Chapter 6

Stationary Stochastic Processes.

6.1 Ergodic Theorems.

A stationary stochastic process is a collection $\{\xi_n : n \in Z\}$ of random variables with values in some space $(X, \mathcal{B})$ such that the joint distribution of $(\xi_{n_1}, \cdots, \xi_{n_k})$ is the same as that of $(\xi_{n_1+n}, \cdots, \xi_{n_k+n})$ for every choice of $k \geq 1$, and $n, n_1, \cdots, n_k \in Z$. Assuming that the space $(X, \mathcal{B})$ is reasonable and Kolmogorov's consistency theorem applies, we can build a measure P on the countable product space Ω of sequences $\{x_n : n \in Z\}$ with values in X, defined for sets in the product σ-field $\mathcal{F}$. On the space Ω there is the natural shift defined by $(T\omega)(n) = x_{n+1}$ for ω with $\omega(n) = x_n$. The random variables $x_n(\omega) = \omega(n)$ are essentially equivalent to $\{\xi_n\}$. The stationarity of the process is reflected in the invariance of P with respect to T i.e. $PT^{-1} = P$. We can without being specific consider a space Ω a σ-field $\mathcal{F}$, a one to one invertible measurable map from $\Omega \to \Omega$ with a measurable inverse T^{-1} and finally a probability measure P on $(\Omega, \mathcal{F})$ that is T-invariant i.e $P(T^{-1}A) = P(A)$ for every $A \in \mathcal{F}$. One says that P is an invariant measure for T or T is a measure preserving transformation for P. If we have a measurable map from $\xi : (\Omega, \mathcal{F}) \to (X, \mathcal{B})$, then it is easily seen that $\xi_n(\omega) = \xi(T^n\omega)$ defines a stationary stochastic process. The study of stationary stochastic process is then more or less the same as the study of measure preserving (i.e. probability preserving) transformations.

The basic transformation $T : \Omega \to \Omega$ induces a linear transformation U

on the space of functions defined on Ω by the rule $(Uf)(\omega) = f(T\omega)$. Because T is measure preserving it is easy to see that

$$\int_\Omega f(\omega)\, dP = \int_\Omega f(T\omega)\, dP = \int_\Omega (Uf)(\omega)\, dP$$

as well as

$$\int_\Omega |f(\omega)|^p\, dP = \int_\Omega |f(T\omega)|^p\, dP = \int_\Omega |(Uf)(\omega)|^p\, dP.$$

In other words U acts as an isometry (i.e. norm presrving linear transformation) on the various L_p spaces for $1 \leq p < \infty$ and in fact it is an isometry on L_∞ as well. Moreover the transformation induced by T^{-1} is the inverse of U so that U is also invertible. In particular U is unitary (or orthogoanl)on L_2. This means it presrves the inner product $< \cdot, \cdot >$.

$$< f, g >= \int f(\omega)g(\omega)dP = \int f(T\omega)g(T\omega)dP =< Uf, Ug > .$$

Of course our linear transformation U is very special and satisfies $U1 = 1$ and $U(fg) = (Uf)(Ug)$.

A basic theorem known as the Ergodic theorem asserts that

Theorem 6.1. *For any $f \in L_1(P)$ the limit*

$$\lim_{n\to\infty} \frac{f(\omega) + f(T\omega) + \cdots + f(T^{n-1}\omega)}{n} = g(\omega)$$

exists for almost all ω with respect to P as well as in $L_1(P)$. Moreover if $f \in L_p$ for some p satisfying $1 < p < \infty$ then the function $g \in L_p$ and the convergence takes place in that L_p. Moreover the limit $g(\omega)$ is given by the conditional expectation

$$g(\omega) = E^P[f|\mathcal{I}]$$

where the σ-field $\mathcal{I}$, called the invariant σ-field, is defined as

$$\mathcal{I} = \{A : TA = A\}.$$

Proof. Fisrst we prove the convergence in the various L_p spaces. These are called mean ergodic theorems. The easiest situation to prove is when $p = 2$. Let us define

$$H_0 = \{f : f \in H, Uf = f\} = \{f : f \in H,\ f(T\omega) = f(\omega)\}.$$

Since H_0 contains constants, it is a closed nontrivial subspace of $H = L_2(P)$, of dimension at least one. Since U is unitary $Uf = f$ if and only if $U^{-1}f = U^*f = f$ where U^* is the adjoint of U. The orthogonal complement $H_0^\perp$ can be defined as

$$H_0^\perp = \{g :< g, f >= 0 \ \forall f : U^*f = f\} = \overline{\text{Range}(I-U)H} \ .$$

