2

$M(\mathcal{X})$ and Priors on $M(\mathcal{X})$

2.1 Introduction

As mentioned in Chapter 1, in the nonparametric case the parameter space Θ is typically the set of all probability measures on $\mathcal{X}$. We denote the set of all probability measures on $\mathcal{X}$ by $M(\mathcal{X})$. The cases of interest to us are when $\mathcal{X}$ is a finite set and when $\mathcal{X} = \mathbb{R}$. The Bayesian aspect requires prior distributions on $M(\mathcal{X})$, in other words, probabilities on the space of probabilities. In this chapter we develop some measure-theoretic and topological features of the space $M(\mathcal{X})$ and discuss various notions of convergence on the space of prior distributions.

The results in this chapter, except for the last section, are mainly used to assert the existence of the priors discussed later. Thus, for a reader who is prepared to accept the existence theorems mentioned later, a cursory reading of this chapter would be adequate. On the other hand, for those who are interested in measure-theoretic aspects, a careful reading of this chapter will provide a working familiarity with the measure-theoretic subtleties involved. The last section where formal definitions of consistency are discussed, can be read independently. While we generally consider the case $\mathcal{X} = \mathbb{R}$, most of the arguments would go through when $\mathcal{X}$ is a complete separable metric space.

2.2 The Space $M(\mathcal{X})$

As before, let $\mathcal{X}$ be a complete separable metric space with $\mathcal{B}$ the corresponding Borel σ-algebra on $\mathcal{X}$. Denote by $M(\mathcal{X})$ the space of all probability measures on $(\mathcal{X}, \mathcal{B})$.

As seen in the chapter 1 there are many reasonable notions of convergence on the space $M(\mathcal{X})$, but they are not all equally convenient for our purpose. We begin with a brief discussion of these.

Total Variation Metric. Recall that the total variation metric was defined by

$$\|P - Q\| = 2 \sup_B |P(B) - Q(B)|$$

If p and q are densities of P and Q with respect to some σ-finite measure μ, then $\|P - Q\|$ is just the L_1-distance $\int |p - q|\ d\mu$ between p and q. The total variation metric is a strong metric. If $x \in \mathcal{X}$ and δ_x is the probability degenerate at x, then $U_x = \{P : \|P - \delta_x\| < \epsilon\} = \{P : P(x) > 1 - \epsilon\}$ is a neighborhood of δ_x. Further if $x \neq x'$ then $U_x \cap U_{x'} = \emptyset$. Thus, when $\mathcal{X}$ is uncountable, $\{U_x : x \in \mathcal{X}\}$ is an uncountable collection of disjoint open sets, the existence of which renders $M(\mathcal{X})$ nonseparable. Further, no sequence of discrete measures can converge to a continuous measure and vice versa. These properties make the total variation metric uninteresting when considered on all of $M(\mathcal{X})$.

The total variation metric when restricted to sets of the form L_μ—all probability measures dominated by a σ-finite measure μ—is extremely useful and interesting. In this context we will refer to the total variation as the L_1-metric. It is a standard result that L_μ with the L_1-metric is complete and separable.

Hellinger Metric. This metric was also discussed in Chapter 1. Briefly the Hellinger distance between P and Q is defined by

$$H(P, Q) = \left[\int (\sqrt{p} - \sqrt{q})^2\ d\mu \right]^{1/2}$$

where p and q are densities with respect to μ. The Hellinger metric is equivalent to the L_1 metric. Associated with the Hellinger metric is a useful quantity $A(P, Q)$ called *affinity*, defined as $A(P, Q) = \int \sqrt{p}\sqrt{q}\, d\mu$. The relation $H^2(P^n, Q^n) = 2-2(A(P, Q))^n$, where P^n, Q^n are n-fold product measures, makes the Hellinger metric convenient in the i.i.d. context.

Setwise convergence. The metrics defined in the last section provide corresponding notions of convergence. Another natural way of saying P_n converges to P is to require that $P_n(B) \to P(B)$ for all Borel sets B. A way of formalizing this topology is as follows. Let $\mathcal{F}$ be the class of functions $\{P \mapsto P(B) : B \in \mathcal{B}\}$. On $M(\mathcal{X})$ give the smallest topology that makes the functions in $\mathcal{F}$ continuous. It is easy to see that under this topology, if f is a bounded measurable function, then $P \mapsto \int f \, dP$ is continuous. Sets of the form $\{P : |P(B_i) - P_0(B_i)| < \epsilon_i, \, B_1, B_2, \ldots, B_k \in \mathcal{B}\}$ give a neighborhood base at P_0.

Setwise convergence is an intuitively appealing notion, but it has awkward topological properties that stem from the fact that convergence of $P_n(B)$ to $P(B)$ for sets in an algebra does not ensure the convergence for all Borel sets. We summarize some additional facts as a proposition.

Proposition 2.2.1. *Under setwise convergence:*

(i) $M(\mathcal{X})$ is not separable,

(ii) If P_0 is a continuous measure then P_0 does not have a countable neighborhood base, and hence the topology of setwise convergence is not metrizable.

Proof. (i) $U_x = \{P : P\{x\} > 1 - \epsilon\}$ is a neighborhood of δ_x, and as x varies form an uncountable collection of disjoint open sets.

(ii) Suppose that there is a countable base for the neighborhoods at P_0. Let $\mathcal{B}_0$ be a countable family of sets such that sets of the type

$$\mathcal{U} = \{P : |P(B_i) - P_0(B_i)| < \epsilon_i, \, B_1, B_2, \ldots, B_k \in \mathcal{B}_0\}$$

form a neighborhood base at P_0. It then follows that $P_n(B) \to P(B)$ for all Borel sets B iff $P_n(B) \to P(B)$ for all sets in $\mathcal{B}_0$.

Let $\mathcal{B}_n = \sigma(B_1, B_2, \ldots, B_n)$ where $B_1, B_2, \ldots$ is an enumeration of $\mathcal{B}_0$. Denote by $B_{n1}, B_{n2}, \ldots B_{nk(n)}$ the atoms of $\mathcal{B}_n$. Define P_n to be the discrete measure that gives mass $P_0(B_{ni})$ to x_{ni} where x_{ni} is a point in B_{ni}. Clearly $P_n(B_{mj}) \to P_0(B_{mj})$ for all m_j. On the other hand $P_n(\cup_{i,m}\{x_{mi}\}) = 1$ for all n but $P_0((\cup_{i,m}\{x_{mi}\}) = 0$.

□

These shortcomings persist even when we restrict attention to subsets $M(\mathcal{X})$ of the form L_μ.

Supremum Metric. When $\mathcal{X}$ is $\mathbb{R}$, the Glivenko-Cantelli theorem on convergence of empirical distribution suggests another useful metric, which we call the supremum metric. This metric is defined by

$$d_K(P,Q) = \sup_t |P(-\infty,t] - Q(-\infty,t]|$$

Under this metric $M(\mathcal{X})$ is complete but not separable.

Weak Convergence. In many ways weak convergence is the most natural and useful topology on $M(\mathcal{X})$. Say that

$$P_n \to P \text{ weakly } \text{ or } P_n \stackrel{\text{weakly}}{\to} P \text{ if}$$

$$\int f\, dP_n \to \int f\, dP$$

for all bounded continuous functions f on $\mathcal{X}$. For any P_0 a neighborhood base consists of sets of the form $\cap_1^k\{P : \left|\int f_i\, dP_0 - \int f_i\, dP\right| < \epsilon\}$ where $f_i, i = 1,2,\ldots,k$ are bounded continuous functions on $\mathcal{X}$. One of the things that makes the weak topology so convenient is that under weak convergence $M(\mathcal{X})$ is a complete separable metric space.

The main results that we need with regard to weak convergence are the Portmanteau theorem and Prohorov's theorem given in Chapter 1.

Because $M(\mathcal{X})$ is a complete separable metric space under weak convergence, we define the *Borel σ-algebra* $\mathcal{B}_M$ on $M(\mathcal{X})$ to be the smallest σ-algebra generated by all weakly open sets, equivalently all weakly closed sets. This σ-algebra has a more convenient description as the smallest σ-algebra that makes the functions $\{P \mapsto P(B) : B \in \mathcal{B}\}$ measurable. Let $\mathcal{B}_0$ be the σ-algebra generated by all weakly open sets. Consider all B such that $P \mapsto P(B)$ is $\mathcal{B}_0$-measurable. This class contains all closed sets, and from the π-λ theorem (Theorem 1.2.1) it follows easily that $\mathcal{B}_M$ is the σ-algebra generated by all weakly open sets.

We have discussed two other modes of convergence on $M(\mathcal{X})$: the total variation and setwise convergence. It is instructive to pause and investigate the σ-algebras corresponding to these and their relationship with $\mathcal{B}_M$.

Because these are nonseparable spaces, there is no good acceptable notion of a Borel σ-algebra. In the case of total variation metric, the two common σ-algebras considered are

(i) $\mathcal{B}_o$—the σ-algebra generated by open sets and

(ii) $\mathcal{B}_b$—the σ-algebra generated by open balls.

The σ-algebra $\mathcal{B}_o$ generated by open sets is much larger than $\mathcal{B}_M$. To see this, restrict the σ-algebra to the space of degenerate measures $D_{\mathcal{X}} = \{\delta_x : x \in \mathcal{X}\}$. Then each δ_x is relatively open, and this will force the restriction of $\mathcal{B}_o$ to $D_{\mathcal{X}}$ to be the power set. On the other hand, $\mathcal{B}_M$ restricted to $D_{\mathcal{X}}$ is just the inverse of the Borel σ-algebra on $\mathcal{X}$ under the map $\delta_x \mapsto x$.

Because every open ball is in $\mathcal{B}_M$, so is every set in the σ-algebra generated by these balls. It can be shown that $\mathcal{B}_b$ is properly contained in $\mathcal{B}_M$.

Similar statements hold when we consider the σ-algebras for setwise convergence. The corresponding σ-algebras here would be those generated by open sets and those generated by basic neighborhoods at a point. A discussion of these different σ-algebras can be found in [71].

We next discuss measurability issues on $M(\mathcal{X})$. Following are a few of elementary propositions.

Proposition 2.2.2. *(i) If $\mathcal{B}_0$ is an algebra generating $\mathcal{B}$ then*

$$\sigma\{P \mapsto P(B) : B \in \mathcal{B}_0\} = \mathcal{B}_M$$

(ii) $\sigma\left\{P \mapsto \int f\ dP : f \text{ bounded measurable}\right\} = \mathcal{B}_M$

Proof. (i) Let $\tilde{\mathcal{B}} = \{B : P \mapsto P(B) \text{ is } \mathcal{B}_M \text{ measurable}\}$. Then $\tilde{\mathcal{B}}$ is a σ-algebra and contains $\mathcal{B}_0$. The result now follows from Theorem 1.2.1.

(ii) It is enough to show that $P \mapsto \int f\ dP$ is $\mathcal{B}_M$ measurable. This is immediate for f simple, and any bounded measurable f is a limit of simple functions.

□

Proposition 2.2.3. *Let $f_P(x)$ be a bounded jointly measurable function of (P, x). Then $P \mapsto \int f_P(x)\ dP(x)$ is $\mathcal{B}_M$ measurable.*

Proof. Consider

$$\mathcal{G} = \left\{F \subset M(\mathcal{X}) \times \mathcal{X} \text{ such that } P(F^P) \text{ is } \mathcal{B}_M \text{ measurable}\right\}$$

Here F^P is the P-section $\{x : (P, x) \in F\}$ of F. $\mathcal{G}$ is a λ-system that contains the π-class of all sets of the form $C \times B; C \in \mathcal{B}_M,\ B \in \mathcal{B}$, and by Theorem 1.2.1 is the product σ-algebra on $M(\mathcal{X}) \times \mathcal{X}$. This proves the proposition when $f_P(x) = I_F(P, x)$. The proof is completed by verifying when $f_P(x)$ is simple, and by passing to limits. □

Proposition 2.2.3 can be used to prove the measurability of the set of discrete probabilities.

