A Concise Course in Bayesian Nonparametrics

Hengrui Luo, with Help by Steve.N.MacEachern

Contents

Part 1

Bayesian Theory

CHAPTER 1

Space of Probability Measures

Reference The basic materials are from [**1**]Chap.1-2. For the analytical aspects I refer to [**5**]. The discussion of exchangeability is adapted from [**31**]. Although I provided a brief introduction to the necessary material to the stochastic process, a good reference is Chap XI of [**3**]. A classical reference for $M(\mathcal{X})$ is [**63**]

Let us make a review of the procedure of a standard parametric Bayesian analysis.

We choose a prior $\Pi(\theta)$ from a class of probability distributions, and then integrate the observed information $\mathbf{X} = (X_1, \cdots, X_n)$ via the likelihood function $L(\theta|\mathbf{X}) \coloneqq \Pi(\theta) \cdot f(\mathbf{X}|\theta)$. This likelihood contains all information from prior and the observation. From the likelihood we can derive marginal distribution $m(\mathbf{X}) \coloneqq \int_{\mathcal{X}} L(\theta|\mathbf{X}) d\mathbf{X}$ and via *Bayes Theorem* the posterior distribution $\Pi(\theta|\mathbf{X}) \coloneqq \frac{L(\theta|\mathbf{X})}{m(\mathbf{X})}$.

When the $supp\Pi$ is of finite dimension, we usually deal with it under the name of *parametric model*; when $supp\Pi$ is of infinite dimension, we will make use of nonparametric model.[1] [2]

This procedure leads to following immediate problems:

(1) What kind of priors are reasonable?
   (a) Must its range cover all possible parameter values? i.e. the wholity of Θ.
   (b) Should it be compactly supported? i.e. $supp\Pi$ is compact.
   We leave this problem to next Chapter where we have a closer look into the space $M(\mathcal{X})$, the space consisting of all distributions defined on $\mathcal{X}$.

(2) Does the marginal distribution $m(\mathbf{X})$ always exist?
   (a) Is the likelihood function always integrable? i.e. $L(\theta|\mathbf{X})$ always integrable?
   (b) When it is not integrable, what conclusion we can draw from it? i.e. Is the formal marginal distribtuion also useful somehow?

---

[1] Prof.MacEachern pointed out that the Bayesian inference is not simple inference on posterior mean or mode, but the inference based on posterior distribution along with a decision theoretic framework setup. The criticism he made is that the techniques from machine learning usually produce an inadmissible decision rule. When that happens, we can often improve such a technique by averaging or making it as an approximation to the Bayesian decision rule. For example, the jiggling of λs in the penalized likelihood problems like [**136**] can be regarded as an approximation to a Bayesian procedure.

[2] As indicated by Prof.MacEachern [**12**], inadmissible rules should only be used for (1) accounting the gap caused by data (2) the construction of an admissible rule is not available or too complicated.(Although we know its existence) (3) the computational cost of admissible rule is way too high (even if a mixture use of MCMC and approximation [**139**]) (4) other practical reasons like policy restrictions will prevent an admissible rule to be used.

This problem partially depends on the analytic behavior of the integrand $L(\theta|\mathbf{X})$, and hence on the choice of $\Pi(\theta)$ and $f(\mathbf{X}|\theta)$. So in order to answer this question we must have an understanding of interaction between these two.

(3) How will the posterior distribution depends on the choice of $\Pi(\theta)$ and $f(\theta|\mathbf{X})$? For an answer to this problem, we have to consider approximation and consistency theorems.

From now on we switch to the rigid notations where capital letters P, Q, F means probability measures or c.d.f.[3], and the lower letters p, q, f means Radon-Nikodym derivatives or p.d.f. If not explicitly mentioned, the $f(x) = \frac{dF}{dx}(x)$ w.r.t. the Lebesgue measure is always assumed if the underlying space is $\mathbb{R}^n$. We must be very careful to these notations, things like $dP(x)$[4] must be understood carefully. [5]

1.1. The Space $M(\mathcal{X})$

Points L^1 metric/Hellinger metric.;Weak convergence; $\mathcal{B}_M$ algebra.

We begin by looking at the basic space we are concerning of. We introduce a few metrics on this space, choose the weak topology on it and $\sigma-$algebra on it with some reasons.

The sample space $\mathcal{X}$ is chosen to be a completely separable space [6] [7] If it is metrizable or even there is a metric on it, we call it a *Polish space*.

Among many possible choices $\mathcal{X} = \mathbb{R}$ and $\mathcal{X}$ where there are only finite elements are of major interest.

Definition 1.1. (Space of probability measures) Given a complete separable sample space $\mathcal{X}$ with a *Borel σ-algebra* $\mathcal{B}$ defined on $\mathcal{X}$.

The set of all probability measures P defined on $(\mathcal{X}, \mathcal{B}, P)$ which makes the topological space into a probability space is called the *space of probability measures* over $\mathcal{X}$. We denote it as $M(\mathcal{X})$. We understand the notion $M(\bullet)$ like an operation, like the operation of taking a dual space. When the $\mathcal{X}$ is not a complete separable space, we cannot necessarily call the $\sigma-$algebra a Borel algebra, and denote it only as $\mathcal{B}$.

The subspace L_μ is the collection of all probability measures dominated by a σ-finite measure μ defined on $\mathcal{X}$[8]. This is a closed subspace of $M(\mathcal{X})$ due to the dominant convergence theorem.

---

[3] I do not suggest the usage of c.d.f, and [**1**] abused this notation from place to place.

[4] which means $\frac{dP}{dx}(dx)$ in some literature

[5] One useful comment made by Prof.MacEachern is that we must choose our model carefully using the amount.

[6] Since a second-countable space is separable, therefore any complete topological manifold can serve as a sample space because they are all second-countable.

[7] Completeness ensures that any limit of events will remain within the $\mathcal{X}$. separable ensures that the limit is well defined within the $\mathcal{X}$.

[8] Which is equivalent to say $\nu \ll \mu, \forall \nu \in L_\mu$, ν is absolutely continuous w.r.t. μ. i.e. $\forall B \in \mathcal{B}, \nu(B) = 0 \Rightarrow \mu(B) = 0$

COROLLARY 1.2. *The space $M(M(\mathcal{X}))$ consists of all possible priors for $\forall P = p(x|\theta)dx \in M(\mathcal{X})$, which is to say $M(M(\mathcal{X}))$ is the* space of possible priors *and* space of resulting posteriors.[9] [10]

DEFINITION 1.3. (L^1 metric) The L^1 metric(OR *total variation metric*) defined on $M(\mathcal{X})$ is $L^1(P,Q) = \|P-Q\|_1 = 2sup_{B\in\mathcal{B}}\left|P(B)-Q(B)\right|, \forall P,Q \in M(\mathcal{X})$, the inner term is simply the absolute value or L^1 norm on $\mathbb{R}$.

DEFINITION 1.4. (Hellinger metric) The Hellinger metric defined on $M(\mathcal{X})$ is $H(P,Q) \coloneqq \sqrt{\int_{\mathcal{X}}(\sqrt{\frac{dP}{dx}}-\sqrt{\frac{dQ}{dx}})} = \sqrt{2(1-A(P,Q))}, \forall P,Q \in M(\mathcal{X})$, where $A(P,Q) \coloneqq \int_{\mathcal{X}}\sqrt{\frac{dP}{dx}\frac{dQ}{dx}}dx$ is known as *affinity* between P,Q. This metric is defined for those P,Q which are dominated by Lebesgue or other measure μ defined on $\mathcal{X}$.

The following result justifies the term "affinity" because it endows $(M(\mathcal{X}))^n$ with an affine space structure under the Hellinger metric.

LEMMA 1.5. *([**1**]p.13)In the i.i.d case, $A(P^n,Q^n) = \int_{\mathcal{X}^n}\sqrt{\left(\frac{dP}{dx}\right)^n\left(\frac{dQ}{dx}\right)^n}(dx)^n = A(P,Q)^n$, and therefore $H^2(P^n,Q^n) = 2(1-A(P,Q)^n)$.*

LEMMA 1.6. *(Hellinger metric is equivalent to L^1metric[**1**]p.14)$\|P-Q\|_1 \leq 4H^2(P,Q) \leq 4\|P-Q\|_1$*

DEFINITION 1.7. (Setwise convergence) Let $\{P_n\} \in M(\mathcal{X})$ be a sequence of measures, we say P_n converge to P setwisely if $P_n(B) \to P(B), \forall B \in \mathcal{B}$. Pointwise convergence of P_n to P for all sets in a σ-algebra does not ensure the setwise convergence for all $B \in \mathcal{B}$.

COROLLARY 1.8. *([**1**]p.59)If $\mathcal{X}$ is a complete separable metric space, then $M(\mathcal{X})$ is neither complete NOR separable space under setwise convergence topology.*

DEFINITION 1.9. (Weak convergence) Let $\{P_n\} \in M(\mathcal{X})$ be a sequence of measures, we say P_n converge to P weakly if $\int_{\mathcal{X}} f(x)\frac{dP_n}{dx}(dx) \to \int_{\mathcal{X}} f(x)\frac{dP}{dx}(dx), \forall f$ bounded continuous function on $\mathcal{X}$. Or we can consider the functional $L_P : f \mapsto \int_{\mathcal{X}} f(x)\frac{dP}{dx}(dx)$, then the weak convergence of a sequence of measures $\{P_n\}$ is defined when its corresponding functionals L_{P_n} is weak*-ly[11] convergent w.r.t. the space of bounded continuous functions on $\mathcal{X}$. The $M(\mathcal{X})$ is regarded as a subspace of the "dual space" of the sample space $\mathcal{X}$. [**6**]

THEOREM 1.10. *(Portmanteau Theorem, [**1**]p.12) The followings are equivalent for complete separable metric space $(\mathcal{X},\mathcal{B})$ with a Borel algebra $\mathcal{B}$[12]:*
(1) $P_n \to P$ weakly.
(2) $\int_{\mathcal{X}} f(x)dP_n(x) \to \int_{\mathcal{X}} f(x)dP(x), \forall f$ bounded continuous function on $\mathcal{X}$.
(3)$limsupP_n(F) \leq P(F), \forall F \in \mathcal{B}$ closed.