Clearly if we let

$$A_n f = \frac{f + Uf + \cdots + U^{n-1}f}{n}$$

then $\|A_n f\|_2 \leq \|f\|_2$ for every $f \in H$ and $A_n f = f$ for every n and $f \in H_0$. Therefore for $f \in H_0$, $A_n f \to f$ as $n \to \infty$. On the other hand if $f = (I-U)g$, $A_n f = \frac{g - U^n g}{n}$ and $\|A_n f\|_2 \leq \frac{2\|g\|_2}{n} \to 0$ as $n \to \infty$. Since $\|A_n\| \leq 1$, it follows that $A_n f \to 0$ as $n \to \infty$ for every $f \in H_0^\perp = \overline{\text{Range}(I-U)H}$. (See exercise 6.1). If we denote by π the orthogonal projection from $H \to H_0$, we see that $A_n f \to \pi f$ as $n \to \infty$ for every $f \in H$ establishing the L_2 ergodic theorem.

There is an alternate characterization of H_0. Functions f in H_0 are invariant under T, i.e. have the property that $f(T\omega) = f(\omega)$. For any invariant function f the level sets $\{\omega : a < f(\omega) < b\}$ are invariant under T. We can therefore talk about invariant sets $\{A : A \in \mathcal{F}, T^{-1}A = A\}$. Technically we should allow ourselves to differ by sets of measure zero and one defines $\mathcal{I} = \{A : P(A \,\Delta T^{-1}A) = 0\}$ as the σ-field of almost invariant sets. . Nothing is therefore lost by taking $\mathcal{I}$ to be the σ-field of invariant sets. We can identify the orthogonal projection π as (see Exercise 4.8)

$$\pi f = E^P\{f|\mathcal{I}\}$$

and as the conditional expectation operator, π is well defined on L_p as an operator of norm 1, for all p in the range $1 \leq p \leq \infty$. If $f \in L_\infty$, then $\|A_n f\|_\infty \leq \|f\|_\infty$ and by the bounded convergence theorem, for any p satisfying $1 \leq p < \infty$, we have $\|A_n f - \pi f\|_p \to 0$ as $n \to \infty$. Since L_∞ is dense in L_p and $\|A_n\| \leq 1$ in all the L_p spaces it is easily seen, by a simple approximation argument, that for each p in $1 \leq p < \infty$ and $f \in L_p$,

$$\lim_{n\to\infty} \|A_n f - f\|_p = 0$$

proving the mean ergodic theorem in all the L_p spaces.

We now concentrate on proving almost sure convergence of $A_n f$ to πf for $f \in L_1(P)$. This part is often called the 'individual ergodic theorem' or

'Birkhoff's theorem' . This will be based on an analog of Doob's inequality for martingales. First we will establish an inequality called the maximal ergodic theorem.

Theorem 6.2. (Maximal Ergodic Theorem.) *Let $f \in L_1(P)$ and for $n \geq 1$, let*

$$E_n^0 = \{\omega : \sup_{1 \leq j \leq n} [f(\omega) + f(T\omega) + \cdots + f(T^{j-1}\omega)] \geq 0]\}.$$

Then

$$\int_{E_n^0} f(\omega)\, dP \geq 0$$

Proof. Let

$$\begin{aligned} h_n(\omega) &= \sup_{1 \leq j \leq n} [f(\omega) + f(T\omega) + \cdots + f(T^{j-1}\omega)] \\ &= f(\omega) + \max(0\,,\, h_{n-1}(T\omega)) \\ &= f(\omega) + h_{n-1}^+(T\omega) \end{aligned}$$

where

$$h_n^+(\omega) = \max(0\,,\, h_n(\omega)).$$

On E_n^0, $h_n(\omega) = h_n^+(\omega)$ and therefore

$$f(\omega) = h_n(\omega) - h_{n-1}^+(T\omega) = h_n^+(\omega) - h_{n-1}^+(T\omega).$$

Consequently,

$$\begin{aligned} \int_{E_n^0} f(\omega)\, dP &= \int_{E_n^0} [h_n^+(\omega) - h_{n-1}^+(T\omega)]\, dP \\ &\geq \int_{E_n^0} [h_n^+(\omega) - h_n^+(T\omega)]\, dP \qquad \text{(because } h_{n-1}^+(\omega) \leq h_n^+(\omega)\text{)} \\ &= \int_{E_n^0} h_n^+(\omega)\, dP - \int_{TE_n^0} h_n^+(\omega)\, dP \quad \text{(because of inavraince of } T\text{)} \\ &\geq 0. \end{aligned}$$

The last step follows from the fact that for any integrable function $h(\omega)$, $\int_E h(\omega)\, dP$ is the largest when we take for E the set $E = \{\omega : h(\omega) \geq 0\}$. □

Now we establish the analog of Doob's inequality or maximal inequality, or sometimes referred to as the weaktype $1-1$ inequality.