Proposition 2.2.4. *The set of discrete probabilities is a measurable subset of $M(\mathcal{X})$.*

Proof. If $E = \{(P, x) : P\{x\} > 0\}$ is a measurable set, then setting $f_P(x) = I_E(P, x)$, the set of discrete measures is just $\{P : \int f_P(x)dP = 1\}$ and would be measurable by Proposition 2.2.3. To see that $E = \{(P, x) : P\{x\} > 0\}$ is measurable, we show that $(P, x) \mapsto P\{x\}$ is jointly measurable in (P, x). Consider the set of all a measurable subsets F of $\mathcal{X} \times \mathcal{X}$ such that $(P, x) \mapsto P(F^x)$ is measurable in (P, x). As before, $F^x = \{y : (x, y) \in F\}$. This class contains all Borel sets of the form $B_1 \times B_2$ and is a λ-system, and by Theorem 1.2.1 is the Borel σ-algebra on $\mathcal{X} \times \mathcal{X}$. In particular $(P, x) \mapsto P(F^x)$ is measurable when $F = \{(x, x) : x \in \mathcal{X}\}$ is the diagonal and $E = \{(P, x) : P(F^x > 0)\}$. □

Consider $f_P(x)$ used in Proposition 2.2.4. Then P is continuous iff $\int f_P(x)dP = 0$. It follows that the set of continuous measures is a measurable set.

If μ is a σ-finite measure on $\mathbb{R}$, then L_μ is a measurable subset of $M(\mathcal{X})$. To see this, assume without loss of generality that μ is a probability measure. Let $\mathcal{B}_n$ be an increasing sequence of algebras, with finitely many atoms, whose union generates $\mathcal{B}$. Denote the atoms of $\mathcal{B}_n$ by $B_{n1}, B_{n2}, \ldots B_{k(n)}$, and for any probability measure P, set $f_P(x) = \lim \sum_1^{k(n)} P(B_{ni})/\mu(B_{ni})$ when it exists and 0 otherwise. To complete the argument note that $L_\mu = \{P : \int f_P(x)d\mu = 1\}$.

2.3 (Prior) Probability Measures on $M(\mathcal{X})$

2.3.1 $\mathcal{X}$ Finite

Suppose $\mathcal{X} = \{1, 2, \ldots, k\}$. In this case $M(\mathcal{X})$ can be identified with the $(k-1)$ dimensional probability simplex $S_k = \{p_1, p_2, \ldots, p_k : 0 \le p_i \le 1, \sum p_i = 1\}$. One way of defining a prior on $M(\mathcal{X})$ is by defining a measure on S_k. Any such measure defines the joint distribution of $\{P(A) : A \subset \mathcal{X}\}$, because for any A, $P(A) = \sum_{i \in A} p_i$, where $p_k = 1 - \sum_1^{k-1} p_i$.

An example of a prior distribution on S_k is the uniform distribution—the normalized Lebesgue measure on $\{p_1, p_2, \ldots, p_{k-1} : 0 \le p_i \le 1, \sum p_i \le 1\}$. Another example is the Dirichlet density which is given by

$$\Pi(p_1, p_2, \ldots, p_{k-1}) = \frac{\Gamma(\sum_1^k \alpha_i)}{\prod \Gamma(\alpha_i)} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} \ldots p_{k-1}^{\alpha_{k-1} - 1} (1 - \sum_1^{k-1} p_i)^{\alpha_k - 1}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_k$ are positive real numbers. This density will be studied in greater detail later.

A different parametrization of $M(\mathcal{X})$ yields another method of constructing a prior on $M(\mathcal{X})$. Assume for ease of exposition that $\mathcal{X}$ contains 2^k elements $\{x_1, x_2, \ldots, x_{2^k}\}$. Let

$$B_0 = \{x_1, x_2, \ldots, x_{2^{k-1}}\} \text{ and } B_1 = \{x_{2^{k-1}+1}, x_{2^{k-1}+2}, \ldots, x_{2^k}\}$$

be a partition of $\mathcal{X}$ into two sets. Let B_{00}, B_{01} be a partition of B_0 into two halves and B_{10}, B_{11} be a similar partition of B_1. Proceeding this way we can get partitions $B_{\epsilon_1\epsilon_2\ldots\epsilon_i 0}, B_{\epsilon_1\epsilon_2\ldots\epsilon_i 1}$ of $B_{\epsilon_1\epsilon_2\ldots\epsilon_i}$ where each ϵ_i is 0 or 1 and $i < k$. Clearly, this partition stops at $i = k$.

We next note that the partitions can be used to identify $\mathcal{X}$ with $E_k = \{0,1\}^k$. Any $x \in \mathcal{X}$ corresponds to a sequence $\epsilon_1(x)\epsilon_2(x)\ldots\epsilon_k(x)$ where $\epsilon_i(x) = 0$ if x is in $B_{\epsilon_1(x)\epsilon_2(x)\ldots\epsilon_{i-1}(x)0}$ and 1 if x is in $B_{\epsilon_1(x)\epsilon_2(x)\ldots\epsilon_{i-1}(x)1}$. Conversely, any sequence $\epsilon_1\epsilon_2\ldots\epsilon_k$ corresponds to the point $\cap_1^k B_{\epsilon_1\epsilon_2\ldots\epsilon_i}$. Thus there is a correspondence—depending on the partition—between the set $M(\mathcal{X})$ of probability measures on $\mathcal{X}$ and the set $M(E_k)$ of probability measures on E_k.

Any probability measure on E_k is determined by quantities like

$$y_{\epsilon_1\epsilon_2\ldots\epsilon_k} = P\left(\epsilon_{i+1} = 0 \mid \epsilon_1, \epsilon_2, \ldots, \epsilon_i\right)$$

Specifically, let E_k^* be the set of all sequences of 0 and 1 of length less than k, including the empty sequence $\emptyset$. If $0 \le y_{\underline{\epsilon}} \le 1$ is given for all $\underline{\epsilon} \in E_k^*$, then there is a probability on E_k by

$$P(\epsilon_1\epsilon_2\ldots\epsilon_k) = \prod_{i=1,\epsilon_i=0}^{k} y_{\epsilon_1\epsilon_2\ldots\epsilon_{i-1}} \prod_{i=1,\epsilon_i=1}^{k} (1 - y_{\epsilon_1\epsilon_2\ldots\epsilon_{i-1}})$$

where $i = 1$ corresponds to the empty sequence $\emptyset$. Hence construction of a prior on E_k amounts to a specification of the joint distribution for $\{y_{\underline{\epsilon}} : \underline{\epsilon} \in E_k^*\}$.

A little reflection will show that all we have done is to reparametrize a probability P on $\mathcal{X}$ by

$$P(B_0), P(B_{00}|B_0), P(B_{10}|B_1), \ldots, P(B_{\epsilon_1\epsilon_2\ldots\epsilon_{k-1}0}|B_{\epsilon_1\epsilon_2\ldots\epsilon_{k-1}0})$$

Of interest to us is the case where the $Y_{\underline{\epsilon}}$s, equivalently $P(B_{\underline{\epsilon}0}|B_{\underline{\epsilon}})$s, are all independent. The case when these are independent beta random variables—the Polya tree processes—will be studied in Chapter 3

Yet another method of obtaining a prior distribution on $M(\mathcal{X})$ is via De Finetti' theorem. De Finetti's theorem plays a fundamental role in Bayesian inference, and we refer the reader to [144] for an extensive discussion.

Let $X_1, X_2, \ldots, X_n$ be $\mathcal{X}$-valued random variables. $X_1, X_2, \ldots, X_n$ is said to be exchangeable if $X_1, X_2, \ldots, X_n$ and $X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)}$ have the same distribution for every permutation π of $\{1, 2, \ldots, n\}$. A sequence $X_1, X_1, \ldots$ is said to be exchangeable if $X_1, X_2, \ldots, X_n$ is exchangeable for every n.

Theorem 2.3.1. *[De Finetti] A sequence of $\mathcal{X}$-valued random variables is exchangeable iff there is a unique measure Π on $M(\mathcal{X})$ such that for all n,*

$$\int_{M(\mathcal{X})} \prod_1^n p(x_i)\, d\Pi(p) = Pr\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$$

In general it is not easy to construct Π from the distribution of the X_is. Typically, we will have a natural candidate for Π. By uniqueness, it is enough to verify the preceding equation. On the other hand, given Π, the behavior of $X_1, X_1, \ldots$ often gives insight into the structure of Π.

As an example, let $\mathcal{X} = \{x_1, x_2, \ldots, x_k\}$. Let $\alpha_1, \alpha_2, \ldots, \alpha_k$ be positive integers. Let $\bar{\alpha}(i) = \alpha_i / \sum \alpha_j$. Consider the following urn scheme: Suppose a box contains balls of k- colors, with α_i balls of color i. Choose a ball at random, so that $P(X_1 = i) = \bar{\alpha}(i)$. Replace the ball and add one more of the same color. Clearly, $P(X_2 = j | X_1 = i) = (\alpha_j + \delta_i(j))/(\sum \alpha_i + 1)$ where $\delta_i(j) = 1$ if $i = j$ and 0 otherwise. Repeat this process to obtain $X_3, X_4, \ldots$ Then

(i) $X_1, X_2, \ldots$ are exchangeable; and

(ii) the prior Π for this case is the Dirichlet density on S_k.

2.3.2 $\mathcal{X} = \mathbb{R}$

We next turn to construction of measures on $M(\mathcal{X})$. Because the elements of $M(\mathcal{X})$ are functions on $\mathcal{B}$, $M(\mathcal{X})$ can be viewed as a subset of $[0,1]^{\mathcal{B}}$ where the product space $[0,1]^{\mathcal{B}}$ is equipped with the canonical product σ-algebra, which makes all the coordinate functions measurable. Note that the restriction of the product σ-algebra to $M(\mathcal{X})$ is just $\mathcal{B}_M$. A natural attempt to construct measures on $M(\mathcal{X})$ would be to use Kolomogorov's consistency theorem to construct a probability measure on $[0,1]^{\mathcal{B}}$, which could then be restricted to $M(\mathcal{X})$. However $M(\mathcal{X})$ is not measurable as a subset of $[0,1]^{\mathcal{B}}$, and that makes this approach somewhat inconvenient. To see that $M(\mathcal{X})$ is not measurable, note that singletons are measurable subsets of $M(\mathcal{X})$ but not so in the product space.

When $\mathcal{X} = \mathbb{R}$, distribution functions turn out to be a useful crutch to construct priors on $M(\mathbb{R})$. To elaborate:

(i) Let Q be a dense subset of $\mathbb{R}$ and let $\mathcal{F}^*$ be all real-valued functions on Q such that

(a) F is right-continuous on Q,

(b) F is nondecreasing, and

(c) $\lim_{t\to\infty} = F(t) = 1, \lim_{t\to-\infty} F(t) = 0$.

(ii) Let $\mathcal{F}$ be all real-valued functions on $\mathbb{R}$ such that

(a) F is right-continuous on $\mathbb{R}$,

(b) F is non decreasing, and

(c) $\lim_{t\to\infty} F(x) = 1, \lim_{t\to-\infty} F(x) = 0$.