[9] And all the posteriors will lie in a subspace of $M(\mathcal{X})$.

[10] This space also include the improper priors if we take the Harmonic analysis into account, yet such a consideration is not explored here.

[11] Recall that the strong/weak convergence is about the space X and weak* convergence is about the space X^*.

[12] Only on complete separable metric space we can use $\mathcal{A}$ and $\mathcal{B}$ interchangeably to represent the Borel σ-algebra.

(4)$liminf P_n(U) \geq P(U), \forall U \in \mathcal{B}$ *open.*
(5) $limP_n(B) = P(B), \forall B \in \mathcal{B}$ *with regular boundary* $P(\partial B) = 0$.

DEFINITION 1.11. (Tightness) A collection of measures $\{P_n\} \in M(\mathcal{X})$ on a topological space $\mathcal{X}$ is called *tight* if $\forall \epsilon > 0$ there exists $K_\epsilon \subset_{compact} \mathcal{X}$ such that $|P_n|(\mathcal{X} \backslash K_\epsilon) < \epsilon$. If $\{P_n\}$ is a collection of probability measure then the last condition can also be written as $P_n(K_\epsilon) > 1 - \epsilon$.

The following well-known result told us when a sequence have weak limit on the $M(\mathcal{X})$ when $\mathcal{X}$ is a complete separable metric space.

THEOREM 1.12. *(Prohorov Theorem,* [**1**]*p.13) For a complete separable metric space* $\mathcal{X}$ *with a Borel algebra* $\mathcal{B}$*, every subsequence of* $\{P_n\} \in M(\mathcal{X})$ *has a weakly convergent subsequence iff* $\{P_n\} \in M(\mathcal{X})$ *is tight.*

The following well-known result directly characterized the weak convergence on the $M(\mathcal{X})$ when $\mathcal{X}$ is a complete separable metric space. Loosely speaking, it said that if the weak convergence is consists of well-behaved measures then we can find a sequence of random variables s.t. the distribution of these random variables are exactly the weak convergent sequence. In other words, we find a sequence of random variables in $\{X_n\} \in \mathcal{X}$ to "represent" the weak sequence $\{P_n\} \in M(\mathcal{X})$

THEOREM 1.13. *(Skorokhod Representation Theorem) For a complete separable metric space* $\mathcal{X}$ *with a Borel algebra* $\mathcal{B}$*, if a sequence* $\{P_n\} \in M(\mathcal{X})$ *has a weak limit* P *iff there exists* $\mathcal{X}$*-valued random variable sequence* $\{X_n\}$ *and a random variable* X *defined on some new probability space with probability measure* P_{new} *s.t.* $P_{new} \circ X_n^{-1} = P_n, P_{new} \circ X^{-1} = X$ *and* $P_{new}(\{\omega \in \mathcal{X} : X_n(\omega) \to X(\omega)\}) = 1$.

THEOREM 1.14. *(*[**1**]*p.13)If* $\mathcal{X}$ *is a complete separable metric space, then* $M(\mathcal{X})$ *is a complete separable metric space under weak convergence topology.*

The separability and metriz-ability is the foundation of the sensitive analysis[13]. We want to measure how far two models are and of course we want to take limits of models.

The above results Theorem 1.14 and Corollary 1.8 about $M(\mathcal{X})$ where $\mathcal{X}$ is a complete separable (metric) space can be summarized as below chart.

However, these two topologies can be connected as we shall see in next chapter[14].

Topology	Complete	separable
Set-wise convergence	No	No
Weak convergence	Yes	Yes

Therefore we want to use weak topology on the $M(\mathcal{X})$ for obvious reasons. One word to mention is that the L_μ under L^1metric is also complete separable although the $M(\mathcal{X})$ under L^1 metric is not. So from now on we talked about $M(\mathcal{X})$ with weak topology. Now we decide a $s\sigma-$algebra on $M(\mathcal{X})$. A natural choice is induced by considering the duality, which is to say, the functional $l_f : f \mapsto \int_{\mathcal{X}} f(x) \frac{dP}{dx}(dx)$

[13]Sensitive analysis is analysis of the posterior w.r.t. the choice of priors, and nowadays the sensitive analysis is also done on likelihood ratios.

[14][**65**]

regarded as a functional of P with f measurable on $(\mathcal{X}, \mathcal{B}, P)$. We define the $\sigma-$algebra generated by all these l_f as the desired $\sigma-$algebra $\mathcal{B}_M$ on $M(\mathcal{X})$.

$$BC(\mathcal{X}) \overset{\overset{L_P}{\rightharpoonup}}{\underset{l_f}{\leftharpoondown}} M(\mathcal{X})$$

THEOREM 1.15. *([1]pp.61-62)The collection of discrete/continuous probability measures forms a $\mathcal{B}_M$-measurable set of $M(\mathcal{X})$.*

1.2. The Space $M(M(\mathcal{X}))$

Points Dirichlet density; Polya tree; Stochastic process; Kolmogorov's Extension Theorem; Tail-free priors.

Due to our notations, $M(M(\mathcal{X}))$ is the space consisting of all (prior) measures on $M(\mathcal{X})$. We are to prove that $M(M(\mathcal{X}))$ contains some elements of interest, in fact, it contains a lot. And sometimes the elements in $M(M(\mathcal{X}))$ will be referred as "random measures" but we will not use that term. That term means that we are considering random variables η whose value is taken in $M(\mathcal{X})$, but this is the same as $M(M(\mathcal{X}))$ because such an random variable η should also follow some distribution $\frac{dP_\eta}{d\eta}$. This distribution corresponds to probability measures P_η over $M(\mathcal{X})$, $P_\eta \in M(M(\mathcal{X}))$.

We begin with a theorem describing the weak topology on $M(M(\mathcal{X}))$ when $\mathcal{X} = \mathbb{R}$.

LEMMA 1.16. *([1] p.78) Let $\mathcal{B}_0 = \{B_i, i \in I\}$ be a family of sets closed under intersections that generates the $\mathcal{B}$ for $(\mathcal{X}, \mathcal{B}, P)$. If for every $B_1, B_2, \cdots B_k \in \mathcal{B}_0$, $(P(B_1), P(B_2), \cdots P(B_k))$ has the same distribution under Π_1 and Π_2, $\Pi_i \in M(M(\mathbb{R}))$, then $\Pi_1 = \Pi_2$.*

The $\Pi : M(\mathbb{R}) \to [0,1], P \mapsto [0,1]$ is defined on the base space $M(\mathbb{R})$ and therefore $(P(B_1), P(B_2), \cdots P(B_k))$ should be regarded as a random vector with value in $\mathbb{R}^k$ when $B_1, B_2, \cdots B_k$ is a fixed collection of measurable sets.

DEFINITION 1.17. (Weak convergence on $M(\mathbb{R})$,[1] p.79) A sequence of probability measure $\{\Pi_n\}$ is said to converge weakly to a probability measure Π if $\int \phi(P) d\Pi_n(P) \to \int \phi(P) d\Pi(P)$ for any bounded continuous function on $M(\mathbb{R})$. There is no explicit form for all bounded continuous functions on $M(\mathbb{R})$. Only on a complete separable space can weak convergence be described using the collection of all bounded continuous functions, for a general space $\mathcal{X}$ the weak(or weak*) convergence on $M(M(\mathcal{X}))$ do not have such a description.

THEOREM 1.18. *(Sethuraman-Tiwari,[1] p.79) A family of probability measures $\{\Pi(t, \bullet) : t \in T\}$ on $M(\mathbb{R})$ is tight w.r.t. weak convergence on $M(\mathbb{R})$ iff the family of corresponding expectations $\left\{E_{\Pi(t,\bullet)}(B) : t \in T\right\}$ is tight in $\mathbb{R}$ for $\forall B \in \mathcal{B}$.*

1.2.1. $\mathcal{X}$ Finite. In this case $\mathcal{X} = \{1, 2, \cdots k\}$ and the $M(\mathcal{X})$ is a probability k-simplex whose dimension is $k-1$. This is the *simplex representation*, which is commonly used in computer vision and document classification. This representation leads to the Dirichlet process.

DEFINITION 1.19. (Dirichlet distribution) The finite dimensional Dirichlet distribution with parameter $(p_1, p_2, \cdots, p_k) \in \mathbb{R}^k$ is a density w.r.t. the induced $\mathbb{R}^k$ Lebesgue measure defined on the probability k-simplex with value $\frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdots p_{k-1}^{\alpha_{k-1}} (1 - \sum_{i=1}^{k-1} p_i)^{\alpha_k}$ at $(\alpha_1, \alpha_2, \cdots, \alpha_k) \in k$-simplex.