Lemma 6.3. *For any $f \in L_1(P)$, and $\ell > 0$, denoting by E_n the set*

$$E_n = \{\omega : \sup_{1\le j\le n} |(A_j f)(\omega)| \ge \ell\}$$

we have

$$P\big[E_n\big] \le \frac{1}{\ell}\int_{E_n} |f(\omega)|\, dP.$$

In particular

$$P\big[\omega : \sup_{j\ge 1} |(A_j f)(\omega)| \ge \ell\big] \le \frac{1}{\ell}\int |f(\omega)|\, dP.$$

Proof. We can assume without loss of generality that $f \in L_1(P)$ is nonnegative. Apply the lemma to $f - \ell$. If

$$E_n = \{\omega : \sup_{1\le j\le n} \frac{[f(\omega) + f(T\omega) + \cdots + f(T^{j-1}\omega)]}{j} > \ell\},$$

then

$$\int_{E_n} [f(\omega) - \ell]\, dP \ge 0$$

or

$$P[E_n] \le \frac{1}{\ell}\int_{E_n} f(\omega)\, dP.$$

We are done. □

Given the lemma the proof of the almost sure ergodic theorem follows along the same lines as the proof of the almost sure convergence in the martingale context. If $f \in H_0$ it is trivial. For $f = (I-U)g$ with $g \in L_\infty$ it is equally trivial because $\|A_n f\|_\infty \le \frac{2\|g\|_\infty}{n}$. So the almost sure convergence is valid for $f = f_1 + f_2$ with $f_1 \in H_0$ and $f_2 = (I-U)g$ with $g \in L_\infty$. But such functions are dense in $L_1(P)$. Once we have almost sure convergence for a dense set in $L_1(P)$, the almost sure convergence for every $f \in L_1(P)$ follows by routine approximation using Lemma 6.3. See the proof of Theorem 5.7. □

Exercise 6.1. For any bounded linear transformation A on a Hilbert Space H, show that the closure of the range of A, i.e $\overline{\text{Range}\, A}$ is the orthogonal complement of the null space $\{f : A^* f = 0\}$ where A^* is the adjoint of A.

Exercise 6.2. Show that any almost invariant set differs by a set of measure 0 from an invariant set i.e. if $P(A\,\Delta T^{-1}A) = 0$ then there is a $B \in \mathcal{F}$ with $P(A\Delta B) = 0$ and $T^{-1}B = B$.

Although the ergodic theorem implies a strong law of large numbers for any stationary sequence of random variables, in particular a sequence of independent identically distributed random variables, it is not quite the end of the story. For the law of large numbers, we need to know that the limit πf is a constant, which will then equal $\int f(\omega)\, dP$. To claim this, we need to know that the invariant σ-field is trivial or essentially consists of the whole space Ω and the empty set Φ. An invariant measure P is said to be ergodic for the transformation T, if every $A \in \mathcal{I}$ i.e every invariant set has measure 0 or 1. Then every invariant function is almost surely a constant and $\pi f = E\big[f|\mathcal{I}\big] = \int f(\omega)\, dP$.

Theorem 6.4. *Any product measure is ergodic for the shift.*

Proof. Let A be an invariant set. Then A can be approximated by sets A_n in the σ-field corresonding to the coordinates from $[-n, n]$. Since A is invariant $T^{\pm 2n} A_n$ will approximate A just as well. This proves that A actually belongs to the tail σ-field, the remote past as well as the remote future. Now we can use Kolmogorov's $0-1$ law (Theorem 3.15), to assert that $P(A) = 0$ or 1. □

6.2 Structure of Stationary Measures.

Given a space $(\Omega, \mathcal{F})$ and a measurable transformation T with a measurable inverse T^{-1}, we can consider the space $\mathcal{M}$ of all T-invariant probability measures on $(\Omega, \mathcal{F})$. The set $\mathcal{M}$, which may be empty, is easily seen to be a convex set.

Exercise 6.3. Let $\Omega = Z$, the integers, and for $n \in Z$, let $Tn = n + 1$. Show that $\mathcal{M}$ is empty.

Theorem 6.5. *A probability measure $P \in \mathcal{M}$ is ergodic if and only if it is an extreme point of $\mathcal{M}$.*