(iii) $M(\mathbb{R}) = \{P : P \text{ is a probability measure on } \mathbb{R}\}$

There is a natural 1-1 correspondence between these three sets: Let $\phi_1 : M(\mathbb{R}) \mapsto \mathcal{F}$ be the function that takes a probability measure P to its distribution function $F_P(t) = P(-\infty, t]$ and let $\phi_2 : \mathcal{F} \to \mathcal{F}^*$ be the function that maps a distribution function to its restriction on Q. These maps are 1-1, onto, and bi-measurable. Thus any probability measure on $\mathcal{F}^*$ can be transferred to a probability on $\mathcal{F}$ and then to $M(\mathbb{R})$. A prior on $\mathcal{F}^*$ only involves the distributions of

$$(F(t_1), F(t_2) - F(t_1), \ldots, F(t_k) - F(t_{k-1}))$$

for t_is in Q. However, because any $F(t)$ is a limit of $F(t_n), t_n \in Q$, the distributions of quantities like $(F(t_1), F(t_2) - F(t_1), \ldots, F(t_k) - F(t_{k-1}))$ for t_i-real can be recovered, at least as limits. On the other hand since a general Borel set B has no simple description in terms of intervals, one can assert the existence of a distribution for $P(B)$ that is compatible with the prior on $\mathcal{F}^*$, but it may not be possible to arrive at anything resembling an explicit description of the distribution.

It is convenient to use the notation $\mathcal{L}(\cdot|\Pi)$ to stand for the distribution or law of a quantity under the distribution Π.

Theorem 2.3.2. *Let Q be a countable dense subset of $\mathbb{R}$. Suppose for every k and every collection $t_1 < t_2 < \ldots < t_k$ with $\{t_1, t_2, \ldots, t_k\} \subset Q$, $\Pi_{t_1,t_2,\ldots,t_k}$ is a probability measure on $[0,1]^k$ which is a specification of a distribution of $((F(t_1), F(t_2), \ldots, F(t_k))$ such that*

(i) if $\{t_1, t_2, \ldots, t_k\} \subset \{s_1, s_2, \ldots, s_l\}$ then the marginal distribution on $(t_1, t_2, \ldots, t_k)$ obtained from $\Pi_{s_1,s_2,\ldots,s_l}$ is $\Pi_{t_1,t_2,\ldots,t_k}$;

(ii) *if* $t_1 < t_2$ *then* $\Pi_{t_1,t_2}\{F(t_1) \le F(t_2)\} = 1$;

(iii) *if* $(t_{1n}, t_{2n}, \ldots, t_{kn}) \downarrow (t_1, t_2, \ldots, t_k)$ *then* $\Pi_{(t_{1n},t_{2n},\ldots,t_{kn})}$ *converges in distribution to* $\Pi_{(t_1,t_2,\ldots,t_k)}$; *and*

(iv) *if* $t_n \downarrow -\infty$ *then* $\Pi_{t_n} \to 0$ *in distribution and if* $t_n \uparrow \infty$ *then* $\Pi_{t_n} \to 1$ *in distribution.*

then there exists a probability measure Π *on* $M(\mathbb{R})$ *such that for every* $t_1 < t_2 < \ldots < t_k$, *with* $\{t_1, t_2, \ldots, t_k\} \subset Q$,

$$\mathcal{L}\left((F(t_1), F(t_2), , \ldots, F(t_k))\,|\Pi\right) = \Pi_{t_1,t_2,\ldots,t_k}.$$

Proof. By the Kolomogorov consistency theorem (i) ensures the existence of a probability measure Π on $[0,1]^Q$ with $\Pi_{(t_1,t_2,\ldots,t_k)}$ as marginals. We will argue that $\Pi(\mathcal{F}^*) = 1$

Suppose $\mathcal{F}_1^* = \cap_{t_i<t_j}\left\{F \in [0,1]^Q : F(t_i) \le F(t_j)\right\}$. Because Q is countable by (ii), $\Pi(\mathcal{F}_1^*) = 1$.

Next, fix t and a sequence t_n in Q decreasing to t. On $\mathcal{F}_1^*, F(t_n)$ as a function of F is decreasing in n and hence has a limit. If $F^*(t) = \lim_n F(t_n)$ then $F^*(t) \ge F(t)$ and by assumption (iii) $E_\Pi F^*(t) = E_\Pi F(t)$, so that $F^*(t) = F(t)$ a.e. Π. Consequently

$$\Pi\{F \in \mathcal{F}_1^* : \text{ F is right-continuous at } t\} = 1$$

and the countability of Q yields

$$\Pi\{F : F \text{ is monotone and } F \text{ is right-continuous at all } t \in Q\} = 1$$

A similar argument shows that with Π probability 1, for F in $\mathcal{F}_1^*, \lim_{t\to\infty} = F(t) = 1$, and $\lim_{t\to-\infty} F(t) = 0$. This shows that $\Pi(\mathcal{F}^*) = 1$.

Thus we have established the existence of a probability measure on $\mathcal{F}^*$. Using the discussion preceding the theorem this prior can be lifted to all of $M(\mathbb{R})$. □

The assumptions of Theorem 2.3.2 require specification of finite-dimensional distribution only for t_is in Q and the conclusion also involves only the finite dimensional distributions for t_is in Q. It is easy to see that if one starts with $\Pi_{(t_1,t_2,\ldots,t_k)}$ with t_i's real and satisfying the conditions of Theorem 2.3.2 then one would get a Π for which the marginals are $\Pi_{(t_1,t_2,\ldots,t_k)}$ for t_is real.

A convenient way of specifying the distribution of $(F(t_1), F(t_2), \ldots, F(t_k))$ for $t_1 < t_2 < \ldots, t_k$, is by specifying the distribution, say $\Pi'_{t_1,t_2,\ldots,t_k}$, of

$$(F(t_1), F(t_2) - F(t_1), \ldots, F(t_k - F(t_{k-1}))$$

The convenience arises from the fact that $(-\infty, t_1], (t_1, t_2], \ldots, (t_k, \infty)$ can be thought of as $k+1$ cells and $(p_1, p_2, \ldots, p_{k+1})$ as the corresponding multinomial probabilities. Note that $\Pi'_{t_1,t_2,\ldots,t_k}$ is a probability measure on $S_k = \{(p_1, p_2, \ldots, p_k : p_i \geq 0, \sum_1^k p_i \leq 1\}$. If the specifications of the collection $\Pi'_{t_1,t_2,\ldots,t_k}$ satisfy assumptions (ii),(iii), and (iv) of Theorem 2.3.2, then so would the collection $\Pi_{t_1,t_2,\ldots,t_k} = \mathcal{L}((p_1, p_1 + p_2, \ldots, \sum_1^k p_i)|\Pi'_{t_1,t_2,\ldots,t_k})$. These observations give the following easy variant of Theorem 2.3.2.

Theorem 2.3.3. *Suppose that for every k and every collection $t_1 < t_2 < \ldots < t_k$ with $\{t_1, t_2, \ldots, t_k\} \subset \mathbb{R}$, $\Pi_{t_1,t_2,\ldots,t_k}$ is a probability measure on $S_k = \{(p_1, p_2, \ldots, p_k) : p_i \geq 0, \sum_1^k p_i \leq 1\}$ such that*

(i) if $\{t_1, t_2, \ldots, t_k\} \subset \{s_1, s_2, \ldots, s_l\}$ then the marginal distribution on $(t_1, t_2, \ldots, t_k)$ obtained from $\Pi_{s_1,s_2,\ldots,s_l}$ is $\Pi_{t_1,t_2,\ldots,t_k}$;

(ii) if $(t_{1n}, t_{2n}, \ldots, t_{kn}) \to (t_1, t_2, \ldots, t_k)$ then $\Pi_{(t_{1n},t_{2n},\ldots,t_{kn})}$ converges in distribution to $\Pi_{(t_1,t_2,\ldots,t_k)}$; and

(iii) if $t_n \downarrow -\infty$ then $\Pi_{t_n} \to 0$ in distribution and if $t_n \uparrow \infty$ then $\Pi_{t_n} \to 1$ in distribution.

then there exists a probability measure Π on $\mathcal{F}$ (equivalently on $M(\mathbb{R})$ such that for every $t_1 < t_2 < \ldots < t_k$, with $\{t_1, t_2, \ldots, t_k\} \subset \mathbb{R}$,

$$\mathcal{L}\left((F(t_1), F(t_2) - F(t_1), \ldots, F(t_k) - F(t_{k-1}))\,|\Pi\right) = \Pi_{t_1,t_2,\ldots,t_k}$$

Suppose $(B_1, B_2, \ldots, B_k)$ is a collection of disjoint subsets of $\mathbb{R}$; the next theorem shows that if the distribution of $P(B_1), P(B_2), \ldots, P(B_k)$ are themselves prescribed consistently then the prior Π would have the prescribed marginal distribution for $(P(B_1), P(B_2), \ldots, P(B_k))$.

Theorem 2.3.4. *Suppose for each collection of disjoint Borel sets $(B_1, B_2, \ldots, B_k)$ a distribution $\Pi_{B_1,B_2,\ldots,B_k}$ is assigned for $(P(B_1), P(B_2), \ldots, P(B_k))$ such that*

(i) $\Pi_{B_1,B_2,\ldots,B_k}$ is a probability measure on k-dimensional probability simplex S_k and if $A_1, A_2, \ldots, A_l$ is another collection of disjoint Borel sets whose elements are

union of sets from $(B_1, B_2, \ldots, B_k)$ *then*

$$\Pi_{A_1,A_2,\ldots,A_l} = \textit{distribution of} \left(\sum_{B_i \subset A_1} P(B_i), \sum_{B_i \subset A_2} P(B_i), \ldots, \sum_{B_i \subset A_l} P(B_i) \right)$$

(ii) if $B_n \downarrow \emptyset$*; and* $\Pi_{B_n} \to 0$ *in distribution,*

(iii) $P(\mathbb{R}) \equiv 1$.

Then there exists a probability measure Π *on* $M(\mathbb{R})$ *such that for any collection of disjoint Borel sets* $(B_1, B_2, \ldots, B_k)$*, the marginal distribution of* $(P(B_1), \ldots, P(B_k))$ *under* Π *is* $\Pi_{B_1,B_2,\ldots,B_k}$.

Remark 2.3.1. Given $\Pi_{B_1,B_2,\ldots,B_k}$ as earlier, we can extend the definition to obtain $\Pi_{A_1,A_2,\ldots,A_m}$ for any collection (not necessarily disjoint) of Borel sets $A_1, A_2, \ldots, A_m$. Toward this, let $B_1 = A_1, B_i = A_i - \cup_{j<i} A_j$, and define $\Pi_{A_1,A_2,\ldots,A_m}$ as the distribution of $(P(B_1, P(B_1) + P(B_2) + \ldots, \sum_1^m P(B_j))$ under $\Pi_{B_1,B_2,\ldots,B_m}$. The following proof shows that the marginal distribution under Π of $(P(A_1), P(A_2), \ldots, P(A_k))$ of any collection of Borel sets is $\Pi_{A_1,A_2,\ldots,A_k}$.

Proof. As in the Theorem 2.3.3 start with partitions of the form $B_i = (t_{i-1}, t_i]$ for $i = 1, 2, \ldots, k$; and let Π be the measure obtained on $\mathcal{F}$. Let ϕ_2 be the map from $\mathcal{F}$ to $M(\mathbb{R})$ defined by $\phi_2(F) = P_F$, where P_F is the probability measure corresponding to F. It is easy to see that this map is 1-1 and measurable. We will continue to denote by Π the induced measure on $M(\mathbb{R})$.