Dirichlet distribution is the most common density when the observations are exchangeable. In fact, using the Theorem 1.31, we yield the prior of a finite exchangeable sample:

PROPOSITION 1.20. *([**1**] p.64) Let $X_1, X_2, \cdots X_n$ be $\mathcal{X}$-valued random variables where $\mathcal{X}$ is finite. If these random variables are exchangeable, then their prior must be a Dirichlet density.*

After defining the Dirichlet distribution for the finite dimensional case we can easily extend it to an infinite case $M(\mathcal{X})$. So taking the limit of dimension k of a finite k-Dirichlet distribution will give us some sort of ∞-Dirichlet distributions which are known as Dirichlet process later.

Another "parameterization" of $M(\mathcal{X})$ is the *binary expansion representation*[15] with k-decimals like $0.1111111, 0.1010101$. This is widely used in real variable theory, coding theory, Brownian motion and cartography. The elements in $M(M(\mathcal{X}))$ can be regarded as measure on the 2-adic k-cube $\{0,1\}^k$. When we let $k \to \infty$ as we did for the Dirichlet distribution, this representation leads to the Polya tree process, which is a more general prior than Dirichlet process because we can not only exchange random variables but also assign a more flexible p_i's as we did for a Dirichlet process.

The following subsection will construct priors making use of both representations.

1.2.2. $\mathcal{X} = \mathbb{R}$**.** In following, we introduce two ways of constructing priors on $M(\mathcal{X})$. One way is to use Dirichlet density which allows us to prescribe probabilities on a finite collection of categories(or their unions) of resulting prior; the other way is to use binary expansions which allows a tail-free structure of resulting prior.

To make definitions more understandable, we gave a brief introduction to the *stochastic process* here, among which the *Markov chain* mentioned later is a special case.

DEFINITION 1.21. (Stochastic process) On a probability space $(\mathcal{X}, \mathcal{B}, P)$ and a measurable space $(S, \mathcal{N}, P')$, an S-valued *stochastic process* is a collection of S-valued random variables on $\mathcal{X}$ indexed by a *totally ordered* set T, which is sometimes known as the *time variable*.

We use the notation $\{X(t, \bullet) : T \times \mathcal{X} \to S, t \in T\}$ OR simply $X(t, \bullet)$ when the underlying $T, (\mathcal{X}, \mathcal{B}, P)$ is clear. The space $\mathcal{X}$ is called the *sample space*; T is called the *time space*; S is called the *state space*. [16]

---

[15] Also known as 2-adic expansion, dyadic expansion or Cantor splitting.

[16] An interesting application of stochastic process is brought to me by Prof.MacEachern. The edge point or break point problem in the linear model theory is generally difficult to handle. We have no idea whether there is an "authentic" break point or not. Details can be founded in Sec.8 of [**125**].

This idea is sometimes referred as the principle of parsimony, or simply *Occam's razor*.

DEFINITION 1.22. (Random measure)A *random measure* on $(\mathcal{X}, \mathcal{B}, P)$ is defined at a fixed time $t \in T$ for the stochastic process $X(t, \bullet)$.

It is not hard to see Markov chain is a stochastic process with T and S countable. And in the following discussions, we want to extend $\mathcal{X}$ into uncountable cases so basically S will be chosen as $[0, 1]$.

Let $F(t) \coloneqq P(T \in (-\infty, t])$ in the following construction.

THEOREM 1.23. *(Kolmogorov's Extension Theorem) Let $T \subset \mathbb{R}$, for a string $\omega = (t_1, t_2, \cdots, t_k), t_i \in T$ we define a probability measure ν_ω on $((\mathbb{R}^n, \mathcal{B}^n, P))^k$ such that v_ω satisfy following "consistency conditions":*

(i) $v_\omega(B_1, B_2, \cdots B_k) = \nu_{\pi(\omega)}(B_{\pi(1)}, B_{\pi(2)}, \cdots B_{\pi(k)}) \forall \pi \in Perm(k), B_i \in \mathcal{B}^n$

(ii)$v_\omega(B_1, B_2, \cdots B_k) = \nu_{\bar{\omega}}(B_1, B_2, \cdots B_k, \mathbb{R}^n, \cdots \mathbb{R}^n) \forall \bar{\omega} = (\omega, t_{k+1}, \cdots t_N), B_i \in \mathcal{B}^n$

Then there exists a stochastic process $X(t, \bullet)$ defined on a probability space $(\mathcal{X}, \mathcal{B}, P)$with values in $\mathbb{R}^n$ that $X : T \times \mathcal{X} \to \mathbb{R}^n$ and $v_\omega(B_1, B_2, \cdots B_k) = P(X_{t_1} \in B_1, \cdots X_{t_k} \in B_k), B_i \in \mathcal{B}^n$.

We knew from Egorov's Theorem that once we define a real valued (Lebesgue) measurable function on a dense set of $\mathbb{R}^n$ or complete separable metric space $\mathcal{X}$, then we can determine it up uniformly to a zero (Lebesgue) measure set. Therefore to define a probability measure on $M(\mathcal{X})$ it suffices to define a probability measure on a dense set $\mathcal{F}^* = [0, 1]^{\mathcal{X}}$ of $M(\mathcal{X})$. This extension theorem told us that we can always find a probability space to "adapt" a finite sequence as marginal distributions. Thus, we yield Theorem 2.3.2 in [**1**]. And we can also yield variants:

THEOREM 1.24. *(Construct a prior on $M(\mathbb{R})$ using Dirichlet density,*[**1**] *p.67) Let $t_1 < t_2 < \cdots < t_k$ be a sequence $\omega = (t_1, t_2, \cdots, t_k)$ in $\mathbb{R}$, Π_ω is a probability measure on the $S_k \coloneqq \{(p_1, p_2, \cdots p_k) : p_i \geq 0, \sum_i p_i \leq 1\}$ such that*

(i)If $lim_{n\to\infty}(t_{1n}, t_{2n}, \cdots t_{kn}) \to \omega$, then $lim_{n\to\infty}\Pi_{(t_{1n}, t_{2n}, \cdots t_{kn})} \to \Pi_\omega$

(ii) $\int_{\mathbb{R}^{N-k}} \frac{d\Pi_{\bar{\omega}}}{dx}(\omega, t_{k+1}, \cdots t_N)(dt_{k+1} \cdots dt_N) = \Pi_\omega$.

(iii)$lim_{t\downarrow-\infty}\Pi_t = 0, lim_{t\uparrow\infty}\Pi_t = 1$

then there exists a probability measure Π on $M(\mathbb{R})$ such that for every $t_1 < t_2 < \cdots < t_k$, $\omega = (t_1, t_2, \cdots, t_k)$ in $\mathbb{R}$ the probability measure of

$$(F(t_1), F(t_2) - F(t_1), \cdots F(t_k) - F(t_{k-1})) =_d \Pi_\omega$$

A concrete constructive version of this existence theorem is Theorem 4.22 in later chapters. The construction is very similar to the Lusin's theorem in real variable theory[**5**].

DEFINITION 1.25. (Tail-free measure, [**38**] Def.2 [**40**]Def.3.2.[17]) A probability measure P on $(\mathbb{R}, \mathcal{B}, P)$ is said to be tail-free with respect to a *binary partition*[18] of $(\mathbb{R}, \mathcal{B}, P)$ if the corresponding $\sigma-$algebras of events.

$\{B_0\}$,$\{B_{00}|B_0, B_{10}|B_1\}$,$\{B_{000}|B_{00}, B_{010}|B_{01}, B_{100}|B_{10}, B_{110}|B_{11}\} \cdots$

are mutually independent under P. If a prior measure on $M(\mathcal{X})$ is tail-free as a probability measure, then we call it a *tail-free prior*. Similarly, we can define a tail-free measure with respect to any *nested partition* defined on a sample probability space $(\mathcal{X}, \mathcal{B})$ as long as the nested hierachy is well-defined and non-intersecting.

[17][**1**] p.71 is a misleading typo.

[18]*i.e. $B_{\epsilon 0} \cup B_{\epsilon 1} = B_\epsilon$ with $B_{\epsilon *} \subset B_\epsilon$*

According to [**40, 57**], this concept can be extended to complete separable $\mathcal{X}$ as well.

A random measure is said to be tail-free w.r.t $(s, +\infty) \subset T = \mathbb{R}$ if for all $s = t_0 < t_1 < \cdots < t_k$ there exists nonnegative random variables $Y_1, Y_2, \cdots Y_k$ which are independent of $\{F(t), \forall t \leq s\}$ such that the joint distribution of the random vector

$$(F(t_1), F(t_2), \cdots F(t_k)) =_d$$
$$\left(F(s) + [1 - F(s)] \left[1 - \prod_{j \leq 1} (1 - Y_j) \right], \right.$$
$$F(s) + [1 - F(s)] \left[1 - \prod_{j \leq 2} (1 - Y_j) \right],$$
$$\left. \cdots F(s) + [1 - F(s)] \left[1 - \prod_{j \leq k} (1 - Y_j) \right] \right)$$

. Intuitively, this means that the joint distribution of tail measure sequence $(F(t_1), F(t_2), \cdots F(t_k))$ is independent of previous events.

An alternative way of stating that Dirichlet process is tail-free is by studying its inducing structure of a Dirichlet process.

