, so that the prior cannot depend on data”, or that “the posterior is the experimenter’s updated belief after the experiment”, or that “Non-Bayesians violate the likelihood principle”, they are all referring to one or the other of the formal justifications below. Not all these justifications are compatible, and none can really claim to be uniformly superior, so it is important to know upon what arguments you are resting your interpretation of the Bayesian procedure.

3.1 Because you abide by the likelihood principle

The “formal” likelihood principle (Berger and Wolpert, 1988) is a semi-formal justification that deals primarily with the estimation problem. We say *half-formal* because some of the notions it deals with are left without rigorous mathematical definition, although this is not the main weak point (Berger and Wolpert, 1988, Discussions).

3.1.1 The formal LP

Consider two statistical experiments

$$E_i = (\mathcal{X}_i, \mathcal{F}_i, \{p_i(\cdot|\vartheta), \vartheta \in \Theta\}), \quad i = 1, 2.$$

Assume that for some realizations \mathbf{y}_1 and \mathbf{y}_2 ,

$$p_1(\mathbf{y}_1|\cdot) \propto p_2(\mathbf{y}_2|\cdot).$$

Note that we are using bold characters to insist on \mathbf{y}_1 and \mathbf{y}_2 being vectors concatenating (possibly many, of arbitrary dimension) observations, the label $i = 1, 2$ is only there to indicate which experiment we consider. In particular, \mathbf{y}_1 and \mathbf{y}_2 may differ in dimension.

Now, assuming that there exists a quantity $\text{Ev}(E, x)$ that encapsulates the “evidence on θ arising from E and x ”, the formal LP principle is the requirement that

$$\text{Ev}(E_1, \mathbf{y}_1) = \text{Ev}(E_2, \mathbf{y}_2).$$

As a corollary, $\text{Ev}(E, \mathbf{y})$ can depend on x solely through $p(\mathbf{y}|\cdot)$.

3.1.2 SEU satisfies the LP

Letting $\mathcal{S} = \mathcal{Y}^n \times \Theta$, SEU satisfies the LP as long as the joint distribution over states has either p_1 or p_2 as its conditional of \mathbf{y} given θ . Indeed, let

$$p_i(s_i) = p_i(\mathbf{y}_i, \theta) = p_i(\mathbf{y}_i|\theta)p(\theta) = \mathbb{Z}p_i(\theta|\mathbf{y}_i), \quad i = 1, 2.$$

Note how we use a common prior. Then for $a : \mathcal{S} \rightarrow \mathcal{X}$,

$$\int L(a, s_1) \frac{p_1(\mathbf{y}_1|\theta)p(\theta)}{\mathbb{Z}} d\theta \propto \int L(a, s_2) \frac{p_2(\mathbf{y}_2|\theta)p(\theta)}{\mathbb{Z}} d\theta,$$

so that the posterior expected losses are the same in both experiments, and Bayes actions coincide. However, note that full expected utilities are different in general,

$$\int L(a, s_1) p_1(\mathbf{y}_1|\theta)p(\theta) d\mathbf{y}_1 d\theta \neq \int L(a, s_2) p_2(\mathbf{y}_2|\theta)p(\theta) d\mathbf{y}_2 d\theta.$$

3.1.3 The stopping rule principle

The same kind of computations shows that SEU with a particular choice of joint distribution is immune to data-dependent stopping rules. This can also be seen as a consequence of the LP (Berger and Wolpert, 1988), but we stick to SEU with some conditions on its joint distribution of states for simplicity.

Assume that we want to model the following inference problem. We collect data one item at a time, independently from some distribution $y_i|\theta$, until the first $n \in \mathbb{N}$ such that $y_1, \dots, y_n \in A_n$, and then you want to estimate θ . We model this by $\mathcal{S} = \Theta \times \cup_{n \geq 1} \mathcal{Y}^n$, and decide to take a joint distribution p such that y_1, y_2, \dots are independent given θ , just like we assume the data generating mechanism works. Then the Bayes action $a^* = a_{g^*}$ minimizes

$$\begin{aligned} \mathbb{E}L(a, s) &= \mathbb{E} \left[L(a, s) \sum_n \mathbb{1}_{\{N=n\}} \right] \\ &= \sum_n \mathbb{E} [L(a, s) \mathbb{1}_{\{N=n\}}] \\ &= \sum_n \int L(a, (\theta, y_{1:n})) \mathbb{1}_{\{y_{1:n} \in A_n\}} \prod_{k < n} \mathbb{1}_{\{y_{1:k} \notin A_k\}} p(y_{1:n}|\theta) p(\theta) dy_{1:n} d\theta. \\ &= \sum_n \int dy_{1:n} \mathbb{1}_{\{y_{1:n} \in A_n\}} \prod_{k < n} \mathbb{1}_{\{y_{1:k} \notin A_k\}} \int L(a, (\theta, y_{1:n})) p(y_{1:n}|\theta) p(\theta), \end{aligned}$$

where we used the monotone convergence theorem and Fubini’s theorem (assuming, e.g., that the loss is bounded). So, to find the minimizer g^* defined on $\cup_n \mathcal{Y}^n$ of the overall expected loss, it is enough, for each n , to define $g^*(y_{1:n})$ as the usual Bayes rule for fixed n , i.e. as the minimizer of the inner integral. In other words, as long as the prior $p(\theta)$ does not depend on data, the Bayes decision is immune to data-dependent stopping rules: just act as if there were no stopping rule.

3.1.4 Pros and cons of the LP

- The LP is compelling to many (Berger and Wolpert, 1988), but it has its downsides.
- Being Bayesian is not the only way to abide by the LP.
- I am personally uncomfortable with the stopping rule principle, probably because my frequentist intuition is still too strong.
- It is hard to make fully formal: is $\text{Ev}(E, x)$ even meaningful? See answer by LeCam to (Berger and Wolpert, 1988).
- It assumes we want to specify a likelihood, this prevents model-free Bayesianism.
- It separates the roles of the likelihood and the prior. For LP-abiding Bayesians, **the prior is not allowed to depend on data.**

3.2 Because you place coherence above all things: subjective Bayes

The literature on the foundations of subjective Bayes is rich, and we refer to (Parmigiani and Inoue, 2009) for entry points. A major milestone was obtained by **Sav53**, building on work of **VoMo** and **Ram**. Savage gave a list of properties (the so-called *Savage axioms* of coherence) that a binary relation \succ on the action space \mathcal{A} should satisfy in order for it to have an essentially unique expected utility representation. More precisely, \succ satisfies the Savage axioms if and only if there is a bounded utility function $u : \mathcal{S} \rightarrow \mathbb{R}_+$ and a (finitely additive) probability measure p on \mathcal{S} such that for $a, b \in \mathcal{A}$,

$$a \succ b \Leftrightarrow \int u(a(s))dp(s) > \int u(b(s))dp(s);$$

u is unique up to affine transformations. The pair (u, p) thus characterizes the behaviour of a Savage-abiding decision maker, whose preferred action is the one minimizing the expected utility. This is the sense of the word *subjective* in SEU: the probability p , as well as the utility u , live in the mind of the DM. This is also the sense in which p can be interpreted as the DM's degree of belief. Finally, note that there is no constraint on the probability p : any pair (u, p) corresponds to a coherent decision-maker. In other words, Savage tells you that to be coherent you have to follow SEU, but it does not help you choose a joint model over states, let alone a prior.

Savage's result is beautiful, in that it builds a probability measure p and a utility function u over states from a coherent ranking of actions. Following Savage, there is no need to assume that the phenomenon we are studying is probabilistic, or to disentangle different forms of uncertainty: a coherent decision maker *must* have a utility and a probability in mind! We now take a closer look at Savage's axioms as an example of prescriptive subjective Bayesian framework, and examine some common criticisms.

3.2.1 A closer look at the axioms

Since axiom numbers 'Px' in the original paper of Savage, 1954 are often referred to as such, I use them as labels. Yet I introduce axioms in the same order as Parmigiani and Inoue, 2009.

An important thing to notice before looking at his axioms, is that Savage considers the action set \mathcal{A} to be all functions from \mathcal{S} to \mathcal{Z} . In particular, constant actions, i.e. actions of the type $s \mapsto z$ for some $z \in \mathcal{Z}$, play a crucial role in Savage's derivation, even if they do not correspond to actual actions available to the decision maker (DM) in most cases. In the sequel, we abusively denote by z the constant action $s \mapsto z$.

Axiom P1 (Preference). \succ is complete and transitive.

Since \succ is complete, we can thus check whether $z_1 \succ z_2$ for any two constant actions z_1, z_2 .

Axiom P5 (No total indifference). $\exists z_1, z_2 \in \mathcal{Z}$ such that $z_1 \succ z_2$.

■ **Example 3.1** If we take binary classification as a running example, where $\mathcal{Z} = \{0, 1\}$, the only two constant actions are the “ideal” classifier 0 that is always right and the one that is always wrong. Most of us will likely elicit $0 \succ 1$, to say that we prefer being always right to being always wrong. Note that none of the two constant actions usually belong to the “practically accessible” actions $\{a_g : s \mapsto y - g(x; x_{1:n}, y_{1:n}), g \in \mathcal{G}\}$. ■

Now for $a, b \in \mathcal{A}$, $T \in \mathcal{S}$, define action a_T^b by $a_T^b(s) = a(s)1_{s \in T} + b(s)1_{s \in T^c}$.

Axiom P2 (Sure Thing principle). $\forall a, b, h_1, h_2 \in \mathcal{A}$ and $\forall T \subset \mathcal{S}$,

$$a_T^{h_1} \succ b_T^{h_1} \Leftrightarrow a_T^{h_2} \succ b_T^{h_2}.$$

Here, Savage directly formulates that if two acts coincide on part of \mathcal{S} , then preference should only depend on their values where they differ. In particular, under **P2 (Sure Thing principle)**, we can define a new family of preference relations, called *conditional preferences*. Let $T \subset \mathcal{S}$, and define $a \succ b|T$ by

$$c_T^a \succ c_T^b \tag{3.1}$$

for some (equivalently all) $c \in \mathcal{A}$. This family of conditional preferences will be discussed later, when we consider how to rank actions after some T has been observed.

■ **Example 3.2 — Continuation of Example 3.1.** ■

Define now a null state as a $T \subset \mathcal{S}$ such that $a \sim b|T$ for all $a, b \in \mathcal{A}$. Intuitively, this means that preferences are insensitive to T obtaining.

Axiom P3 (No reduction to conditional indifference). If T is not null, then $a_T^{z_1} \succ a_T^{z_2}|T$ iff $z_1 \succ z_2$.

Note that by **P2 (Sure Thing principle)**, the values of a and b outside T do not matter. Axiom **P3 (No reduction to conditional indifference)** makes sure that no preference among consequences can be reduced to indifference conditional on a non-null event.

Axiom P4 (Separation). Assume that $z_1 \succ z_2$ and $z'_1 \succ z'_2$. Let $T_1, T_2 \subset \mathcal{S}$, and

$$a = z_{2T_1}^{z_1}, b = z_{2T_2}^{z_1}, a' = z_{2T_1}^{z'_1}, b' = z_{2T_2}^{z'_1}.$$

Then $a \succ b \Leftrightarrow a' \succ b'$.

In words, for two actions that are both piecewise constant taking the same set of values, I can change the constant values without altering the preference, as long as I preserve the order between outcomes. Intuitively, if outcomes were money, your willingness to bet on T_1 obtaining rather than T_2 obtaining does not depend on how much money you make/lose in each of these two bets.

We can now define a binary relation $T_1 \succ T_2$ over subsets of \mathcal{S} that we could interpret as “ T_1 is more likely to me than T_2 ”. We say $T_1 \succ T_2$ if for all $z_1, z_2 \in \mathcal{Z}$ such that $z_1 \succ z_2$,

$$z_{2T_1}^{z_1} \succ z_{2T_2}^{z_1}.$$

In the betting metaphor, you’re more willing to bet on T_1 obtaining than T_2 for the same rewards.

These first five axioms are enough to get the following representation.

Theorem 3.2.1 — Qualitative representation. If **P1 (Preference)**, **P2 (Sure Thing principle)**, **P3 (No reduction to conditional indifference)**, **P4 (Separation)**, and **P5 (No total indifference)** hold, then \succ on \mathcal{S} is a qualitative probability, that is

- \succ is negatively transitive: $T_1 \preceq T_2 \preceq T_3$ implies $T_1 \preceq T_3$,
- $\forall R \subset \mathcal{S}, T \succ \emptyset$.
- If $T_1 \cap U = T_2 \cap U = \emptyset$, then $T_1 \succ T_2$ iff $T_1 \cup U \succ T_2 \cup U$.

We could stop here and work with beliefs specified by pairwise rankings of events. But Savage, 1954 goes forward, and keeps adding axioms to get a more usual *quantitative* probability. In particular, we need an additional structural axiom on \mathcal{S} . There are variations of this “partition axiom”, and Savage chooses to embed it in its Archimedean axiom.

Axiom P6 (Archimedean). $\forall a, b \in \mathcal{A}$ such that $a \succ b$, and $\forall z \in \mathcal{Z}$, there exists a finite partition of $\mathcal{S} = T_1 \cup \dots \cup T_M$ such that for all T_i , either $a_{T_i}^z \succ b$ or $a \succ b_{T_i}^z$.

An important consequence of **P6 (Archimedean)** is that \mathcal{S} must be rich enough to be splittable into tiny pieces suiting any pair of actions. This is in general not possible for a discrete space, and we will generally have to use a continuous state space.

Theorem 3.2.2 — Qualitative representation. Under **P1 (Preference)**, **P2 (Sure Thing principle)**, **P3 (No reduction to conditional indifference)**, **P4 (Separation)**, **P5 (No total indifference)**, and **P6 (Archimedean)**, there exists a unique finitely additive probability measure Π on $2^{\mathcal{S}}$ such that

- $T_1 \succ T_2$ iff $\Pi(T_1) > \Pi(T_2)$,
- $\forall T_1 \subset \mathcal{S}$ and $k \in [0, 1]$, there exists $T_2 \subset \mathcal{S}$ such that $\Pi(T_2) = k\Pi(T_1)$.

Now to obtain a full expected utility representation, we need a final axiom.

Axiom P7 (No aversion to risk). $\forall T \subset \mathcal{S}$,

- If $\forall s \in T, a \succ b(s)|T$, then $a \succ b|T$.
- If $\forall s \in T, a(s) \succ b|T$, then $a \succ b|T$.

I call this *no aversion to risk*, since it intuitively means that if you prefer a to all certain consequences of b then you prefer a to b .

Theorem 3.2.3 — Qualitative representation. Under **P1 (Preference)**, **P2 (Sure Thing principle)**, **P3 (No reduction to conditional indifference)**, **P4 (Separation)**, **P5 (No total indifference)**, **P6 (Archimedean)**, and **P7 (No aversion to risk)**, there exists a unique finitely additive probability measure satisfying the results of Theorem 3.2.2. Furthermore, there is a unique (up to positive affine transformations) bounded utility function $u : \mathcal{Z} \rightarrow \mathbb{R}$ such that

$$a \succ b \Leftrightarrow \int u(a(s))d\pi > \int u(b(s))d\pi.$$

Note that the probability measure in Theorem 3.2.3 is finitely additive, and defined on the Boolean algebra $2^{\mathcal{S}}$ of all subsets of \mathcal{S} . In particular, the integrals in Theorem 3.2.3 are *not* Lebesgue integrals, see e.g. Kreps, 1988.

3.2.2 Major criticisms

There are several points that can be raised against using Savage’s result to justify Bayesian learning, which we roughly rank by increasing seriousness. First, the resulting probability measure is only finitely additive: it is hard to bring a σ -algebra like $\Sigma_{\mathcal{S}}$ into the picture with a consensual coherence axiom that does not involve an infinite number of choices from the DM. This is a minor inconvenience, as we can always restrict actions to a set of measurable functions and choose a *bona fide* probability distribution. The price is that we might be unable to represent all coherent behaviours, and there might be interesting statistical stances to be had with finitely additive probability (**<empty citation>**).

A second objection is the boundedness of the utility function, first noticed by (Fishburn, 1970). Common losses such as the squared loss need to be trimmed to fit into the picture without modifying the axioms. At the cost of some less natural axioms, one can however accomodate unbounded loss functions (**Wak93**).

A third objection is the pivotal role played by constant actions in Savage’s (and actually von Neumann and Morgenstern’s) proof. Constant actions are those that assign the same reward $r \in \mathcal{R}$ to all states $s \in \mathcal{S}$. The utility function in approaches derived from **VoMo** is defined by finding the weight in a convex combination of two extreme constant actions; see e.g. (Parmigiani and Inoue, 2009, Chapter 3). But constant actions are not part of the action sets that we considered in Section ?? . In binary classification, for instance, there are two constant actions: the ideal classifier a^* that always predicts the right label, and the worst possible classifier a_* , which consistently predicts the wrong label. The utility function of any reward r is defined in reference to these two idealized classifiers. It would be more satisfying to have a set of axioms that does not give such a key role to *non-physical* classifiers.¹

A fourth and related objection is that the utility in Savage’s result is independent of the state: $u(a(s))$ only depends on the state through $a(s)$. This means that the notation $L(a, s) = -u(a(s))$ is inappropriate, as it lets the user think that the loss of action a can depend on the state s . In textbook ML tasks, this is not much of a problem, but in real applications, this can prove limiting; see (Parmigiani and Inoue, 2009, Chapter XXX).

Maybe the most important weakness of axiomatic constructions like Savage’s is that they do not prescribe what action to choose after some part of the state is observed. Using the conditional SEU is only a natural candidate for what to do, not more. We should not be fooled by the name “conditional expectation”: call it “strange mathematical construction” and then try to think if this is how you want to change your behaviour once you observe an event. There are ways to justify conditioning by means of sequential Dutch books (**<empty citation>**), but none that is not controversial in the general case of an infinite state space. Worse: it is doubtful that this can be done, since in practice, we all refrain from conditioning from time to time. For instance, if you observe that your classifier has poor test error, you change the family \mathcal{G} of classifiers that you consider. Unless you included the choice of \mathcal{G} in your state space, this kind of “external” move, which looks at some calibration property of a subjective model, is not constrained by Savage-like constructions. Modern attempts to build Bayesian statistics with more focus on calibration include the prequential view of **Daw**. One could also argue that PAC-Bayes goes along those lines, by abandoning conditioning in favour of better frequentist properties; more about this in Section ?? .

¹The inclusion of constant actions is also what prevents one from making u depend on s outside of the consequence of the action. One could for instance (mistakenly) argue that, for f a probability density w.r.t. π , the pair $u'(s, z) = u(z)/f(s)$ and $d\pi' = f(s)d\pi$ leads to the same expected utilities as $u(s, z) = u(z)$ and π . But this is only true for some actions: the expected utility of a constant action $s \mapsto (s_0, z_0)$ would be different according to the two pairs. Forcing the rankings to be the same will actually force $f \equiv 1$; along the lines of (Schervish, 2012, Lemma 3.141).

3.3 Because you like coherence and consensus: objective Bayes

3.4 Because you are a Waldian frequentist in disguise

3.4.1 On the consistency of Bayesian estimators

Restrict to parametric Bernstein von-Mises and its consequences.

3.4.2 Complete class theorems

3.4.3 PAC-Bayes statistical learning



Implementing Bayesian machine learning

4	Markov chain Monte Carlo	31
4.1	Basic Monte Carlo	
4.2	The Metropolis-Hastings algorithm	
4.3	Gibbs sampling	
4.4	Combining MCMC kernels	
4.5	Hamiltonian Monte Carlo	
4.6	MCMC practice: convergence diagnostics	
4.7	Alternative methods	
5	Variational Bayes	39
5.1	The evidence lower bound (ELBO)	
5.2	Mean-field inference	
5.3	Gradient-based algorithms	
5.4	Theoretical guarantees	
5.5	An alternative derivation of VB, and generalizations	
5.6	Variants of the ELBO	

4. Markov chain Monte Carlo

Maximizing expected utility requires computing integrals. Numerical integration consists in finding T nodes x_t and weights w_t , such that

$$\mathcal{E}_T(f) \triangleq \int f(x) d\mu(x) - \sum_{t=1}^T w_t f(x_t) = o_{T \rightarrow \infty}(1), \quad \forall f: \mathbb{R}^d \rightarrow \mathbb{R} \in \mathcal{C},$$

where \mathcal{C} is a large class of functions. For any smooth f , Riemann-like (i.e., grid-based) integration leads to $\mathcal{E}_T(f) \sim \sqrt{d}T^{-1/d}$, so that grids become essentially useless beyond $d = 3$. Monte Carlo methods are randomized constructions of nodes and weights. Since $\mathcal{E}_T(f)$ is then random, statements on the error shall be with high probability, in expectation, about the asymptotic fluctuations, etc.

We shall see that Monte Carlo methods (i) can accommodate settings where the target μ in (??) is only available through the evaluation of an unnormalized density $d\mu(x) = \pi(x)dx = \pi_u(x)/Zdx$, as is almost always the case in Bayesian learning; and that (ii) some Monte Carlo methods have an error that scales only polynomially with the dimension of the support of the integrand.

References.

The reference book for basic Monte Carlo methods is (Robert and Casella, 2004). For theoretical results on MCMC, we refer to (Douc, Moulines, and Stoffer, 2014, Chapters 5 to 7). For more recent methods, we shall refer to papers. For demos of MCMC samplers, see <https://chi-feng.github.io/mcmc-demo/>. Finally, we shall not cover deterministic alternatives to Monte Carlo methods in large dimension, see quasi-Monte Carlo methods (Dick and Pillichshammer, 2010).

4.1 Basic Monte Carlo

If we knew how to sample from π , we could take $x_t \sim \pi$ i.i.d., $w_t = 1/T$. Chebyshev's inequality would lead to

$$\mathbb{P}\left(\mathcal{E}_T(f) \geq \alpha \frac{\sigma(f)}{\sqrt{T}}\right) \leq \frac{1}{\alpha^2}, \quad \forall \alpha > 0,$$

as soon as $\sigma(f)^2 := \mathbb{E}_{X \sim \pi}[f(X) - \int f(x)\pi(x)dx]^2 < +\infty$.