Π by construction sits on $M(\mathbb{R})$. What we then need to show is that the marginal distribution of $(P(B_1), P(B_2), \ldots, P(B_k))$ under Π is $\Pi_{B_1,B_2,\ldots,B_k}$.

Step 1 (ii) implies that

$$\text{if } (B_{1n}, B_{2n}, \ldots, B_{kn}) \downarrow (B_1, B_2, \ldots, B_{k1}) \text{ then}$$

$$(P(B_{1n}), P(B_{2n}), \ldots, P(B_{kn})) \to (P(B_1), P(B_2), \ldots, P(B_k)) \text{ in distribution.}$$

To see this,

$$\begin{aligned} &((P(B_{1n}), P(B_{2n}), \ldots, P(B_{kn})) \\ &\quad = (P(B_1) + (P(B_{1n}) - P(B_1)), P(B_2) + (P(B_{2n}) - P(B_2)), \ldots, \\ &\qquad\qquad P(B_k) + (P(B_{kn}) - P(B_k))) \end{aligned}$$

and for each i, $(B_{in} - B_i) \downarrow \emptyset$ and hence $(P(B_{in}) - P(B_i))$ goes to 0 in distribution and hence in probability. As a result, the whole vector

$$((P(B_{1n}) - P(B_1)), (P(B_{2n}) - P(B_2)), \ldots, (P(B_{kn}) - P(B_k))) \downarrow 0 \text{ in probability}$$

Step 2 Denote by $\mathcal{B}_0$ the algebra generated by intervals of the form $(a, b]$. For any $B_1, B_2, \ldots, B_k$, let $\mathcal{L}(P(B_1), P(B_2), \ldots, P(B_k)|\Pi)$ denote the distribution of the vector $(P(B_1), P(B_2), \ldots, P(B_k))$ under Π. Fix k. Let $C_i = (a_i, b_i], i = 2, \ldots, k$. Consider

$$\hat{\mathcal{B}} = \left\{B_1 : \mathcal{L}(P(B_1), P(C_2), \ldots, P(C_k)|\Pi) = \Pi_{(B_1, C_2, \ldots, C_k)}\right\}$$

Then $\hat{\mathcal{B}}$ contains all sets of the form $(a, b]$, is closed under disjoint unions of such sets, and hence contains $\mathcal{B}_0$. In addition, by Step 1 this is a monotone class. So $\hat{\mathcal{B}}$ is $\mathcal{B}$.

Step 3 Now consider

$$\left\{B_2 : \mathcal{L}(P(B_1), P(B_2), P(C_3), \ldots, P(C_k)|\Pi) = \Pi_{(B_1, B_2, C_3, \ldots, C_k)}\right\}$$

From step 2, this class contains all sets of the form $(a, b]$, and their finite disjoint unions and hence contains $\mathcal{B}_0$. Further, it is a monotone class and so is $\mathcal{B}$. Continuing similarly, it follows that for any Borel sets $B_1, B_2, \ldots, B_k$,

$$\mathcal{L}(P(B_1), P(B_2), \ldots, P(B_k)|\Pi) = \Pi_{B_1, B_2, \ldots, B_k}$$

. □

Example 2.3.1. Let α be a finite measure on $\mathbb{R}$. For any partition $(B_1, B_2, \ldots, B_k)$, let $\Pi_{B_1, B_2, \ldots, B_k}$ on S_k be a Dirichlet $(\alpha(B_1), \alpha(B_2), \ldots, \alpha(B_k))$. We will show in Chapter 3 that this assignment satisfies the conditions of Theorem 2.3.4.

Remark 2.3.2. Theorem 2.3.4 on constructing a measure Π on $\mathcal{F}$ through finite-dimensional distribution can be viewed from a different angle. Toward this, for each n, divide the interval $[-2^n, 2^n]$ into intervals of length 2^{-n} and let $-2^n = t_{n1} < t_{n2} < \ldots < t_{nk(n)} = 2^n$ denote the endpoints of the intervals. These define a partition of $\mathbb{R}$ into $k(n) + 1$ cells in an obvious way. Any probability $(p_1, p_2, \ldots, p_{k(n)+1})$ on these $k(n) + 1$ cells corresponds to a distribution function on $\mathbb{R}$, which is constant on each interval and thus any probability $\Pi_{t_{n1}, t_{n2}, \ldots, t_{nk(n)}}$ on $S_{k(n)+1}$ defines a probability measure μ_n on $\mathcal{F}_n$ = all distribution functions, which are constant on the interval $(t_{ni}, t_{ni+1}]$. The consistency assumption on $\Pi_{t_{n1}, t_{n2}, \ldots, t_{nk(n)}}$ shows that the marginal distribution on $\mathcal{F}_n$ obtained from μ_{n+1} is just μ_n. Now it can be shown that

1. $\{\mu_n\}_{n\geq 1}$ is tight as a sequence of probability measures on $\mathcal{F}$. To see this, let $\varepsilon_i \downarrow 0$ and let K_i be a sequence of compact subsets of $\mathbb{R}$. Then

   $$\{P : P(K_i) \geq 1 - \varepsilon_i \text{ for all } i\}$$

   is a compact subset of $M(\mathbb{R})$. What is needed to show tightness is that given δ, there is a set of the form given earlier with μ_n measure greater than $1-\delta$ for all n. Use assumptions (i) and (iii) of Theorem 2.3.4 and show that for each i, you can get an n_i such that for all $n, \mu_n\{F : F(t_{n_i 0}) > \varepsilon_i$ and $1-F(t_{n_i,k(n_i)}) > \varepsilon_i\} < \delta/2^i$;

2. $\{\mu_n\}$ converges to a measure Π; and

3. Π satisfies the conclusions of Theorem 2.3.4.

2.3.3 Tail Free Priors

When $\mathcal{X}$ is finite, we have seen that by partitioning $\mathcal{X}$ into

$$\{B_0, B_1\}, \{B_{00}, B_{01}, B_{10}, B_{11}\}, \ldots$$

and reparametrizing a probability by $P(B_0), P(B_{00}|B_0)\ldots$, we can identify measures on $M(\mathcal{X})$ with E_k—the set of sequences of 0s and 1s of length k. Tail free priors arise when these conditional probabilities are independent. In this section we extend this method to the case $\mathcal{X}$ =$\mathbb{R}$.

Let E be all infinite sequences of 0s and 1s, i.e., $E = \{0,1\}^{\mathbf{N}}$. Denote by E_k all sequences $\epsilon_1, \epsilon_2, \ldots, \epsilon_k$ of 0s and 1s of length k, and let $E^* = \cup_k E_k$ be all sequences of 0s and 1s of finite length. We will denote elements of E^* by $\underline{\epsilon}$.

Start with a partition

$$\underline{\mathcal{T}}_0 = \{B_0, B_1\}$$

of $\mathcal{X}$ into two sets. Let

$$\underline{\mathcal{T}}_1 = \{B_{00}, B_{01}, B_{10}, B_{11}, \}$$

where B_{00}, B_{01} is a partition of B_0 and B_{10}, B_{11} is a partition of B_1. Proceeding this way,let $\underline{\mathcal{T}}_n$ be a partition consisting of sets of the form $B_{\underline{\epsilon}}$, where $\underline{\epsilon} \in E_n$ and further $B_{\underline{\epsilon}1}$, $B_{\underline{\epsilon}0}$ is a partition of $B_{\underline{\epsilon}}$.

We assume that we are given a sequence of partitions $\underline{\mathcal{T}} = \{\mathcal{T}_n\}_{n\geq 1}$ constructed as in the last paragraph such that the sets $\{B_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}$ generate the Borel σ-algebra.

Definition 2.3.1. A prior Π on $M(\mathbb{R})$ is said to be *tail free* with respect to $\underline{\mathcal{T}} = \{\mathcal{T}_n\}_{n\geq 1}$ if rows in

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

are independent.

To turn to the construction of tail free priors on $M(\mathbb{R})$, start with a dense set of numbers Q, like the binary rationals in $(0,1)$, and write it as $Q = \{a_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}$ such that for any $\underline{\epsilon}$ $\underline{\epsilon}0 < \underline{\epsilon} < \underline{\epsilon}1$ and construct the following sequence of partitions of $\mathbb{R}$: $\underline{\mathcal{T}}_0 = \{B_0, B_1\}$ is a partition of $\mathbb{R}$ into two intervals, say

$$B_0 = (-\infty, a_0], B_1 = (a_0, \infty)$$

Let $\underline{\mathcal{T}}_1 = \{B_{00}, B_{01}, B_{10}, B_{11},\}$, where

$$B_{00} = (-\infty, a_{00}], B_{01} = (a_{00}, a_0]$$

and

$$B_{10} = (a_0, a_{01}], B_{11} = (a_{01}, \infty)$$

Proceeding this way, let $\underline{\mathcal{T}}_n$ be a partition consisting of sets of the form $B_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n}$, where $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are 0 or 1 and further $B_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 0}, B_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n 1}$ is a partition of $B_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n}$.

The assumption that Q is dense is equivalent to the statement that the sequence of partitions $\underline{\mathcal{T}} = \{\mathcal{T}_n\}_{n\geq 1}$ constructed as in the last paragraph are such that the sets $\{B_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}$ generate the Borel σ-algebra.

For each $\underline{\epsilon} \in E^*$, let $Y_{\underline{\epsilon}}$ be a random variable taking values in $[0,1]$. If we set $Y_{\underline{\epsilon}} = P(B_{\underline{\epsilon}0}\,|B_{\underline{\epsilon}})$, then for each k, $\{Y_{\underline{\epsilon}} : \underline{\epsilon} \in \cup_{i\leq k} E_i\}$ define a joint distribution for $P(B_{\underline{\epsilon}}) : \underline{\epsilon} \in E_k$. By construction, these are consistent. In order for these to define a prior on $M(\mathbb{R})$ we need to ensure that the continuity condition (ii) of Theorem 2.3.2 holds.

Theorem 2.3.5. *If $Y_{\underline{\epsilon}} = P(B_{\underline{\epsilon}0}\,|B_{\underline{\epsilon}})$, where $Y_{\underline{\epsilon}} : \underline{\epsilon} \in E^*$ is a family of $[0,1]$ valued random variables such that*

(i)

$$Y \perp \{Y_0, Y_1\} \perp \{Y_{00}, Y_{01}, Y_{10}, Y_{11}\} \perp \ldots$$

(ii) for each $\epsilon \in E^*$,

$$Y_{\epsilon 0} Y_{\epsilon 00} Y_{\epsilon 000} \ldots = 0 \text{ and } Y_1 Y_{11} \ldots = 0 \tag{2.1}$$

then there exists a tail free prior Π *on* $M(\mathbb{R})$ *(with respect to the partition under consideration) such that* $Y_\epsilon = P(B_{\epsilon 0}|B_\epsilon)$.