Proof. A point of a convex set is extreme if it cannot be written as a nontrivial convex combination of two other points from that set. Suppose $P \in \mathcal{M}$ is not extremal. Then P can be written as nontrivial convex combination of $P_1, P_2 \in \mathcal{M}$, i.e. for some $0 < a < 1$ and $P_1 \neq P_2$, $P = aP_1 + (1-a)P_2$. We claim that such a P cannot be ergodic. If it were, by definition, $P(A) = 0$ or 1 for every $A \in \mathcal{I}$. Since $P(A)$ can be 0 or 1 only when $P_1(A) = P_2(A) = 0$ or $P_1(A) = P_2(A) = 1$, it follows that for every invariant set $A \in \mathcal{I}$, $P_1(A) = P_2(A)$. We now show that if two invariant measures P_1 and P_2 agree on $\mathcal{I}$, they agree on $\mathcal{F}$. Let $f(\omega)$ be any bounded $\mathcal{F}$-measurable function. Consider the function

$$h(\omega) = \lim_{n\to\infty} \frac{1}{n}[f(\omega) + f(T\omega) + \cdots + f(T^{n-1}\omega)]$$

defined on the set E where the limit exists. By the ergodic theorem $P_1(E) = P_2(E) = 1$ and h is $\mathcal{I}$ measurable. Moreover, by the stationarity of P_1, P_2 and the bounded convergence theorem,

$$E^{P_i}[f(\omega)] = \int_E h(\omega) dP_i \qquad \text{for } \; i = 1, 2$$

Since $P_1 = P_2$ on $\mathcal{I}$ and h is $\mathcal{I}$ measurable and $P_i(E) = 1$ for $i = 1, 2$ we see that

$$E^{P_1}[f(\omega)] = E^{P_2}[f(\omega)]$$

Since f is arbitrary this implies that $P_1 = P_2$ on $\mathcal{F}$.

Conversely if P is not ergodic, then there is an $A \in \mathcal{I}$ with $0 < P(A) < 1$ and we define

$$P_1(E) = \frac{P(A \cap E)}{P(A)} \; ; \; P_2(E) = \frac{P(A^c \cap E)}{P(A^c)}.$$

Since $A \in \mathcal{I}$ it follows that P_i are stationary. Moreover $P = P(A)P_1 + P(A^c)P_2$ and hence P is not extremal. □

One of the questions in the theory of convex sets is the existence of sufficiently many extremal points, enough to recover the convex set by taking convex combinations. In particular one can ask if any point in the convex set can be obtained by taking a weighted average of the extremals. The next theorem answers the question in our context. We will assume that our space $(\Omega, \mathcal{F})$ is nice, i.e. is a complete separable metric space with its Borel sets.

Theorem 6.6. *For any invariant measure P, there is a probability measure μ_P on the set $\mathcal{M}_e$ of ergodic measures such that*

$$P = \int_{\mathcal{M}_e} Q \ \mu_P(dQ)$$

Proof. If we denote by P_ω the regular conditional probability distribution of P given $\mathcal{I}$, which exists (see Theorem 4.4) because $(\Omega, \mathcal{F})$ is nice, then

$$P = \int_\Omega P_\omega \, P(d\omega)$$

We will complete the proof by showing that P_ω is an ergodic stationary probability measure for almost all ω with respect to P. We can then view P_ω as a map $\Omega \to \mathcal{M}_e$ and μ_P will be the image of P under the map. Our integral representation in terms of ergodic measures will just be an immediate consequence of the change of variables formula. □

Lemma 6.7. *For any stationary probability measure P, for almost all ω with respect to P, the regular conditional probability distribution P_ω, of P given $\mathcal{I}$, is stationary and ergodic.*

Proof. Let us first prove stationarity. We need to prove that $P_\omega(A) = P_\omega(TA)$ a.e. We have to negotiate carefully through null sets. Since a measure on the Borel σ-field $\mathcal{F}$ of a complete separable metric space is determined by its values on a countable generating field $\mathcal{F}_0 \subset \mathcal{F}$, it is sufficient to prove that for each fixed $A \in \mathcal{F}_0$, $P_\omega(A) = P_\omega(TA)$ a.e. P. Since P_ω is $\mathcal{I}$ measurable all we need to show is that for any $E \in \mathcal{I}$,

$$\int_E P_\omega(A) \, P(d\omega) = \int_E P_\omega(TA) \, P(d\omega)$$

or equivalently

$$P(E \cap A) = P(E \cap TA)$$

This is obvious because P is stationary and E is invariant.

We now turn to ergodicity. Again there is a minefield of null sets to negotiate. It is a simple exercise to check that if, for some stationary measure Q, the ergodic theorem is valid with an almost surely constant limit for the indicator functions $\mathbf{1}_A$ with $A \in \mathcal{F}_0$, then Q is ergodic. This needs to be checked only for a countable collection of sets $\{A\}$. We need therfore only to

check that any invariant function is constant almost surely with respect to almost all P_ω. Equivalently for any invariant set E, $P_\omega(E)$ must be shown almost surely to be equal to 0 or 1. But $P_\omega(E) = \chi_E(\omega)$ and is always 0 or 1. This completes the proof. □

Exercise 6.4. Show that any two distinct ergodic invariant measures P_1 and P_2 are orthogonal on $\mathcal{I}$, i.e. there is an invariant set E such that $P_1(E) = 1$ and $P_2(E) = 0$.