In practice, we never have access to a sampler of π , so we choose an instrumental distribution $q(x)dx$, sample x_t i.i.d. from q . If we can evaluate π , then $w_t = \pi(x_t)/q(x_t)$ leads to an unbiased estimator called the *importance sampling*. Its error can also be controlled by the Chebyshev inequality. But more often than not, we can only evaluate an unnormalized density π_u . An alternative is to then set $w_t \propto \pi_u(x_t)/q(x_t)$ and normalize the weights so that $\sum_t w_t = 1$. This leads to the *self-normalized importance sampling estimator*

$$\hat{I}_T^{\text{NIS}} = \frac{\sum_{t=1}^T \frac{\pi_u(\theta_t)}{q(\theta_t)} f(x_t)}{\sum_{t=1}^T \frac{\pi_u(\theta_t)}{q(\theta_t)}} \rightarrow \frac{\int f(x)\pi_u(x)dx}{\int \pi_u(x)dx} = \int f(x)\pi(x)dx,$$

where the convergence is almost sure (apply the strong law of large numbers to both the numerator and the denominator).

Proposition 4.1.1 — CLT for NIS. The NIS estimator satisfies

$$\sqrt{T} \left(\hat{I}_T^{\text{NIS}} - \int f(x)\pi(x)dx \right) \rightarrow_d \mathcal{N}(0, \sigma_{\text{NIS}}^2(f))$$

where f is such that

$$\sigma_{\text{NIS}}^2(f) \triangleq TBC < \infty.$$

Proof. See exercise sheet. Hint: use the delta method. ■

Unfortunately, while NIS does accomodate unnormalized targets, it does not solve the curse of dimensionality: even when π and q are both Gaussian, only with different covariance matrices, one can show that

$$\log \sigma_{\text{NIS}}^2(f) = \Theta(d).$$

4.2 The Metropolis-Hastings algorithm

Metropolis-Hastings (MH) is the archetypal MCMC algorithm, and is still the main building block of modern MCMC algorithms. The idea is to take the nodes as the truncated history of a Markov chain (X_t) , which we build so as to guarantee that $\mathcal{E}_T(f) \rightarrow 0$. To see how to build (X_t) , remember first the law of large numbers for Markov chains.

Proposition 4.2.1 — LLN for Markov chains; see e.g. Douc, Moulines, and Stoffer, 2014. Let $(X_t)_{t \in \mathbb{N}}$ be a Markov chain with Markov kernel P . If

1. There exists μ s.t.

$$\int d\mu(x)P(x, B) = \mu(B).$$

2. For any A with $\mu(A) > 0$, for any $\theta \in \Theta$,

$$\mathbb{P}_x \left(\sum_{t=0}^{\infty} 1_{\theta_t \in A} = +\infty \right) = 1,$$

then for any initial distribution μ_0 of X_0 , almost surely

$$\frac{1}{T} \sum_{t=1}^T f(\theta_t) \rightarrow \int f d\mu,$$

where $f \in L^1(\mu)$.

The first condition states that π is an *invariant distribution* of the chain: if $X_t \sim \mu$, then $X_{t+1} \sim \mu$. The second condition is called *Harris recurrence*: starting from any x , i.e. $X_0 \sim \delta_x$, then the chain returns an infinite number of times to A almost surely, as soon as A is charged by μ . Intuitively, this makes sure that there is an infinity of nodes on A , so that the integral of f on A is accurately estimated.

The Metropolis-Hastings kernel (MTTR50; Has73) is a Markov kernel that satisfies the assumptions of Proposition 4.2.1 with a user-specified limiting distribution π , all of that using only an unnormalized density of π . The MH kernel reads

$$P_{\text{MH}}(x, x') = \alpha(x, x')q(x'|x) + \delta_x(x') \left[1 - \int \alpha(x, x')q(x'|x) \right] dx',$$

where, for each x , $q(\cdot|x)$ is a probability distribution, and

$$\alpha(x, x') = 1 \wedge \frac{\pi(x') q(x|x')}{\pi(x) q(x'|x)}. \quad (4.1)$$

Lemma 1. P_{MH} leaves π invariant.

Proof. See exercise sheet. Hint: prove first detailed balance, i.e.

$$\pi(x)P_{\text{MH}}(x, x') = \pi(x')P_{\text{MH}}(x', x), \quad x, x' \in \mathcal{X},$$

then integrate both sides. ■

Note that (4.1) is not the only way to satisfy detailed balance; see e.g. Barker's acceptance rate and Peskun's result on the optimality of the MH acceptance rate.

Now that we have shown that π is left invariant, it remains to check Harris recurrence to get the LLN. Intuitively, it is enough that the proposal distribution charges the whole space, so that there is a small chance to move anywhere in a single step. Indeed, Robert and Casella, 2004 prove that if

$$\pi(A) > 0 \Rightarrow (\forall x)q(A|x) > 0,$$

then P_{MH} satisfies the LLN.

■ **Example 4.1 — Logistic regression with pyMC.** ■

4.3 Gibbs sampling

While sampling from π is not accessible in general, it is sometimes possible to sample from the conditionals of some components of x given the rest of the components. Let $[d]$ be partitioned in I_1, \dots, I_p . The MH kernel with proposal

$$q(\theta'|\theta) = \frac{1}{p} \sum_{k=1}^p \pi(\theta'_{I_k} | \theta_{\setminus I_k}) 1_{\theta'_{I_k} = \theta_{\setminus I_k}}, \quad \theta_{\setminus I_k} := (\theta_{I_1}, \dots, \theta_{I_{k-1}}, \theta_{I_{k+1}}, \dots, \theta_{I_d}),$$

has acceptance probability 1. The algorithm that iteratively applies this MH kernel is called the (random scan) Gibbs sampler.

As an example of application, recall the LDA model of ?? 2.2. This is a joint distribution for which we can easily derive the conditionals for the natural partition given by the names of the unobserved variables. To derive them, we shall repeatedly use Bayes's theorem as well as the DAG structure. We start with a general observation: if $x_A \perp x_B | x_C$ in a DAG, then

$$p(x_A | x_B, x_C) = \frac{p(x_A, x_B, x_C)}{p(x_B, x_C)} = \frac{p(x_A, x_B | x_C) p(x_C)}{p(x_B, x_C)} = \frac{p(x_A | x_C) p(x_B | x_C) p(x_C)}{p(x_B | x_C) p(x_C)} = p(x_A | x_C). \quad (4.2)$$

The conditional of π_i

Fix a document index $i \in [N]$. Using (4.2), the conditional of π_i given the rest in LDA is simply

$$\pi_i | \pi_{\setminus i}, \mathbf{z}, \mathbf{y}, B = \pi_i | z_i.$$

Now we recognize a conjugate model, the Dirichlet-categorical (see exercise sheet), so that

$$\pi_i | z_i \sim \text{Dir} \left((\alpha_k + \sum_{\ell=1}^{L_i} 1_{z_{i\ell}=k})_{k \in [K]} \right).$$

The conditional of B_k

Fix a topic index $k \in [K]$, then by (4.2),

$$B_k | \pi, \mathbf{z}, \mathbf{y}, B_{\setminus k} \sim B_k | \mathbf{z}, \mathbf{y}.$$

Now a trained eye would recognize a Dirichlet-categorical conjugate situation again. Alternately, we can use Bayes' theorem and force terms to appear that we have specified in our model

$$\begin{aligned} p(B | \mathbf{z}, \mathbf{y}) &\propto p(B, \mathbf{z}, \mathbf{y}) \\ &= p(\mathbf{y} | B, \mathbf{z}) p(\mathbf{z} | B) p(B). \end{aligned}$$

Note that $\mathbf{z} \perp B$ in the DAG, so that $p(\mathbf{z} | B) = p(\mathbf{z})$ is independent of B . We thus obtain

$$\begin{aligned} p(B | \mathbf{z}, \mathbf{y}) &\propto \prod_{i=1}^N \prod_{\ell=1}^{L_i} \prod_{k=1}^K B_{kv}^{1_{z_{i\ell}=k} 1_{y_{i\ell}=v}} \prod_{k=1}^K 1_{\Delta_V}(B_{k:}) \prod_{v=1}^V B_{kv}^{\gamma_v-1} \\ &= \prod_{k=1}^K \left[1_{\Delta_V}(B_{k:}) \prod_{i=1}^N \prod_{\ell=1}^{L_i} B_{kv}^{1_{z_{i\ell}=k} 1_{y_{i\ell}=v}} \prod_{v=1}^V B_{kv}^{\gamma_v-1} \right]. \end{aligned}$$

The conjugacy should now be obvious, as well as the conditional independence of the rows of B . We see in particular that

$$B_k | \mathbf{z}, \mathbf{y} \sim \text{Dir} \left((\gamma_v + \sum_{i=1}^N \sum_{\ell=1}^{L_i} 1_{z_{i\ell}=k} 1_{y_{i\ell}=v})_{v \in [V]} \right).$$

The conditional of $z_{i\ell}$

Fix $i \in [N]$ and $\ell \in [L_i]$. Then by (4.2) and the DAG structure,

$$z_{i\ell} | \pi, \mathbf{z}_{\setminus i\ell}, \mathbf{y}, B \sim z_{i\ell} | \pi_i, y_{i\ell}, B.$$

Using Bayes' theorem, we can again force terms to appear that we have specified in our model.

$$\begin{aligned} p(z_{i\ell} = k | \pi_i, y_{i\ell}, B) &\propto p(z_{i\ell} = k, \pi_i, y_{i\ell}, B) \\ &= p(y_{i\ell} | z_{i\ell} = k, \pi_i, B) p(z_{i\ell} = k, B | \pi_i) p(\pi_i). \end{aligned}$$

Upon noting that $y_{i\ell} \perp \pi_i | z_{i\ell} = k, B$ and $z_{i\ell} = k \perp B | \pi_i$, and keeping only terms that contain $z_{i\ell}$, it comes

$$p(z_{i\ell} = k | \pi_i, y_{i\ell}, B) \propto p(y_{i\ell} | z_{i\ell} = k, B) p(z_{i\ell} = k | \pi_i) \propto B_{ky_{i\ell}} \pi_{ik}.$$

To sample from the conditional of $z_{i\ell}$, it is thus enough to draw a new index $k \in [K]$ with probability

$$\frac{B_{ky_{i\ell}} \pi_{ik}}{\sum_{k=1}^K B_{ky_{i\ell}} \pi_{ik}}.$$

4.4 Combining MCMC kernels

Note that if P_1 and P_2 both leave π invariant, then so does the compound kernel

$$P_1 P_2(x, x') = \int P_1(x, \xi) P_2(\xi, x') d\xi.$$

In particular, one can apply two MH kernels with different proposals one after the other, and still get a valid kernel, in the sense that it leaves the same target invariant. Another basic example is the systematic scan Gibbs sampler; see exercise sheet. Alternately, we can combine an MH and a Gibbs kernel, to use the fact that we can sample some of the components of the state conditionally on the others.

■ **Example 4.2 — Hierarchical logistic regression.** TBC ■

Another famous example of compound kernel is Hamiltonian Monte Carlo.

4.5 Hamiltonian Monte Carlo

Check out Chi Feng's MCMC gallery¹. In particular, run MH and the Gibbs sampler on the banana-shaped Gaussian, and see how difficult it is for the corresponding chains to move along the banana. Intuitively, it would be more efficient to have an MCMC kernel that somehow follows the level lines of the target π . Inspired by numerical schemes for solving differential equations in mechanics, R. M. Neal, 2000 introduced HMC, an MCMC kernel that approximately follows the level lines of an augmented target with marginal π . The HMC kernel and its adaptive variants like NUTS are now the default choice in most probabilistic programming languages, such as PyMC and Stan.

While sometimes handwavingly introduced as an instance MH, we believe that the practical popularity of HMC justifies understanding it in its own right. We closely follow Bou-Rabee and Sanz-Serna, 2018; and temporarily adopt their notational conventions: the variable over which we wish to integrate is $q \in \mathbb{R}^d$, while $x \in \mathcal{X}$ denotes a generic variable, later taken to be $x = (p, q) \in \mathbb{R}^{2d}$. Note that vanilla HMC is limited to continuous variables.

¹<https://chi-feng.github.io/mcmc-demo>

4.5.1 An abstract variant of Metropolis-Hastings

Let S be a linear involution of $\mathcal{X} \subset \mathbb{R}^{2d}$, such that $\eta \circ S = \eta$ for some (possibly unnormalized) PDF η . Let further $\Phi : \mathbb{R}^{2d} \rightarrow \mathbb{R}^{2d}$ be a C^1 -diffeomorphism² that is reversible w.r.t. to S , that is, $S \circ \Phi = \Phi^{-1} \circ S$. Now let

$$\alpha(x) \triangleq 1 \wedge \frac{\eta(\Phi(x))}{\eta(x)} |\Phi'(x)|, \quad (4.3)$$

and consider the Markov kernel

$$P_{aHMC}(x, A) = \alpha(x) 1_{\Phi(x) \in A} + (1 - \alpha(x)) 1_{S(x) \in A}.$$

Algorithmically, this “abstract” HMC kernel corresponds to accepting $\Phi(x)$ with probability $\alpha(x)$, and otherwise setting the new Markov state to $S(x)$.

Proposition 4.5.1 P_{aHMC} leaves η invariant.

Proof. Left as an exercise. ■

4.5.2 An augmented target

Consider the PDF on \mathbb{R}^{2d} defined by $\tilde{\pi}(q, p) = \frac{1}{2} \mathcal{N}(p|0, M) \times \pi(q)$. Clearly, the p -marginal is Gaussian, while the q -marginal is π . If we manage to obtain an MCMC chain for $\tilde{\pi}$, i.e. a chain with a Markov kernel that leaves $\tilde{\pi}$ invariant, then simply discarding the p -component of every realization will yield a chain that is invariant w.r.t. π . This is an example of *augmentation* of the state space: unlike for the collapsed Gibbs sampling of Section ??, we augment the dimensionality of the problem in the hope to make sampling easier. Our hope is justified here by the fact that we know how to efficiently move in \mathbb{R}^{2d} along the level lines of the augmented density $\tilde{\pi}$: this is where Hamiltonian dynamics come into play.

4.5.3 Hamiltonian dynamics

Let $H(q, p) = \log \tilde{\pi}(q, p)$, and consider the differential equation

$$\frac{d}{dt} \begin{pmatrix} x \\ p \end{pmatrix} = J \nabla H(q, p), \quad \text{where } J = \begin{pmatrix} 0_d & -I_d \\ I_d & 0_d \end{pmatrix}. \quad (4.4)$$

In particular, if ∇H is Lipschitz, which we shall always assume in this section, the Cauchy-Lipschitz theorem yields a unique solution to (4.4) passing through (q_0, p_0) at $t = 0$, which we denote by $t \mapsto \phi_t(q_0, p_0)$. We shall further assume that ϕ_t is well defined for all t , and call ϕ_t the *Hamiltonian flow*. By definition, the Hamiltonian flow preserves the Hamiltonian, since

$$\frac{d}{dt} H(\phi_t(q_0, p_0)) = \nabla H(\phi_t(q_0, p_0))^T J \nabla H(\phi_t(q_0, p_0)) = 0,$$

J being skew-symmetric. In other words, the flow ϕ_t follows level lines of H .

²While we prefer to keep Φ generic at this stage, you can intuitively think of Φ as moving along a level line of η .

4.5.4 Ideal HMC and numerical HMC

The ideal HMC is the concatenation of two kernels. Given (q_n, p_n) , first resample p from its conditional under $\tilde{\pi}$; i.e. $p' \sim \mathcal{N}(0, M)$. This obviously leaves $\tilde{\pi}$ invariant, as in Gibbs sampling. Then set $(q_{n+1}, p_{n+1}) = \phi_T(q_n, p')$. In words, one step of the corresponding Markov chain consists of sampling a random momentum variable from the corresponding conditional, and then following the Hamiltonian flow ϕ_t up to time $T > 0$. Note that, unlike Gibbs sampling, this second step changes both variables.

One can formulate the intuition that, since the second step just follows a level line of H , the ideal HMC kernel leaves $\tilde{\pi}$ invariant. This is indeed the case (Bou-Rabee and Sanz-Serna, 2018), but since ϕ_t is usually not available in closed form, the ideal HMC kernel cannot be implemented, and we will skip the proof of its invariance.

In practice, one has to approximate the Hamiltonian flow ϕ_t , and there is a large literature in numerical analysis on the subject, with integrators showcasing many interesting properties. In terms of notation, denote by h a stepsize parameter, and $n = \lfloor T/h \rfloor$, so that we think of the numerical integrator $\psi_h^n(x, p)$ as an approximation to $\phi_T(x, p)$. There exists numerical integrators ϕ_h^n that are (i) C^1 diffeomorphisms, (ii) are reversible w.r.t. to momentum flip, and are volume-preserving, i.e. $|\det(\psi_h^n)'(q, p)| = 1$. Since the momentum flip preserves $\tilde{\pi}$, we can replace the second step of the ideal HMC algorithm by the abstract HMC algorithm of Section 4.5.1, with $\eta = \tilde{\pi}$, $\Phi = \phi_h^n$ and S the momentum flip. Because the numerical integrator is volume-preserving, the acceptance probability becomes suprisingly simple, as

$$\alpha_{HMC}((q, p), (q', p')) = 1 \wedge e^{-H(q, p) - H(q', p')}.$$

Intuitively, the acceptance step compensates for the fact that we did not exactly follow the level lines of H .

One common numerical integrator satisfying all the required properties is the leapfrog (aka velocity Verlet in (Bou-Rabee and Sanz-Serna, 2018)) integrator defined as $\psi_h^n = \psi_h \circ \dots \circ \psi_h$, where $(p', q') = \psi_h(p, q)$ is defined by

$$\begin{aligned} p_{1/2} &= p + \frac{h}{2} \nabla \log \pi(q) \\ q' &= q + hM^{-1} p_{1/2} \\ p' &= p_{1/2} + \frac{h}{2} \nabla \log \pi(q'); \end{aligned}$$

see (Bou-Rabee and Sanz-Serna, 2018, Section 3) for more information, and <https://chi-feng.github.io/mcmc-demo/> for an interactive demo.

4.5.5 On the ergodicity of HMC

4.5.6 An ubiquitous variant: NUTS

NUTS for auto-tuning, etc.

4.6 MCMC practice: convergence diagnostics

Convergence diagnostics. Discuss the output of pymc.

4.7 Alternative methods

4.7.1 Other randomized quadratures

More kernels, SMC samplers, PDMPs

4.7.2 Deterministic quadrature in large dimensions

Herding QMC, randomized QMC

5. Variational Bayes

Monte Carlo methods are randomized numerical quadratures that approximate the target measure $d\mu(x) = \pi(x)dx$ by a random discrete measure with a finite number of atoms. Alternately, one might try to approximate μ by directly minimizing some distance between μ and a candidate approximation $q(x)dx$ in some parametrized subset $\mathcal{Q} \subset \mathcal{M}_1$ of non-atomic candidate approximations. Since our ultimate goal is to approximate expected utilities with respect to μ , see Section ??, one may wish to find a $q \in \mathcal{Q}$ that minimizes the worst-case error we would make by replacing π with q in the integration of some class \mathcal{F} of functions. Formally, we might wish to minimize

$$q \mapsto d_{\mathcal{F}}(\pi, q) \triangleq \sup_{f \in \mathcal{F}} \left| \int f(x) \pi(x) dx - \int f(x) q(x) dx \right| \quad (5.1)$$

When \mathcal{F} is a reproducing kernel Hilbert space and \mathcal{Q} is the set of weighted measures with N atoms, this leads to kernel herding; see Section 4.7.

In this chapter, we rather consider minimizing the KL divergence between q and π , i.e.

$$q^* \in \arg \min_{q \in \mathcal{Q}} \text{KL}(q, \pi) := \mathbb{E}_q \log \frac{q(x)}{\pi(x)} = \int q(x) \log \frac{q(x)}{\pi(x)} dx. \quad (5.2)$$

Once (5.2) has been solved, one can use the resulting q^* as a plug-in replacement for the target π . Unfortunately, the (reverse) KL divergence does not have the form (5.1), and in general, q^* will not be guaranteed to lead to a controlled integration error. Actually, the combined use of the reverse KL and of $\mathcal{Q} \subsetneq \mathcal{M}_1$ usually leads an optimal q^* with a smaller support than π . This also makes it difficult to use q^* as a proposal distribution in importance sampling.

While replacing (5.1) by (5.2) might seem disappointing at first sight, our sacrifice will not be in vain. We shall see in Sections 5.1 to 5.3 that minimizing (5.2) can be computationally tractable in settings where no efficient Monte Carlo method has been proposed yet, in particular in very high dimensions such as when fitting a deep neural network. There is research in frequentist theoretical results that support using (5.2), generalized formulations of Bayesian inference that justify (5.2), and case-specific empirical demonstrations that q^* retains some of the desirable properties of the

posterior; we survey some of these results in Section ???. Closing the gap between VB and Bayes is an exciting current research topic.

5.1 The evidence lower bound (ELBO)

Remember that the density π is often known only through the evaluation of an unnormalized version π_u , i.e., $\pi_u(\theta) = Z\pi(\theta), \forall \theta$. If we are to carry out (5.2), we thus need to know how to express $KL(q, \pi)$ using only π_u .

Lemma 2. Let $J(q) := \int q(\theta) \log \frac{q(\theta)}{\pi_u(\theta)} d\theta$. Then

$$J(q) = KL(q, \pi) - \log Z. \quad (5.3)$$

Proof. Left as an exercise. ■

Two remarks are in order. First, since Z does not depend on q , the optimization problem in (5.2) is equivalent to $\min J(q)$. Letting $L(q) = -J(q)$, (5.2) is further equivalent to $\max L(q)$. Second and the nonnegativity of the KL divergence implies that

$$L(q) \leq \log Z. \quad (5.4)$$

In Bayesian inference,

$$\pi_u(\theta) = p(y_{1:N}|\theta)p(\theta),$$

so that $Z = p(y_{1:N})$. Furthermore, (5.4) says that $L(q)$ is a lower bound for the (logarithm of the) evidence, shortened in ELBO. Most VB algorithms in the literature are cast as maximizing the ELBO.

5.2 Mean-field inference

The most common variational family is the so-called *mean-field approximation*. If you need to approximate a posterior over parameters $\theta \in \mathbb{R}^d$ and latent variables z_1, \dots, z_N , this means taking

$$\mathcal{Q} = \left\{ \theta \mapsto \prod_{d=1}^D q_d(\theta_d) \prod_{i=1}^N q_i(z_i) \right\}. \quad (5.5)$$

In other words, we approximate π with a separable PDF. Note that (5.5) only specifies the structure of the variational approximations. This is enough to derive the abstract form of the VB updates, which we do in the remainder of this section. In practice, though we further make explicit parametric choices for the individual factors, as we shall see in Section 5.2.1.