Proof. As noted earlier we need to verify condition (ii) of Theorem 2.3.2. In the current situation it amounts to showing that if $\underline{\epsilon}^o = \epsilon_1^0 \epsilon_2^0 \ldots \epsilon_k^0$ and as $n \to \infty$, $a_{\underline{\epsilon}_n}$ decreases to $a_{\underline{\epsilon}^0}$, then the distribution of $F\left(a_{\underline{\epsilon}_n}\right)$ converges to $F\left(a_{\underline{\epsilon}^0}\right)$. Because any sequence of $a_{\underline{\epsilon}}$ decreasing to $a_{\underline{\epsilon}^0}$ is a subsequence of $a_{\underline{\epsilon}^0 1}, a_{\underline{\epsilon}^0 10}, a_{\underline{\epsilon}^0 100}, \cdots$,

$$F\left(a_{\underline{\epsilon}^0 10\ldots 0}\right) = F\left(a_{\underline{\epsilon}^0}\right) + P(B_{\underline{\epsilon}^0 10\ldots 0})$$

and

$$P(B_{\underline{\epsilon}^0 1,0\ldots 0}) = P(B_{\underline{\epsilon}^0})(1 - Y_{\underline{\epsilon}^0}) Y_{\underline{\epsilon}^0 1} Y_{\underline{\epsilon}^0 10} \ldots$$

the result follows from (ii). □

These discussions can be usefully and elegantly viewed by identifying $\mathbb{R}$ with the space of sequences of 0s and 1s.

As before, let E be $\{0,1\}^{\mathbf{N}}$. Any probability on E gives rise to the collection of numbers $\{y_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}$, where $y_{\epsilon_1 \epsilon_2 \ldots \epsilon_n} = P(\epsilon_{n+1} = 0 | \epsilon_1 \epsilon_2 \ldots \epsilon_n)$. Conversely, setting $y_{\epsilon_1 \epsilon 2 \ldots \epsilon_n} = P(\epsilon_{n+1} = 0 | \epsilon_1 \epsilon_2 \ldots \epsilon_n)$, any set numbers $\{y_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}$, with $0 \le y_{\underline{\epsilon}} \le 1$ determines a probability on E. In other words,

$$P(\epsilon_1 \epsilon_2 \ldots \epsilon_k) = \prod_{i=1, \epsilon_i = 0}^{k} y_{\epsilon_1 \epsilon_2 \ldots \epsilon_{i-1}} \prod_{i=1, \epsilon_i = 1}^{k} (1 - y_{\epsilon_1 \epsilon_2 \ldots \epsilon_{i-1}}) \tag{2.2}$$

Hence, to define a prior on $M(E)$, we need to specify a joint distribution for $\{Y_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}$, where each $Y_{\underline{\epsilon}}$ is between 0 and 1.

As in the finite case, we want to use the partitions $\underline{\mathcal{T}} = \{\mathcal{T}_n\}_{n \ge 1}$ to identify $\mathbb{R}$ with sequences of 0s and 1s. and Let $x \in \mathbb{R}$. $\phi(x)$ is the function that sends x to the sequence $\underline{\epsilon}$ in E, where

$$\begin{aligned} \epsilon_1(x) &= 0 \quad \text{if } x \in B_0 & \epsilon_1(x) &= 1 \quad \text{if } x \in B_1 \\ \epsilon_i(x) &= 0 \quad \text{if } x \in B_{\epsilon_1, \epsilon_2, \ldots, \epsilon_{i-1} 0} & \epsilon_i(x) &= 1 \quad \text{if } x \in B_{\epsilon_1 \epsilon_2 \ldots \epsilon_{i-1} 1} \end{aligned}$$

Because each $\underline{\mathcal{T}}_n$ is a partition of $\mathbb{R}$, ϕ defines a function from $\mathbb{R}$ into E. ϕ is 1-1, measurable but not onto E. The range of ϕ will not contain sequences that are eventually 0. This is another way of saying that with binary expansions we consider the expansion with 1 in the tails rather than 0s. If $D = \{\underline{\epsilon} \in E : \epsilon_i = 0 \text{ for all } i \geq n \text{ for some } n\} \cup \{\underline{\epsilon} : \epsilon_i = 1 \text{ for all } i\}$, then ϕ is 1-1, measurable from $\mathbb{R}$ onto $D^c \cap E$. Further, ϕ^{-1} is measurable on $D^c \cap E$. Thus, as before, the set of probability measures $M(\mathbb{R})$ can be identified with $M^0(E)$—the set of probability measures on E that give mass 0 to D. This reduces the task of defining a prior on $M(\mathbb{R})$ to one of defining a prior on $M^0(E)$.

The condition $P(D) = 0$ gets translated to

$$y_{\underline{\epsilon}0}(y_{\underline{\epsilon}00})\ldots = 0 \text{ for all } \underline{\epsilon} \in E^* \text{ and } y_1 y_{11} \ldots = 0 \tag{2.3}$$

As before, defining a prior on $M(\mathbb{R})$, equivalently on $M^0(E)$, amounts to defining $\{Y_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}$ such that (2.3) is satisfied almost surely. Satisfying (2.3) almost surely corresponds to condition (ii) in Theorem 2.3.5.

A useful way to specify a prior on $M(E)$ is by having $Y_{\underline{\epsilon}}$ for $\underline{\epsilon}$ of different lengths be mutually independent, which yields tail free priors. In Chapter 3, we return to this construction, to develop Polya tree priors.

Tail free prior are conjugate in the sense that if the prior is tail free, then so is the posterior. To avoid getting lost in a notational mess we first state an easy lemma.

Lemma 2.3.1. *Let $\xi_1, \xi_2, \ldots, \xi_k$ be independent random vectors (not necessarily of the same dimension) with joint distribution $\mu = \prod_1^k \mu_i$. Let J be a subset of $\{1, 2, \ldots, k\}$ and let μ^* be the probability with*

$$\frac{d\mu^*}{d\mu} = C \prod_{j \in J} \xi_j$$

Then $\xi_1, \xi_2, \ldots, \xi_k$ are independent under μ^*.

Proof. Clearly $C = \prod_{j \in J} [\int \xi_j d\mu_j]^{-1}$. Further,

$$\text{Prob}(\xi_i \in B_i : 1 \leq i \leq k) = \int_{(\xi_i \in B_i : 1 \leq i \leq k)} C[\prod_{j \in J} \xi_j] d\mu$$

$$= \prod_{i \notin J} \mu_i(B_i) \prod_{j \in J} \frac{\int_{B_j} \xi_j d\mu_j}{\int \xi_j d\mu_j}$$

□

Theorem 2.3.6. *Suppose Π is a tail free prior on $M(\mathbb{R})$ with respect to the sequence of partitions $\{\underline{\mathcal{T}}_k\}_{k\geq 1}$. Given P, let $X_1, X_2, \ldots, X_n$ be,i.i.d. P; then the posterior is also tail free with respect to $\{\underline{\mathcal{T}}_k\}_{k\geq 1}$.*

Proof. We will prove the result for $n = 1$; the general case follows by iteration. Consider the posterior distribution given $\underline{\mathcal{T}}_k$. Because $\{B_{\underline{\epsilon}} : \underline{\epsilon} \in E_k\}$ are the atoms of $\underline{\mathcal{T}}_k$, it is enough to find the posterior distribution given $X \in B_{\underline{\epsilon}'}$ for each $\underline{\epsilon}' \in E_k$.

Let $\underline{\epsilon}' = \epsilon_1\epsilon_2 \ldots \epsilon_k$. Then the likelihood of $P(B_{\underline{\epsilon}'})$ is

$$\prod_1^k P(B_{\epsilon_1,\epsilon_2,\ldots,\epsilon_j} | B_{\epsilon_1,\epsilon_2,\ldots,\epsilon_{j-1}})$$

so that the posterior density of $\{P(B_{\underline{\epsilon}1} | B_{\underline{\epsilon}})\}$ with respect to Π is

$$C \prod_{i=1,\epsilon_i=0}^{n} P(B_{\epsilon_1\epsilon_2\ldots\epsilon_i} | B_{\epsilon_1\epsilon_2\ldots\epsilon_{i-1}}) \prod_{i=1,\epsilon_i=1}^{n} (1 - P(B_{\epsilon_1\epsilon_2\ldots\epsilon_i} | B_{\epsilon_1\epsilon_2\ldots\epsilon_{i-1}})$$

From Lemma 2.3.1

$$\{P(B_{\underline{\epsilon}1} | B_{\underline{\epsilon}}) : \underline{\epsilon} \in E_1\}, \{P(B_{\underline{\epsilon}1} | B_{\underline{\epsilon}}) : \underline{\epsilon} \in E_2\}, \ldots, \{P(B_{\underline{\epsilon}1} | B_{\underline{\epsilon}}) : \underline{\epsilon} \in E_{k-1}\}$$

are independent under the posterior.

In particular if $m < k$, independence holds for

$$\{P(B_{\underline{\epsilon}1} | B_{\underline{\epsilon}}) : \underline{\epsilon} \in E_1\}, \{P(B_{\underline{\epsilon}1} | B_{\underline{\epsilon}}) : \underline{\epsilon} \in E_2\}, \ldots, \{P(B_{\underline{\epsilon}1} | B_{\underline{\epsilon}}) : \underline{\epsilon} \in E_{m-1}\}.$$

Letting $k \to \infty$, an application of the martingale convergence theorem gives the conclusion for the posterior given X_1. □

In this section we have discussed two general methods of constructing priors on $M(\mathbb{R})$. There are several other techniques for obtaining nonparametric priors. There are priors that arise from stochastic processes. If f is the sample path of a stochastic process then $\hat{f} = k^{-1}(f)e^f$ yields a random density when $k(f) = Ee^f$ is finite. We will study a method of this kind in the context of density estimation. Or one can look at expansions of a density using some orthogonal basis and put a prior on the coefficients. A class of priors called *neutral to the right priors*, somewhat like tail free priors, will be studied in Chapter 10 on survival analysis.

2.4 Tail Free Priors and 0-1 Laws

Suppose Π is a prior on $M(\mathbb{R})$ and $\{B_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}$ is a set of partitions as described in the last section. To repeat, for each n, $\underline{\mathcal{T}}_n = \{B_{\underline{\epsilon}} : \underline{\epsilon} \in E_n\}$ is a partition of $\mathbb{R}$ and $B_{\underline{\epsilon}0}$, $B_{\underline{\epsilon}1}$ is a partition of $B_{\underline{\epsilon}}$. Further $\mathcal{B} = \sigma\{B_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}$. Unlike the last section it is not required that $B_{\underline{\epsilon}}$ be intervals. The choice of intervals as sets in the partition played a crucial role in the construction of a probability measure on $M(\mathbb{R})$. Given a probability measure on $M(\mathbb{R})$, the following notions are meaningful, even if the $B_{\underline{\epsilon}}$ are not intervals.

For notational convenience, as before, denote by $Y_{\underline{\epsilon}} = P(B_{\underline{\epsilon}0}|B_{\underline{\epsilon}})$. Formally, $Y_{\underline{\epsilon}}$ is a random variable defined on $M(\mathbb{R})$ with $Y_{\underline{\epsilon}}(P) = P(B_{\underline{\epsilon}0}|B_{\underline{\epsilon}})$. Recall that Π is said to be tail free with respect to the partition $\underline{\mathcal{T}} = \{\mathcal{T}_n\}_{n\geq 1}$ if

$$Y \perp \{Y_0, Y_1\} \perp \{Y_{00}, Y_{01}, Y_{10}, Y_{11}\} \perp \ldots$$

Theorem 2.4.1. *Let λ be any finite measure on $\mathbb{R}$, with $\lambda(B_{\underline{\epsilon}}) > 0$ for all $\underline{\epsilon}$. If $0 < Y_{\underline{\epsilon}} < 1$ for all $\underline{\epsilon}$ then*

$$\Pi\{P : P << \lambda\} = 0 \text{ or } 1$$

Proof. Assume without loss of generality that λ is a probability measure.