THEOREM 1.26. *(D_α is tail-free,* [**55**] *Prop.2) Let D_α be a Dirichlet process defined on $M(\mathcal{X}, \mathcal{B})$ and $B \in \mathcal{B}$ such that $D_\alpha(B) = m$, the conditional distribution of $\frac{1}{m} D_\alpha$ restricted to $M(B, \mathcal{B} \mid_B)$ is still a Dirichlet process with concentration measure $\alpha \mid_B$.*

Next we can construct a prior on $M(\mathbb{R})$ with tail-free structure other than Dirichlet process.

THEOREM 1.27. *(Construct a prior on $M(\mathbb{R})$ using binary expansion,* [**1**] *p.71) Let B_ϵ be a binary partition of $(\mathbb{R}, \mathcal{B}, P)$ and $Y_\epsilon = P(B_{\epsilon 0} | B_\epsilon)$ be a family of $[0, 1]$-valued random variables such that*

(i)$\sigma \{Y_{\epsilon'}\}, \epsilon'$ with different lengths are mutually independent.

(ii) $Y_{\epsilon' 0} \cdot Y_{\epsilon' 00} \cdot Y_{\epsilon' 000} \cdots = 0 = Y_1 Y_{11} Y_{111} \cdots$.

then there exists a tail free *probability measure Π on $M(\mathbb{R})$ w.r.t. B_ϵ.*

We will revisit the notion and motivation of tail-free priors when we talked about the Polya tree in later chapters, for now we can regard it as a technical requirement which allows us to use Kolmogorov's extension theorems as Freedman suggested in [**38**].[19]

THEOREM 1.28. *(Kolmogorov 0-1 Law) For independent $\sigma-$subalgebras $\mathcal{B}_n \subset \mathcal{B}$ on $(\mathcal{X}, B, P)$, $P(\cap_{i=1}^{\infty} \sigma \left(\cup_{j=i}^{\infty} \mathcal{B}_j \right)) = 0 \, OR \, 1$*

COROLLARY 1.29. *(0-1 Law,* [**1**] *p.75) Suppose Π to be a tail-free prior on $M(\mathbb{R})$ w.r.t. binary partition B_ϵ. Let λ be any finite measure on $(\mathbb{R}, \mathcal{B})$ with $\lambda(B_\epsilon) > 0, \forall \epsilon$.*

If $Y_\epsilon = P(B_{\epsilon 0} | B_\epsilon), 0 < Y_\epsilon < 1, \forall \epsilon$ then $\Pi \{P \ll \lambda\} = 0 \, OR \, 1$.

[19] Actually, the Dirichlet process is also tail-free from the expression of $(F(t_1), F(t_2), \cdots F(t_k))$ in above theorems, for details see Theorem 3.17 and its implication that the tail-free prior leads to the tail-free posterior.

1.3. Exchangeability

Points exchangeability; de Finetti's theorem; Markov chain; ergodic decomposition theorem.

We have a brief touch on de Finetti's subjective theory, see how it applies to simple exchangeable random variable sequence. We explain how they can be regarded as special cases of ergodic theorem. If one wish, we can introduce the Bayesian framework without touching the exchangeability, yet my personal favor told me that we should know Ergodic theory somehow.

1.3.1. de Finetti's Subjective Theory. De Finetti's argument that Kolmogorov's axiomatic system of probability is equivalent to a *coherent prevision* in a gambler's mind is striking. According to de Finetti, the seemingly objective axiomatic definition of a probability measure[20] is no more than a coherent prevision function, which is so subjective by definition.

However, instead of refusing to use probability, de Finetti himself developed a system based on the principle of coherence. And this radical frequentist derived the exchangeability theorem which becomes the foundation of Bayes theory nowadays.

Of course there are more than one motivations of Bayes theory [21] [22], one is that if we based our decisions on a set of rational axioms then we must make an optimal decision as a Bayes decision or limit of Bayes decisions; the other one is from the modeling perspective[23] that we might fit a model w.r.t. a certain decision theoretic framework and in that way the Bayes estimators are elicited naturally.[24]

Definition 1.30. (Exchangeability) A sequence of random variables $\{X_i\}_{i=1}^n, X_i : \mathcal{X} \to \mathbb{R}$ on probability space $(\mathcal{X}, \mathcal{B}, P)$ is said to be *exchangeable* w.r.t. the probability measure P if $P(X_i \in B_i) = P(X_{\pi(i)} \in B_i), \forall \pi \in Perm(n), B_i \in \mathcal{B}$[25]. The exchangeability of an infinite sequence$\{X_i\}_{i=1}^\infty$ is understood correspondingly.

Theorem 1.31. *(De Finetti's Theorem) If* $\{X_i\}_{i=1}^n$ *is exchangeable w.r.t.* P *and only take values in* $\{0, 1\}$*, then there exists a probability measure* $\Pi \in M(M(\mathcal{X}))$

---

[20] De Finetti's arguement is concerning about finite countability instead of countably additivity. De Finetti himself was against the assumption of countably additivity and assumes only finite additivity. However, the problem of *Non-conglomerability* for finite-valued, finitely additive probability makes the finitely additivity less preferable than countably additivity as pointed out in [**41**]. In fact, not every finitely additive measure can be extended to countably additive measure on a given probability space. The philosophical problem of whether we should accept coutability axiom is pursued further in decision framework in [**42, 70**]. In this work the author argued that we can make a better decision if we do not know information at all, if the model probability is only finitely additive. This anti-intuitive fact about finitely additive measures urged us to accept countably additive measures instead.

[21] For a detailed discussion about justifications of Bayes theory, refer to the Chap.2 of [**2**]

[22] A more formal way of justification of Bayes principle is one famous result due to D.Basu saying that conditionality principle and sufficiency principle must lead to the likelihood principle. And the likelihood principle is the ground of Bayes analysis.

[23] [**69**]

[24] Sometimes we must assume that some (hyper)parameters are fixed in the modelling procedure, which does not harm our identities of being Bayesians.

[25] $P(X_i \in B_i) := \int_{B_i} X_i(dx) \frac{dP}{dx}(dx)$

on $M(\mathcal{X}) \cong (p, 1-p) \cong (0,1)$ *such that*

$$P\left(X_i = x_i, i = 1, \cdots, n\right) = \int_0^1 \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \frac{d\Pi}{d\theta}(d\theta) = \int_{M(\mathcal{X})} \prod_{i=1}^n P\left(X_i = x_i | p\right) \frac{d\Pi}{dp}(dp)$$

. i.e. With Π *as a prior measure on* θ *we can regard* $\{X_i\}$ *as Bernoulli i.i.d samples.*

Following theorem requires further notions. And these theorems provide ways of decomposing the joint sample measure (likelihood) into individual measures. Such a kind of decompositions are special cases of ergodic decomposition.

THEOREM 1.32. *(Regazzini's Theorem* [**31**]*) If* $\{X_i\}_{i=1}^{\infty}$ *is exchangeable w.r.t.* P *and take values on* $\mathbb{R}$*, then there exists a probability measure* $\Pi \in M(M(\mathcal{X}))$ *on* $M(\mathcal{X})$ *such that*

$$P\left(X_i \in B_i, i = 1, \cdots, n\right) = \int_{M(\mathcal{X})} \prod_{i=1}^n P\left(X_i \in B_i | p\right) \frac{d\Pi}{dp}(dp)$$

$\forall B_i \in \mathcal{B}$ *where* $P\left(X_i \in B_i | p\right)$ *should be regarded as a function of* $p \in M(\mathcal{X})$*.*

A second think on the notion of exchangeability will lead us to think that an infinite sequence of exchangeable random variables (equivalently their corresponding probability measures) are exactly the subspace of $M(\mathcal{X})$ which is invariant under the infinite permutation group S_∞. As long as we regard the study of exchangeability as the study of invariant subspace of S_∞, we can have a unifying view through the rest of this section.

1.3.2. Ergodic Theory. Here we present the ergodic theory in a rather abstract way, for a more intuitive construction with applications in stochastic processes, see [**14**] Chap.9 Sec.5-6. They introduce the most simple case of ergodic theorem.

Ergodic theory is the study of the invariant subspaces of $M(\mathcal{X})$ under some sort of measurable transformation group. Usually statisticians care only the invariant subspace of a single measurable transformation T[26] while mathematicians will care the invariant subspace of a transformation family $\{T_\tau\}$ with a group structure on it.

Loosely speaking, if the law of large numbers is the main large sample result in frequentists' world; the de Finetti's theorem is the main large sample result in Bayesians' world, then the ergodic theorems generalized all these into some more general spaces. We are mostly concerned about the case where $\mathcal{X}$ is complete separable or Polish.

This analogy is based on the fact that all these three results are describing the average behavior of a sequence of random variables in a long run. Its proof can be found in [**34, 35**].

To do that, we extend the i.i.d. random variable sequence a bit.

DEFINITION 1.33. (Markov chain) A *Markov chain* is a sequence of random variables $\{X_n\}$ defined on $(\mathcal{X}, \mathcal{B}, P)$ along with a transitional function $T : \mathcal{X} \times \mathcal{B} \to \mathbb{R}$. The *transitional function* should satisfy that

[26] Like the transitional function of a Markov chain, the invariant measure π is actually the induced measure on the T-invariant subspace.

(i) $T(\bullet, A), \forall A \in \mathcal{B}$ is a measurable function on $(\mathcal{X}, \mathcal{B}, P)$.
(ii) $T(x, \bullet), \forall x \in \mathcal{X}$ is a probability measure on $(\mathcal{X}, \mathcal{B}, P)$.
(iii)$X_n \sim T(X_{n-1}, \bullet)$ holds for the random variable sequence $\{X_n\}$.