The whole motivation of the mean-field variational family is that if your target has simple conditionals, coordinate-wise optimization of the ELBO is easy. Indeed, write $x = (\theta, z) \in \mathbb{R}^p$ and let $1 \leq i \leq p$. Writing $q(x) = q_i(x_i)q_{\setminus i}(x_{\setminus i})$, and keeping track only of the additive terms that depend on q_i , it comes

$$\begin{aligned} L(q) &= \iint q_i(x_i)q_{\setminus i}(x_{\setminus i}) [\log \pi_u(x) - (\log q_i(x_i) + \log q_{\setminus i}(x_{\setminus i}))] dx_i dx_{\setminus i} \\ &\propto \int q_i(x_i) \left[\int q_{\setminus i}(x_{\setminus i}) \log \pi_u(x) dx_{\setminus i} \right] dx_i - \int q_i(x_i) \log q_i(x_i) dx_i \\ &= -KL(q_i, \phi_i), \end{aligned}$$

where

$$\phi_i(x_i) = \exp \left[\int q_{\setminus i}(x_{\setminus i}) \log \pi_u(x) dx_{\setminus i} \right] \quad (5.6)$$

is an unnormalized PDF. By a fundamental property of the KL, the ELBO L is thus maximized by setting $q_i \propto \phi_i$. The bottleneck is thus to be able to compute ϕ_i in (5.6). Like in deriving conditionals in Gibbs sampling, this is the part where conjugate distributions play a role. In practice, the choice of \mathcal{Q} is often made so that this step is easy, as we shall see in Section 5.2.1.

Finally, note that taking the variational family to be (5.5) is akin to assuming independence of all variables under the posterior. Combined with the fact that reverse KL (5.2) penalizes q^* putting a lot of mass where π does not, this often implies a gross underestimation of the support of the target (and thus of posterior uncertainty), along with the built-in ignorance of posterior correlations; see Figure ???. While separability makes algorithmic derivations easier, we thus usually rather aim for “as separable as required by computation”. In other words, if, for modeling reasons, you believe that there is correlation under π of some subset of the variables, say $x_i, i \in I$, you should try to keep these variables correlated in your variational approximation, by rather defining

$$\mathcal{Q} = \left\{ \theta \mapsto q_I(x_I) \prod_{i \notin I} q_i(x_i) \right\},$$

for some nonseparable q_I . K. Murphy, 2012, Chapter 23 calls this *structured mean-field*.

5.2.1 Mean-field VB for LDA

Recall the latent Dirichlet allocation model from Section ??, for which

$$\log p(y, z, \pi, B) \quad (5.7)$$

$$\begin{aligned} &= \sum_{i=1}^N \left[\log p(\pi_i | \alpha) + \sum_{\ell=1}^{L_i} \left(\log p(z_{i\ell} | \pi_i) + \log p(y_{i\ell} | z_{i\ell}, B) \right) \right] + p(B | \gamma) \\ &\propto \sum_{i=1}^N \left[\sum_{k=1}^K \alpha_k \log \pi_{ik} + \sum_{\ell=1}^{L_i} \left(\sum_{k=1}^K 1_{z_{i\ell}=k} \log \pi_{ik} + \sum_{v=1}^V \sum_{k=1}^K 1_{y_{i\ell}=v} 1_{z_{i\ell}=k} \log b_{kv} \right) \right] \\ &\quad + \sum_{k=1}^K \sum_{v=1}^V \gamma_v \log b_{kv}. \end{aligned} \quad (5.8)$$

We want to fit a mean-field approximation

$$\mathcal{Q} = \left\{ \prod_{i=1}^N \left[\text{Dir}(\pi_i | \tilde{\pi}_i) \prod_{\ell=1}^{L_i} \text{Cat}(z_{i\ell} | \tilde{z}_{i\ell}) \right] \prod_{k=1}^K \text{Dir}(B_k | \tilde{B}_k) \right\}.$$

Tilded variables parametrize the variational approximation q , and optimizing over q will thus be implemented as an optimization over these parameters. As we shall see, the Dirichlet distributions are chosen to make the following computations easy thanks to conjugacy.

To implement VB, we need to compute (5.6) for every coordinate, that is, we need to integrate the log joint distribution (5.8) with respect to all variables but one, for every choice of that singled out variable.

Singling out π_i .

We start by singling out $\pi_i \in \Delta_K$ for some $1 \leq i \leq N$, denoting the corresponding expectation by $\mathbb{E}_{\setminus \pi_i}$. We are confident that we shall be able to identify the functional form (and thus the normalization constant) of the resulting distribution, and thus we do not keep track of additive variable that do not imply π_i . This yields

$$\begin{aligned} \mathbb{E}_{\setminus \pi_i} \log p(y, z, \pi, B) &\propto \sum_{k=1}^K \alpha_k \log \pi_{ik} + \sum_{\ell=1}^{L_i} \mathbb{E}_{z_{i\ell}} \sum_{k=1}^K 1_{z_{i\ell}=k} \log \pi_{ik} \\ &= \sum_{k=1}^K \alpha_k \log \pi_{ik} + \sum_{\ell=1}^{L_i} \sum_{k=1}^K \tilde{z}_{i\ell k} \log \pi_{ik} \end{aligned}$$

and we recognize the log PDF of a Dirichlet distribution in π_i , with parameters

$$\tilde{\pi}_i \triangleq \left(\alpha_k + \sum_{\ell=1}^{L_i} \tilde{z}_{i\ell k} \right)_{1 \leq k \leq K}.$$

Singling out $z_{i\ell}$.

To compute $\mathbb{E}_{\setminus z_{i\ell}} \log p(y, z, \pi, B)$, we need to be able to compute expectation of log weights w.r.t. a Dirichlet distribution.

Lemma 3. Let $\Psi(\cdot) := \Gamma'(\cdot)/\Gamma(\cdot)$ be the digamma function. Let $\tilde{\eta} \in \Delta_M$ be a probability distribution over $\{1, \dots, M\}$. Then, for $m \in \{1, \dots, M\}$,

$$\mathbb{E}_{\text{Dir}(\eta|\tilde{\eta})} \log \eta_m = \Psi(\tilde{\eta}_m) - \Psi(\|\tilde{\eta}\|_1) \triangleq \Psi_m(\tilde{\eta}).$$

Proof. Left as an exercise. ■

Now we derive

$$\begin{aligned} \mathbb{E}_{\setminus z_{i\ell}} \log p(y, z, \pi, B) &\propto \sum_{k=1}^K 1_{z_{i\ell}=k} \mathbb{E}_{\pi_i} \log \pi_{ik} + \sum_{v=1}^V \sum_{k=1}^K 1_{y_{i\ell}=v} 1_{z_{i\ell}=k} \mathbb{E}_{B_{k:}} \log b_{kv} \\ &= \sum_{k=1}^K 1_{z_{i\ell}=k} \Psi_k(\tilde{\pi}_i) + \sum_{v=1}^V \sum_{k=1}^K 1_{y_{i\ell}=v} 1_{z_{i\ell}=k} \Psi_v(\tilde{B}_{k:}) \end{aligned}$$

We recognize a categorical distribution with parameters

$$\tilde{z}_{i\ell} \propto \left(\exp \left[\Psi_k(\tilde{\pi}_i) + \Psi_{y_{i\ell}}(\tilde{B}_{k:}) \right] \right)_{1 \leq k \leq K}.$$

Once again, the normalization constant can be guessed after doing the computation, since necessarily

$$\tilde{z}_{i\ell} = \left(\frac{\exp \left[\Psi_k(\tilde{\pi}_i) + \Psi_{y_{i\ell}}(\tilde{B}_{k:}) \right]}{\sum_{k=1}^K \exp \left[\Psi_k(\tilde{\pi}_i) + \Psi_{y_{i\ell}}(\tilde{B}_{k:}) \right]} \right)_{1 \leq k \leq K}.$$

Singling out $B_{k:}$.

In the same vein,

$$\mathbb{E}_{\setminus B_{k:}} \log p(y, z, \pi, B) \propto \sum_{i=1}^N \sum_{\ell=1}^{L_i} \sum_{v=1}^V 1_{y_{i\ell}=v} \mathbb{E}_{z_{i\ell}} 1_{z_{i\ell}=k} \log b_{kv} + \sum_{v=1}^V \gamma_v \log b_{kv}$$

and we recognize a Dirichlet with parameters

$$\tilde{B}_k \triangleq \left(\gamma_v + \sum_{i=1}^N \sum_{\ell=1}^{L_i} 1_{y_{i\ell}=v} \tilde{z}_{i\ell k} \right)_{1 \leq v \leq V}.$$

This concludes the derivation of VB for LDA.

5.2.2 Mean-field VB for marginal LDA

As an exercise, derive the updates for the marginalized LDA model of Section ??; see K. Murphy, 2012, Chapter 27.3 for the solution.

5.2.3 VB generalizes the EM algorithm

TBC

5.3 Gradient-based algorithms

An alternate approach to finding a simple \mathcal{Q} leading to closed-form updates is to directly run a gradient algorithm on the ELBO (5.4). Take $\mathcal{Q} = \{q(\cdot|\phi), \phi \in \Phi\}$, where for all x , $\phi \mapsto q(x|\phi)$ is differentiable. Then, assuming the necessary regularity conditions, Paisley, D. M. Blei, and M. I. Jordan, 2012 note that gradient of the ELBO can be rewritten using the so-called *score function trick* as

$$\begin{aligned} \nabla_{\phi} L(q) &= \nabla_{\phi} \mathbb{E}_{x \sim q(\cdot|\phi)} \log \frac{\pi_u(x)}{q(x|\phi)} \\ &= \int \log \pi_u(x) \nabla_{\phi} q(x|\phi) dx + \nabla_{\phi} H[q(\cdot|\phi)]. \\ &= \mathbb{E}_{x \sim q(\cdot|\phi)} [\log \pi_u(x) \nabla_{\phi} \log q(x|\phi)] + \nabla_{\phi} H[q(\cdot|\phi)]. \end{aligned}$$

The first term can be estimated by vanilla Monte Carlo. The entropy term can usually be differentiated in closed form; if not, it can be estimated by vanilla Monte Carlo as well. Overall, we can plug an unbiased estimator of the gradient of the ELBO in any stochastic gradient algorithm. More often than not, the ELBO as a function of ϕ is not convex, though, and one has to be happy with searching for local optima. Moreover, vanilla Monte Carlo estimators of (??) have been reported to have high variance even in simple models (Paisley, D. M. Blei, and M. I. Jordan, 2012).

Variance reduction for ELBO gradients has been a field of active research. Paisley, D. M. Blei, and M. I. Jordan, 2012 propose to use control variates, while D. P. Kingma and M. Welling, 2014 and a large body of follow-up work propose *reparametrization tricks* that work as follows. Assume that there exists a (deterministic) smooth and invertible function f such that $f(\epsilon, \phi) \sim q(\cdot|\phi)$ whenever $\epsilon \sim p(\epsilon)$, with ϵ easy to sample. Now rewrite

$$\nabla_{\phi} L(q) = \nabla_{\phi} \mathbb{E}_{\epsilon \sim p} \log \pi_u(f(\epsilon, \phi)) + \nabla_{\phi} H[q(\cdot|\phi)].$$

This time the gradient can be passed under the integral in the first term, without relying on the score function trick, and we obtain

$$\nabla_{\phi} L(q) = \mathbb{E}_{\epsilon \sim p} \nabla_{\phi} \log \pi_u(f(\epsilon, \phi)) + \nabla_{\phi} H[q(\cdot|\phi)].$$

As long as we can compute gradients of $\log \pi_u$, we can compute the gradient in the expectation using the chain rule. This suggests a second vanilla Monte Carlo estimator, drawing $\varepsilon_i \sim p$ i.i.d. In practice, the resulting estimator has been found to have much lower variance (Rezende, Mohamed, and Wierstra, 2014), like in variational auto-encoders (D. P. Kingma and M. Welling, 2014). I haven't seen a completely convincing explanation why and when variance reduction happens with the reparametrization trick in general, though. Finally, note that we again assumed that the entropy of q could be differentiated in closed form, but the entropy term can also be treated using the reparametrization trick if needed. We shall do so in the next section, for the sake of illustration.

5.3.1 VB for deep networks

One of the hot applications of gradient-based VB is for Bayesian deep learning, which has generated a huge literature in a short amount of time; see e.g. recent NeurIPS tutorials and workshops for pointers. For instance, Blundell et al., 2015 proceed as follows. We consider networks as generative models, so consider the softmax (classification) or squared (regression) loss. A network thus corresponds to a likelihood $p(\mathbf{y}|\mathbf{w})$. We take a prior $p(\mathbf{w})$ for the weights, and want to fit $q(\mathbf{w}|\phi)$ to the posterior $\pi(\mathbf{w}) \propto p(\mathbf{y}|\mathbf{w})p(\mathbf{w}) = \pi_u(\mathbf{w})$. The gradient of the reparametrized ELBO writes

$$\begin{aligned}\nabla_\phi L(q(\cdot|\phi)) &= \mathbb{E}_\varepsilon \nabla_\phi \log \frac{\pi_u(f(\varepsilon, \phi))}{q(f(\varepsilon, \phi)|\phi)} \\ &\approx \frac{1}{N_\varepsilon} \sum_{i=1}^{N_\varepsilon} \nabla_\phi \log \frac{\pi_u(f(\varepsilon_i, \phi))}{q(f(\varepsilon_i, \phi)|\phi)}.\end{aligned}$$

Now notice that $\log \pi_u(\mathbf{w}) = \log p(\mathbf{w}) + \sum_{i=1}^{N_y} \log p(y_i|\mathbf{w})$, so that one can further uniformly draw (with or without replacement) a minibatch of data points B , and further obtain an unbiased estimator

$$\nabla_\phi L(q(\cdot|\phi)) \approx \frac{N_y}{N_\varepsilon |B|} \sum_{i=1}^{N_\varepsilon} \sum_{y \in B} \nabla_\phi \left[\frac{1}{N_y} \log p(f(\varepsilon_i, \phi)) + \log p(y|f(\varepsilon_i, \phi)) - \frac{1}{N_y} \log q(f(\varepsilon_i, \phi)|\phi) \right].$$

Note that following Blundell et al., 2015, we do not assume that the entropy can be differentiated in closed form, but the method applies *mutatis mutandis*. Note also that it is not obvious that replacing the entropy by its closed-form would reduce the variance of the estimator.

Now the key argument is that the gradient inside the sum can be computed using the chain rule, backpropagation, and the (assumed known) gradient of q . As an example, assume $\mathbf{w} \in \mathbb{R}^d$, $\phi = (\mu, \sigma) \in \mathbb{R}^{d+1}$, so that $f(\phi, \varepsilon) = \mu + \sigma \varepsilon \sim \mathcal{N}(\mu, \sigma I_d)$. For $y \in B$, Let $F(\mathbf{w}) = \log p(y|\mathbf{w}) + \log p(\mathbf{w})$. The gradient of F is provided by backpropagation. Now the chain rule yields

$$\begin{aligned}\nabla_\phi (F(f(\varepsilon, \cdot)))(\phi_0) &= J_{f(\varepsilon, \cdot)}(\phi_0)^T \nabla F(f(\varepsilon, \phi_0)) \\ &= (I_d \quad \varepsilon)^T \nabla F(f(\varepsilon, \phi_0)).\end{aligned}$$

5.4 Theoretical guarantees

5.5 An alternative derivation of VB, and generalizations

We defined VB as minimizing $q \mapsto KL(q, \pi)$ over \mathcal{Q} . There is an equivalent definition that sheds some light on the relationship of VB with Bayes, and leads to natural generalizations of VB; see (Knoblauch, Jewson, and Damoulas, 2022) and references therein.

We consider the problem of inference for simplicity. Given a loss function ℓ , a divergence D , and a set of distributions \mathcal{Q} , consider the optimization problem $P(\ell, D, \mathcal{Q})$ defined by

$$q^* \in \arg \min_{q \in \mathcal{Q}} \mathbb{E}_q \sum_{i=1}^N \ell(\theta, x_i) + D(q(\theta) d\theta, p(\theta) d\theta). \quad (P(\ell, D, \mathcal{Q}))$$

A classical result going back at least to (Donsker and Varadhan, 1975) implies that taking $\ell(\theta, x_i) = -\log p(y_{1:N}|\theta)$, $D = \text{KL}$, and $\mathcal{Q} = \mathcal{M}_1$ leads to $q^*(\theta) = \pi(\theta) \propto p(y_{1:N}|\theta)p(\theta)$. Moreover, taking $\mathcal{Q} \subsetneq \mathcal{M}_1$ leads to VB, i.e. $q^* \in \arg \min \text{KL}(q, \pi)$. In that sense, VB corresponds to the same optimization problem as Bayes, but with a constraint on where the “posterior” may lie. Knoblauch, Jewson, and Damoulas, 2022 examine justifications behind what they call the *optimization-centric* view on Bayesian inference, and how the choices of ℓ and D in $(P(\ell, D, \mathcal{Q}))$ impacts the results of VB. In particular, when the prior has a structure that does not represent what we expect from a posterior, like when using an uncorrelated Gaussian prior on the weights of a neural network, while we expect large correlations across layers and multimodality from a posterior. Knoblauch, Jewson, and Damoulas, 2022 demonstrate that empirical performance in regression benefits from replacing the KL by Renyi’s α -divergence. Interestingly, this benefit depends on α being in a certain finite range, which allows both a weaker dependence to the prior, and the support of the resulting q^* not collapsing to a single point.

5.6 Variants of the ELBO

TBC: unbiased estimators of the marginal likelihood + Jensen leads to another bound, can approximate it by importance sampling, and even sophisticated SIS variants (Thin, Doucet, Teh).

Bayesian nonparametrics

6	Random functions: Gaussian processes	49
6.1	Introduction and definitions	
6.2	Examples of Gaussian processes	
6.3	Reproducing kernel Hilbert space	
7	Random probability measures: Dirichlet processes and the like	53
7.1	Dirichlet process	
7.2	Mixtures and model-based clustering	
7.3	Priors beyond the Dirichlet process	
8	Asymptotic frequentist properties	75
8.1	Introduction	
8.2	Posterior consistency	
8.3	Concentration rates	

6. Random functions: Gaussian processes

References.

The reference textbook on Gaussian processes is Rasmussen and Williams (2006), <https://gaussianprocess.org/gpml/chapters/RW.pdf>. See also Chapter 18 on Gaussian processes of K. P. Murphy (2023), <https://probml.github.io/pml-book/book2.html>. For a comprehensive treatment of Gaussian processes seen as priors, we refer to Chapter 11 and Appendix I of Ghosal and Van der Vaart (2017). For more recent methods, we shall refer to papers. For demos, see the scikit-learn page on Gaussian processes https://scikit-learn.org/stable/modules/gaussian_process.html. See also the integrated nested Laplace approximation (INLA) for a related method (Rue, Martino, and Chopin, 2009).

6.1 Introduction and definitions

We start with definitions and basic properties for Gaussian processes.

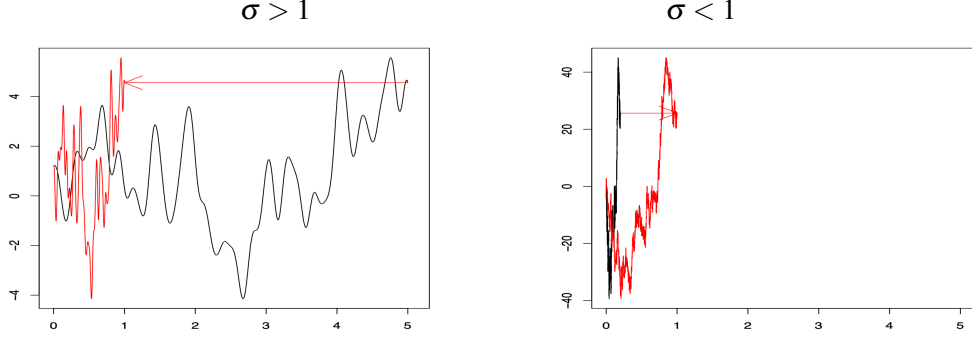
Definition 6.1.1 A Gaussian process is a stochastic process $W = (W_t : t \in T)$ indexed by an arbitrary set T such that the vector $(W_{t_1}, \dots, W_{t_k})$ has a multivariate normal distribution, for every $t_i \in T$ and $k \in \mathbb{N}$. A Gaussian process W indexed by \mathbb{R}^d is called:

- self-similar of index α if $(W_{\sigma t} : t \in \mathbb{R}^d)$ is distributed like $(\sigma^\alpha W_t : t \in \mathbb{R}^d)$, for every $\sigma > 0$, and
- stationary if $(W_{t+h} : t \in \mathbb{R}^d)$ has the same distribution of $(W_t : t \in \mathbb{R}^d)$, for every $h \in \mathbb{R}^d$.

Vectors $(W_{t_1}, \dots, W_{t_k})$ are called marginals, and their distributions marginal distributions or finite-dimensional distributions. Since a multivariate normal distribution is characterized by its mean vector and covariance matrix, the finite-dimensional distributions are determined by the mean function and covariance kernel, defined by

$$\mu(t) = \mathbb{E}(W_t), \quad K(s, t) = \text{Cov}(W_s, W_t), \quad s, t \in T.$$

If $W = (W_t : t \in \mathbb{R}^d)$ is a Gaussian process with covariance kernel K , then the process $(W_{\sigma t} : t \in \mathbb{R}^d)$ is another Gaussian process, with covariance kernel $K(\sigma s, \sigma t)$, for any $\sigma > 0$. A scaling factor $\sigma > 1$ shrinks the sample paths, whereas a factor $\sigma < 1$ stretches them.



From Ghosal and Van der Vaart (2017)

6.2 Examples of Gaussian processes

■ **Example 6.1 Random series.** If $Z_1, \dots, Z_m \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and a_1, \dots, a_m are [deterministic] functions, then the *Random series* $W_t = \sum_{i=1}^m a_i(t) Z_i$ defines a Gaussian process with $\mu(t) = 0$ and $K(s, t) = \sum_{i=1}^m a_i(t) a_i(s)$. ■

■ **Example 6.2 Brownian motion (or Wiener process).** The *Brownian motion* is the zero-mean Gaussian process, say on $[0, \infty)$, with continuous sample paths and covariance function $K(s, t) = \min(s, t)$.

Let B_t be a Brownian motion. It is:

- Stationarity: $\forall s < t, B_t - B_s \sim \mathcal{N}(0, t - s)$.
- Independent increments: $\forall s < t, (B_t - B_s) \perp\!\!\!\perp (B_u, u \leq s)$.