Exercise 6.5. Let $(\Omega, \mathcal{F}) = ([0,1), \mathcal{B})$ and $Tx = x + a \pmod 1$. If a is irrational there is just one invariant measure P, namely the uniform distribution on $[0,1)$. This is seen by Fourier Analysis. See Remark 2.2.

$$\int e^{i\,2\,n\,\pi\,x} dP = \int e^{i\,2\,n\,\pi\,(Tx)} dP = \int e^{i\,2\,n\,\pi\,(x+a)} dP = e^{i\,2\,n\,\pi\,a} \int e^{i\,2\,n\,\pi\,x} dP$$

If a is irrational $e^{i\,2\,n\,\pi\,a} = 1$ if and only if $n = 0$. Therefore

$$\int e^{i\,2\,n\,\pi\,x} dP = 0 \qquad \textit{for } n \neq 0$$

which makes P uniform. Now let $a = \frac{p}{q}$ be rational with $(p,q) = 1$, i.e. p and q are relatively prime. Then, for any x, the discrete distribution with probabilities $\frac{1}{q}$ at the points $\{x, x+a, x+2a, \ldots, x+(q-1)a\}$ is invariant and ergodic. We can denote this distribution by P_x. If we limit x to the interval $0 \leq x < \frac{1}{q}$ then x is uniquely determined by P_x. Complete the example by determining all T invariant probability distributions on $[0,1)$ and find the integral representation in terms of the ergodic ones.

6.3 Stationary Markov Processes.

Let $\pi(x, dy)$ be a transition probability function on $(X, \mathcal{B})$, where X is a state space and $\mathcal{B}$ is a σ-field of measurable subsets of X. A stochastic process with values in X is a probability measure on the space $(\Omega, \mathcal{F})$, where Ω is the space of sequences $\{x_n : -\infty < n < \infty\}$ with values in X, and $\mathcal{F}$ is the product σ-field. The space $(\Omega, \mathcal{F})$ has some natural sub σ-fields. For any two integers $m \leq n$, we have the sub σ-fields, $\mathcal{F}_n^m = \sigma\{x_j : m \leq j \leq n\}$ corresponding to information about the process during the time interval $[m, n]$. In addition we have $\mathcal{F}_n = \mathcal{F}_n^{-\infty} = \sigma\{x_j : j \leq n\}$ and $\mathcal{F}^m = \mathcal{F}_\infty^m = \sigma\{x_j : j \geq m\}$ that

correspond to the past and future. P is a Markov Process on $(\Omega, \mathcal{F})$ with transition probability $\pi(\cdot, \cdot)$, if for every n, $A \in \mathcal{B}$ and P-almost all ω,

$$P\big\{x_{n+1} \in A | \mathcal{F}_n\big\} = \pi(x_n\,, A)$$

Remark 6.1. Given a π, it is not always true that P exists. A simple but illuminating example is to take $X = \{0, 1, \cdots, n, \cdots\}$ to be the nonnegative integers and define $\pi(x, x+1) = 1$ and all the process does is move one step to the right every time. Such a process if it had started long time back will be found nowhere today! So it does not exist. On the other hand if we take X to be the set of all integers then P is seen to exist. In fact there are lots of them. What is true however is that given any initial distribution μ and initial time m, there exist a unique process P on $(\Omega, \mathcal{F}^m)$, i.e. defined on the future σ-field from time m on, that is Markov with transition probability π and satisfies $P\{x_m \in A\} = \mu(A)$ for all $A \in \mathcal{B}$.

The shift T acts naturally as a measurable invertible map on the product space Ω into itself and the notion of a stationary process makes sense. The following theorem connects stationarity and the Markov property.

Theorem 6.8. *Let the transition probability π be given. Let P be a stationary Markov process with transition probability π. Then the one dimensional marginal distribution μ, which is independent of time because of stationarity and given by*

$$\mu(A) = P\big\{x_n \in A\big\}$$

is π invariant in the sense that

$$\mu(A) = \int \pi(x, A)\mu(dx)$$

for every set $A \in \mathcal{B}$. Conversely given such a μ, there is a unique stationary Markov process P with marginals μ and transition probability π.

Exercise 6.6. Prove the above Theorem. Use Remark 4.7.

Exercise 6.7. If P is a stationary Markov process on a countable state space with transition probaility π and invariant marginal distribution μ, show that the time reversal map that maps $\{x_n\}$ to $\{x_{-n}\}$ takes P to another stationary Markov process Q, and express the transition probability $\hat{\pi}$ of Q, as explicitly as you can in terms of π and μ .