Let $T^k(x, \bullet)$ denote the probability measure defined on $(\mathcal{X}, \mathcal{B}, P)$ with initial value x at *time* k. The value of this Markov chain at time k is X_k and the recorded sequence $\omega = X_1, X_2, \cdots, X_k$ is called a *string* of length k OR *path* by time k. The length OR time of a string ω is sometimes denoted as $|\omega|$.

A usual notation which is frequently used in discussion of a Markov chain is τ_x, which is the time at which the chain reach value x.

An i.i.d. random variable sequence can be regarded as a special Markov chain with identical conditional distributions. i.e. $T(x, \bullet)$ is the same for all $x \in \mathcal{X}$. Following concepts, although formalized in setting of Markov chains with discrete time variable, can also be generalized into a general stochastic process with a continuous time variable and continuous states since they are not specifically focus on the at-most countable property provided by a Markov chain.[27]

Definition 1.34. (Measure-preserving transformations, [**34, 32**]) Consider a P-measurable transformation $T : \mathcal{X} \to \mathcal{X}$ from a measure space $(\mathcal{X}, \mathcal{B}, P)$ to itself, we call T a *measure-preserving transformation* if T^{-1} exists and measurable w.r.t. P. A set $E \in \mathcal{B}$ is called a *T-invariant set* if $T(E) = E$. We can equivalently say that P is an invariant measure w.r.t. T, in this way we can justify our previous definition of invariant measure π w.r.t. a Markov chain, which is actually an invariant measure w.r.t. the transitional function $T(\bullet, A) : \mathcal{X} \to \mathcal{X}$ that is always P-measurable by definition of the transitional function of a Markov chain.[28]

Definition 1.35. (Invariant measure, [**34, 32**]) A probability measure π on $(\mathcal{X}, \mathcal{A}, P)$ is said to be *invariant* w.r.t. a Markov chain $\{X_n, T\}$ if $\pi(A) = \int_{\mathcal{X}} T(x, A) \frac{d\pi}{dx}$ holds for $\forall A \in \mathcal{A}$ and the corresponding density $\frac{d\pi}{dx}$ is called the *stationary distribution* for this chain. For a Markov chain, if there exists an invariant measure for this chain, then we call this a *stationary chain*.

Definition 1.36. (Ergodic measure) Consider a P-measurable transformation T from a measure space $(\mathcal{X}, \mathcal{B}, P)$ to itself, T is called an *ergodic transformation* if for any *T-invariant set* $E \in \mathcal{B}$, $P(E) = 0\, OR\, 1$. An **invariant** probability measure π w.r.t. a Markov chain $\{X_n, T\}$is called an *ergodic measure* w.r.t. T if T is an ergodic transformation on $(\mathcal{X}, \mathcal{B}, \pi)$.

Definition 1.37. (T-invariant space of measures) Given the definition of invariant measure w.r.t. a P-measurable transformation T, we can talk about the collection of probability measures $P' \in M(\mathcal{X})$ s.t. T is also P'-measurable and P' are T-invariant measures. The collection of all these P's are denoted as $M_T(\mathcal{X})$.

---

[27] If we prefer to write a Markov chain in notation of stochastic process, we can write $X(t, \bullet)$ and take the time space as $T = \left\{id = T^0, T^1 \cdots\right\}$. This divergence of notation is just a historical issue and we should know both sets of notations. Conversely we can describe a stochastic process using the transitional function as we did for Markov chain, $\tau : S^{\mathcal{X}} \to S^{\mathcal{X}}, X_t \mapsto X_{\tau(t)}$. We want to emphasize that $\tau \in T$ "acts" on the ordered set T because sometimes we want to see the group structure of T.

[28] These synonyms of terminology is due to two different ways of approaching stochastic process. One is the more pure mathematical way which emphasize the invariance; the other is more probabilistic considering the measure-preserving stuff.

COROLLARY 1.38. *The $M_T(\mathcal{X})$ is a convex set.*

By noticing that non-extreme points within a convex set are the only points that cannot be written as a convex combination of two other points in the convex set, we can derive:

THEOREM 1.39. *The ergodic measures w.r.t. T are precisely the extreme points of the convex set $M_T(X)$*[29]*.*

> "The ergodic decomposition theorem states, roughly, that most stationary processes of interest can be viewed as an average or mixture (or composite) of ergodic processes. In other words, most stationary sources are the result of nature randomly choosing a particular source from a class of ergodic sources at time minus infinity, and then sending that particular ergodic source forever." ([**35**])

DEFINITION 1.40. (Empirical probability measure) An *empirical probability measure* ν_ω on $(\mathcal{X}, \mathcal{A}, P)$ defined by a string $\omega = X_1, X_2, \cdots, X_k$ of length k is $\nu_\omega(A) \coloneqq \frac{1}{k}\sum_{i=1}^{k} I_A(X_i) = \frac{1}{k}\sum_{i=1}^{k} \delta_{X_i}(A), \forall A \in \mathcal{A}$.

THEOREM 1.41. *(Ergodic decomposition Theorem* [**35**]*) Assume that the space $\mathcal{X}$ of $(\mathcal{X}, \mathcal{A}, P)$ is countable, then the empirical probability measure ν_ω on $\mathcal{X}$ defined by a string $\omega \in \mathcal{X} - N$ generated by a stationary Markov chain $\{X_n, T\}$ satisfy*

(1)(Ergodic w.r.t π) $N \subset \mathcal{X}, \pi(N) = 0$,

(2)(Average decomposition)$\forall A \in \mathcal{A}, \pi(A) = \int_{\mathcal{X}-N} v_\omega(A) d\pi(\omega)$

, where π is the invariant measure w.r.t. $\{X_n, T\}$.

Further, $\forall \omega_1, \omega_2$, the induced probability measure $\nu_{\omega_1}, \nu_{\omega_2}$ are either identical or mutually singular. So the decomposition is unique up to singularity.

COROLLARY 1.42. *Any two ergodic measures w.r.t. T are either **a.e. identical***[30] *OR **singular**.*

This corollary is a direct observation that if ν_1, ν_2 are not mutually singular. Then on $\forall E \in \mathcal{B}$ either $\nu_1(E) > 0, \nu_2(E) > 0$ or $\nu_1(E) = \nu_2(E) = 0$. If there is a set $F \in \mathcal{B}$ such that $\nu_1(F) > 0, \nu_2(F) > 0$. But ergodic measure is defined as $\nu_1(F) = \int_{\mathcal{X}} I_F \circ \nu_{1\omega} \frac{d\nu_1}{dx} = \int_{\mathcal{X}} I_F \circ \nu_{2\omega} \frac{d\nu_2}{dx} = \nu_2(F)$ the empirical measure on the same set of the same sequence should be the same, the ergodic theorem applies one the first and the third equality.

Examine the definition of a stationary Markov chain we know that by asserting a Markov chain to be stationary we assume there is such an invariant measure π on which the chain's measure $T(x, \bullet)$ can be *averaged* to yield π. Now we know from the above theorem that except for a stationary measure zero set N, the empirical measure ν_ω can also be averaged on π to yield π. [31] This is like saying $\nu_\omega \to T(\omega, \bullet)$ w.r.t. the invariant measure π as the length of ω goes to infinity.[32]

---

[29] A point in a convex set S is *extreme* if it is in the interior of S with a set dimension not higher than S. i.e. the vertices of S.

[30] w.r.t. a common dominating probability measure P.

[31] Notice that the countability assumption of $\mathcal{X}$ is not essential if we have a complete separable space $\mathcal{X}$, we can apply this decomposition on its countable dense subset $\mathcal{X}_0$ and the zero measure set $\mathcal{X} - \mathcal{X}_0$ is negligible after integration over $\mathcal{X} - N = \mathcal{X} - \mathcal{X}_0 + \mathcal{X}_0 - N$.

[32] Try to consider this as sort of law of large numbers.

There is another place where the ergodic theory comes into play[33], the Markov Chain Monte Carlo algorithms. Basically speaking, the aperiodicity is to ensure the convergence in variational norm rather than average variational norm; the irreducible condition is to ensure the law of large numbers.

THEOREM 1.43. *(Convergence of MCMC* [**33**][34]*) Assume that*

(Existence of invariant measure) *The Markov chain* $\{X_n, T\}$ *has an invariant probability measure* π *on* $\mathcal{X}$ *(or equivalently has a* stationary probability distribution $\frac{d\pi}{dx}$*).*

(1) (Doeblin-Doob) If the $\{X_n, T\}$ satisfies Doeblin condition
$\exists k_0 \in \mathbb{Z}, \epsilon > 0, T^{k_0}(x, \bullet) \geq \epsilon \cdot \nu(\bullet), \forall x \in \mathcal{X}$ for *some probability measure* ν *on* $\mathcal{X}$
then there exists an *unique* invariant measure π such that $sup_{C \in \mathcal{A}} |T^n(x, C) - \pi(C)| \leq (1-\epsilon)^{\frac{n}{k_0}-1}, \forall x \in \mathcal{X}$.

(2) (Nummelin-Tierney) If the $\{X_n, T\}$ satisfies
   (a) the $\pi - irreducible$ condition,
   $\forall A \in \mathcal{A}, s.t. \pi(A) > 0$ is accessible[35] from $\forall x \in \mathcal{X}$.
   (b) aperiodic condition
   g.c.d$\{k | \exists \epsilon_k > 0, s.t. T^k(x, \bullet) \geq \epsilon_k \pi(\bullet), \forall x \in \mathcal{X}\} = 1$
   then the invariant measure π such that $sup_{C \in \mathcal{A}} |T^n(x, C) - \pi(C)| \to 0, \forall x \in \mathcal{X}$.