Thus it is a Lévy process (a stochastic process with independent, stationary increments).

- Self-similar of index $1/2$.

■ **Example 6.3 Ornstein–Uhlenbeck.** The standard *Ornstein–Uhlenbeck process* with parameter $\theta > 0$ is a mean-zero, stationary GP with time set $T = [0, \infty)$, continuous sample paths, and covariance function

$$K(s, t) = (2\theta)^{-1} \exp(-\theta|t - s|).$$

■

The standard Ornstein–Uhlenbeck process with parameter $\theta > 0$ can be constructed from a Brownian motion B through the relation

$$W_t = (2\theta)^{-1/2} \exp(-\theta t) B_{e^{2\theta t}}.$$

Mandt, Hoffman, and David M. Blei (2017) describes a relationship between [fixed learning rate] stochastic gradient descent (SGD) and Markov chain Monte Carlo (MCMC) through the Ornstein–Uhlenbeck process.

■ **Example 6.4 Square exponential.** GP with covariance function (a.k.a. radial basis function kernel)

$$K(s, t) = \exp\left(-\frac{\|t - s\|^2}{2\ell^2}\right).$$

Parameter ℓ is called the *characteristic length-scale*. ■

■ **Example 6.5 Fractional Brownian motion.** The *fractional Brownian motion* (fBm) with *Hurst parameter* $\alpha \in (0, 1)$ is the mean zero Gaussian process $W = (W_t : t \in [0, 1])$ with continuous sample paths and covariance function

$$K(s, t) = \frac{1}{2} (s^{2\alpha} + t^{2\alpha} - |t - s|^{2\alpha}).$$

- $\alpha = 2$ yields the standard Brownian motion. ■

■ **Example 6.6 Kriging.** For a given Gaussian process $W = (W_t : t \in T)$ and fixed, distinct points $t_1, \dots, t_m \in T$, the conditional expectations $W_t^* = \mathbb{E}[W_t | W_{t_1}, \dots, W_{t_m}]$ define another Gaussian process. ■

Properties of Kriging:

- If W has continuous sample paths, then so does W^* .
- In that case the process W^* converges to W when $m \rightarrow \infty$ and the interpolating points (t_1, \dots, t_m) grow dense in T .

6.3 Reproducing kernel Hilbert space

To every Gaussian process corresponds a Hilbert space, determined by its covariance kernel. This space determines the support and shape of the process, and therefore is crucial for the properties of the Gaussian process as a prior.

■ **Definition 6.3.1** A *Hilbert space* is an inner product space that is complete wrt the distance function induced by the inner product.

For a Gaussian process $W = (W_t : t \in T)$, let $\overline{\text{lin}}(W)$ be the closure of the set of all linear combinations $\sum_I \alpha_i W_{t_i}$ in the L_2 -space of square-integrable variables. The space $\overline{\text{lin}}(W)$ is a Hilbert space.

■ **Definition 6.3.2** The *reproducing kernel Hilbert space* (RKHS) of the mean-zero, Gaussian process $W = (W_t : t \in T)$ is the set \mathbb{H} of all functions $z_H : T \rightarrow \mathbb{R}$ defined by $z_H(t) = \mathbb{E}(W_t H)$, for H ranging over $\overline{\text{lin}}(W)$. The corresponding inner product is

$$\langle z_{H_1}, z_{H_2} \rangle_{\mathbb{H}} = \mathbb{E}(H_1 H_2).$$

The reproducing kernel Hilbert space has the following properties:

- Correspondance $z_H \leftrightarrow H$ is an isometry (by def of inner product), so the definition is well-posed (the correspondence is one-to-one), and H is indeed a Hilbert space.
- Function corresponding to $H = \sum_I \alpha_i W_{s_i}$ is $z_H =$
- For any $s \in T$, function $K(s, \cdot)$ is in RKHS \mathbb{H} associated with $H = W_s$.

Reproducing formula: For a general function $z_H \in \mathbb{H}$ we have

$$\langle z_H, K(s, \cdot) \rangle_{\mathbb{H}} = \mathbb{E}(H W_s) = z_H(s).$$

That is to say, for any function $h \in \mathbb{H}$,

$$h(t) = \langle h, K(t, \cdot) \rangle_{\mathbb{H}}.$$

7. Random probability measures: Dirichlet process

References

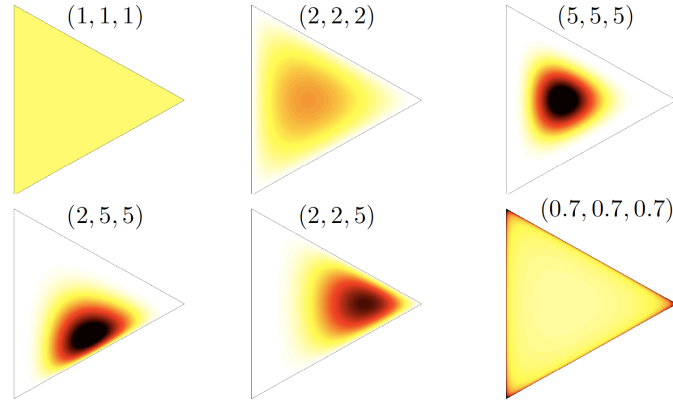
One of the first textbooks on Bayesian nonparametrics, with a focus on the Dirichlet process, is Ghosh and Ramamoorthi (2003). The book Hjort et al. (2010) is a collection of chapters contributed by renowned researchers in the field. For a treatment of discrete random structures beyond the Dirichlet process, refer to Chapter 3 of Hjort et al. (2010), Chapter 14 of Ghosal and Van der Vaart (2017) and Appendix J of Ghosal and Van der Vaart (2017) (on completely random measures).

7.1 Dirichlet process

Definition 7.1.1 Dirichlet distribution The *Dirichlet distribution* on the simplex Δ_K is a probability distribution with parameter $\alpha = (\alpha_1, \dots, \alpha_K)$ with $\alpha_j > 0$ and density function, for $x = (x_1, \dots, x_K) \in \Delta_K$,

$$f(x; \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^K x_i^{\alpha_i - 1}.$$

The Dirichlet distribution is conjugate for the multinomial distribution.



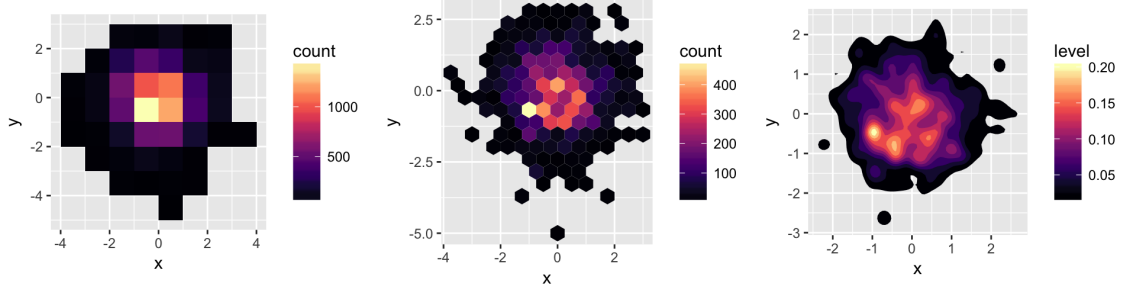
[Image by Y.W. Teh]

The Dirichlet process plays a central role as a Bayesian nonparametric prior (Ferguson, 1973).

Definition 7.1.2 Dirichlet process A Dirichlet process on the space \mathcal{Y} is a random process P such that there exist $\alpha > 0$ (precision parameter) and P_0 (base/centering distribution) such that for any finite partition $\{A_1, \dots, A_k\}$ of \mathcal{Y} , the random vector $(P(A_1), \dots, P(A_k))$ is Dirichlet distributed

$$(P(A_1), \dots, P(A_k)) \sim \text{Dir}(\alpha P_0(A_1), \dots, \alpha P_0(A_k))$$

Notation: $P \sim \text{DP}(\alpha, P_0)$



Proposition 7.1.1 Let $P \sim \text{DP}(\alpha, P_0)$ then for every measurable sets A, B we have

$$\begin{aligned} \mathbb{E}[P(A)] &= P_0(A), \\ \text{Var}[P(A)] &= \frac{P_0(A)(1 - P_0(A))}{1 + \alpha}, \\ \text{Cov}(P(A), P(B)) &= \frac{P_0(A \cap B) - P_0(A)P_0(B)}{1 + \alpha}. \end{aligned}$$

Proof. We will make use of $p(A) \sim \text{Beta}(\alpha P_0(A), \alpha(1 - P_0(A)))$. From this we obtain

$$\mathbb{E}(p(A)) = \frac{\alpha P_0(A)}{\alpha P_0(A) + 1 - P_0(A)} = P_0(A)$$

and

$$\text{Var}(p(A)) = \frac{\alpha^2 P_0(A)(1 - P_0(A))}{\alpha^2(\alpha + 1)}.$$

We derive the covariance term in two cases, firstly taking into consideration the one with $A \cap B = \emptyset$. In that case any space Ω may be decomposed into three sets:

$$\Omega = \{A, B, (A \cup B)^c\}.$$

Using de Morgan's law the last can be written as $(A \cup B)^c = A^c \cap B^c =: C$. Therefore we may write a joint probability vector

$$(P(A), P(B), P(A^c \cap B^c)) \sim \text{Dir}(\alpha P_0(A), \alpha P_0(B), \alpha P_0(C))$$

and hence $\text{Cov}(P(A), P(B)) = -P_0(A)P_0(B)/(1 + \alpha)$. In the more general case one may decompose

$$\begin{aligned} A &= (A \cap B) \cup (A \cap B^c) \\ B &= (B \cap A) \cup (B \cap A^c), \end{aligned}$$

so that

$$\text{Cov}(P(A), P(B)) = \text{Cov}(P(A \cap B) + P(A \cap B^c), P(B \cap A) + P(B \cap A^c))$$

and so forth using the linearity of covariance. ■

Marginalizing out the DP Property $\mathbb{E}[P(A)] = P_0(A)$ can be written equivalently as

$$\mathbb{E}(P(A)) = P_0(A) = \int P(A) d\text{DP}(P).$$

A Dirichlet process model can be constructed as a two level sampling model:

$$\begin{cases} P \sim \text{DP}(\alpha, P_0) \\ X|P \sim P, \end{cases}$$

i.e. we sample a probability measure P from the Dirichlet process and then given P , we sample random variables X_i .

Marginalizing out P , we obtain the marginal distribution of X :

$$X \sim P_0.$$

Posterior distribution Let $X_{1:n} := (X_1, \dots, X_n)$ be sampled from the hierarchical model

$$\begin{cases} P \sim \text{DP}(\alpha, P_0) \\ X_{1:n}|P \stackrel{\text{iid}}{\sim} P. \end{cases}$$

This model is usually used as a building block in a larger hierarchical model, e.g. mixture models, graphs, etc.

Theorem 7.1.2 — DP posterior distribution. The DP is conjugate, with posterior equal to

$$P|X_{1:n} \sim \text{DP}(\alpha P_0 + \sum_{i=1}^n \delta_{X_i}).$$

The predictive distribution, called Pólya urn or Blackwell–MacQueen scheme, is given by

$$P(X_{n+1}|X_{1:n}) = \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{i=1}^n \delta_{X_i}.$$

Proof. The posterior distribution of $a = (a_1, \dots, a_k) = (P(A_1), \dots, P(A_k))$ depends on the observations only via their cell counts $N = (N_1, \dots, N_k)$, $N_j = \#\{i : X_i \in A_j\}$ (it comes from *tail-free* property), so

$$a|X_{1:n} \sim a|N_{1:k}.$$

The prior and model are

$$\begin{cases} a \sim \text{Dir}_k(\alpha P_0(A_1), \dots, \alpha P_0(A_k)) \\ N|P \sim \text{Multinom}_k(a). \end{cases}$$

This results in the posterior of form

$$\begin{aligned} p(a|N) &\propto a_1^{\alpha P_0(A_1) + N_1 - 1} \dots a_k^{\alpha P_0(A_k) + N_k - 1} \\ &= \text{Dir}_k(\alpha P_0(A_1) + N_1, \dots, \alpha P_0(A_k) + N_k). \end{aligned}$$

■

Combinatorial properties: Number of distinct values

Assume that the base measure P_0 is non-atomic. Then with probability 1:

$$X_i \notin \{X_1, \dots, X_{i-1}\} \Leftrightarrow X_i \sim P_0.$$

Let $D_i = \mathbb{I}(X_i \text{ is a new value})$ and let's denote $K_n = \sum_{i=1}^n D_i$, a number of distinct values X_1, \dots, X_n with distribution $\mathcal{L}(K_n)$.

Proposition 7.1.3 — Asymptotics for K_n . Random variables D_i are distributed i.i.d. with respect to *Bernoulli*($\alpha/(\alpha + i - 1)$). Therefore for fixed α and for $n \rightarrow \infty$ we have:

- i) $\mathbb{E}K_n \sim \alpha \log n \sim \text{Var}(K_n)$
- ii) $K_n / \log(n) \xrightarrow{a.s.} \alpha$
- iii) $(K_n - \mathbb{E}K_n) / \text{sd}(K_n) \rightarrow N(0, 1)$
- iv) $d_{\text{TV}}(\mathcal{L}(K_n), \text{Poisson}(\mathbb{E}K_n)) = o(1/\log(n))$ where

$$d_{\text{TV}}(P, Q) = \sup |P(A) - Q(A)|$$

over measurable partition A

Proof. i) $\mathbb{E}K_n = \sum_{i=1}^n \frac{\alpha}{\alpha + i - 1}$ and $\text{Var}(K_n) = \sum_{i=1}^n \frac{\alpha(i-1)}{(\alpha + i - 1)^2}$.

ii) Since D_i 's are \mathbb{I} one may use Kolmogorov law of strong numbers and

$$\sum_{i=1}^{\infty} \frac{\text{Var}(D_i)}{(\log i)^2} = \sum_{i=1}^{\infty} \frac{\alpha(i-1)}{(\alpha + i - 1)^2 (\log i)^2} < \infty$$

by e.g. the fact that $\sum_i (1/i(\log i)^2)$ converges.

iii) By Lindeberg central limit theorem.

iv) This is implied from Chein–Stein approximation.

■

A central limit theorem for independent random variables (possibly not identically distributed).

Theorem 7.1.4 — Lindeberg central limit theorem. Suppose X_i are i.i.d. such that $\mathbb{E}X_i = \mu_i$ and $\text{Var}X_i = \sigma_i^2 < \infty$. Define $Y_i = X_i - \mu_i$, $T_n = \sum_{i=1}^n Y_i$, $s_n^2 = \text{Var}(T_n) = \sum_{i=1}^n \sigma_i^2$. Then provided that

$$\forall \varepsilon > 0 \quad \frac{1}{s_n^2} \sum_{i=1}^n \mathbb{E}(Y_i^2 \mathbb{I}(|Y_i| > \varepsilon s_n)) \xrightarrow{n \rightarrow \infty} 0 [\text{Lindeberg condition}],$$

we have the central limit theorem: $T_n/s_n \xrightarrow{d} N(0, 1)$.

Combinatorial properties: Distribution of distinct values We have now the limits of K_n and we know its approximate distribution $\mathcal{L}(K_n)$. The exact distribution of K_n is:

Proposition 7.1.5 — Distribution of K_n . If P_0 is non-atomic then

$$P(K_n = k) = \mathfrak{C}_n(k) n! \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \quad (7.1)$$

where

$$\mathfrak{C}_n(k) = \frac{1}{n!} \sum_{S \in \mathfrak{J}_n(k)} \prod_{j \in S} j \quad (7.2)$$

and $\mathfrak{J}_n(k) = \{S \subset \{1, \dots, n-1\}, |S| = n-k\}$.

Recall the definition of the Gamma function $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$.

Let us consider when we may deal with events $K_n = k$: we have two cases

$$\begin{cases} K_{n-1} = k-1 \text{ and } X_n \text{ is a new value} \\ K_{n-1} = k \text{ and } X_n \text{ is not a new value.} \end{cases}$$

This results in

$$p_n(k, \alpha) := P(K_n = k | \alpha) = \frac{\alpha}{\alpha + n - 1} p_{n-1}(k-1, \alpha) + \frac{n-1}{\alpha + n - 1} p_{n-1}(k, \alpha). \quad (7.3)$$

Now let us remark that $\mathfrak{C}_n(k) = p_n(k, \alpha = 1)$. Therefore

$$\mathfrak{C}_n(k) = \frac{1}{n} \mathfrak{C}_{n-1}(k-1) + \frac{n-1}{n} \mathfrak{C}_{n-1}(k). \quad (7.4)$$

By induction over n : first we check case $n = 1$:

$$p_1(1, \alpha) = \mathfrak{C}_1(1) \frac{\alpha}{\alpha} = \mathfrak{C}_1(1).$$

To check case $n > 1$ we use (7.1) and then (7.3):

$$\begin{aligned} p_n(k, \alpha) &= \frac{\alpha}{\alpha + n - 1} p_{n-1}(k-1, \alpha) + \frac{n-1}{\alpha + n - 1} p_{n-1}(k, \alpha) \\ &= \frac{\alpha}{\alpha + n - 1} \mathfrak{C}_{n-1}(k-1) (n-1)! \alpha^{k-1} \frac{\Gamma(\alpha)}{\Gamma(\alpha + n - 1)} + \\ &\quad + \frac{n-1}{\alpha + n - 1} \mathfrak{C}_{n-1}(k) (n-1)! \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n - 1)} \\ &= \frac{\alpha^k}{\alpha + n - 1} (n-1)! \frac{\Gamma(\alpha)}{\Gamma(\alpha + n - 1)} n \left(\frac{1}{n} \mathfrak{C}_{n-1}(k-1) + \frac{n-1}{n} \mathfrak{C}_{n-1}(k) \right) \\ &= \mathfrak{C}_n(k) n! \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \end{aligned}$$

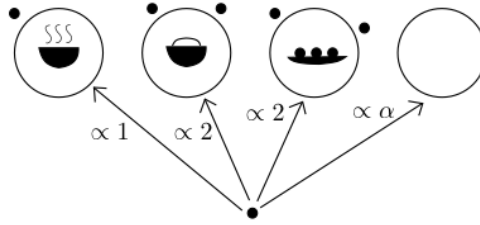
which proves property (7.1).

To prove (7.2) let us define a polynomial $A_n(s)$ as $A_n(s) = \sum_{k=1}^{\infty} \mathfrak{C}_n(k) s^k$. Then using (7.4) polynomial $A_n(s)$ can be written as

$$\begin{aligned} A_n(s) &= \sum_{k=1}^{\infty} \left(\frac{1}{n} \mathfrak{C}_{n-1}(k-1) + \frac{n-1}{n} \mathfrak{C}_{n-1}(k) \right) s^k \\ &= \frac{1}{n} (s A_{n-1}(s) + (n-1) A_{n-1}(s)) = \frac{s+n-1}{n} A_{n-1}(s) \\ &= \dots = A_1(s) \prod_{j=2}^n \frac{s+j-1}{j} = \frac{s(s+1) \dots (s+n-1)}{n!}. \end{aligned}$$

Last equality implies from the fact that $\mathfrak{C}_1(k) = \mathbf{1}\{k=1\}$ and hence $A_1(s) = s$. Checking terms after the expansion finishes the proof of (7.2).

Combinatorial properties: Chinese Restaurant process A culinary metaphor of the random partition induced by the DP. Customers join tables with probability proportional to n_j , the number of clients already sitting, or sit at new table with probability proportional to α .



Proposition 7.1.6 — Chinese Restaurant process. A random sample $X_{1:n}$ from a DP with precision parameter α induces a partition of $\{1, \dots, n\}$ into k sets of sizes n_1, \dots, n_k with probability

$$p(n_1, \dots, n_k) = p(\{n_1, \dots, n_k\}) = \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \prod_{j=1}^k \Gamma(n_j).$$

Proof. We use the Pólya urn scheme slightly changed by using n_1, \dots, n_k

$$P(X_{n+1}|X_{1:n}) = \frac{\alpha}{\alpha+n} P_0 + \frac{1}{\alpha+n} \sum_{j=1}^k n_j \delta_{X_j^*}.$$

By exchangeability, the distribution of $\{n_1, \dots, n_k\}$ does not depend on the order of the observations. Let's compute $p(n_1, \dots, p_k)$ as the probability of one draw where the first table consists of first n_1 observations etc.

To proceed, let us use Pólya urn scheme: we denote $\bar{n}_j = \sum_{i=1}^j n_i$ and hence $\bar{n}_k = n$, the total number of observations. We can observe the following pattern: first ball open new table, following $n_j - 1$ ones fill in that table and so forth. That quantity can be rewritten as

$$\frac{\alpha^k}{\alpha(\alpha+1) \dots (\alpha+n-1)} \prod_{j=1}^k (n_j - 1)!,$$

where one can rewrite both terms using Gamma function $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$: the first term can be written as

$$\frac{\alpha^k}{\alpha(\alpha+1) \dots (\alpha+n-1)} = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)},$$

while the second one as $(n_j - 1)! = \Gamma(n_j)$.

One should remark that for ordered partitions we have

$$\bar{p}(n_1, \dots, n_k) = \frac{p(n_1, \dots, n_k)}{k!}.$$

■

Combinatorial properties: Ewens sampling formula

Ewens sampling formula (ESF), presented originally by Ewens (1972), is the distribution of multiplicities $m = (m_1, \dots, m_n)$, m_ℓ is the number of groups of size ℓ . Also known as allelic partitions in population genetics, when there is no selective difference between types: null hypothesis in non Darwinian theory. See also Antoniak (1974).

Proposition 7.1.7 — Ewens sampling formula. The distribution of the multiplicities (m_1, \dots, m_n) induced by a DP is

$$p(m_1, \dots, m_n) = \frac{\alpha^k}{\alpha_{(n)}} \frac{n!}{\prod_{\ell=1}^n \ell^{m_\ell} m_\ell!}.$$

Notation $n_{(k)} := n(n-1) \cdots (n-k+1)$.

Proof. Two steps: 1) Compute probability of particular sequence of X_1, \dots, X_n in given class (m_1, \dots, m_n) , note that all such sequences are equally likely and 2) multiply obtained quantity by the number of such sequences.