Let $Z_0 = Y, Z_1 = \{Y_0, Y_1\}, Z_2 = \{Y_{00}, Y_{01}, Y_{10}, Y_{11}\}, \ldots$. By assumption, $Z_1, Z_2, \ldots$ are independent random vectors. The basic idea of the proof is to show that $L(\lambda) = \{P : P << \lambda\}$ is a tail set with respect to the Z_is. The Kolmogorov $0-1$ law then yields the conclusion. In the next two lemmas it is shown that for each n, $L(\lambda)$ depends only on $Z_n, Z_{n+1}, \ldots$ and is hence a tail set. □

Lemma 2.4.1. *When $P(B_{\underline{\epsilon}}) > 0$, define $P(\cdot|B_{\underline{\epsilon}})$ to be the probability $P(A|B_{\underline{\epsilon}}) = P(A \cap B_{\underline{\epsilon}})/P(B_{\underline{\epsilon}})$. Define $\lambda(\cdot|B_{\underline{\epsilon}})$ similarly. Fix n; then*

$$L(\lambda) = \{P : P(\cdot|B_{\underline{\epsilon}}) << \lambda(\cdot|B_{\underline{\epsilon}}) \text{ for all } \underline{\epsilon} \in E_n \text{ such that } P(B_{\underline{\epsilon}}) > 0\}$$

Proof. Because

$$P(A) = \sum_{\underline{\epsilon}\in E_n} P(A|B_{\underline{\epsilon}})P(B_{\underline{\epsilon}}) \text{ and } \lambda(A) = \sum_{\underline{\epsilon}\in E_n} \lambda(A|B_{\underline{\epsilon}})\lambda(B_{\underline{\epsilon}})$$

the result follows immediately.

□

Lemma 2.4.2. *Let* $\underline{\mathbf{Y}} = \{Y_{\underline{\epsilon}}(P) : \underline{\epsilon} \in E^*, P \in M(\mathbb{R})\}$. *The elements* $\underline{\mathbf{y}}$ *of* $\underline{\mathbf{Y}}$ *are thus a collection of conditional probabilities arising from a probability. Conversely any element* $\underline{\mathbf{y}}$ *of* $\underline{\mathbf{Y}}$ *gives rise to a probability which we denote by* $P_{\underline{\mathbf{y}}}$. *Then for each* $\underline{\epsilon} \in E_n$, *for all* $A \in \mathcal{B}$, *and for every* $\underline{\mathbf{y}}$ *in* $\underline{\mathbf{Y}}$

$$P_{\underline{\mathbf{y}}}(A|B_{\underline{\epsilon}}) \text{ depends only on } Z_n, Z_{n+1}, \ldots$$

Proof. Let

$$\mathcal{B}_0 = \Big\{A : \text{ for all } \underline{\mathbf{y}},\ P_{\underline{\mathbf{y}}}(A|B_{\underline{\epsilon}}) \text{ depends only on } Z_n, Z_{n+1}, \ldots\Big\}$$

Because $0 < Y_{\underline{\epsilon}} < 1$ for all $\underline{\epsilon} \in E^*$, $P_{\underline{y}}\,(B_{\underline{\epsilon}}) > 0$ for all $\underline{\epsilon} \in E^*$. Hence $\mathcal{B}_0$ contains the algebra of finite disjoint unions of elements in $\{B_{\underline{\epsilon}'} : \underline{\epsilon}' \in \cup_{m>n} E_m\}$ and is a monotone class. Hence $\mathcal{B}_0 = \mathcal{B}$.

□

Remark 2.4.1. Let Π be tail free with respect to $\underline{\mathcal{T}} = \{\mathcal{T}_n\}_{n\geq 1}$ such that $0 < Y_{\underline{\epsilon}} < 1$; for all $\underline{\epsilon} \in E^*$. Argue that P is discrete iff $P(.|B_{\underline{\epsilon}})$ is discrete for all $\underline{\epsilon} \in E_n$. Now use the Kolmogorov 0-1 law to conclude that $\Pi\{P : P \text{ is discrete }\} = 0$ or 1.

The next theorem, due to Kraft, is useful in constructing priors concentrated on sets like $L(\lambda)$.

Let $\Pi, \{B_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}, \{Y_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}$ be as in the Theorem 2.4.1, and, as before given any realization $\underline{y} = \{y_{\underline{\epsilon}} : \underline{\epsilon} \in E^*\}$, let $P_{\underline{y}}$ denote the corresponding probability measure on $\mathbb{R}$.

Theorem 2.4.2. *Let* λ *be a probability measure on* $\mathbb{R}$ *such that* $\lambda(B_{\underline{\epsilon}}) > 0$ *for all* $\underline{\epsilon} \in E^*$. *Suppose*

$$f^n_{\underline{y}}\,(x) = \sum_{\underline{\epsilon}\in E_n} \frac{P_{\underline{y}}\,(B_{\underline{\epsilon}})}{\lambda(B_{\underline{\epsilon}})} I_{B_{\underline{\epsilon}}}(x) = \sum_{\underline{\epsilon}\in E_n} \frac{\prod_{i=1,\epsilon_i=0}^{k} y_{\epsilon_1\epsilon_2\ldots\epsilon_{i-1}} \prod_{i=1,\epsilon_i=1}^{k}(1-y_{\epsilon_1\epsilon_2\ldots\epsilon_{i-1}})}{\lambda(B_{\underline{\epsilon}})}$$

If $\sup_n E_\Pi \left[f^n_{\underline{y}}\,(x)\right]^2 \leq K$ *for all* x *then* $\Pi\{P : P << \lambda\} = 1$

Proof. For each $\underline{y} \in \underline{\mathbf{Y}}$, by the martingale convergence theorem $f^n_{\underline{y}}$ converges almost surely λ to a function $f_{\underline{y}}$. Consider the measure $\Pi \times \lambda$, which is the joint distribution of $\underline{y}$ and x, on $\prod_{\underline{\epsilon}\in E^*} Y_{\underline{\epsilon}} \times \mathbb{R}$.

Because for each $\underline{y}$, $f^n_{\underline{y}} \to f_{\underline{y}}$ a.s λ, we have $f^n_{\underline{y}} \to f_{\underline{y}}$ a.s $\Pi \times \lambda$. Further, under our assumption $\left\{ f^n_{\underline{y}}(x) : n \geq 1 \right\}$ is uniformly integrable with respect to $\Pi \times \lambda$ and hence $E_{\Pi\times\lambda}\left| f^n_{\underline{y}}(x) - f_{\underline{y}}(x)\right| \to 0$. Now for each $\underline{y}$, by Fatou's lemma, $E_\lambda f_{\underline{y}} \leq 1$. On the other hand, $E_{\Pi\times\lambda} f^n_{\underline{y}}(x) = 1$ for all n, and by the L_1-convergence mentioned earlier, $E_{\Pi\times\lambda} f_{\underline{y}}(x) = 1$. Thus $E_\lambda f_{\underline{y}} = 1$ a.e. π and this shows $\pi\{L(\lambda)\} = 1$. □

The next theorem is an application of the last theorem. It shows how, given a probability measure λ, by suitably choosing both the partitions and the parameter of the $Y_{\underline{\epsilon}}$s , we can obtain a prior that concentrates on $L(\lambda)$.

Theorem 2.4.3. *Let λ be a continuous probability distribution on $\mathbb{R}$. Denote by F the distribution function of λ and construct a partition as follows:*

$$B_0 = F^{-1}(0, 1/2] \qquad B_1 = F^{-1}(1/2, 1]$$
$$B_{00} = F^{-1}(0, 1/4], B_{01} = F^{-1}(1/4, 1/2] \quad B_{10} = F^{-1}(1/2, 3/4], B_{11} = F^{-1}(3/4, 1]$$

and in general

$$B_{\epsilon_1,\epsilon_2,\ldots,\epsilon_n} = F^{-1}\left(\sum_1^n \frac{\epsilon_i}{2^n}, \sum_1^n \frac{\epsilon_i}{2^n} + \frac{1}{2^n}\right]$$

Suppose $E(Y_{\underline{\epsilon}}) = 1/2$ for all $\underline{\epsilon} \in E^*$ and $\sup_{\underline{\epsilon}\in E_n} V(Y_{\underline{\epsilon}}) \leq b_n$, with $\sum b_n < \infty$. Then the resulting prior satisfies $\Pi(L(\lambda)) = 1$.

Proof. $\lambda(B_{\underline{\epsilon}}) > 0$, because $\lambda(B_{\underline{\epsilon}0}|B_{\underline{\epsilon}}) = 1/2$, for all $B_{\underline{\epsilon}}$. Fix x. If $x \in B_{\epsilon_1\epsilon_2,\ldots\epsilon_n}$, then

$$f^n_Y(x) = \prod_{i=0}^n \frac{Y^{1-\epsilon_i}_{\epsilon_1\epsilon_2,\ldots\epsilon_{i-1}}(1 - Y_{\epsilon_1\epsilon_2,\ldots\epsilon_{i-1}})^{\epsilon_i}}{1/2}$$

and

$$E[f^n_Y(x)]^2 = \prod_o^n 4E\left[[Y^2_{\epsilon_1\epsilon_2,\ldots\epsilon_{i-1}}]^{1-\epsilon_i}[(1 - Y_{\epsilon_1\epsilon_2,\ldots\epsilon_{i-1}})^2]^{\epsilon_i}\right]$$
$$\leq \prod_0^n 4a_i$$

where $a_i = \max\left(EY^2_{\epsilon_1\epsilon_2,\ldots\epsilon_{i-1}}, E(1 - Y_{\epsilon_1\epsilon_2,\ldots\epsilon_{i-1}})^2\right)$. Now

$$EY^2_{\epsilon_1\epsilon_2,\dots\epsilon_{i-1}} = V(Y_{\epsilon_1\epsilon_2,\dots\epsilon_{i-1}}) + (1/2)^2 \le b_i + 1/4$$

and

$$E\left(1 - Y_{\epsilon_1\epsilon_2,\dots,\epsilon_{i-1}})^2\right) \le b_i + 1/4$$

Thus $\prod_1^n 4a_i \le \prod_1^n (1+4b_i)$ converges, because $\sum b_n < \infty$. □

2.5 Space of Probability Measures on $M(\mathbb{R})$

We next turn to a discussion of probability measures on $M(\mathbb{R})$. To get a feeling for what goes on we begin by asking when are two probability measures Π_1 and Π_2 equal?

Clearly $\Pi_1 = \Pi_2$ if for any finite collection $B_1, B_2, \dots, B_k$ of Borel sets,

$$(P(B_1), P(B_2), \dots, P(B_k))$$

has the same distribution under both Π_1 and Π_2. This is an immediate consequence of the definition of $\mathcal{B}_M$.

Next suppose that $(C_1, C_2, \dots, C_k)$ are Borel sets. Consider all intersections of the form

$$C_1^{\epsilon_1} \cap C_2^{\epsilon_2} \cap \dots \cap C_k^{\epsilon_k}$$

where $\epsilon_i = 0, 1$, $C_i^1 = C_i$ and $C_i^0 = C_i^c$. These intersections would give rise to a partition of $\mathcal{X}$, and since every C_i can be written as a union of elements of this partition, the distribution of $(P(C_1), P(C_2), \dots, P(C_k))$ is determined by the joint distribution of the probability of elements of this partition. In other words, if the distribution of $(P(B_1), P(B_2), \dots, P(B_k))$ under Π_1 and Π_2 are the same for every finite disjoint collection of Borel sets then $\Pi_1 = \Pi_2$. Following is another useful proposition.