[33]According to Prof.MacEachern, the application of ergodic theorem on the proof of convergence of MCMC is more familiar to statisticians than the ergodic decomposition theorem stated above. But actually these are the same thing, we treat each step of MCMC as transitional function T sending X_n to X_{n+1}. Then the ergodic decomposition theorem will tell us that $\pi(A) = \int_{\mathcal{X}-N} v_\omega(A) d\pi(\omega)$. This expression means that the empirical measure ν_ω for a chain ω given by MCMC will approximate the stationary measure, which is the limit status of MCMC, by averaging over the stationary measure.

[34]There are some problems when actually put MCMC along with Metropolis-Hastings algorithm into practice, one of the most common one is that the parameter space is not supported on an unbounded domain. ***How to conduct the metropolis step when the support of prior(parameter space) is bounded?*** There are following solutions if the parameter space is bounded instead of the whole parameter space.

(1) Individual sampler. We used another random walk generator like truncated version of normal random walk to replace the. The simulation result showed that if we directly used such a sampler then acceptance rate is oddly low.
(2) Modify the model. We can modify the model using the transformation $\eta = log\tau^2$ and the model likelihood becomes $\underbrace{(e^\eta)^{-2-1} \cdot exp\left(\frac{2}{e^\eta}\right) \left|\frac{\partial e^\eta}{\partial \eta}\right|}_{\tau^2 = e^\eta \, prior}$ with Jacobian of the transformation, now the parameter η is supported on the whole $\mathbb{R}$ and the metropolis step can be conducted using the usual normal random walk because $\eta \in \mathbb{R}$. This method is suggested by Prof.K.Calder.
(3) Drop the random walk when it goes out of the support of prior. That is to say, when the normal random walk goes out of the support of the prior we simply rejected the proposal value and stay with the old value of the parameter. This is adopted by my solution above.
(4) Extended the model. This is a remark from [**24**], we can sometimes extend the problem to allow a larger parametric space if the support if actually too small to put a normal random walk on it.

[35]$A \in \mathcal{B}$ is accessible from $x \in \mathcal{X}$ if the probability of the event that reaching time t of A is finite is greater than zero. That is to say, the expected time of reaching A is finite.

(3) (Athreya-Doss-Sethuraman) If the $\{X_n, T\}$ satisfies
   (a) the mild $\pi - irreducible$ condition for some $A \in \mathcal{A}$,
   $\forall A \in \mathcal{A}, s.t. \pi(A\, is\, accessible\, from\, x \in A) = 1\ \forall x \in \mathcal{X}$ exists.
   (b) aperiodic condition on this A
   g.c.d$\big\{k|\exists \epsilon_k > 0, s.t. T^k(x, \bullet) \geq \epsilon_k \pi(\bullet), \forall x \in A\big\} = 1$
   then the invariant measure π such that $sup_{C\in\mathcal{A}} |T^n(x, C) - \pi(C)| \to 0, \forall x \in \mathcal{D}$ where $\mathcal{D} \subset \mathcal{X}$ s.t. $\pi(\mathcal{D}) = 1$.

A typical procedure should first verify the existence of a stationary distribution/invariant measure, and then try to use results from ergodic theory.

In summary, the ergodic theory can be used to describe the behavior of a sequence of random variables in the long run.

In exchangeable setting, it can be used to decompose the underlying probability measure. In MCMC setting, it can be used to guarantee the convergence of simulation sequence. To improve the mixing performance of MCMC, a universally effective way if to block and reduce the dimension of the parameter space(a framework known as PCG sampler)[**161**]; a model-specific way is to use an independent sampler to yield a primary guess to facilitate later simulations. However, all these methods are based on the basic agreement that the ergodic measure cannot be altered while we blocked the parameters or using the estimates from previous independent sampler.

CHAPTER 2

Nonsubjective Priors

Reference The basic materials are from [**2**]Chap.5-6. For the measure theory I refer [**3**]. The first and probably the most important discussion of conjugate priors is [**21**]. Some examples in practice can be found in [**24**], the literature of choosing priors for specific statistical models is abundant especially for the multivariate cases. A authoritative definition and nice introduction of reference prior can be found in[**20**]. The discussion of Jeffereys principle is mainly borrowed from [**2**] and Chap.8 of [**1**]. The right-invariant Haar priors are borrowed from [**74**].

After considering the space $M(\mathcal{X})$ and $M(M(\mathcal{X}))$, we review some classical constructions of priors and see in the abstract view that what they actually are. These priors are usually called "nonsubjective" because they are constructed under some kind of invariant principle so resulting posteriors seem more "objective". The motivation of finding a nonsubjective prior is to avoid the criticism to Bayesian method by saying that the choice of priors seems too subjective to be reasonable.

Not all priors are good enough for a specific Bayesian analysis. Suppose we have a Bernoulli prior, then the Bayes estimation will often be very bad because it usually gives a values in $(0, 1)$ while the only possible choices of parameter is $\{0, 1\}$as indicated by the Bernoulli prior. Therefore we must consider following possibilities:

(1) Is the prior $\Pi(\theta)$ too informative?
    If so, we might want to consider choosing a "nonsubjective prior".
(2) Is the parameter space Θ too restrictive?
    If so, we might want to extend Θ and thus use Dirichlet priors.
(3) Should we consider other statistical analysis?
    That is out of scope of this note. However, if such a suspect is not groundless, we should look closer into the dataset first.

It is a beneficial work to look through how all these nonsubjective priors are derived, not only of historical interest, but also see how they were criticized.

Unlike previous and other chapters, historically the nonsubjective priors are mostly proposed in setting of parametric Bayesian statistics, so they aimed at providing convenient tools for parametric models. Examples in [**2, 24**] showed that how these priors, when properly chosen, can greatly simplify the Bayes model.

However, the problem of using nonsubjective priors in nonparametric setting is that we are facing $M(\mathcal{X})$ whose dimension is usually ∞. If we are in a parametric setting, then the $M(\mathcal{X})$ can be identified, at least up to an isomorphism, as the parameter space Θ whose dimension is usually finite.

By admitting $dim M(\mathcal{X}) = \infty$, we have to overcome the identifiability problem as well as the high-dimensional problem before we adapt nonsubjective priors to nonparametric setting. Our main concern is still convenience of model building.

2.1. Conjugate Priors

Points Conjugate prior; exponential family; characterization

The conjugate priors, formally can be regarded as a mimic of the likelihood function yet the essence of using conjugate priors is more than that. In a regular exponential family P, the following result asserted that to use a conjugate prior is equivalent to assuming a shrinkage estimator under the L^2 loss.

In following discussions, the Θ is a finite dimensional parameter space if nothing further is assumed.

Definition 2.1. (Conjugate priors, [**21**]) Let $\Theta \subset \mathbb{R}^d$ be an open set and $(\Theta, \mathcal{B}_\Theta)$ is equipped with the induced Lebesgue measure on Θ, an *exponential family* is a family of probability measures defined on $(\mathcal{X}, \mathcal{B}, P)$ with densities $\frac{dP_\theta}{dx} = exp(x \cdot \theta - M(\theta))$

A conjugate prior π for the exponential family P_θ can be defined as $\frac{d\pi}{d\theta} \propto exp(nx \cdot \theta - nM(\theta)), \forall \theta \in \Theta, n \in \mathbb{N}, x \in \mathcal{X}$.

Generally, a *conjugate prior* π or a sample measure P defined on $(\mathcal{X}, \mathcal{B}, P)$ will make the posterior measure $\pi|X_1, X_2 \cdots X_n \sim P$ in the same class of measures as π, with possibly different parameters.

By definition, the Dirichlet distribution is the conjugate prior of any finite exchangeable sampling; the Dirichlet process is the conjugate prior of any countable exchangeable sampling[1].

Clever readers may notice that the definition of conjugate priors is parallel to stationary measure of a Markov chain. Again we encourage reader to have a Markov chain in mind, which is a more general case than an i.i.d. random variable sequence. However, this parallel imply something more subtle, that is Bayesian procedure of yielding posterior is actually a built-in feature of ergodic decomposition.

However, a special result which is of central concern should be mentioned for using conjugate prior in parametric settings.

Theorem 2.2. *(Characterization of conjugate priors for exponential families,* [**21**] *Theorem 2,3,4) Let $\Theta \subset \mathbb{R}^d$ be an open set and $(\Theta, \mathcal{B}_\Theta)$ is equipped with the induced Lebesgue measure on Θ.*

If $\theta \sim \pi, X \sim P_\theta$ where π is the conjugate prior for an exponential family P_θ, then $E(\nabla_\theta M(\theta)|X) = X$.

Conversely, if $\theta \sim \pi, X \sim P_\theta$ where $\pi \in M(M(\mathcal{X}))$ is not degenerated and $E(\nabla_\theta M(\theta)|X) = aX + b$ for some constants $a, b \in \mathbb{R}$ then π must be a conjugate prior for this exponential family.

This means that as long as we use $E(\nabla_\theta M(\theta)|X)$ from posterior of $\pi \mid X$ as an estimator, which is always true when the loss function is chosen to be L^2, then the linearity of this estimator implies conjugacy of π and vice versa.