- 1) Consider a sequence X_1, \dots, X_n such that X_1, \dots, X_{m_1} occur each only once, then the next m_2 occur only twice and so on. This sequence has probability which may be obtained by the Pólya Urn scheme in the same fashion as CRP:

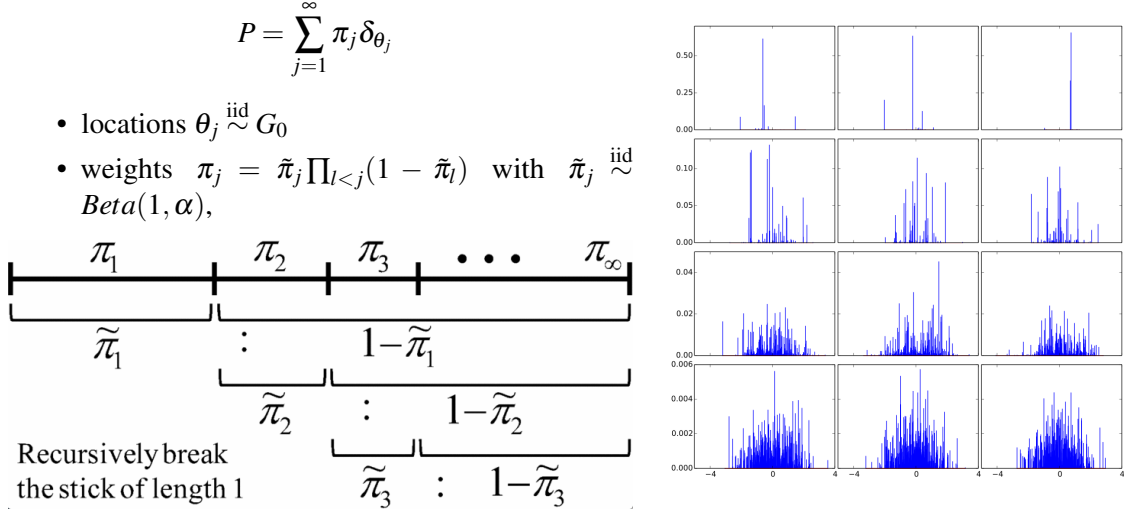
$$\frac{\alpha^{m_1} (\alpha \cdot 1)^{m_2} \cdots (\alpha \cdot 1 \cdots (n-1))^{m_n}}{\alpha_{(n)}} = \frac{\alpha^k}{\alpha_{(n)}} \prod_{\ell=1}^n ((\ell-1)!)^{m_\ell}.$$

- 2) Number of sequences X_1, \dots, X_n with frequencies (m_1, \dots, m_n) is a number of ways of putting n distinct objects into bins, so called multinomial coefficient. Since ordering of the m_ℓ bins of frequency ℓ is irrelevant, divide by $m_\ell!$:

$$\frac{1}{\prod_{\ell=1}^n (m_\ell)!} \binom{n}{1 \times \#m_1, 2 \times \#m_2, \dots, n \times \#m_n} = \frac{n!}{\prod_{\ell=1}^n m_\ell! (\ell!)^{m_\ell}}$$

To finish one needs to multiply results obtained in 1) and 2). ■

Stick-breaking representation The DP has almost surely discrete realizations (Sethuraman, 1994)



Stick-breaking representation

A constructive representation of the DP due to Sethuraman (1994).

Theorem 7.1.8 — Stick-breaking. If $V_1, V_2, \dots \stackrel{\text{iid}}{\sim} \text{Be}(1, \alpha)$ and $\phi_1, \phi_2, \dots \stackrel{\text{iid}}{\sim} P_0$ are i.i.d. variables, then define $p_1 = V_1$ and

$$p_j = V_j \prod_{1 \leq l < j} (1 - V_l)$$

then

$$P = \sum_{i=1}^{\infty} p_i \delta_{\phi_i} \sim \text{DP}(\alpha, P_0).$$

Lemma 4. For independent $\phi \sim P_0$ and $V \sim \text{Be}(1, \alpha)$ the DP is the only solution of the distributional equation

$$P \sim V \delta_{\phi} + (1 - V)P,$$

where $P \sim \text{DP}(\alpha, P_0)$.

Proof. 1) The weights (p_1, p_2, \dots) need to form a probability vector. The leftover mass at stage j is

$$1 - \left(\sum_{i=1}^j p_i \right) = \prod_{i=1}^j (1 - V_i) =: R_j.$$

One may notice that R_j is decreasing and for every j we have $R_j \in [0, 1]$, hence we obtain almost sure convergence which is equivalent with convergence in mean. Therefore

$$\mathbb{E} R_j = \mathbb{E} \prod_{i=1}^j (1 - V_i) = \prod_{i=1}^j \mathbb{E}(1 - V_i) = \left(\frac{\alpha}{\alpha + 1} \right)^j \rightarrow 0.$$

So (p_1, \dots) is a probability vector almost surely and P is a probability measure almost surely.

2) Now one may write

$$P = p_1 \delta_{\phi_1} + \sum_{j=2}^{\infty} p_j \delta_{\phi_j} = V_1 \delta_{\phi_1} + (1 - V_1) \sum_{j=1}^{\infty} \tilde{p}_j \delta_{\tilde{\phi}_j},$$

where $\tilde{p}_j = \frac{p_{j+1}}{1-V_1} = V_{j+1} \prod_{l=2}^j (1 - V_l)$ and $\tilde{\phi}_j = \phi_{j+1}$, then (\tilde{p}_j) and $(\tilde{\phi}_j)$ satisfy the same distributional definitions as (p_j) and (ϕ_j) , hence $\tilde{P} \sim P$ and so P is solution of the Lemma equation (??) whose only solution is the DP. ■

DP as a normalized Gamma process The DP can be obtained by normalizing a Gamma process. It is a generic way to obtain random probability measures from almost surely finite random measures. Let us restrict to $\mathcal{Y} = \mathbb{R}$.

Definition 7.1.3 Gamma process on \mathbb{R}_+ is a process $(S(u) : u \geq 0)$ with independent increments satisfying

$$\forall u_1 : 0 \leq u_1 \leq u_2 : \quad S(u_2) - S(u_1) \stackrel{\text{ind}}{\sim} Ga(u_2 - u_1, 1).$$

This ensures that the process has non-decreasing right continuous sample path $u \mapsto S(u)$.

Theorem 7.1.9 For every $\alpha > 0$ and for every cumulative distribution function G , a random cumulative distribution function such that

$$F(t) = \frac{S(\alpha G(t))}{S(\alpha)}$$

is the distribution of a $DP(\alpha, G)$.

Proof. For any set of t_i satisfying $-\infty = t_0 < t_1 < \dots < t_k = \infty$ we have

$$S(\alpha G(t_i)) - S(\alpha G(t_{i-1})) \sim Ga(\alpha G(t_i) - \alpha G(t_{i-1}), 1).$$

Use property that if $Y_i \stackrel{\text{ind}}{\sim} Ga(\alpha_i, 1)$ then $(Y_1, \dots, Y_n) / \sum_i Y_i \sim \text{Dir}_n(\alpha_1, \dots, \alpha_n)$ to obtain

$$(F(t_1) - F(t_0), \dots, F(t_k) - F(t_{k-1})) \sim \text{Dir}_k(\alpha G(t_1) - \alpha G(t_0), \dots, \alpha G(t_k) - \alpha G(t_{k-1})).$$

Hence the definition of DP holds for every partition in intervals. These form a measure determining class, so that the definition holds for every partition in general. ■

Definition via the Pólya urn scheme A Pólya sequence with parameter αP_0 is a sequence of random variables X_1, \dots, X_n whose joint distribution satisfies

$$X_1 \sim P_0, \quad X_{n+1} | X_1, \dots, X_n \sim \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{i=1}^n \delta_{X_i}.$$

Theorem 7.1.10 If X_1, X_2, \dots is a Pólya sequence then exists random probability measure P such that $X_i | P \stackrel{\text{iid}}{\sim} P$ and $P \sim DP(\alpha, P_0)$.

Proof. We can consider Pólya sequence as an outcome of Pólya urn, we see that it is exchangeable. By de Finetti theorem exists such probability measure P such that $X_i | P \stackrel{\text{iid}}{\sim} P$. So far we have proved existence of the DP and know that DP generates a Pólya sequence. Since the RPM given by de Finetti's theorem is unique this proves that $P \sim DP(\alpha, P_0)$. ■

7.2 Mixtures and model-based clustering

A parametric approach Mixture model with K components

$$G = \sum_{k=1}^K \pi_k \delta_{\phi_k}$$

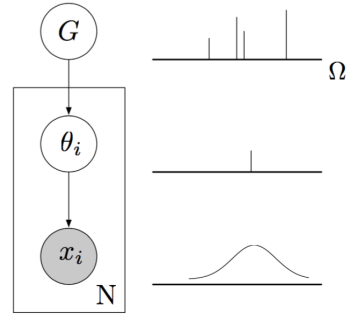
δ_{ϕ_k} is a point mass at ϕ_k .

G is to be understood as a K -faceted dice. The mixture density is:

$$p(X|\pi, \phi) = \sum_{k=1}^K \pi_k p(x|\phi_k)$$

Then

$$\begin{aligned} \theta_i &\sim G \\ x_i &\sim p(x|\theta_i) \end{aligned}$$

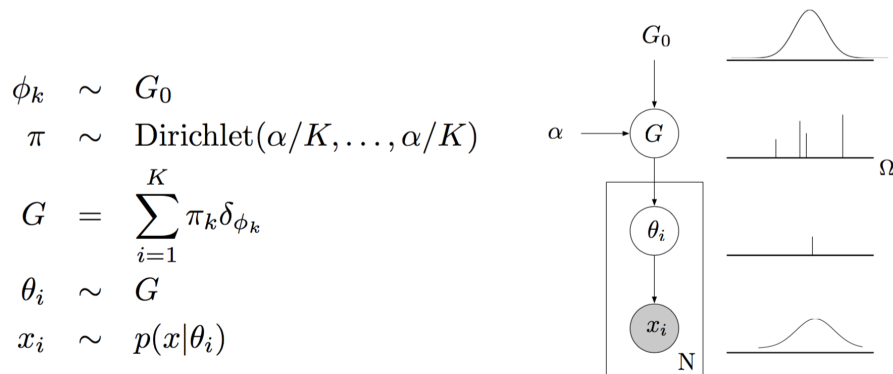


A Bayesian parametric approach Bayesian mixture models with K components

We need a distribution over the probability measure (aka dice) G , that is a distribution over weights or classes $\pi = (\pi_1, \dots, \pi_K)$ and over mean and covariance (for 2-dimensional data) $\phi_k = (\mu_k, \Sigma_k)$

- $\pi \sim \text{Dirichlet}(\alpha/K, \dots, \alpha/K)$
- $(\mu_k, \Sigma_k) \sim \text{Normal} \times \text{Inverse-Wishart}$

This makes $G = \sum_{k=1}^K \pi_k \delta_{\phi_k}$ a random dice



$$\begin{aligned} \phi_k &\sim G_0 \\ \pi &\sim \text{Dirichlet}(\alpha/K, \dots, \alpha/K) \\ G &= \sum_{i=1}^K \pi_k \delta_{\phi_k} \\ \theta_i &\sim G \\ x_i &\sim p(x|\theta_i) \end{aligned}$$

Choosing K There are several options for choosing K

- Model selection with information criteria: AIC, BIC, or cross-validation, etc

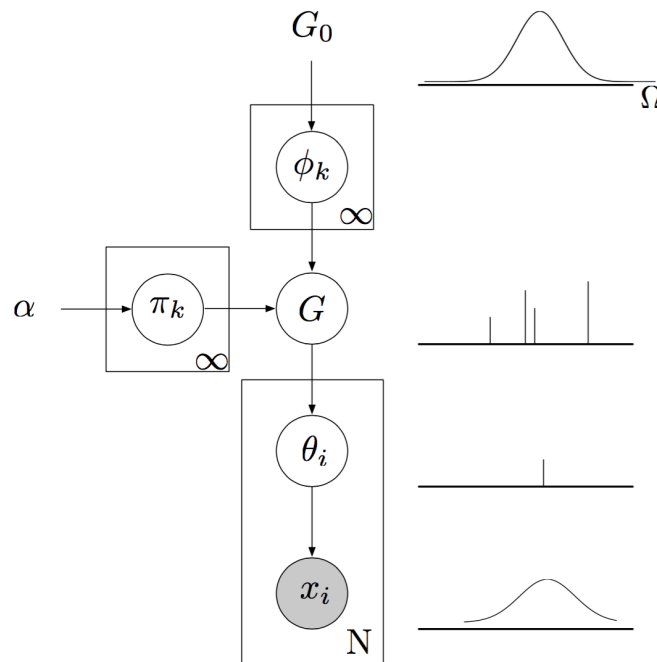
- Hierarchical model, with a prior on K
- Be nonparametric, and let K get large... possibly infinite.

A Bayesian nonparametric approach Bayesian nonparametric mixture models

We now move to G being an infinite sum $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$.

We need a distribution over this infinite dice G , that is exactly what the Dirichlet process does. It is parameterized by the precision parameter α and the base measure G_0 .

- $\pi = (\pi_1, \pi_2, \dots) \sim \text{GEM}(\alpha)$
- $\phi_k \sim G_0$



Posterior sampling Markov chain Monte Carlo (MCMC) methods:

- Marginal methods: marginalizing over the posterior DP P , and sampling using the posterior Pólya urn scheme (easy in conjugate case, see Radford M Neal, 2000)
- Conditional methods: sampling a finite but sufficient number of parameters (Ishwaran and James, 2001). BNPdensity R package (Arbel et al., 2021).

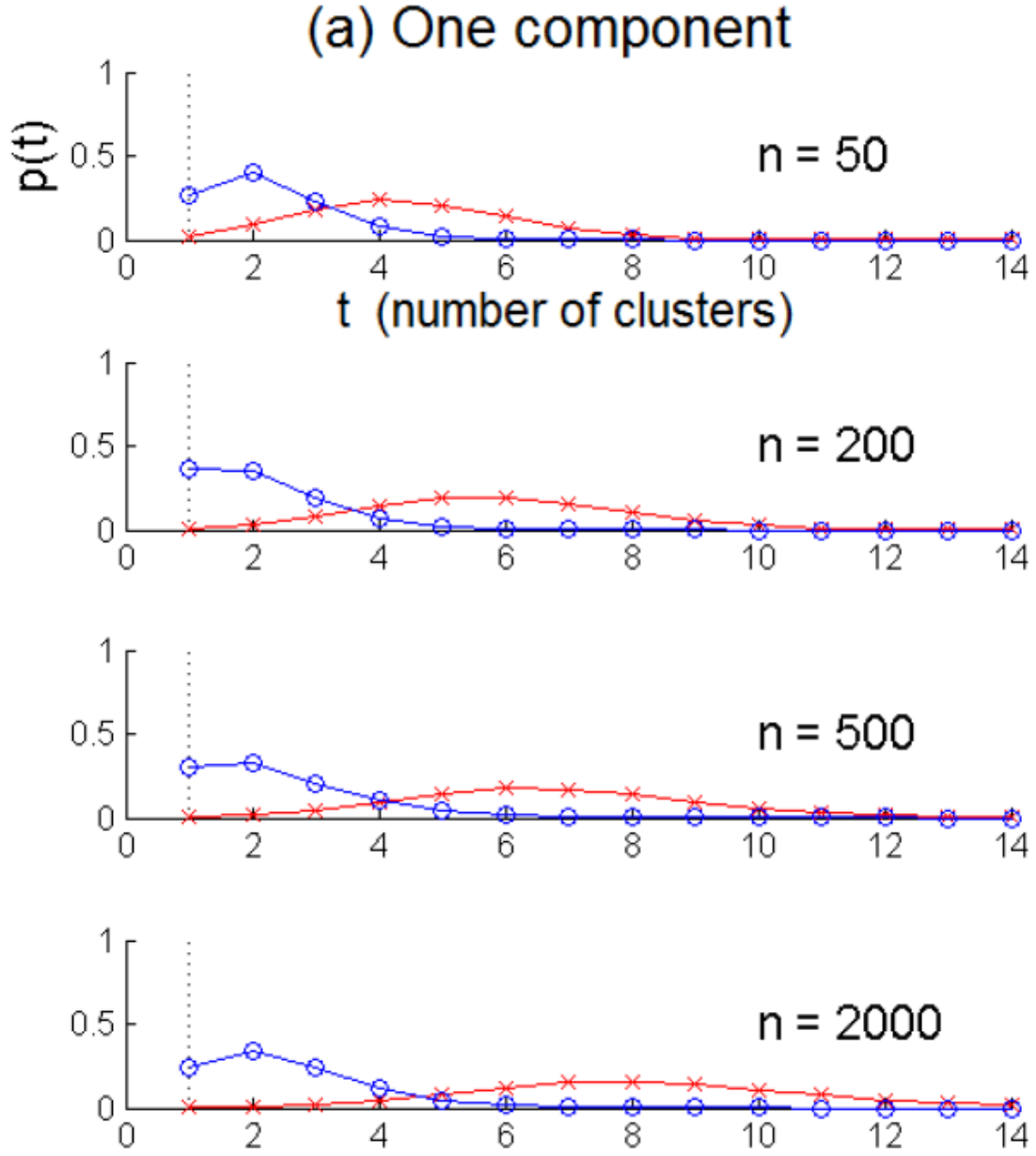
Variational approximations (David M Blei, Michael I Jordan, et al., 2006)

Warning on interpretation of K_n

Consider a simple DP mixture model with

- Gaussian base measure,
- Gaussian kernel,
- where data are sampled iid from some distribution.

Then the posterior on K_n is inconsistent (Miller and Harrison, 2013).



From Miller and Harrison (2013) (here K_n is denoted T_n):

Theorem 4.1. *If $X_1, X_2, \dots \in \mathbb{R}$ are i.i.d. from any distribution with $\mathbb{E}|X_i| < \infty$, then with probability 1, under the standard normal DPM with $\alpha = 1$ as defined above, $p(T_n = 1 \mid X_{1:n})$ does not converge to 1 as $n \rightarrow \infty$.*

Theorem 5.1. *If $X_1, X_2, \dots \sim \mathcal{N}(0, 1)$ i.i.d. then*

$$p(T_n = 1 \mid X_{1:n}) \xrightarrow{\text{Pr}} 0 \quad \text{as } n \rightarrow \infty$$

under the standard normal DPM with concentration parameter $\alpha = 1$.

But there is some hope...

Bayesian decision theory From decision theory: a Bayes estimator minimizes a posterior expected loss.

$$\hat{a}_L = \arg \inf_{a \in A} \mathbb{E}_{\pi(\theta)}[L_a(\theta)].$$

Examples with Euclidean parameter spaces:

- L^2 , squared loss \rightarrow posterior mean
- L^1 , absolute loss \rightarrow posterior median
- 0 – 1 loss \rightarrow mode a posteriori (MAP)

Deriving an optimal clustering

The posterior expected loss of clustering c' , denoted by $L(c')$, is obtained by averaging the loss with respect to posterior weight

$$L(c') = \sum_{c \in \mathcal{A}_n} L(c, c') p(c|x),$$

and the decision is taken by choosing the best

$$\hat{c} = \arg \min_{c' \in \mathcal{A}_n} \sum_{c \in \mathcal{A}_n} L(c, c') p(c|x)$$

Several losses have been considered:

- 0-1 loss (Rajkowski, 2019),
- Binder loss (Dahl, 2006),
- Variation of information (Wade and Ghahramani, 2018).

Simplest loss: L_{0-1}

$$\begin{aligned} L_{0-1}(c') &= \sum_{c \in \mathcal{A}_n} L_{0-1}(c, c') p(c|x) = \sum_{c \in \mathcal{A}_n, c \neq c'} p(c|x), \\ &= 1 - p(c'|x) \end{aligned}$$

which is to say that the expected loss of c' is all the posterior mass except that of c' . So that it is easily minimized at the value c' which has maximum posterior weight:

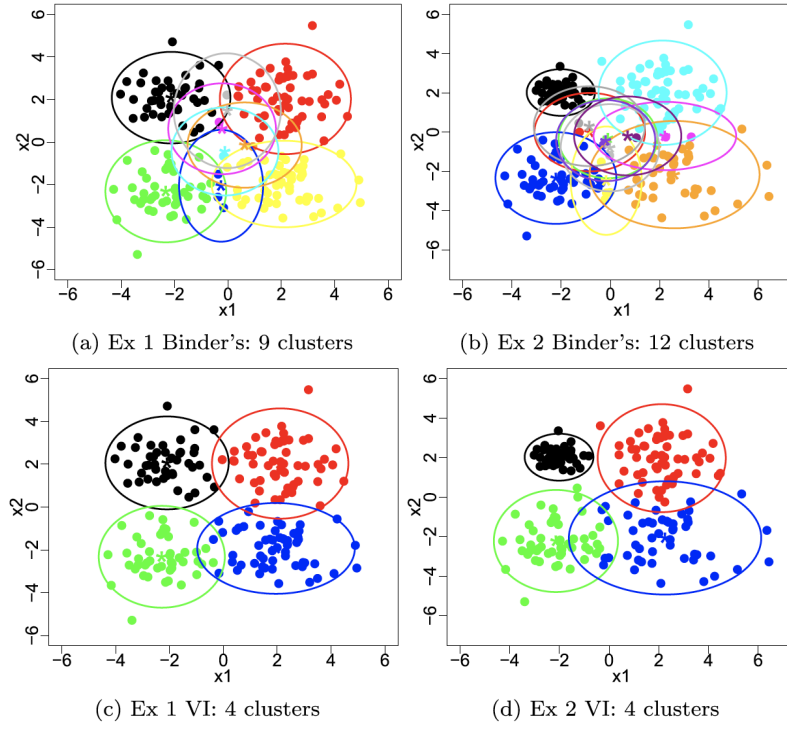
$$\hat{c} = \arg \min_{c' \in \mathcal{A}_n} L_{0-1}(c') = \arg \max_{c' \in \mathcal{A}_n} p(c'|x) := MAP.$$

Negative results by Rajkowski (2019) show that the mode a posteriori (MAP) is inconsistent.