Exercise 6.8. If μ is an invariant measure for π, show that the conditional expectation map $\Pi : f(\cdot) \to \int f(y)\,\pi(\,\cdot\,, dy)$ induces a contraction in $L_p(\mu)$ for any $p \in [1, \infty]$. We say that a Markov process is reversible if the time reversed process Q of the previous example coincides with P. Show that P corresponding to π and μ is reversible if and only if the corresponding Π in $L_2(\mu)$ is self-adjoint or symmetric.

Since a given transition probability π may in general have several invariant measures μ, there will be several stationary Markov processes with transition probability π. Let $\mathcal{M}$ be the set of invariant probability measures for the transition probabilities $\pi(x, dy)$ i.e.

$$\widetilde{\mathcal{M}} = \left\{ \mu : \mu(A) = \int_X \pi(x\,, A)\, d\mu(x) \;\text{ for all }\; A \in \mathcal{B} \right\}$$

$\widetilde{\mathcal{M}}$ is a convex set of probability mesures and we denote by $\widetilde{\mathcal{M}}_e$ its (possibly empty) set of extremals. For each $\mu \in \widetilde{\mathcal{M}}$, we have the corresponding stationary Markov process P_μ and the map $\mu \to P_\mu$ is clearly linear. If we want P_μ to be an ergodic stationary process, then it must be an extremal in the space of all stationary processes. The extremality of $\mu \in \widetilde{\mathcal{M}}$ is therfore a necessary condition for P_μ to be ergodic. That it is also sufficient is a little bit of a surprise. The following theorem is the key step in the proof. The remaining part is routine.

Theorem 6.9. *Let μ be an invariant measure for π and $P = P_\mu$ the corresponding stationary Markov process. Let $\mathcal{I}$ be the σ-field of shift invariant subsets on Ω. To within sets of P measure 0, $\mathcal{I} \subset \mathcal{F}_0^0$.*

Proof. This theorem describes completely the structure of nontrivial sets in the σ-field $\mathcal{I}$ of invariant sets for a stationary Markov process with transition probability π and marginal distribution μ. Suppose that the state space can be partitioned nontrivially i.e. with $0 < \mu(A) < 1$ into two sets A and A^c that satisfy $\pi(x, A) = 1$ a.e μ on A and $\pi(x, A^c) = 1$ a.e μ on A^c. Then the event

$$E = \{\omega : x_n \in A \;\text{ for all }\; n \in Z\}$$

provides a non trivial set in $\mathcal{I}$. The theorem asserts the converse. The proof depends on the fact that an invariant set E is in the remote past $\mathcal{F}_{-\infty}^{-\infty} = \cap_n \mathcal{F}_n^{-\infty}$ as well as in the remote future $\mathcal{F}_\infty^\infty = \cap_m \mathcal{F}_\infty^m$. See the proof of Theorem 6.4. For a Markov process the past and the future are

conditionally independent given the present. See Theorem 4.9. This implies that

$$P\big[E|\mathcal{F}_0^0\big] = P\big[E \cap E|\mathcal{F}_0^0\big] = P\big[E|\mathcal{F}_0^0\big]\, P\big[E|\mathcal{F}_0^0\big]$$

and must therfore equal either 0 or 1. This in turn means that corresponding to any invariant set $E \in \mathcal{I}$, there exists $A \subset X$ that belongs to $\mathcal{B}$, such that $E = \{\omega : x_n \in A \text{ for all } n \in Z\}$ up to a set of P measure 0. If the Markov process starts from A or A^c, it does not ever leave it. That means $0 < \mu(A) < 1$ and

$$\pi(x, A^c) = 0 \;\text{ for }\; \mu \;\text{ a.e }\; x \in A \;\text{ and }\; \pi(x, A) = 0 \;\text{ for }\; \mu \;\text{ a.e }\; x \in A^c$$

□

Remark 6.2. One way to generate markov processes with multiple invariant measures is to start with two markov processes with transition probabilities $\pi_i(x_i, dy_i)$ on X_i and invariant measures μ_i, and consider $X = X_1 \cup X_2$. Define

$$\pi(x, A) = \begin{cases} \pi_1(x, A \cap X_1) & \text{if } x \in X_1 \\ \pi_2(x, A \cap X_2) & \text{if } x \in X_2 \end{cases}$$

Then any one of the two processes can be going on depending on which world we are in. Both μ_1 and μ_2 are invariant measures. We have combined two distinct possibilities into one. What we have shown is that when we have multiple invariant measures they essentially arise in this manner.

Remark 6.3. We can therefore look at the convex set of measures μ that are π invariant, i.e. $\mu\Pi = \mu$. The extremals of this convex set are precisely the ones that correspond to ergodic stationary processes and they are called ergodic or extremal invariant measures. If the set of invariant probability measures is nonempty for some π, then there are enough extremals to recover arbitrary invariant measure as an integral or weighted average of extremal ones.