---

[1] Refer to Theorem 4.6 in later chapters

2.2. Jeffereys Priors

Points Jeffereys principle; Marginalization paradox; Infinite-dimension uniform priors

2.2.1. Jeffereys Principle. The Jeffereys principle is usually used to choose the so-called "non-informative" prior. Such priors are usually relatively "flat" and heavy-tailed according to the definition. This is probably the most popular choice of priors in practice.

DEFINITION 2.3. (Jeffereys prior, [**1**] p.222) Let $\Theta \subset \mathbb{R}^d$ be an open set, a uniform distribution on Θ should be proportional to $I(\boldsymbol{\theta}) = \left(\frac{\partial log \frac{dP_\theta}{d\boldsymbol{\chi}}}{\partial \theta_i \partial \theta_j} \right)_{ij}$ associated with the population measure P_θ indexed by the parameter $\theta \in \Theta$. The Fisher-Rao metric is $\rho(\boldsymbol{d\theta}) \coloneqq \sum_{i,j} \frac{\partial l}{\partial \theta_i \partial \theta_j} \cdot d\theta_i d\theta_j$. If the model is multivariate normal, then the Rao-Fisher metric is also called Mahalanobis metric.

Consider a bijective smooth mapping ψ on manifold Θ, then the Fisher-Rao metric under this transformation becomes $\boldsymbol{d\psi} = detI(\boldsymbol{\theta})\boldsymbol{d\theta}$ and the Jeffereys prior is defined to be proportional to $detI(\boldsymbol{\theta}), \forall \boldsymbol{\theta} \in \Theta$ on the parameter space.

In nonparametric setting, the Jeffereys principle is not directly applicable because the elements in $M(\mathcal{X})$ are not always differentiable, let alone admits change of variables. However, we can restrict ourselves to $\mathcal{C}^k(\mathcal{X}) \subset M(M(\mathcal{X}))$ where $\mathcal{C}^k(\mathcal{X})$ are those probability measures whose Radon-Nikodym derivatives w.r.t. Lebesgue measure are k-continuous in the weak topology of $M(\mathcal{X})$. There are yet still obstacles because the elements in $M(M(\mathcal{X}))$ takes their value in $M(\mathcal{X})$ so we have to define an explicit form of elements in $M(M(\mathcal{X}))$ itself.[2]

However, Jeffereys principle is useful even in nonparametric setting if we consider the $M(\mathcal{X})$ as a manifold and Jeffereys principle simply saying that we should choose a consistent prior under the Fisher-Rao metric on this manifold[3]. This is very pleasing a principle since it allows us to specify a prior measure on a infinite dimensional manifold $M(\mathcal{X})$ and change the coordinate charts on this manifold smoothly.

2.2.2. Marginalization Paradox. The Marginalization Paradox is about the convergence behavior of a sequence of posterior marginal distributions when the sample size grows [4]. The solution is to use the correct type of convergence.

This paradox arise when we try to use an improper prior under the guidance of Jeffereys principle.

DEFINITION 2.4. (Formal posterior) Let $p(x \mid \theta)$ be the density function for some statistical model, if $\pi(\theta)$ is a possibly improper prior then the posterior $\pi(\theta \mid x) \propto p(x \mid \theta)\pi(\theta)$ is called a formal posterior because it is obtained via the usage of formal Bayes principle.

---

[2]Some people choose the concentration measure α in Dirichlet process to be a nonsubjective measuer in order to use this notion. See [**1**] p.221

[3]This metric is the only invariant metric change consistently under smooth transformations on the manifolds among the CAPHF category. This result is the Cencov theorem.[**60**]

[4][**59**]

If we regard this notion in background of Markov chain, an improper prior means that we have to make ourselves to believe the resulting stationary measure is in a "chaotically diffused" status. But the point here is that using formal Bayes posterior defined above as the ground of inference will lead to the marginalization paradox.

There is another guidance which may lead to the use of an improper prior. Consider the Hellinger metric we introduced in previous chapter, we can define the Hellinger distance between two priors in $M(M(\mathcal{X}))$ as "entropy". If we require the prior π to be the prior measure with the maximum entropy w.r.t. some base point π_0, then it is a common case that we will have to admit an improper prior. This choice of prior according to the maximum entropy will always lead to a unique Gibbs measure[5] we will mention later. And a Gibbs measure is defined to be improper.

2.2.3. Jeffereys-Lindley Paradox. But Jeffereys priors is not necessarily good when we are dealing with a large sample and try to make Bayes inference based on the posterior obtained from the Jeffereys prior and the the sample.

The Jeffereys-Lindley Paradox is about the asymptotic behavior of the posterior rejection probability when the sample size grows and the Jeffereys prior is chosen. When the sample size grows the decision rule based on p-value and Bayesian posterior probability will differ greatly and hence leads to completely different decisions.

The solution is to consider the Pitman's alternatives[6] by choosing a different prior under H_1 or the rescaled prior emphasizing more on the alternative hypothetical (parametric) space[7].The rescaled prior will put more prior belief on the alternative hypothesis and therefore result in a less significant posterior probability[8].

This is not a problem specifically associated with Bayesian methods, when the sample size grows, almost every hypothesis testing will face the same problem that its p-value will become very low and it is hard to tell whether to reject H_0 or not based on this small p-value. [9]

We pointed out this paradox here because it shows us sometimes the Jeffereys prior is not necessarily good for inference problem.

2.2.4. Uniform Priors for Infinite Dimensional Spaces. Although usually identical to the uniform priors in the location parameter of a location family, Jeffereys priors do not necessarily coincide with uniform priors when imposed on a general parameter or for infinite dimensional spaces

Definition 2.5. (Sieve, [**1**] p.223) Let $\epsilon_i, i = 1, 2, \cdots$ be a real sequence converging to 0. A sieve associated with ϵ_i on $M(\mathcal{X})$ w.r.t. a metric ρ on $M(\mathcal{X})$ is a

[5][**58**] Sec.7

[6]In which the alternative hypothesis H_1 is not fixed but depending on sample size. For example, the alternative for testing the normal unknown mean is $H_0 : \theta = \theta_0$ v.s. $H_1 : \theta = \theta_0 + \frac{1}{\sqrt{n}}$ instead of $H_1 : \theta \neq \theta_0$

[7][**2**]

[8]For hypothesis testing problems, there is a method named nonsubjective Bayes factor which involves in modifying the Bayes factor instead of priors in order to yield a reasonable result. See [**2**] Sec 6.7.

[9]One frequentist's solution is widely adopted, that is to use false discovery rate as an alternative criterion based on p-value adjustment, the co-called Benjamini-Hochberg procedure [**62**].

series of collection $\mathcal{S}_i \coloneqq \left\{P_i^{(1)}, P_i^{(2)}, \cdots P_i^{(n_i)}\right\}, i = 1, 2 \cdots$ of finite sets with maximal cardinality in $M(\mathcal{X})$ such that $\rho(P_i^{(m)}, P_i^{(n)}) > \epsilon_i, \forall m \neq n, P_i^{(\bullet)} \in \mathcal{S}_i$. Denote the maximal cardinality as $D(\epsilon_i, M(\mathcal{X}), \rho)$.

In practice, the ρ is usually chosen as Hellinger metric. A single $\mathcal{S}_i$ is sometimes referred to as ϵ_i-net on $M(\mathcal{X})$. And the notion of sieve can be defined on a general space more than specific $M(\mathcal{X})$.

One thing to be mentioned in study of consistency results is that the concept of net is equivalent to the concept of filter we introduce later in order to set up the playground of stochastic processes. The net has the advantage of intuition in extending the convergence concept into topological space category; the filter has the advantage of being a dual to the ring ideal and thus introduce the algebraic characterization of the object naturally. [**64**]

Now if we can define a prior for a sieve with ϵ_i properly chosen according to the sample size, then we can yield a prior on $M(\mathcal{X})$. This actually leads to the Blackwell's Polya's urn scheme definition of Dirichlet process in later chapters.

Another approach is to take the weak limit of Π_i, where Π_i are sequence of uniform measures on $\mathcal{S}_i$. This actually leads to the Ferguson's extension definition of Dirichlet process in later chapters.

A third approach is to consider the index i as a hyper parameter and put a hierarchical prior on it picking index i(OR equivalently Π_i) with a preassigned probability λ_i. This actually leads to the Sethuraman's constructive definition of Dirichlet process OR Polya tree process construction in later chapters.

Definition 2.6. (ϵ-net, [**1**] p.224) Let K be a compact metric space with a metric ρ. A finite subset $S \subset K$ is called *ϵ-dispersed set* if $\rho(x, y) \geq \epsilon, \forall x \neq y \in S$. A maximal ϵ-dispersed set w.r.t. inclusion of sets is called an *ϵ-net*. An ϵ-net with the maximal cardinality is called an *ϵ-lattice*. Thus a $\{\epsilon_i\}$-sieve is a collection of $\{\epsilon_i - lattices\}$. The ϵ-net does not have to be an ϵ-lattice because a maximal set w.r.t. set inclusion does not have to be unique and thus may have different cardinalities. Also, the ϵ-lattice does not have to be unique.

Definition 2.7. (Packing number) The cardinality $D(\epsilon, K, \rho)$ of a ϵ-lattice in K w.r.t. ρ is called the *packing number of* K w.r.t the metric ρ. Because we are always assuming K to be compact in definition of ϵ-net, the packing number must be finite.