Variation of information Variation of information (VI) by Meilă (2007) for cluster comparison. From information theory, compares information in two clusterings with information shared between the two clusterings:

$$VI(c, \hat{c}) = H(c) + H(\hat{c}) - 2\mathcal{I}(c, \hat{c})$$

Variation of information Wade and Ghahramani (2018) compare Binder and VI:



Variation of information Wade and Ghahramani (2018) also provide credible balls around the estimated clustering, based on Hasse diagram:

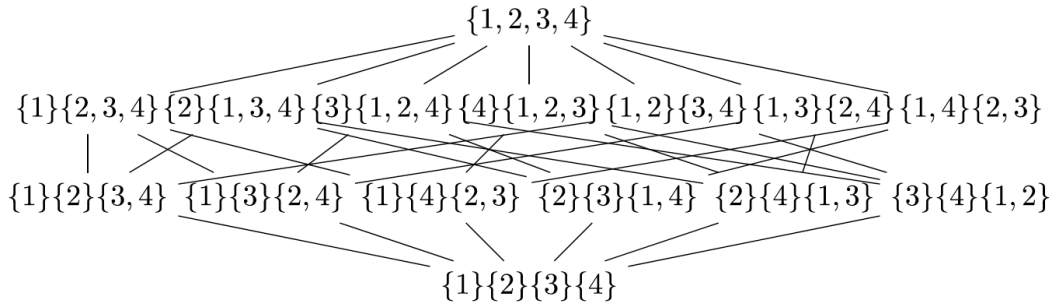
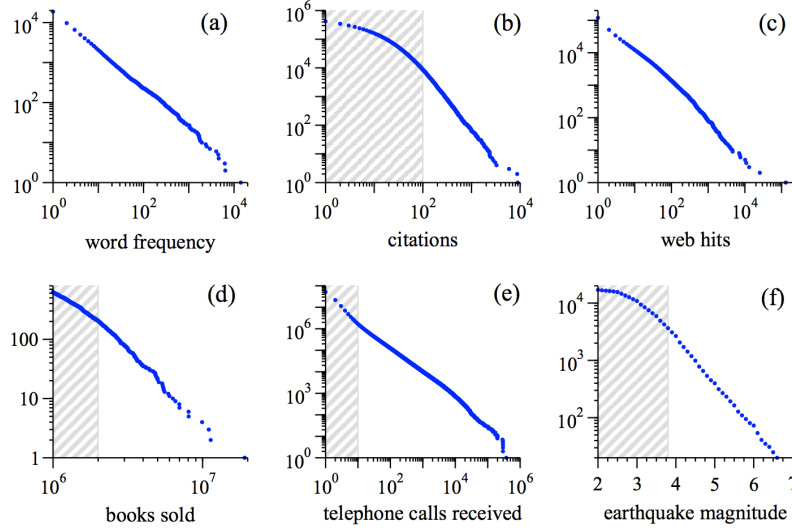


Figure 1: Hasse diagram for the lattice of partitions with a sample of size $N = 4$. A line is drawn from \mathbf{c} up to $\hat{\mathbf{c}}$ when \mathbf{c} is covered by $\hat{\mathbf{c}}$.

7.3 Priors beyond the Dirichlet process

Need for a power-law for K_n

Clauset, Shalizi, and Newman (2009) and Newman (2005) show that “Power-law distributions occur in many situations of scientific interest and have significant consequences for our understanding of natural and man-made phenomena”.



[Image from Newman (2005)]

Hence the need to depart from $K_n \sim \alpha \log n$ induced by a Dirichlet process.

Chinese restaurant process

Consider discrete data $X_1, \dots, X_n | P \stackrel{\text{iid}}{\sim} P$, and $P \sim \mathcal{Q}$

Features $k_n \leq n$ unique values $X_1^*, \dots, X_{k_n}^*$ with resp. frequencies n_1, \dots, n_{k_n}

Discrete random probability measures are characterized by predictive distr.

Dirichlet process by Ferguson (1973): $P \sim DP(\alpha, G_0)$

$$P[X_{n+1} \in \cdot | X_1, \dots, X_n] = \frac{\alpha}{\alpha + n} G_0(\cdot) + \frac{1}{\alpha + n} \sum_{j=1}^{k_n} n_j \delta_{X_j^*}(\cdot)$$

Log rate for number of clusters $k_n \asymp \alpha \log n$

Product form exchangeable partition probability function

$$p(n_1, \dots, n_{k_n}) = \alpha^{k_n} \frac{\Gamma(\alpha)}{\Gamma(\alpha + k_n)} \prod_{j=1}^{k_n} (n_j - 1)!$$

Pitman–Yor process by Pitman and Yor (1997): $P \sim PY(\sigma, \alpha, G_0)$, $\sigma \in (0, 1)$

$$P[X_{n+1} \in \cdot | X_1, \dots, X_n] = \frac{\alpha + \sigma k_n}{\alpha + n} G_0(\cdot) + \frac{1}{\alpha + n} \sum_{j=1}^{k_n} (n_j - \sigma) \delta_{X_j^*}(\cdot)$$

Power law rate for number of clusters $k_n \asymp S n^\sigma$

Product form exchangeable partition probability function

$$p(n_1, \dots, n_{k_n}) = \frac{\prod_{i=1}^{k_n-1} (\alpha + i\sigma)}{(\alpha + 1)_{(n-1)}} \prod_{j=1}^{k_n} (1 - \sigma)_{(n_j-1)}$$

Gibbs-type processes by Pitman (2003): $P \sim \text{Gibbs}(\sigma, (V_{n,k})_{n,k}, G_0)$, $\sigma < 1$

$$P[X_{n+1} \in \cdot | X_1, \dots, X_n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} G_0(\cdot) + \frac{V_{n+1,k_n}}{V_{n,k_n}} \sum_{j=1}^{k_n} (n_j - \sigma) \delta_{X_j^*}(\cdot)$$

$$\text{Rate for number of clusters } k_n \asymp \begin{cases} K \text{ random variable a.s. finite if } \sigma < 0 \\ \alpha \log n \text{ if } \sigma = 0 \\ S n^\sigma \text{ if } \sigma \in (0, 1), (S \text{ random variable}). \end{cases}$$

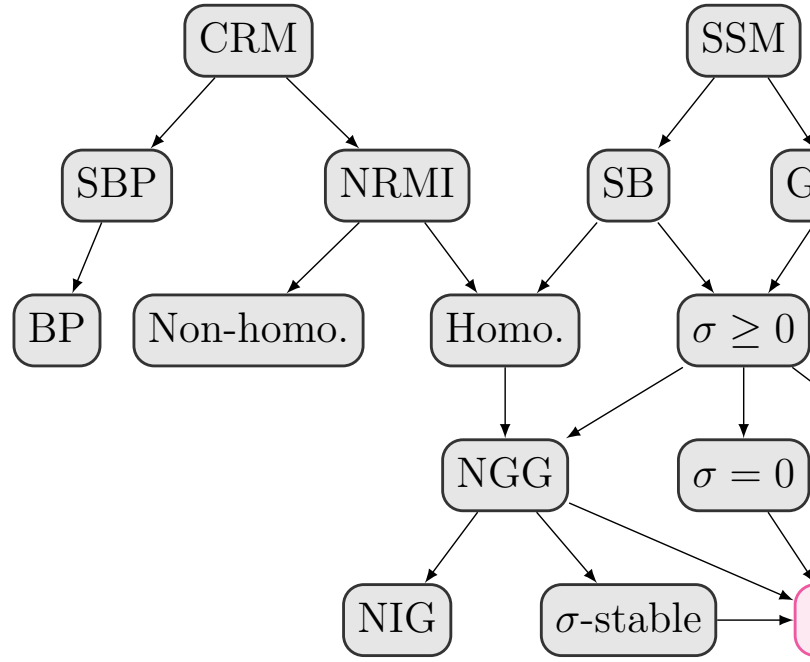
Product form exchangeable partition probability function

$$p(n_1, \dots, n_{k_n}) = V_{n,k_n} \prod_{j=1}^{k_n} (1 - \sigma)_{(n_j-1)}$$

Beyond the DP from predictive function viewpoint

A discrete random probability measure P can be classified in 3 main categories according to $P[X_{n+1} \text{ is "new"} | X_n]$

- 1) $P[X_{n+1} \text{ is "new"} | X_n] = f(n, \text{model parameters})$
 \iff depends on n but not on k_n and (n_1, \dots, n_{k_n})
 \iff Dirichlet process (Ferguson, 1973);
- 2) $P[X_{n+1} \text{ is "new"} | X_n] = f(n, k_n, \text{model parameters})$
 \iff depends on n and k_n but not on (n_1, \dots, n_{k_n})
 \iff Gibbs-type prior (Pitman, 2003);
- 3) $P[X_{n+1} \text{ is "new"} | X_n] = f(n, k_n, (n_1, \dots, n_{k_n}), \text{model parameters})$
 \iff depends on n , k_n and (n_1, \dots, n_{k_n})
 \iff tractability issues



Tree of discrete random probability measures

7.3.1 Pitman–Yor process

Proposition 7.3.1 — Pitman Sampling formula. The multiplicities (m_1, \dots, m_n) in $X_1, \dots, X_n | P \stackrel{\text{iid}}{\sim} P$, $P \sim PY(\sigma, \alpha P_0)$ have distribution

$$p(m_1, \dots, m_n) = \frac{n!}{(1 + \alpha)_{(n-1)}} (\alpha + \sigma) \cdots (\alpha + (k-1)\sigma) \prod_{\ell=1}^n \frac{1}{m_\ell!} \left(\frac{(1 - \sigma)_{(\ell-1)}}{\ell!} \right)^{m_\ell}$$

Proof. Same technique as for the Dirichlet process Ewen sampling formula. ■

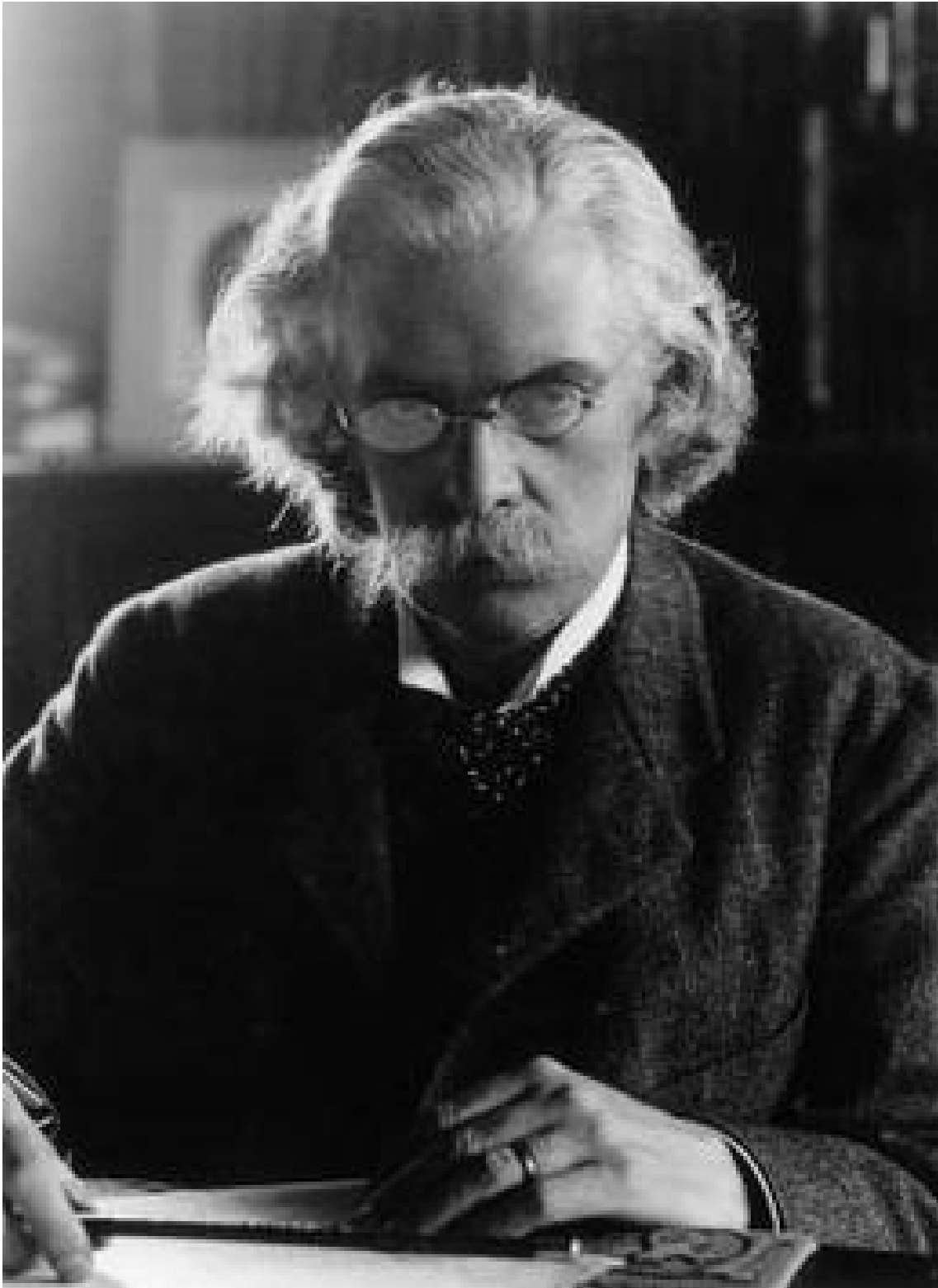
Proposition 7.3.2 — Power law and σ -diversity. For $\sigma > 0$ we have the almost sure convergence

$$n^{-\sigma} K_n \rightarrow S_{\sigma, \alpha},$$

where $S_{\sigma, \alpha}$ is called σ -diversity of the PY, whose density is a polynomially tilted Mittag–Leffler density (ML):

$$g_{\sigma, \alpha}(x) \propto x^{\alpha/\sigma} g_\alpha(x),$$

and g_α is ML density.



[Image: Wikipedia]

Theorem 7.3.3 — Stick breaking representation for PY. If $V_j \stackrel{\text{ind}}{\sim} \text{Be}(1 - \sigma, \alpha + j\sigma)$ and

$p_1 = V_1$, $p_j = V_j \prod_{l < j} (1 - V_l)$ and further we have $\phi_j \stackrel{\text{iid}}{\sim} P_0$ then

$$P = \sum_{j=1}^{\infty} p_j \delta_{\phi_j} \sim PY(\sigma, \alpha P_0).$$

Proposition 7.3.4 — Moments of PY. If $P \sim PY(\sigma, \alpha P_0)$, then for every measurable sets A, B we have

- 1) $\mathbb{E}[P(A)] = P_0(A)$,
- 2) $\mathbb{E}[P(A)P(B)] = (1 - \sigma)/(1 + \alpha)P_0(A \cap B) + (\alpha + \sigma)/(1 + \alpha)P_0(A)P_0(B)$,
- 3) $\text{Cov}[P(A), P(B)] = (1 - \sigma)/(1 + \alpha)(P_0(A \cap B) - P_0(A)P_0(B))$.

Proof. 1) We use the stick-breaking representation:

$$\mathbb{E}P(A) = \sum_j \mathbb{E}p_j \mathbb{E}\delta_{\phi_j} = \sum_j \mathbb{E}(p_j)P_0(A) = P_0(A)\mathbb{E}(\sum_j p_j) = P_0(A).$$

- 2) Let $X_1, X_2 | P \stackrel{\text{iid}}{\sim} P$, then

$$\mathbb{E}(P(A)P(B)) = P(X_1 \in A, X_2 \in B) = P(X_1 \in A)P(X_2 \in B | X_1 \in A).$$

Lets investigate two terms above: from 1) we know that $P(X_1 \in A) = P_0(A)$. We know the predictive of PY:

$$X_2 | X_1 \sim \frac{\alpha + \sigma}{\alpha + 1} P_0 + \frac{1 - \sigma}{\alpha + 1} \delta_{X_1},$$

and hence

$$P(X_2 \in B | X_1 \in A) = \frac{\alpha + \sigma}{\alpha + 1} P_0(B) + \frac{1 - \sigma}{\alpha + 1} P_{0A}(B),$$

when we used notation $P_{0A}(B) = P_0(B|A) = P_0(A \cap B)/P_0(A)$ for a conditional measure.

- 3) It is straightforward combination of 1) and 2). ■

Unlike the DP, PY is not conjugate under incoming independent samples. However, the posterior can be explicit.

Theorem 7.3.5 — Posterior distribution of PY. If $P \sim PY(\sigma, \alpha P_0)$ then the posterior of P based on observations $X_{1:n} | P \stackrel{\text{iid}}{\sim} P$ has the distribution of the random probability measure

$$(1 - q_n)P_n + q_n \sum_{j=1}^{K_n} p_j^* \delta_{X_j^*},$$

where $X_{1:n}^*$ are the K_n distinct values in $X_{1:n}$, frequencies are referred to as n_1, \dots, n_{K_n} and

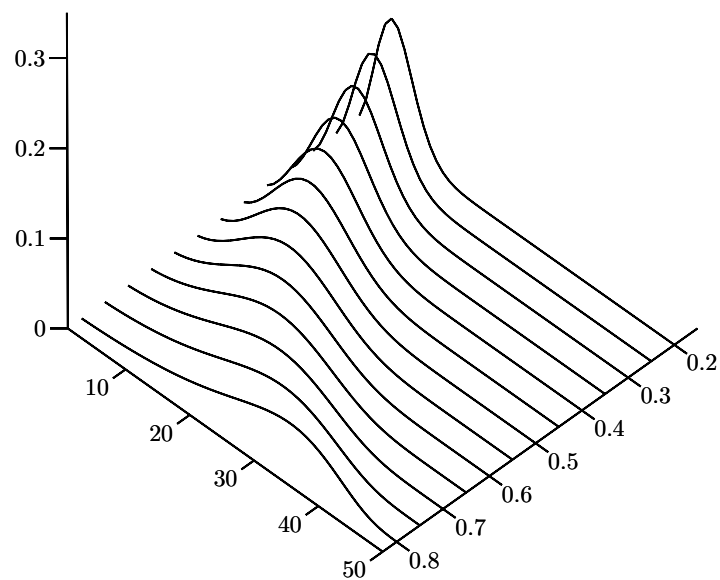
- $q_n \sim \text{Beta}(n - K_n \sigma, \alpha + K_n \sigma)$,
- $(p_1^*, \dots, p_{K_n}^*) \sim \text{Dir}_{K_n}(n_1 - \sigma, \dots, n_{K_n} - \sigma)$,
- $P_n \sim PY(\sigma, (\alpha + \sigma K_n)P_0)$.

Impact of the stability parameter σ Prior distribution of the number of clusters k_n

- α controls the location (as for the DP)

- σ controls the flatness (or variability)

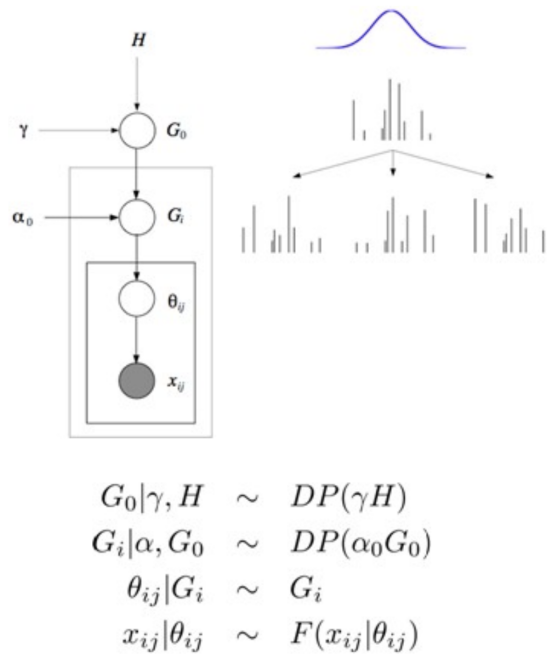
Example with $n = 50$, $\alpha = 1$ and $\sigma = 0.2, 0.3, \dots, 0.8$



[Image by De Blasi et al. (2015)]

Hierarchical Dirichlet process

A nonparametric version of **Latent dirichlet allocation** (blei2003latent) due to Teh et al. (2006)



[Image by M. Jordan]

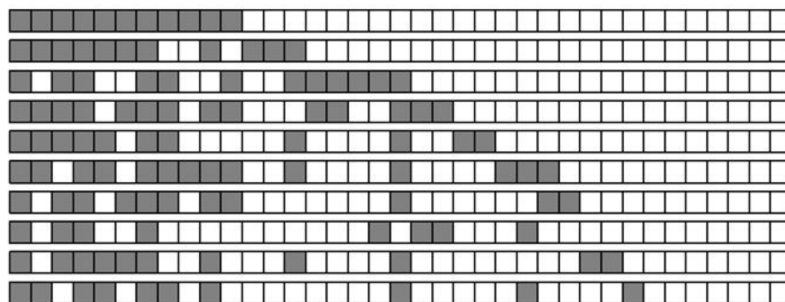
Associated partition distr. called Chinese Restaurant Franchise.

Indian Buffet process

Feature allocation model by Ghahramani and Griffiths (2006), where observations may share several features. Generative model is as follows

- first customer samples $\text{Poisson}(\gamma)$ dishes
- second customer chooses every dish of first customer *wp* $1/2$, plus $\text{Poisson}(\gamma/2)$ new dishes
- ...
- *i*th step: K dishes have been sampled, each by n_1, \dots, n_K customers; *i*th customer chooses *j*th dish *wp* n_j/i , plus $\text{Poisson}(\gamma/i)$ new dishes.

Log growth: $K_n \sim \text{Poisson}(\gamma \log n)$.



[Image by M. Jordan]

8. Asymptotic frequentist properties

References.

- Ghosh and Ramamoorthi (2003)
- Hjort et al. (2010)
- Ghosal and Van der Vaart (2017)

8.1 Introduction

Why Asymptotics

- Construction of a prior on a nonparametric space is difficult
- We cannot hope to cover all the space of density (for example) with our prior (the prior does not have full support)
- We need to check that our inference is not completely off!

Parametric setting We have the celebrated Bernstein–von Mises theorem that implies that the effect of the prior on the posterior inference vanishes when the amount of information grows.

This is not true anymore in a nonparametric setting!

Why asymptotics?

A first order approximation is to consider the asymptotic setting:

- Adopt a Frequentist point of view: “There exists a *true* parameter θ_0 , and we study the posterior distribution with data generated w.r.t. θ_0 .”
- Ideally, the posterior distribution will concentrate around θ_0 when $n \rightarrow \infty$.

8.2 Posterior consistency

Consistency

Setting:

- $\forall n \in \mathbb{N}$, let X^n be some observations in a sample space $\{\mathcal{X}^n, \mathcal{A}^n\}$ with distribution P_θ
- $\theta \in \Theta$ with (Θ, d) a (semi-)metric space

Let Π be a prior distribution on Θ and $\Pi(\cdot|X^n)$ a version of its posterior distribution.