Exercise 6.9. Show that any two distinct extremal invariant measures μ_1 and μ_2 for the same π are orthogonal on $\mathcal{B}$.

Exercise 6.10. Consider the operator Π on the $L_p(\mu)$ spaces corresponding to a given invariant measure. The dimension of the eigenspace $f : \Pi f = f$ that corresponds to the eigenvalue 1, determines the extremality of μ. Clarify this statement.

Exercise 6.11. Let P_x be the Markov process with stationary transition probability $\pi(x, dy)$ starting at time 0 from $x \in X$. Let f be a bounded measurable function on X. Then for almost all x with respect to any extemal invariant measure ν,

$$\lim_{n\to\infty} \frac{1}{n}[f(x_1) + \cdots + f(x_n)] = \int f(y)\nu(dy)$$

for almost all ω with respect to P_x.

Exercise 6.12. We saw in the earlier section that any stationary process is an integral over stationary ergodic processes. If we represent a stationary Markov Process P_μ as the integral

$$P_\mu = \int R\, Q(dR)$$

over stationary ergodic processes, show that the integral really involves only stationary Markov processes with transition probability π, so that the integral is really of the form

$$P_\mu = \int_{\widetilde{\mathcal{M}_e}} P_\nu\, Q(d\nu)$$

or equivalently

$$\mu = \int_{\widetilde{\mathcal{M}_e}} \nu\, Q(d\nu).$$

Exercise 6.13. If there is a reference measure α such that $\pi\,(x\,,\,dy)$ has a density $p(x, y)$ with respect to α for every α, then show that any invariant measure μ is absolutely continuous with respect to α. In this case the eigenspace $f : \Pi f = f$ in $L_2(\mu)$ gives a complete picture of all the invariant measures.

The question of when there is at most one invariant measure for the Markov process with transition probability π is a difficult one. If we have a density $p(x, y)$ with respect to a reference measure α and if for each x, $p(x, y) > 0$ for almost all y with respect to α, then there can be atmost one inavriant measure. We saw already that any invariant measure has a density with respect to α. If there are at least two invariant mesaures, then there are at least two ergodic ones which are orthogonal. If we denote by f_1 and

f_2 their densities with respect to α, by orthogonality we know that they are supported on disjoint ivariant sets, A_1 and A_2. In particular $p(x,y) = 0$ for almost all x on A_1 in the support of f_1 and almost all y in A_2 with respect to α. By our positivity assumption we must have $\alpha(A_2) = 0$, which is a contradiction.

6.4 Mixing properties of Markov Processes.

One of the questions that is important in the theory of Markov Processes is the rapidity with which the memory of the initial state is lost. There is no unique way of assessing it and depending on the circumstances this could happen in many differerent ways at many different rates. Let $\pi^{(n)}(x\,, dy)$ be the n step transition probability. The issue is how the measures $\pi^{(n)}(x\,, dy)$ depend less and less on x as $n \to \infty$. Suppose we measure this dependence by

$$\rho_n = \sup_{x,y \in X} \sup_{A \in \mathcal{B}} |\pi^{(n)}(x\,, A) - \pi^{(n)}(y\,, A)|$$

then the following is true.

Theorem 6.10. *Either* $\rho_n \equiv 1$ *for all* $n \geq 1$, *or* $\rho_n \leq C\theta^n$ *for some* $0 \leq \theta < 1$

Proof. From the Chapman-Kolmogorov equations

$$\pi^{(n+m)}(x\,, A) - \pi^{(n+m)}(y\,, A) = \int \pi^{(m)}(z\,, A)[\pi^{(n)}(x\,, dz) - \pi^{(n)}(y\,, dz)]$$

If $f(x)$ is a function with $|f(x) - f(y)| \leq C$ and $\mu = \mu_1 - \mu_2$ is the difference of two probability measures with $\|\mu\| = \sup_A |\mu(A)| \leq \delta$, then it is elementary to estimate, using $\int c d\mu = 0$,

$$|\int f\, d\mu| = \inf_c |\int (f - c) d\mu| \leq 2 \inf_c \{\sup_x |f(x) - c|\} \|\mu\| \leq 2\frac{C}{2}\delta = C\delta$$

It follows that the sequence ρ_n is submultiplicative, i.e.

$$\rho_{m+n} \leq \rho_m \rho_n$$

Our theorem follows from this property. As soon as some $\rho_k = a < 1$ we have

$$\rho_n \leq [\rho_k]^{[\frac{n}{k}]} \leq C\theta^n$$

with $\theta = a^{\frac{1}{k}}$. □

Although this is an easy theorem it can be applied in some contexts.

Remark 6.4. If $\pi(x\,, dy)$ has density $p(x, y)$ with respect to some reference measure α and $p(x, y) \geq q(y) \geq 0$ for all y with $\int q(y)d\alpha \geq \delta > 0$, then it is elementary to show that $\rho_1 \leq (1 - \delta)$.