Definition 2.8. (Covering number) The *covering number* $N(\epsilon, K, \rho)$ *of the space* K w.r.t. the metric ρ is the **minimum** number of balls $B_\epsilon(x) \coloneqq \{y : \rho(x, y) < \epsilon\}$ needed to cover K. The cardinality of a maximal set is always well-defined since there cannot be two maximal coverings with different number of balls.

These two numbers are the measurement of the size of the model K under investigation. Here we should regard model as a subset of compact K.

Corollary 2.9. *([**1**], p.224)* $N(\epsilon, K, \rho) \leq D(\epsilon, K, \rho) \leq N(\frac{\epsilon}{2}, K, \rho)$

Definition 2.10. (ϵ-probability) The ϵ-probability $P_\epsilon(S) \coloneqq \frac{D(\epsilon, S, \rho)}{D(\epsilon, K, \rho)}, \forall S \subset K$. It is a probability measure defined on the compact space K and every $\epsilon \in \mathbb{R}$. Therefore the compactness of K will result a weak converging measure subsequence for every measure sequence $P_{\epsilon_i}, \epsilon_i \to 0$ by Prohov Theorem in previous chapter.

DEFINITION 2.11. (Uniformizable space) If the weak limit is for the whole sequence P_{ϵ_i}, or equivalently each weakly convergent subsequence has the same weak limit, then K is called *uniformizable*. This term means that the weak convergence is uniform on K. The resulting weak limit P_0 is called the *uniform probability* on K.

DEFINITION 2.12. (Convergence Determining Class, [**65**]) For a class/collection $\mathcal{U}$ of Borel sets on a probability space $(\mathcal{X}, \mathcal{B}, P)$. We pick those sets $B \in \mathcal{U}$ such that $P \mid_B$ is a continuous measure when restricted on it and denote the collection of these sets as $\mathcal{U} \mid_P$. We called $\mathcal{U}$ a convergence-determining class (CDC) if for all measures $P_n, P \in M(\mathcal{X})$, $P_n(A) \to P(A), \forall A \in \mathcal{U} \mid_P \Rightarrow P_n \to_w P$.

In a separable metric space $\mathcal{X}$, finite intersections of all open balls centered on a dense subset form a CDC. This is a intermediate notion connecting the setwise convergence and the weak convergence we discussed before.

THEOREM 2.13. *(Dembski Theorem,* [**1**] *p.224 Theorem 8.3.1) Let K be a compact metric space with metric ρ. The following assertions hold:*

(1) If K is a uniformizable with uniform probability P_0 then $P_0(S) = lim_{\epsilon \to 0} P_\epsilon(S), \forall S \subset K$ and $P_0(\partial S) = 0$ where ∂S is the topological boundary of S.

(2) If $P_0(S) = lim_{\epsilon \to 0} P_\epsilon(S), \forall S \subset K$ exists for some convergence-determining class in K, then K is uniformizable.

This result serves as a tool in devising an explicit expression of the setwise limit notion of ϵ-probability in terms of weak limits. Setwise topology is not tractable yet it is convenient when we try to describe some phenomena where weak convergence is way too strong.

LEMMA 2.14. *(*[**19**] *p.81 Theorem 7.6) For a parametric model indexed by $\theta \in \Theta \subset \mathbb{R}^d$ we have following inequality for Hellinger metric is controlled by eigenvalues of $I(\theta)$ via trace, i.e.*

$H(P_\theta, P_{\theta+h}) \leq \frac{|h|^2_{\mathbb{R}^d}}{4} \int_0^1 trI(\boldsymbol{\theta} + sh)ds$ where $I(\boldsymbol{\theta})$ is the Fisher information matrix. As $h \to 0$, $H(P_\theta, P_{\theta+h})^2$ behaves like $h^2 \cdot I(\boldsymbol{\theta})$

The trace of Fisher's information matrix should be regarded as the mean curvature over this Fisher manifold. Loosely speaking, if we assign the Fisher information matrix as a Riemannian metric on the parameter space Θ to make it into a Riemannian manifold, then the Lemma simply can be regarded as a geometric result. So the Fisher information matrix is the first fundamental form of this Fisher manifold and the trace is the volume element of a curve on this manifold. A clever reader may notice that if we integrate to yield the RHS of the inequality above, we may get a neighborhood based on the volume element of the Fisher manifold Θ. If the family happen to be a regular exponential family then this is saying that a Hellinger neighborhood is contained in a geodesic neighborhood under the Rao-Fisher metric.

That is to say, the Hellinger metric is controlled by the integral over the mean curvature the same direction. For interested readers, I highly recommend that you read [**78**] for the elegant framework of information geometry.

THEOREM 2.15. *(Limit of ϵ-probability* [**1**] *p.225, Theorem 8.4.1.) For a fixed compact subset $K \subset \Theta$ the parameter space, then for all $Q \subset K$ with $volume(\partial Q) = 0$ we have*

$lim_{\epsilon\to 0}P_\epsilon(Q) = \frac{\int_Q \sqrt{detI(\boldsymbol{\theta})}d\boldsymbol{\theta}}{\int_K \sqrt{detI(\boldsymbol{\theta})}d\boldsymbol{\theta}}$ *where the* $I(\boldsymbol{\theta})$ *is the Fisher information matrix w.r.t.* $\boldsymbol{\theta}$.

Therefore the Jeffereys measure on Θ can be defined as $\mu(Q) \propto \int_Q \sqrt{detI(\boldsymbol{\theta})}d\boldsymbol{\theta}$. The Jeffereys prior measure can be extended to the situation where Θ(OR $M(\mathcal{X})$) is σ-compact by intersecting with the K_i compact supports sequence. It can be regarded as a uniform prior imposed on the unit volume of a general $M(\mathcal{X})$ manifold. Also it has sharp consistency properties, which we will explore in full details in next chapter.

Now assume that the model $\mathcal{P}$ under consideration is totally bounded for Hellinger metric. The following result is a result that can be understood geometrically under Amari's α-connection. Or we can say it is the geometric compatibility that brought the consistency behavior of Jeffereys prior.

THEOREM 2.16. *(Consistency behavior of Jeffereys prior,* [**1**] *p.229) Let* $\mathcal{P} \subset M(\mathcal{X})$ *be a family of measures, which is compact when metricized by Hellinger metric. Let* ϵ_n *be a positive sequence satisfying* $\sum_{n=1}^{\infty}\sqrt{n}\epsilon_n < \infty$. *Let* $\mathcal{P}_n$ *be a* ϵ_n*-net in* $\mathcal{P}$ *and* μ_n *be the uniform probability associated with this* ϵ_n*-net on compact* $\mathcal{P}$.

Define $\mu = \sum_{n=1}^{\infty}\lambda_n\mu_n, \sum_{n=1}^{\infty}\lambda_n = 1, \forall\lambda_n > 0$. *If for* $\forall\beta > 0$ *we have* $lim_{n\to\infty}e^{n\beta}\frac{\lambda_n}{D(\epsilon_n,\mathcal{P}_n,H)} = \infty$*then the posterior measure based on the prior mixture of uniform probabilities* μ *and i.i.d. observations* $X_1, X_2\cdots$ *is strongly consistent at* $\forall p \in \mathcal{P}$.

Another more subtle result is how fast the posterior converge to the "true" sampling probability.

THEOREM 2.17. *(Convergence rate of Jeffereys prior,* [**1**] *p.231) Let* $\mathcal{P} \subset M(\mathcal{X})$ *be a family of measures, which is compact when metrized by metric* ρ. *Let* ϵ_n *be a positive sequence satisfying* $lim_{n\to\infty}\epsilon_n \to 0, lim_{n\to\infty}n\epsilon_n^2 \to \infty$. *Let* $\mathcal{P}_n$ *be a subset of* $\mathcal{P}$ *such that for some* $C > 0$

(1) $D(\epsilon_n, \mathcal{P}_n, \rho) \le e^{n\epsilon_n^2}$[10]

(2)$\Pi_n(\mathcal{P} - \mathcal{P}_n) \le e^{-n\epsilon_n^2(C+4)}$

(3) $\Pi_n(\left\{P : -E\left(log\frac{dP}{dP_0} \mid P_0\right) \le \epsilon_n^2, -E\left(\left[log\frac{dP}{dP_0}\right]^2 \mid P_0\right) \le \epsilon_n^2\right\}) \ge e^{-n\epsilon_n^2 C}$

then $\Pi_n\left(\{P : d(P, P_0) \ge M\epsilon_n \mid X_1, X_2\cdots X_n\}\right) \to 0$ *a.e.* P_0^n *for some* M *large enough where* Π_n *is the posterior measure based on sample of size* n; P_0 *is the underlying sampling measure.*

A sequence of convergence rate ϵ_n satisfying (1) above for $\mathcal{P}_n = \mathcal{P}$ and $d = H$ Hellinger metric will be regarded as "optimal" rate of convergence for estimators of P relative to Hellinger metric under model $\mathcal{P}$. Since the convergence rate not only involve in the prior but also the likelihood function, we must also require that the likelihood ratios be controlled, so the condition (1) is actually controlling the likelihood ratio as well. The most general version of convergence rate result involves a form of control imposed on priors called envelope.

[10]This requirement is equivalent to restriction on $N(\epsilon_n, \mathcal{P}_n, \rho)$ via Corollary2.9. We can also define the $logN(\epsilon_n, \mathcal{P}_n, \rho)$ as sort of entropy, thus this is a restriction on the size of model as well as the entropy. ϵ_n are chosen as the minimal solution of the equation $logN(\epsilon, \mathcal{P}_n, \rho) \le n\epsilon^2$.