Proposition 2.5.1. *Let $\mathcal{B}_0 = \{B_i : i \in I\}$ be a family of sets closed under finite intersection that generates the Borel σ-algebra $\mathcal{B}$ on $\mathcal{X}$. If for every $B_1, B_2, \dots, B_k$ in $\mathcal{B}_0$, $(P(B_1), P(B_2), \dots, P(B_k))$ has the same distribution under Π_1 and Π_2, then $\Pi_1 = \Pi_2$.*

Proof. Let $\mathcal{B}_M^0 = \{E \in \mathcal{B}_M : \Pi_1(E) = \Pi_2(E)\}$. Then $\mathcal{B}_M^0$ is a λ-system. For any J finite subset of I, by our assumption Π_1 and Π_2 coincide on the σ-algebra $\mathcal{B}_M^J$—the σ-algebra generated by $\{P(B_j) : j \in J\}$ and hence $\mathcal{B}_M^J \subset \mathcal{B}_M^0$. Further the union of $\mathcal{B}_M^J$ over all finite subsets of I forms a π-system. Because these also generate $\mathcal{B}_M$, $\mathcal{B}_M^0 = \mathcal{B}_M$. □

Remark 2.5.1. A convenient choice of $\mathcal{B}_0$ is the collection of all open balls, all closed balls, etc. When $\mathcal{X} = \mathbb{R}$ a very useful choice is the collection $\{(-\infty, a] : a \in Q\}$, where Q is a dense set in $\mathbb{R}$.

As noted earlier $M(\mathbb{R})$ when equipped with weak convergence becomes a complete separable metric space with $\mathcal{B}_M$ as the Borel σ-algebra. Thus a natural topology on $M(\mathbb{R})$ is the weak topology arising from this metric space structure of $M(\mathbb{R})$. Formally, we have the following definitions.

Definition 2.5.1. A sequence of probability measure $\{\Pi\}_n$ on $M(\mathbb{R})$ is said to *converge weakly* to a probability measure Π if

$$\int \phi(P)\ d\Pi_n \to \int \phi(P)\ d\Pi$$

for all bounded continuous functions ϕ on $M(\mathbb{R})$.

Note that continuity of ϕ is with respect to the weak topology on $M(\mathbb{R})$. If f is a bounded continuous function on $\mathbb{R}$ then $\phi(P) = \int f dP$ is bounded and continuous on $M(\mathbb{R})$. However in general there is no clear description of all the bounded continuous functions on $M(\mathbb{R})$. If $\mathcal{X}$ is compact metric, then the following description is available.

If $\mathcal{X}$ is compact metric then, by Prohorov's theorem, so is $M(\mathcal{X})$ under weak convergence. It follows from the Stone-Weirstrass theorem that the set of all functions of the form

$$\sum \prod_{j=1}^{k_i} \phi_{f_{i,j}}^{r_i}$$

where $\phi_{f_{i,j}}^{r_i}(P) = \int f_{i,j}(x) dP(x)$ with $f_{i,j}(x)$ continuous on $\mathcal{X}$, is dense in the space of all continuous functions on $M(\mathcal{X})$.

The following result is an extension of a similar result in Sethuraman and Tiwari [149].

Theorem 2.5.1. *A family of probability measures $\{\Pi_t : t \in T\}$ on $M(\mathbb{R})$ is tight with respect to weak convergence on $M(\mathbb{R})$ iff the family of expectations $\{E_{\Pi_t} : t \in T\}$, where $E_{\Pi_t}(B) = \int P(B)\ d\Pi_t(P)$, is tight in $\mathbb{R}$.*

Proof. Let $\mu_t = E_{\Pi_t}$. Fix $\delta > 0$. By the tightness of $\{\mu_t : t \in T\}$, for every positive integer d there exists a sequence of compact sets K_d in $\mathbb{R}$, such that $\sup_t \mu_t(K_d^c) \leq 6\delta/(d^3\pi^2)$.

For $d = 1, 2, \ldots,$ let, $M_d = \{P \in M(\mathbb{R}) : P(K_d^c) \le 1/d\}$, and let $M = \cap_d M_d$. Then, by the pormanteau and Prohorov theorems, M is a compact subset of $M(\mathbb{R})$, in the weak topology. Further, by Markov's inequality,

$$\begin{aligned}
\Pi_n(M_d^c) &\le dE_{\Pi_t}(P(K_d^c)) \\
&= d\mu_t(K_d^c) \\
&\le \frac{6\delta}{d^2\pi^2}
\end{aligned}$$

Hence, for any $t \in T$, $\Pi_t(M) \le \sum_d 6\delta/(d^3\pi^2) = \delta$. This proves that $\{\mu_t\}_{t\in T}$ is tight. The converse is easy. □

Theorem 2.5.2. *Suppose $\Pi, \Pi_n, n \ge 1$ are probability measures on M. If any of the following holds then Π_n converges weakly to Π.*

(i) For any $(B_1, B_2, \ldots, B_k)$ of Borel sets

$$\mathcal{L}_{\Pi_n}(P(B_1), P(B_2), \ldots, P(B_k)) \to \mathcal{L}_{\Pi}(P(B_1), P(B_2), \ldots, P(B_k))$$

(ii) For any disjoint collection $(B_1, B_2, \ldots, B_k)$ of Borel sets

$$\mathcal{L}_{\Pi_n}(P(B_1), P(B_2), \ldots, P(B_k)) \to \mathcal{L}_{\Pi}(P(B_1), P(B_2), \ldots, P(B_k))$$

(iii) For any $(B_1, B_2, \ldots, B_k)$ where for $= i = 1, 2, \ldots, k$, $B_i = (a_i, b_i]$,

$$\mathcal{L}_{\Pi_n}(P(B_1), P(B_2), \ldots, P(B_k)) \to \mathcal{L}_{\Pi}(P(B_1), P(B_2), \ldots, P(B_k))$$

(iv) For any $(B_1, B_2, \ldots, B_k)$ where for $= i = 1, 2, \ldots, k$, $B_i = (a_i, b_i]$,a_i, b_i rationals,

$$\mathcal{L}_{\Pi_n}(P(B_1), P(B_2), \ldots, P(B_k)) \to \mathcal{L}_{\Pi}(P(B_1), P(B_2), \ldots, P(B_k))$$

(v) For any $(B_1, B_2, \ldots, B_k)$ where for $= i = 1, 2, \ldots, k$, $B_i = (-\infty, t_i]$,

$$\mathcal{L}_{\Pi_n}(P(B_1), P(B_2), \ldots, P(B_k)) \to \mathcal{L}_{\Pi}(P(B_1), P(B_2), \ldots, P(B_k))$$

(vi) For any $(B_1, B_2, \ldots, B_k)$ where for $= i = 1, 2, \ldots, k$,$B_i = (-\infty, t_i]$, t_i rationals

$$\mathcal{L}_{\Pi_n}(P(B_1), P(B_2), \ldots, P(B_k)) \to \mathcal{L}_{\Pi}(P(B_1), P(B_2), \ldots, P(B_k))$$

Proof. Because (vi) is the weakest, we will show that (vi) implies $\Pi_n \overset{weakly}{\to} \Pi$. Note that for all rationals t, $E_{\Pi_n}(P(-\infty, t)) \to E_{\Pi}(P(-\infty, t))$ and hence E_{Π_n} converges weakly to E_{Π}. By the Theorem 2.5.1 this shows that $\{\Pi_n\}$ is tight. If Π^* is the limit of any subsequence of $\{\Pi_n\}$, then it follows, using Proposition 2.5.1, that $\Pi^* = \Pi$. □

Remark 2.5.2. Note that $\Pi_n \overset{weakly}{\to} \Pi$ does not imply any of the preceding. The modifications are easy, however. For example (i) would be changed to "For any $(B_1, B_2, \ldots, B_k)$ of Borel sets such that $(P(B_1), P(B_2), \ldots, P(B_k))$ is continuous a.e Π."

We have considered other topologies on $M(\mathbb{R})$ namely, total variation, setwise convergence and the supremum metric. It is tempting to consider the weak topologies on probabilities on $M(\mathbb{R})$ induced by these topologies. But as we have observed, these topologies possess properties that make the notion of weak convergence awkward to define and work with. Besides, the σ-algebras generated by these topologies, via either open sets or open balls do not coincide with $\mathcal{B}_M$ [57]. Our interests do not demand such a general theory. Our chief interest is when the limit measure Π is degenerate at P_0, and in this case we can formalize convergence via weak neighborhoods of P_0.

When $\Pi = \delta_{P_0}$, $\Pi_n \overset{weakly}{\to} \delta_{P_0}$ iff $\Pi_n(U) \to \Pi(U)$ for every open neighborhood U. Because weak neighborhoods of P_0 are of the form $U = \{P : \left|\int f_i \, dP - \int f_i \, dP_0\right|\}$, weak convergence to a degenerate measure δ_{P_0} can be described in terms of continuous functions of $\mathbb{R}$ rather than those on $M(\mathbb{R})$ and can be verified more easily. The next proposition is often useful when we work with weak neighborhoods of a probability P_0 on $\mathbb{R}$.

Proposition 2.5.2. *Let Q be a countable dense subset of $\mathbb{R}$. Given any weak neighborhood U of P_0 there exist $a_1 < a_2 \ldots < a_n$ in Q and $\delta > 0$ such that*

$$\{P : |P[a_i, a_{i+1}) - P_0[a_i, a_{i+1})| < \delta \text{ for } 1 \le i \le n\} \subset U$$

Proof. Suppose $U = \{P : |\int f dP - \int f dP_0| < \epsilon\}$, where f is continuous with compact support. Because Q is dense in $\mathbb{R}$ given δ there exist $a_1 < a_2 \ldots < a_n$ in Q such that $f(x) = 0$ for $x \le a_1$, $x \ge a_n$, and $|f(x) - f(y)| < \delta$ for x $\in [a_i, a_{i+1}], 1 \le i \le n-1$. Then the function f^* defined by

$$f^*(x) = f(a_i) \text{ for } x \in [a_i, a_{i+1}), i = 1, 2, \ldots n-1$$

satisfies $\sup_x |f^*(x) - f(x)| < \delta$.

For any P, $\int f^* dP = \sum f(a_i)P[a_i, a_{i+1})$,

$$|\int f^* dP - \int f^* dP_0| < ck\delta \text{ where } c = \sup_x |f(x)|$$

In addition, if P is in U then we have

$$|\int f dP - \int f dP_0| < 2\delta + ck\delta$$

Thus with $B_i = [a_i, a_{i+1}]$ for small enough δ,$\{P : |P(B_i) - P_0(B_i)| < \delta\}$ is contained in U. The preceding argument is easily extended if U is of the form

$$\{P : |\int f_i dP - \int f_i dP_0| \leq \epsilon_i, 1 \leq i \leq k, f_i \text{ continuous with compact support}\}$$

□

Following is another useful proposition.

Proposition 2.5.3. *Let* $U = \{F : \sup_{-\infty<x<\infty} |F_0(s) - F(x)| < \epsilon\}$ *be a supremum neighborhood of a continuous distribution function* F_0. *Then* U *contains a weak neighborhood of* F_0.