Definition 8.2.1 — Consistency. The posterior distribution $\Pi(\cdot|X^n)$ is said to be weakly consistent at θ_0 if for all $\varepsilon > 0$

$$\Pi(d(\theta, \theta_0) > \varepsilon | X^n) \xrightarrow[n \rightarrow \infty]{P_{\theta_0}} 0.$$

If the convergence is almost sure, then the posterior is said to be strongly consistent.

Point estimators

Naturally one will hope that posterior consistency implies that some summary of the posterior location would be a consistent estimator.

Theorem 8.2.1 Let $\Pi(\cdot|X^n)$ be a posterior distribution on Θ and suppose that it is consistent at θ_0 relative to a metric d on Θ . For $\alpha \in (0, 1)$, define $\hat{\theta}_n$ as the centre of the smallest ball containing at least α of the posterior mass. Then

$$d(\hat{\theta}_n, \theta_0) \xrightarrow[n \rightarrow \infty]{P_{\theta_0}, \text{ or } P_{\theta_0} \text{ a.s.}} 0.$$

Extra notes

Take $\alpha = 1/2$ for simplicity and consistency in probability. Define $B(\theta, r)$ the closed ball of radius r centred around θ , and let

$$\hat{r}(\theta) = \inf\{r, \Pi(B(\theta, r)|X^n) \geq 1/2\}$$

(and inf over the empty set is ∞). Now let $\hat{\theta}_n$ be such that

$$\hat{r}(\hat{\theta}_n) \leq \inf_{\theta \in \Theta} \hat{r}(\theta) + 1/n$$

Consistency implies that $\Pi(B(\theta_0, \varepsilon)|X^n) \rightarrow 1$ so $\hat{r}(\theta_0) \leq \varepsilon$ with probability tending to 1. Furthermore, $\hat{r}(\hat{\theta}_n) \leq \hat{r}(\theta_0) + 1/n$ thus $\hat{r}(\hat{\theta}_n) \leq \varepsilon + 1/n$ with probability tending to 1.

In addition, $B(\theta_0, \varepsilon) \cap B(\hat{\theta}_n, \hat{r}(\hat{\theta}_n)) \neq \emptyset$ otherwise $\Pi(B(\theta_0, \varepsilon) \cup B(\hat{\theta}_n, \hat{r}(\hat{\theta}_n))|X^n) = \Pi(B(\theta_0, \varepsilon)|X^n) + \Pi(B(\hat{\theta}_n, \hat{r}(\hat{\theta}_n))|X^n) \rightarrow 1 + 1/2$. So we have

$$d(\theta_0, \hat{\theta}_n) \leq \hat{r}(\hat{\theta}_n) + \varepsilon \leq 2\varepsilon + 1/n$$

with probability that goes to 1.

- If Θ is a vector space, then one might want to use the posterior mean.
- But... weak convergence to a Dirac does not imply convergence of moments.
- Consistency of the posterior mean requires additional assumptions such as boundedness of posterior moments in probability or a.s. for some $p > 1$ would be sufficient.

Point estimator

Under some assumption on the space Θ and on the metric d one can show consistency of the posterior mean through consistency of the posterior distribution.

Theorem 8.2.2 — Posterior mean. Assume that the balls of the metric space (Θ, d) are convex. Suppose that for any sequence $\theta_{1,n}, \theta_{2,n}$ in Θ and $\lambda_n \rightarrow 0$

$$d(\theta_{1,n}, (1 - \lambda_n)\theta_{1,n} + \lambda_n\theta_{2,n}) \rightarrow 0$$

Then consistency of the posterior distribution implies consistency of the posterior mean.

Extra notes

Let $\varepsilon > 0$ and write $\hat{\theta}_n = \int \theta \Pi(d\theta|X^n)$. We decompose

$$\hat{\theta}_n = \int_{B(\theta_0, \varepsilon)} \theta \Pi(d\theta|X^n) + \int_{B(\theta_0, \varepsilon)^c} \theta \Pi(d\theta|X^n) = \theta_{1,n}(1 - \lambda_n) + \lambda_n\theta_{2,n}$$


where $\theta_{1,n} = \int_{B(\theta_0, \varepsilon)} \theta \frac{\Pi(d\theta|X^n)}{\Pi(B(\theta_0, \varepsilon)|X^n)}$, $\lambda_n = \Pi(B(\theta_0, \varepsilon)|X^n)$ and similarly for $\theta_{2,n}$ on the complement of $B(\theta_0, \varepsilon)$. Using Jensen inequality we have

$$d(\theta_{n,1}, \theta_0) \leq \int_{B(\theta_0, \varepsilon)} d(\theta, \theta_0) \frac{\Pi(d\theta|X^n)}{\Pi(B(\theta_0, \varepsilon)|X^n)} \leq \varepsilon$$

In addition we have

$$d(\hat{\theta}_n, \theta_0) \leq d(\theta_{n,1}, \theta_0) + d(\theta_{n,1}, \theta_{1,n}(1 - \lambda_n) + \lambda_n\theta_{2,n}).$$

Using the fact that $\lambda_n \rightarrow 0$ since the posterior is consistent, we have the desired result.

 For the condition on d to hold, one can assume it to be convex and uniformly bounded.

A first consistent posterior

■ **Example 8.1 — Dirichlet process.** Assume the following model

$$\begin{aligned} X_1, \dots, X_n &\stackrel{iid}{\sim} P, \\ P &\sim \text{DP}(M\alpha) \end{aligned}$$

Consider the semi-metric $d_A(P, Q) = |P(A) - Q(A)|$ for some measurable event A on Θ , then $\Pi(\cdot|X^n)$ is strongly consistent at any P_0 for d_A . ■

From this result, we can easily obtain consistency under the weak topology. We could also obtain stronger consistency using Glivenko–Cantelli theorem.

Extra notes

Consider $\Pi(|P(A) - P_0(A)| \geq \varepsilon|X^n)$ which calls for applying Markov inequality. Properties of the Dirichlet process imply that

$$P|X^n \sim \text{DP}(M\alpha + n\mathbb{P}_n),$$

thus

$$P(A)|X^n \sim \text{Beta}(M\alpha(A) + n\mathbb{P}_n(A), M\alpha(A^c) + n\mathbb{P}_n(A^c)).$$

We thus have

$$\mathbb{E}(P(A)|X^n) = \frac{M}{M+n} \alpha(A) + \frac{n}{M+n} \mathbb{P}_n(A) := \bar{P}(A)$$

$$\text{var}(P(A)|X^n) = \frac{\bar{P}(A)\bar{P}(A^c)}{1+n+M} \leq \frac{1}{4(1+n+M)}.$$

Markov inequality gives

$$\Pi(|P(A) - P_0(A)| \geq \varepsilon | X^n) \leq \frac{1}{\varepsilon^2} (|\bar{P}(A) - P_0(A)|^2 + \text{var}(P(A)|X^n))$$

$$\rightarrow 0 [P_0, a.s.]$$

using the law of large numbers on $\mathbb{P}(A)$.

Bayesian modelling perspective

From a Bayesian point of view, a Dirac measure at θ_0 corresponds to perfect knowledge of the parameter.

- Prior and posterior distributions model our knowledge about the parameter.
- Consistency thus implies that when the amount of information grows, we tend towards perfect knowledge of the parameter.

A validation of Bayesian methods

The frequentist setting where there exists a *true* parameter θ_0 that generates the data can be seen as an idealized set-up.

- An experimenter feeds a Bayesian with some data using the same data-generating mechanism.
- When the number of observation grows, a Bayesian should be able to pin-point the data-generating mechanism, whatever their prior.
- <2> A prior that does not lead to a consistent posterior should not be used.

Robustness

Two Bayesians walk into a bar... with *almost* the same prior, then their posterior inference should not differ that much.

- Let Π_1 be the prior of Bayesian number 1
- Bayesian number 2 uses an “ ε -corrupted” prior $\Pi_2 = (1 - \varepsilon)\Pi_1 + \varepsilon\delta_{p_0}$ for some $p_0 \in \Theta$

The posterior of Bayesian number 2 is consistent at p_0 (to be seen later), now what if Π_1 is not consistent at p_0 ?

Let d_W be the metric for the weak topology, then $d_W(\Pi_1(\cdot|X^n), \Pi_2(\cdot|X^n))$ would not go to 0.

Extra notes

There exists some $\varepsilon_0 > 0$ such that

$$\Pi_{n,1}(B(\theta_0, \varepsilon_0)|X^n) \not\rightarrow 0$$

Thus

$$|\Pi_{n,1}(B(\theta_0, \varepsilon_0)|X^n) - \Pi_{n,2}(B(\theta_0, \varepsilon_0)|X^n)| \not\rightarrow 0$$

since $\Pi_{n,2}(B(\theta_0, \varepsilon_0)|X^n) \rightarrow 0$.

8.2.1 Doob's Theorem

Doob's Theorem

Can one get general conditions on the prior to ensure that it is consistent?

- A first answer: Doob's Theorem
- The posterior is consistent at every θ Π -a.s.

Consider the case of *i.i.d.* observations

Theorem 8.2.3 — Doob's Theorem. Let $\{\mathcal{X}^n, P_\theta, \Theta\}$ be a statistical model where $\{\mathcal{X}^n, \mathcal{A}^n\}$ is a Polish space with Borel σ -field and Θ a Borel subset of a Polish space. Suppose that the map $\theta \mapsto P_\theta(A)$ is Borel measurable for every $A \in \mathcal{A}$ and $\theta \mapsto P_\theta$ is one-to-one. Then for any prior distribution Π on Θ , if $X_1, \dots, X_n \stackrel{iid}{\sim} P_\theta$, $\theta \sim \Pi$, the posterior is strongly consistent at any θ Π -a.s.

Doob's Theorem

Some remarks on Doob's Theorem

- The conditions of the theorem are extremely weak
- And no conditions on the prior
- However this is only true Π -almost surely.
- Note: the Π -null set can be quite big! we can be happy with this result only if we are confident that the parameters are on the support of the prior. In general no one can be sure that the parameter generating the data inside the support of the prior, this is a real problem in fact in general the support of the prior can be quite thin.
An extreme example is the case where the prior is a Dirac on some parameter θ_0 . Then Doob's theorem still holds.

8.2.2 Schwartz approach

Setting

Setting Doob's approach is not enough to show consistency of the posterior. For simplicity we focus on the **density estimation** setting.

- Θ is the set of probability density functions on \mathcal{X} w.r.t. a common dominating measure ν . We denote the parameter p (instead of θ) and P the associated probability measure.
- Observations follow $X_1, \dots, X_n \stackrel{iid}{\sim} p$, and $p \sim \Pi$.

Considering **density estimation** makes things easier without being too simplistic. The same results can be extended to **nonparametric regression**.

Hypotheses

KL property

To achieve consistency, we do not want to require that the true parameter p_0 is **inside** the support of Π . However we still require **some prior mass near p_0** .

Definition 8.2.2 — Kullback–Leibler. Let p and p_0 be two p.d.f. with respect to a common

measure such that $p_0 \ll p$. Then the Kullback–Leibler divergence between p and p_0 is

$$\text{KL}(p, p_0) = \int p_0 \log(p_0/p) d\nu.$$

KL property

Definition 8.2.3 — KL property. We say that a prior distribution Π satisfies the **Kullback–Leibler property** at p_0 if for every $\varepsilon > 0$,

$$\Pi(p : \text{KL}(p, p_0) \geq \varepsilon) > 0$$

We note $p_0 \in \text{KL}(\Pi)$ and alternatively will say that p_0 is in the KL-support of Π .

This extends quite a lot the parameters at which the posterior can be consistent.

Existence of tests

The other requirement is that the parameter set is not too complex.

Definition 8.2.4 — Exponentially consistent tests. We say that a sequence of tests ϕ_n for $H_0 : p = p_0$ versus $H_1 : p \in U^c$ is exponentially consistent if

$$P_0^n(\phi_n) \lesssim e^{-Cn}, \sup_{p \in U^c} P^n(1 - \phi_n) \lesssim e^{-Cn}$$

A test is understood as a measurable map $\mathcal{X}^n \rightarrow [0, 1]$ and the corresponding statistic $\phi_n(X_1, \dots, X_n)$. ϕ_n is interpreted as the probability that the null is rejected.

Extra notes

The existence of tests means that we can differentiate between p_0 and parameter in U^c .

It is enough to have uniformly consistent sequence of test

$$P_0(\phi_n) \rightarrow 0, \sup_{p \in U^c} P(1 - \phi_n) \rightarrow 0.$$

Since the test is uniformly consistent then there exists $k \in \mathbb{N}$ such that $P_0^k(\phi_k) \leq 1/4$, $P^k(1 - \phi_k) \leq 1/4$. Now for n large, write $n = mk + r$. Slice $X^n = (X_1, \dots, X_n)$ into m sub-sample of size k $X_l^n = (X_{(l-1)k+1}, \dots, X_{lk})$ and define $Y_{l,n} = \phi_k(X_l^n)$. Now create a new test $\psi_n = \mathcal{I}\{\bar{Y}_m > 1/2\}$. We have for every $p \in U^c$, $P(1 - Y_j) \leq 1/4$

$$P(\psi_n) = P(\bar{Y} \leq 1/2) = P(1 - \bar{Y} \geq 1/2) =$$

$$P(1 - \bar{Y} \geq 1/2) \leq e^{-2m/16} \lesssim e^{-Cn}$$

Using Hoeffding inequality: $P(\bar{X} - \mathbb{E}(X) \geq \varepsilon) \leq \exp\{-2\varepsilon^2 m\}$.

Schwartz Theorem

Theorem 8.2.4 — Schwartz Theorem. Let Π be a prior distribution on Θ such that $p_0 \in \text{KL}(\Pi)$. Let U be a neighbourhood of p_0 such that there exists an exponentially consistent sequence of tests for p_0 against U^c . Then

$$\Pi(U^c | X^n) \rightarrow 0 [P_{0a.s.}].$$

This theorem is not due to Herman Schwarz (without t!), nor to Laurent Schwartz the Fields Medalist! But to Lorraine Schwartz, former student of Lucien Le Cam.

Extra notes

$$\Pi(U^c|X^n) = \frac{\int_{U^c} \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p)}{\int_{\mathcal{P}} \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p)} := \frac{N_n}{D_n}.$$

We first show $\liminf D_n e^{n\varepsilon} / \Pi(KL(p, p_0) > \varepsilon) \geq 1$, $P_0[a.s.]$. Let $\Pi_0(\cdot) = \Pi(\cdot \cap KL(p, p_0) > \varepsilon) / \Pi(KL(p, p_0) > \varepsilon)$. Then

$$\begin{aligned} \log(D_n) &\geq \log\left(\int_{KL(p, p_0) > \varepsilon} \frac{p}{p_0}(X_i) d\Pi_0(p)\right) + \log(\Pi(KL(p, p_0) < \varepsilon)) \\ &\geq \int_{KL(p, p_0) > \varepsilon} \log\left(\prod_{i=1}^n \frac{p}{p_0}(X_i)\right) d\Pi_0(p) + \log(\Pi(KL(p, p_0) < \varepsilon)) \\ &= \sum_{i=1}^n \int \log \frac{p}{p_0}(X_i) d\Pi_0(p) + \log(\Pi(KL(p, p_0) < \varepsilon)) \end{aligned}$$

The law of large numbers implies

$$\frac{1}{n} \sum_{i=1}^n \int \log \frac{p}{p_0}(X_i) d\Pi_0(p) \rightarrow P_0 \int \frac{p}{p_0}(X_i) d\Pi_0(p), \quad P_0[a.s.]$$

which is $-\int KL(p, p_0) d\Pi_0(p) > -\varepsilon$. Thus

$$\liminf D_n e^{n\varepsilon} / \Pi(KL(p, p_0) > \varepsilon) \geq 1, \quad P_0[a.s.]$$

For n large enough we have the following $P_0[a.s.]$

$$\begin{aligned} \Pi(U^c|X^n) &\leq \phi_n + (1 - \phi_n) \frac{N_n}{D_n} \\ &\leq \phi_n + (1 - \phi_n) N_n e^{\varepsilon n} \Pi(KL(p, p_0) > \varepsilon) \end{aligned}$$

Furthermore we have that

$$\begin{aligned} P_0^n N_n (1 - \phi_n) &= P_0^n \int_{U^c} (1 - \phi_n) \prod_{i=1}^n \frac{p}{p_0}(X_i) \Pi(dp) \\ &= \int_{U^c} P^n (1 - \phi_n) \Pi(dp) \leq e^{-Cn} \end{aligned}$$

We thus get $P_0 \Pi(U^c|X^n) \leq e^{-C'n}$ for $\varepsilon < C$ and for $C' = C - \varepsilon$. Using Borel–Cantelli we get that $\Pi(U^c|X^n) \rightarrow 0$ $P_0[a.s.]$.

Schwartz Theorem

- Need to test away all densities in U^c
- Might not be possible for strong neighbourhood of p_0 (L_1 metrics)

Extension of Schwartz theorem The idea is that not *all* functions in U^c matters and we can discard function with very low prior probabilities.

Theorem 8.2.5 The results of the previous theorem are still valid if we replace the assumption

on the existence of tests by:

$$\Theta_n \subset \Theta$$

$$\Pi(\Theta_n) \leq e^{-Cn}, P_0^n \phi_n \leq e^{-Cn}, \sup_{p \in U^c \cap \Theta_n} P(1 - \phi_n) \leq e^{-Cn}$$

Existence of tests

Existence of tests

Schwartz' theorem requires the existence of exponentially consistent tests.

- We can differentiate between θ_0 and U^c
- The model is not too complex

Question When do such tests exist?

Let's see the example of iid observations.

Sketch of the proof

- Cannot directly construct test against $U^c = \{p, d(p, p_0) > \varepsilon\}$...
- Construct an exponentially consistent test against a generic ball that is at least at distance ε
- Cover U^c with N of these balls, and construct a test from the N corresponding tests.

Consistency under Entropy bound

Consistency under Entropy bound

We combine the preceding results to get general conditions <2->on the prior and <2->on the model, that ensure consistency.

Theorem 8.2.6 The posterior is strongly consistent relative to the L_1 distance at every p_0 in the KL-support of the prior if for every $\varepsilon > 0$ there exist Θ_n such that for $C > 0$ and $0 < c < 1/2$

$$\Pi(\Theta_n^c) \leq e^{-Cn}, \log N(\varepsilon, \Theta_n, \|\cdot\|_1) \leq cn\varepsilon_n^2,$$

for n large enough.

8.3 Concentration rates

Definition

Contraction rates are a refinement of posterior consistency.

- How fast posterior concentrates its mass around the true parameter
- Helps to see how much the prior influences the posterior

Definition 8.3.1 Let ε_n be a positive sequence. The posterior contracts at the rate ε_n at θ_0 if for any $M_n \rightarrow \infty$

$$\Pi(d(\theta, \theta_0) > M_n \varepsilon_n | X^n) \xrightarrow[n \rightarrow \infty]{P_{\theta_0}} 0$$

If all the experiments share the same probability space and the convergence is $P_{\theta_0}[a.s]$ we say that the posterior contracts in the strong sense.

Remarks

- Any slower rate than ε_n also fits the definition so we will say a posterior contraction rate
- We will naturally try to find the fastest possible rate!

Regarding M_n

- The sequence M_n plays virtually no role in the posterior rate. In many cases it can be fixed to a constant M .
- For finite dimensional models M_n must be allowed to grow to obtain the usual $n^{-1/2}$ rate in smooth models.

Consequences of posterior contraction

Point Estimator

- Let $\hat{\theta}_n$ = centre of the smallest ball that contains at least $1/2$ of the posterior mass.
- Assume that the posterior contracts at θ_0 with rate ε_n for the metric d

Then $d(\hat{\theta}_n, \theta) = O_P(\varepsilon_n)$ in P_0 probability (or a.s. if strong contraction).

Consequences of posterior contraction

Posterior mean If the metric d is bounded and $\theta \mapsto d^s(\theta, \theta_0)$ is convex for some $s \geq 1$ then the posterior mean $\tilde{\theta}_n$ satisfies

$$d(\tilde{\theta}_n, \theta_0) \leq M_n \varepsilon_n + \|d\|_\infty^{1/s} \Pi_n(d(\theta, \theta_0) \geq M_n \varepsilon_n | X^n)^{1/s}.$$

- First term is the dominating term
- The second term is exponentially small in general

Some first Examples - Parametric models

- Let $X_1, \dots, X_n | \theta \stackrel{iid}{\sim} \mathcal{B}(\theta)$, and $\theta \sim \text{Beta}(\alpha, \beta)$. The posterior contracts at a rate $n^{-1/2}$
- Let $X_1, \dots, X_n | \theta \stackrel{iid}{\sim} \mathcal{U}([0, \theta])$ and $\pi(\theta) \propto \theta^{-a}$. The posterior contracts at a rate n^{-1} .

Parametric regular models In fact for all regular finite dimensional models the Bernstein von-Mises theorem implies a posterior rate of $n^{-1/2}$.

Nonparametric example: Dirichlet Process

- $X_1, \dots, X_n | P \stackrel{iid}{\sim} P$
- $P \sim \text{DP}(M\alpha)$ for α a probability measure on \mathcal{X} .

The posterior distribution is $P | X^n \sim \text{DP}(M\alpha + n\mathbb{P}_n)$.

Local semi-metric¹ For a measurable set A , let $d(P, Q) = |P(A) - Q(A)|$. The posterior distribution is consistent at P_0 at a rate $n^{-1/2}$.

Global metric For ν a σ -finite measure and F and G two c.d.f. let $d(F, G) = \|F - G\|_\nu^2 = \int (F(t) - G(t))^2 d\nu(t)$. The posterior contracts at rate $n^{-1/2}$ at P_0 for this metric.

¹ie $d(P, Q) \geq 0$ but might be 0 for $P \neq Q$.

Nonparametric example: White Noise

Consider the following model for W_t a white noise

$$X_t = f(t) + n^{-1/2}W_t.$$

Projecting this model onto the Fourier basis if $f \in L_2$, we have the equivalent formulation

$$X_{i,n} = \theta_i + n^{-1/2}\epsilon_i, \quad i \in \mathbb{N}^*$$

$\theta \in \ell_2(\mathbb{L})$. Assume the following prior

$$\theta_i \stackrel{ind.}{\sim} \mathcal{N}(0, i^{-2\alpha-1}).$$

If $\theta_0 \in \mathcal{S}_\beta^{2,2}$ then the posterior contracts at θ_0 at the rate $n^{-\min(\alpha, \beta)/(2\alpha+1)}$.