Proof. Choose $-\infty = x_0 < x_1 < x_2 < \ldots < x_k = \infty$ such that $F(x_{i+1}) - F(x_i) < \epsilon/4$ for $i = 1, \ldots, k-1$. Consider

$$W = \{F : |F(x_i) - F_0(x_i)| < \epsilon/4\}, i = 1, 2, \ldots, k$$

If $x \in (x_{i-1}, x_i)$,

$$\begin{aligned} |F(x) - F_0(x)| \leq & |F(x_{i-1}) - F_0(x_i)| \vee |F(x_i) - F_0(x_{i-1})| \\ \leq & |F(x_{i-1}) - F_0(x_{i-1})| + |F_0(x_{i-1}) - F_0(x_i)| \\ & + |F(x_i) - F_0(x_i)| + |F_0(x_{i-1}) - F_0(x_i)| \end{aligned}$$

which is less than ϵ if $F \in W$. □

2.6 De Finetti's Theorem

Much of classical statistics has centered around the conceptually simplest setting of independent and identically distributed observations. In this case, $X_1, X_2, \ldots$ are a sequence of i.i.d. random variables with an unknown common distribution P. In the parametric case, P would be constrained to lie in a parametric family, and in the general nonparametric situation P could be any element of $M(\mathbb{R})$. The Bayesian framework in this case consists of a prior Π on the parameter set $M(\mathbb{R})$; given P the $X_1, X_2, \ldots$ is modeled as i.i.d. P. In a remarkable theorem, De Finetti showed that a minimal judgment of exchangeability of the observation sequence leads to the Bayesian formulation discussed earlier.

In this section we briefly discuss De Finetti's theorem. A detailed exposition of the theorem and related topics can be found in Schervish [144] in the section on De Finetti's theorem and the section on Extreme models.

As before, let $X_1, X_2, \ldots$ be a sequence of $\mathcal{X}$ -valued random variables defined on $\Omega = \mathbf{R}^\infty$.

Definition 2.6.1. Let μ be a probability measure on $\mathbb{R}^\infty$. The sequence $X_1, X_2, \ldots$ is said to be *exchangeable* if, for each n and for every permutation g of $\{1, \ldots, n\}$, the distribution of $X_1, X_2, \ldots, X_n$is the same as that of $X_{g(1)}, X_{g(2)}, \ldots, X_{g(n)}$.

Theorem 2.6.1 (De Finetti). *Let μ be a probability measure on $\mathbb{R}^\infty$. Then $X_1, X_2, \ldots$ is exchangeable iff there is a unique probability measure Π on $M(\mathbb{R})$ such that for all n and for any Borel sets $B_1, B_2, \ldots, B_n$,*

$$\mu\{X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n\} = \int_{M(\mathbb{R})} \prod_1^n P(B_i)\ d\Pi(P) \tag{2.4}$$

Proof. We begin by proving the theorem when all the X_is take values in a finite set $\mathcal{X} = \{1, 2, \ldots, k\}$. This proof follows Heath and Sudderth [95].

So let $\mathcal{X} = \{1, 2, \ldots, k\}$ and μ be a probability measure on $\mathcal{X}^\infty$ such that $X_1, X_2, \ldots$ is exchangeable. For each n, let $T_n(X_1, X_2, \ldots, X_n) = (r_1, r_2, \ldots, r_k)$, where $r_j = \sum_{i=1}^n I_{\{j\}}(X_i)$ is the number of occurrences of js in $X_1, X_2, \ldots, X_n$. Let μ_n^* denote the distribution of $T_n/n = (r_1/n, r_2/n, \ldots, r_k/n)$ under μ. μ_n^* is then a discrete probability measure on $M(\mathcal{X})$ supported by points of the form $(r_1/n, r_2/n, \ldots, r_k/n)$, where for $j = 1, 2, \ldots, k$, $r_j \geq 0$ is an integer and $\sum r_j = n$. Because $M(\mathcal{X})$ is compact, there is a subsequence $\{n_i\}$ that converges to a probability measure Π on $M(\mathcal{X})$. We will argue that Π satisfies (2.4).

Because $X_1, X_2, \dots, X_n$ is exchangeable, it is easy to see that the conditional distribution of $X_1, X_2, \dots, X_n$ given T_n is also exchangeable. In particular, the conditional probability given $T_n(X_1, X_2, \dots, X_n) = (r_1, r_2, \dots, r_k)$ is just the uniform distribution on $T_n^{-1}(r_1, r_2, \dots, r_k)$. In other words, the conditional distribution of $X_1, X_2, \dots, X_n$ given $T_n = (r_1, r_2, \dots, r_k)$ is the same as the distribution of n successive draws from an urn containing n balls with r_i of color i, for $i = 1, 2, \dots, k$.

Fix m and $n > m$. Then, given $T_n(X_1, X_2, \dots, X_n) = (r_1, r_2, \dots, r_k)$, the conditional probability that

$$\left(X_1 = 1, \dots, X_{s1} = 1, X_{s_1+1} = 2, \dots, X_{s_1+s_2} = 2, \dots, X_{m-s_{k-1}+1} = k, \dots, X_m = k\right)$$

is $(r_1)_{s_1}(r_2)_{s_2}\dots(r_k)_{s_k}/(n)_m$, where for any real a and integer b, $(a)_b = \prod_0^{b-1}(a-i)$.

Because

$$\sum_{(r_1,r_2,\dots,r_k),\sum r_j = n} \frac{(r_1)_{s_1}(r_2)_{s_2}\dots(r_k)_{s_k}}{(n)_m}\mu\left(\frac{T_n}{n} = (\frac{r_1}{n}, \frac{r_2}{n}, \dots, \frac{r_k}{n})\right)$$
$$= \int_{M(\mathcal{X})} \frac{(p_1 n)_{s_1}(p_2 n)_{s_2}\dots(p_k n)_{s_k}}{(n)_m}\ d\mu_n^*(p_1, p_2, \dots, p_k)$$

As $n \to \infty$ the sequence of functions

$$\frac{(p_1 n)_{s_1}(p_2 n)_{s_2}\dots(p_k n)_{s_k}}{(n)_m}$$

converges uniformly on $M(\mathcal{X})$to $\prod p_j^{s_j}$ so that by taking the limit through the subsequence $\{n_i\}$, the probability of

$$(X_i = 1, 1 \le i \le s_1; X_i = 2, s_1 + 1 \le i \le s_1 + s_2, \dots, X_i = k, m - s_{k-1} + 1 \le i \le m)$$

is

$$\int_{M(\mathcal{X})} \prod p_j^{s_j}\ d\Pi(p_1, p_2, \dots, p_k) \tag{2.5}$$

Uniqueness is immediate because if Π_1, Π_2 are two probability measures on $M(\mathcal{X})$ satisfying (2.5) then it follows immediately that they have the same moments.

To move on to the general case $\mathcal{X} = \mathbb{R}$, let $B_1, B_2, \dots, B_k$ be any collection of disjoint Borel sets in $\mathbb{R}$. Set $B_0 = \left(\cup_1^k B_i\right)^c$.

Define $Y_1, Y_2, \ldots$ by $Y_i = j$ if $X_i \in B_j$. Because $X_1, X_2, \ldots$ is exchangeable, so are $Y_1, Y_2, \ldots$. Since each Y_i takes only finitely many values, we use what we have just proved and writing $X_i \in B_j$ for $Y_i = j$, there is probability measure $\Pi_{B_1,B_2,\ldots,B_k}$ on $\{p_1, p_2, \ldots, p_k : p_j \geq 0, \sum p_j \leq 1\}$ such that for any m,

$$\mu\left(X_1 \in B_{i1}, X_2 \in B_{i2}, \ldots, X_m \in B_{im}\right) = \int \prod_1^m P(B_{ij})\, d\Pi_{B_1,B_2,\ldots,B_k}(P) \qquad (2.6)$$

where $i1, i2, \ldots, im$ are all elements of $\{0, 1, 2, \ldots, k\}$ and $P(B_0) = 1 - \sum_1^k P(B_i)$.

We will argue that these $\Pi_{B_1,B_2,\ldots,B_k}$s satisfy the conditions of Theorem 2.3.4.

If $A_1, A_2, \ldots, A_l$ is a collection of disjoint Borel sets such that B_i are union of sets from $A_1, A_2, \ldots, A_l$ then the distribution of $P(B_1), P(B_2), \ldots, P(B_k)$ obtained from $P(A_1), P(A_2), \ldots, P(A_l)$ and $\Pi_{B_1,B_2,\ldots,B_k}$ both would satisfy (2.5). Uniqueness then shows that both distributions are same.

If $(B_{1n}, B_{2n}, \ldots, B_{kn}) \to (B_1, B_2, \ldots, B_k)$ then (2.6) again shows that moments of $\Pi_{B_{1n},B_{2n},\ldots,B_{kn}}$ converges to the corresponding moment of $\Pi_{B_1,B_2,\ldots,B_k}$.

It is easy to verify the other conditions of Theorem 2.3.4. Hence there exists a Π with $\Pi_{B_1,B_2,\ldots,B_k}$s as marginals. It is easy to verify that Π satisfies (2.4). □

De Finetti's theorem can be viewed from a somewhat general perspective. Let $\mathcal{G}_n$ be the group of permutations on $\{1, 2, \ldots, n\}$ and let $\mathcal{G} = \cup \mathcal{G}_n$. Every $g \in \mathcal{G}$ induces in a natural way a transformation on $\Omega = \mathcal{X}^\infty$ through the map, if, say g in $\mathcal{G}_n$, then $(x_1, \ldots, x_n, \ldots) \mapsto (x_{g(1)}, \ldots, x_{g(n)}, \ldots)$. It is easy to see that the set of exchangeable probability measures is the same as the set of probability measures on Ω that are invariant under $\mathcal{G}$. This set is a convex set, and De Finetti's theorem asserts that the set of extreme points of this convex set is $\{P^\infty : P \in M(\mathcal{X})\}$ and that every invariant measure is representable as an average over the set of extreme points. This view of exchangeable measures suggests that by suitably enlarging $\mathcal{G}$ it would be possible to obtain priors that are supported by interesting subsets of $M(\mathcal{X})$. Following is a simple, trivial example.

Example 2.6.1. Let $H = \{h, e\}$, where $h(x) = -x$ and $e(x) = x$. Set $\mathcal{H} = \cup H^n$. If $(h_1, h_2, \ldots, h_n)) \in H^n$, then the action on Ω is defined by $(x_1, x_2, \ldots, x_n) \mapsto (h(x_1), h(x_2), \ldots, h(x_n)$. Then an exchangeable probability measure μ is $\mathcal{H}$ invariant iff it is a mixture of symmetric i.i.d. probability measures. To see this by De Finetti's theorem

$$\mu(A) = \int P^\infty(A) d\Pi(P)$$

Because by $\mathcal{H}$ invariance $\mu(X_1 \in A, X_2 \in -A) = \mu(X_1 \in A, X_2 \in A)$, it is not hard to see that $E_\Pi(P(A) - P(-A))^2 = 0$. Letting A run through a countable algebra generating the σ-algebra on $\mathcal{X}$, we have the result.

More non trivial examples are in Freedman [68]

Sufficiency provides another frame through which De Finetti's theorem can be usefully viewed. The ideas leading to such a view and the proofs involve many measure-theoretic details. Most of the interesting examples involve invariance and sufficiency in some form. We do not discuss these aspects here but refer the reader to the excellent survey in Schervish [144], the paper by Diaconis and Freedman [[44]] and Fortini, Ladelli, and Regazzini [67].

To use DeFinetti's theorem to construct a specific prior on $M(\mathbb{R})$, we need to know what to expect from the prior in terms of the observables $X_1, X_2, \ldots, X_n$. Although this method of assigning a prior is attractive from a philosophical point of view, it is not easy to either describe explicitly an exchangeable sequence or identify a prior, given such a sequence. We will not pursue this aspect here.