8.3.1 General theorem in the iid case

General theorem

- Result similar to Schwartz theorem?
- We focus on the case of i.i.d observations $X_1, \dots, X_n \stackrel{iid}{\sim} P$
- The parameter set Θ is the set of probability densities with respect to a common dominating measure μ .

Let Π_n be a sequence of priors. We study the sequence of posterior distributions $\Pi_n(\cdot | X^n)$ under the assumption that the data are generated from P .

We follow the same steps as for Schwartz' Theorem:

- Existence of tests to separate p_0 from the complement of balls
- KL condition: the prior puts enough mass on neighbourhood of p_0

Define $V_{2,0}$, the 2nd KL variation

$$V_2 = P_0 \left(\log^2 \left(\frac{p_0}{P}(X) \right) \right),$$

and define two KL neighbourhoods as

$$B_0(p_0, \epsilon) = \{p, \text{KL}(p_0, p) \leq \epsilon^2\},$$

$$B_2(p_0, \epsilon) = \{p, \text{KL}(p_0, p) \leq \epsilon^2, V_2(p_0, p) \leq \epsilon^2\}.$$

Theorem 8.3.1 — Ghosal, Ghosh and van der Vaart. Let $d \leq h$ be a metric on Θ for which balls are convex, and let $\Theta_n \subset \Theta$. The posterior contracts at a rate ϵ_n for all ϵ_n such that $n\epsilon_n^2 \rightarrow \infty$ and such that for positive constants c_1, c_2 and any $\underline{\epsilon}_n \leq \epsilon_n$

$$\log N(\epsilon_n, \Theta_n, d) \leq c_1 n \epsilon_n^2,$$

$$\Pi_n(B_{2,0}(p_0, \epsilon_n^2)) \geq e^{-c_2 n \epsilon_n^2}$$

$$\Pi(\Theta_n^c) \leq e^{-(c_2+3)n \epsilon_n^2}$$

- The KL condition can be refined, but the idea is basically the same
- Entropy condition is useful for the existence of tests

- Entropy condition can be replaced by a local entropy, which is more like a *dimension of Θ_n*

Interpretation Assume that d and KL are equivalent

- We need $e^{n\varepsilon_n^2}$ balls to cover Θ_n .
- If the prior spread evenly the mass on these balls, we have $e^{-Cn\varepsilon_n^2}$ mass on each of these balls thus KL condition is satisfied
- If the spread is uneven, then KL condition might not be satisfied for some p_0 .

8.3.2 Non iid observations

General result

General observations

- The previous theorem can be generalized to other models (like regression for instance)
- But we have to be careful with the metric we use, and the existence of test is not guaranteed!
- To be general we will have to assume that we can test away parameters

Existence of tests Let d_n and e_n be two semi-metrics on Θ . For $\varepsilon > 0$, and for all $\theta_1 \in \Theta$ such that $d_n(\theta_0, \theta_1) > \varepsilon$ there exists ϕ_n

$$P_{\Theta_0}^n \phi_n \leq e^{-Kn\varepsilon^2}, \quad \sup_{\theta, d_n(\theta, \theta_1) \leq \xi\varepsilon} P_{\theta}^n(1 - \phi_n) \leq e^{-Kn\varepsilon^2}$$

General theorem Define the following KL-neighbourhood

$$V_{k,0}(f, g) = \int f |\log(f/g) - \text{KL}(f, g)|^k d\mu$$

$$B_n(\theta_0, \varepsilon, k) = \left\{ \theta \in \Theta \mid \text{KL}(p_{\theta_0}^n, p_{\theta}^n) \leq n\varepsilon^2, V_{k,0}(p_{\theta_0}^n, p_{\theta}^n) \leq n^{k/2} \varepsilon^k \right\}$$

General theorem

Theorem 8.3.2 Let d_n and e_n be two semi-metrics on Θ , such that tests exists, $\varepsilon_n \rightarrow 0$, $n\varepsilon_n^2 \rightarrow \infty$, $k > 1$, $\Theta_n \subset \Theta$ such that for sufficiently large $j \in \mathbb{N}$

$$\sup_{\varepsilon \geq \varepsilon_n} \log N \left(\frac{1}{2} \xi \varepsilon, \{ \theta \in \Theta_n \mid d_n(\theta_0, \theta) \leq \varepsilon \}, e_n \right) \leq n\varepsilon_n^2$$

$$\frac{\Pi_n(\theta \in \Theta_n, j\varepsilon_n \leq d_n(\theta, \theta_0) \leq 2j\varepsilon_n)}{\Pi_n(B_n(\theta_0, \varepsilon_n, k))} \leq e^{Kn\varepsilon_n^2 j^2 / 2}$$

$$\frac{\Pi_n(\Theta_n^c)}{\Pi_n(B_n(\theta_0, \varepsilon_n, k))} \leq e^{-2n\varepsilon_n}$$

then $P_{\theta_0}^n \Pi_n(d_n(\theta_0, \theta) \geq M_n \varepsilon_n) = o(1)$

Independent Observations

Independent observations

- Assume that the measure $P_{\theta}^n = \bigotimes_{i=1}^n P_{i,\theta}$ on some product space $\bigotimes_{i=1}^n \{\mathcal{X}_i, \mathcal{A}_i\}$.
- Assume that each measures $P_{i,\theta}$ are absolutely continuous w.r.t μ_i

- Define the Root average Hellinger distance

$$d_{n,H}(\theta, \theta') = \left(\frac{1}{n} \sum_{i=1}^n \int (\sqrt{dP_{i,\theta}} - \sqrt{dP_{i,\theta'}})^2 \right)^{1/2}$$

Lemma 5. *For all here exists tests ϕ_n such that*

$$P_{\theta_0}^n \phi_n \leq e^{-nd_{n,H}(\theta_0, \theta_1)}, \quad P_{\theta}^n \leq e^{-nd_{n,H}(\theta_0, \theta_1)}$$

for all θ such that $d_{n,H}(\theta, \theta_1) \leq \frac{1}{18} d_{n,H}(\theta_0, \theta_1)$.

Independent observations

We can also simplify the KL condition in this case. Note that

$$KL(p_{\theta_0}^n, p_{\theta}^n) = \sum_{i=1}^n KL(p_{i,\theta_0}, p_{i,\theta})$$

Furthermore for the KL-variation term we have that

$$V_{k,0}(p_{\theta_0}^n, p_{\theta}^n) \leq n^{k/2} C_k \frac{1}{n} \sum_{i=1}^n V_{k,0}(p_{i,\theta_0}, p_{i,\theta})$$

Thus the KL condition can be re-written

$$B_n^*(\theta_0, \varepsilon, k) = \left\{ \theta, \frac{1}{n} \sum_{i=1}^n KL(p_{i,\theta_0}, p_{i,\theta}) \leq \varepsilon^2, \frac{1}{n} \sum_{i=1}^n V_{k,0}(p_{i,\theta_0}, p_{i,\theta}) \leq C_k \varepsilon^2 \right\}$$

Example: Nonparametric Regression using Splines

NP Regression with splines

Consider the model

$$X_i = f(z_i) + \varepsilon_i, \quad i = 1, \dots, n$$

where $\varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and the $z_i \in \mathbb{L}$ are known fixed covariates. For simplicity σ^2 is also assume to be known. Let $\mathbb{P}_n^z = \frac{1}{n} \sum_{i=1}^n \delta_{z_i}$ and $\|\cdot\|_n$ the $L_2(\mathbb{P}_n^z)$ norm

Lemma 6. *We have the following results*

$$KL(P_{f,i}, P_{g,i}) = \frac{1}{2\sigma^2} (f(z_i) - g(z_i))^2$$

$$V_{0,2}(P_{f,i}, P_{g,i}) = \frac{1}{\sigma^2} (f(z_i) - g(z_i))^2$$

Assume that $f_0 \in \mathcal{H}(\alpha, L)$ such that $\|f_0\|_\infty \leq H$, then the $d_{n,H}^2$ and $\|\cdot\|_n^2$ are equivalent.

Spline prior Consider $(B_j)_{j=1}^J$ the B-splines basis with J equally spaced nodes, and consider

$$f_\beta(\cdot) = \sum_{j=1}^J \beta_j B_j(\cdot)$$

and induce a prior on f by choosing a prior on β , $\beta_j \stackrel{iid}{\sim} g$.

Approximation techniques with splines gives us that for $\beta^* \in \mathbb{L}^J$ the coefficient of the projection of f_0 in $\text{Span}(B_j)$,

$$\|f_{\beta^*} - f_0\|_\infty \leq J^{-\alpha} \|f_0\|_\alpha$$

We also need to impose conditions on the design. Let Σ_n be such that $\Sigma_{n,i,j} = \int B_i B_j d\mathbb{P}_n^z$. We assume that

$$J^{-1} \|\beta\|^2 \asymp \beta' \Sigma_n \beta$$

so that

$$\|f_{\beta_1} - f_{\beta_2}\|_n \asymp \sqrt{J} \|\beta_1 - \beta_2\|$$

We can thus perform calculations in terms of the Euclidean norm of the coefficients.

Theorem 8.3.3 Assume that g is a standard Gaussian distribution, and assume that $J = J_n \asymp n^{1/(2\alpha+1)}$, then the posterior contracts at a rate $\varepsilon_n = n^{-\alpha/(2\alpha+1)}$.

- This is the minimax rate, in addition this rate is uniform over all bounded $\mathcal{H}(\alpha, L)$ functions.
- Some condition can be relaxed, in particular, g could be any distribution such that for every β^* such that $\|\beta^*\|_\infty \leq C$ $\Pi(\|\beta - \beta^*\| \leq \varepsilon) \geq e^{-cJ \log(1/\varepsilon)}$. Some log factor may appear in the rate.
- The boundedness condition could also be dropped by considering likelihood ratio tests for $\|\cdot\|_n$ norm.



Bayesian deep learning

9	Deep Gaussian processes and variational autoencoders	91
9.1	Introduction	
9.2	Deep Gaussian processes	
9.3	Variational autoencoders	
10	Bayesian neural networks	93
10.1	Introduction	
10.2	Prior distributions	
10.3	Posterior sampling algorithms	
10.4	Generalization	
	Bibliography	95



9. Deep Gaussian processes and variational autoe

9.1 Introduction

9.2 Deep Gaussian processes

9.3 Variational autoencoders

10. Bayesian neural networks

10.1 Introduction

10.2 Prior distributions

10.2.1 Connection prior/initialization

10.2.2 Neural-network Gaussian process (NN-GP)

10.2.3 Neural tangent kernel (NTK)

10.2.4 Edge of Chaos

10.2.5 Unit priors get heavier with depth

10.2.6 Other priors

10.3 Posterior sampling algorithms

10.3.1 MCMC

10.3.2 Laplace approximation

10.3.3 Monte Carlo dropout

The dropout technique can be reinterpreted as a form of approximate Bayesian variational inference (Gal and Ghahramani, 2016; Diederik P Kingma, Salimans, and Max Welling, 2015). Gal and Ghahramani, 2016 build a connection between dropout and the Gaussian process representation, while Diederik P Kingma, Salimans, and Max Welling (2015) propose a way to interpret Gaussian dropout. They develop a *variational dropout* where each weight of a model has its individual dropout rate. *Sparse variational dropout*, proposed by Molchanov, Ashukha, and Vetrov (2017), extends *variational dropout* to all possible values of dropout rates and leads to a sparse solution.

The approximate posterior is chosen to factorize either over rows or over individual entries of the weight matrices. The prior usually factorizes in the same way. Therefore, performing dropout can be used as a Bayesian approximation. However, as noted by Duvenaud et al. (2014), dropout has no regularization effect on infinitely-wide hidden layers. Nalisnick, Hernández-Lobato, and Smyth (2019) propose a Bayesian interpretation of regularization via multiplicative noise, with dropout being the particular case of Bernoulli noise. They find that noise applied to hidden units ties the scale parameters in the same way as the ARD algorithm (Radford M Neal, 1996), a well-studied shrinkage prior.

Let us describe Monte Carlo dropout in more details. The idea is simple and consists in performing random sampling at test time. Instead of turning off the dropout layers at test time (as is usually done), hidden units are randomly dropped out according to a Bernoulli(p) distribution. Repeating this operation M times provides M versions of the MAP estimate of the network parameters w^m , $m = 1, \dots, M$ (where some units of the MAP are dropped), yielding an approximate posterior predictive in the form of the equal-weight average:

$$p(y|x, \mathcal{D}^n) \approx \frac{1}{M} \sum_{m=1}^M p(y|x, w^m). \quad (10.1)$$

Monte Carlo dropout captures some uncertainty from out-of-distribution (OOD) inputs, but is nonetheless incapable of providing valid posterior uncertainty. Indeed, Monte Carlo dropout changes the Bayesian model under study, which modifies also the properties of the approximate Bayesian inference performed. Specifically, Folgoc et al., 2021 shows that the Monte Carlo dropout posterior predictive (10.1) assigns zero probability to the true model posterior predictive distribution.

10.3.4 Variational inference

10.3.5 Expectation propagation

10.3.6 SGD-based methods

10.3.7 Last-layer methods

10.3.8 Deep ensembles

10.3.9 Cold posteriors

10.4 Generalization

10.4.1 PAC-Bayes

Bibliography

- [1] Charles E Antoniak. “Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems”. In: *The Annals of Statistics* (1974), pages 1152–1174 (cited on page 59).
- [2] Julyan Arbel et al. “BNPdensity: Bayesian nonparametric mixture modeling in R”. In: *Australian & New Zealand Journal of Statistics* 63 (3 2021), pages 542–564. DOI: [10.1111/anzs.12342](https://doi.org/10.1111/anzs.12342). eprint: [2110.10019](https://arxiv.org/abs/2110.10019) (cited on page 63).
- [3] J. O. Berger and R. L. Wolpert. *The likelihood principle: A review, generalizations, and statistical implications*. Volume 6. Institute of Mathematical Statistics, 1988 (cited on pages 21–23).
- [4] David M Blei, Michael I Jordan, et al. “Variational inference for Dirichlet process mixtures”. In: *Bayesian analysis* 1.1 (2006), pages 121–144 (cited on page 63).
- [5] C. Blundell et al. “Weight uncertainty in neural networks”. In: *International conference on machine learning*. PMLR. 2015, pages 1613–1622 (cited on page 44).
- [6] N. Bou-Rabee and J. M. Sanz-Serna. “Geometric integrators and the Hamiltonian Monte Carlo method”. In: *Acta Numerica* 27 (2018), pages 113–206 (cited on pages 35, 37).
- [7] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. “Power-law distributions in empirical data”. In: *SIAM review* 51.4 (2009), pages 661–703 (cited on page 66).
- [8] David B Dahl. “Model-based clustering for expression data via a Dirichlet process mixture model”. In: *Bayesian inference for gene expression and proteomics* (2006), pages 201–218 (cited on page 65).
- [9] Pierpaolo De Blasi et al. “Are Gibbs-type priors the most natural generalization of the Dirichlet process?” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2 (2015), pages 212–229 (cited on page 72).
- [10] J. Dick and F. Pilichshammer. *Digital Nets and Sequences. Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, 2010 (cited on page 31).

- [11] M. D. Donsker and S. R. S. Varadhan. “Asymptotic evaluation of certain Markov process expectations for large time, I”. In: *Communications on Pure and Applied Mathematics* 28.1 (1975), pages 1–47 (cited on page 45).
- [12] R. Douc, É. Moulines, and D. Stoffer. *Nonlinear time series*. Chapman-Hall, 2014 (cited on pages 31, 32).
- [13] David Duvenaud et al. “Avoiding pathologies in very deep networks”. In: *International Conference on Artificial Intelligence and Statistics*. 2014 (cited on page 94).
- [14] Warren J Ewens. “The sampling theory of selectively neutral alleles”. In: *Theoretical population biology* 3.1 (1972), pages 87–112 (cited on page 59).
- [15] T.S. Ferguson. “A Bayesian analysis of some nonparametric problems”. In: *The Annals of Statistics* 1.2 (1973), pages 209–230. ISSN: 0090-5364 (cited on pages 54, 67, 68).
- [16] P. C. Fishburn. *Utility theory for decision making*. Technical report. Research analysis corp. McLean VA, 1970 (cited on page 26).
- [17] Loic Le Folgoc et al. “Is MC Dropout Bayesian?” In: *arXiv preprint arXiv:2110.04286* (2021) (cited on page 94).
- [18] Yarín Gal and Zoubin Ghahramani. “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning”. In: *International Conference on Machine Learning*. 2016 (cited on page 93).
- [19] Zoubin Ghahramani and Thomas L Griffiths. “Infinite latent feature models and the Indian buffet process”. In: *Advances in neural information processing systems*. 2006, pages 475–482 (cited on page 73).
- [20] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Volume 44. Cambridge University Press, 2017 (cited on pages 49, 50, 53, 75).
- [21] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian Nonparametrics*. New York: Springer, 2003 (cited on pages 53, 75).
- [22] Nils Lid Hjort et al. *Bayesian nonparametrics*. Volume 28. Cambridge University Press, Apr. 2010. URL: <http://www.cambridge.org/us/academic/subjects/statistics-probability/statistical-theory-and-methods/bayesian-nonparametrics> (cited on pages 53, 75).
- [23] H. Ishwaran and L.F. James. “Gibbs sampling methods for stick-breaking priors”. In: *Journal of the American Statistical Association* 96.453 (2001), pages 161–173. ISSN: 0162-1459 (cited on page 63).
- [24] D. P. Kingma and M. Welling. “Stochastic gradient VB and the variational auto-encoder”. In: *Second International Conference on Learning Representations, ICLR*. Volume 19. 2014, page 121 (cited on pages 43, 44).
- [25] Diederik P Kingma, Tim Salimans, and Max Welling. “Variational dropout and the local reparameterization trick”. In: *Advances in Neural Information Processing Systems*. 2015 (cited on page 93).
- [26] J. Knoblauch, J. Jewson, and T. Damoulas. “An optimization-centric view on Bayes’ rule: Re-viewing and generalizing variational inference”. In: *Journal of Machine Learning Research* 23.132 (2022), pages 1–109 (cited on pages 44, 45).
- [27] D. Kreps. *Notes on the Theory of Choice*. Westview press, 1988 (cited on page 25).
- [28] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. “Stochastic Gradient Descent as Approximate Bayesian Inference”. In: *J. Mach. Learn. Res.* 18.1 (Jan. 2017), pages 4873–4907. ISSN: 1532-4435 (cited on page 50).

- [29] Marina Meilă. “Comparing clusterings—an information based distance”. In: *Journal of Multivariate Analysis* 98.5 (2007), pages 873–895 (cited on page 65).
- [30] Jeffrey W Miller and Matthew T Harrison. “A simple example of Dirichlet process mixture inconsistency for the number of components”. In: *Advances in neural information processing systems*. 2013, pages 199–206 (cited on pages 63, 64).
- [31] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. “Variational dropout sparsifies deep neural networks”. In: *International Conference on Machine Learning*. 2017 (cited on page 93).
- [32] K. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012 (cited on pages 17, 41, 43).
- [33] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL: <http://probml.github.io/book2> (cited on page 49).
- [34] Eric Nalisnick, José Miguel Hernández-Lobato, and Padhraic Smyth. “Dropout as a structured shrinkage prior”. In: *International Conference on Machine Learning*. 2019 (cited on page 94).
- [35] R. M. Neal. “Markov chain sampling methods for Dirichlet process mixture models”. In: *Journal of Computational and Graphical Statistics* 9 (2000), pages 249–265 (cited on page 35).
- [36] Radford M Neal. *Bayesian learning for neural networks*. Springer Science & Business Media, 1996 (cited on page 94).
- [37] Radford M Neal. “Markov chain sampling methods for Dirichlet process mixture models”. In: *Journal of computational and graphical statistics* 9.2 (2000), pages 249–265 (cited on page 63).
- [38] Mark EJ Newman. “Power laws, Pareto distributions and Zipf’s law”. In: *Contemporary physics* 46.5 (2005), pages 323–351 (cited on pages 66, 67).
- [39] J. Paisley, D. M. Blei, and M. I. Jordan. “Variational Bayesian Inference with Stochastic Search”. In: *International Conference on Machine Learning (ICML)*. 2012 (cited on page 43).
- [40] G. Parmigiani and L. Inoue. *Decision theory: principles and approaches*. Volume 812. John Wiley & Sons, 2009 (cited on pages 15, 23, 26).
- [41] Jim Pitman. “Poisson-Kingman partitions”. In: *Lecture Notes-Monograph Series* (2003), pages 1–34 (cited on page 68).
- [42] Jim Pitman and Marc Yor. “The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator”. In: *The Annals of Probability* 25.2 (1997), pages 855–900 (cited on page 67).
- [43] Łukasz Rajkowski. “Analysis of the maximal a posteriori partition in the Gaussian Dirichlet Process Mixture Model”. In: *Bayesian Analysis* 14.2 (2019), pages 477–494 (cited on page 65).
- [44] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. DOI: [10.1.1.86.3414](https://doi.org/10.1.1.86.3414) (cited on page 49).
- [45] D. J. Rezende, S. Mohamed, and D. Wierstra. “Stochastic backpropagation and approximate inference in deep generative models”. In: *International conference on machine learning*. PMLR. 2014, pages 1278–1286 (cited on page 44).
- [46] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 2004 (cited on pages 31, 33).

- [47] Håvard Rue, Sara Martino, and Nicolas Chopin. “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations”. In: *Journal of the royal statistical society: Series b (statistical methodology)* 71.2 (2009), pages 319–392 (cited on page 49).
- [48] L. J. Savage. *The Foundations of Statistics*. John Wiley & Sons, 1954 (cited on pages 23, 25).
- [49] M. J. Schervish. *Theory of statistics*. Springer Science & Business Media, 2012 (cited on pages 15, 26).
- [50] Jayaram Sethuraman. “A constructive definition of Dirichlet priors”. In: *Statistica Sinica* 4 (1994), pages 639–650 (cited on page 60).
- [51] Y.W. Teh et al. “Hierarchical Dirichlet processes”. In: *Journal of the American Statistical Association* 101.476 (2006), pages 1566–1581. ISSN: 0162-1459 (cited on page 72).
- [52] Sara Wade and Zoubin Ghahramani. “Bayesian cluster analysis: Point estimation and credible balls (with discussion)”. In: *Bayesian Analysis* 13.2 (2018), pages 559–626 (cited on pages 65, 66).
- [53] A. Wald. *Statistical decision functions*. Wiley, 1950 (cited on page 16).