Remark 6.5. If $\rho_n \to 0$, we can estimate

$$|\pi^{(n)}(x\,, A) - \pi^{(n+m)}(x\,, A)| = |\int [\pi^{(n)}(x\,, A) - \pi^{(n)}(y, A)]\pi^{(m)}(x\,, dy)| \leq \rho_n$$

and conclude from the estimate that

$$\lim_{n\to\infty} \pi^{(n)}(x\,, A) = \mu(A)$$

exists. μ is seen to be an invariant probability measure.

Remark 6.6. In this context the invariant measure is unique. If β is another invariant measure because

$$\beta(A) = \int \pi^{(n)}(x\,, A)\beta(dy)$$

for every $n \geq 1$

$$\beta(A) = \lim_{n\to\infty} \int \pi^{(n)}(x\,, A)\beta(dy) = \mu(A).$$

Remark 6.7. The stationary process P_μ has the property that if $E \in \mathcal{F}_m^{-\infty}$ and $F \in \mathcal{F}_\infty^n$ with a gap of $k = n - m > 0$ then

$$\begin{aligned} P_\mu[E \cap F] &= \int_E \int_X \pi^{(k)}(x_m(\omega), dx)P_x(T^{-n}F)P_\mu(d\omega) \\ P_\mu[E]P_\mu[F] &= \int_E \int_X \mu(dx)P_x(T^{-n}F)P_\mu(d\omega) \\ P_\mu[E \cap F] - P_\mu[E]P_\mu[F] &= \int_E \int_X P_x(T^{-n}F)[\pi^{(k)}(x_m(\omega), dx) - \mu(dx)]P_\mu(d\omega) \end{aligned}$$

from which it follows that

$$|P_\mu[E \cap F] - P_\mu[E]P_\mu[F]| \leq \rho_k \; P_\mu(E)$$

proving an asymptotic independence property for P_μ.

There are situations in which we know that an invariant probability measure μ exists for π and we wish to establish that $\pi^{(n)}(x\,,A)$ converges to $\mu(A)$ uniformly in A for each $x \in X$ but not necessarily uniformly over the starting points x. Uniformity in the starting point is very special. We will illustrate this by an example.

Example 6.1. The Ornstein-Uhlenbeck process is Markov Chain on the state space $X = \mathbf{R}$, the real line with transition probability $\pi(x, dy)$ given by a Gaussian distribution with mean ρx and variance σ^2. It has a density $p(x, y)$ with respect to the Lebesgue measure so that $\pi(x, A) = \int_A p(x, y)dy$.

$$p(x, y) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(y - \rho x)^2}{2\sigma^2}]$$

It arises from the 'auto-regressive' representation

$$x_{n+1} = \rho x_n + \sigma\xi_{n+1}$$

where $\xi_1, \cdots, \xi_n \cdots$ are independent standard Gaussians. The characteristic function of any invariant mesure $\phi(t)$ satisfies, for every $n \geq 1$,

$$\phi(t) = \phi(\rho t) \exp[-\frac{\sigma^2 t^2}{2}] = \phi(\rho^n t) \exp[-\frac{(\sum_{j=0}^{n-1} \rho^{2j})\sigma^2 t^2}{2}]$$

by induction on n. Therefore

$$|\phi(t)| \leq \exp[-\frac{(\sum_{j=0}^{n-1} \rho^{2j})\sigma^2 t^2}{2}]$$

and this cannot be a characteristic function unless $|\rho| < 1$ (otherwise by letting $n \to \infty$ we see that $\phi(t) = 0$ for $t \neq 0$ and therefore discontinuous at $t = 0$). If $|\rho| < 1$, by letting $n \to \infty$ and observing that $\phi(\rho^n t) \to \phi(0) = 1$

$$\phi(t) = \exp[-\frac{\sigma^2 t^2}{2(1 - \rho^2)}]$$

The only possible invariant measure is the Gaussian with mean 0 and variane $\frac{\sigma^2}{(1-\rho^2)}$. One can verify that this Gaussian is infact an invariant measure. If $|\rho| < 1$ a direct computation shows that $\pi^{(n)}(x, dy)$ is a Gaussian with mean $\rho^n x$ and variance $\sigma_n^2 = \sum_{j=0}^{n-1} \rho^{2j}\sigma^2 \to (1 - \rho^2)\sigma^2$ as $n \to \infty$. Clearly there is uniform convergence only over bounded sets of starting points x. This is typical.

6.5 Central Limit Theorem for Martingales.

If $\{\xi_n\}$ is an ergodic stationary sequence of random variables with mean zero then we know from the ergodic theorem that the mean $\frac{\xi_1 \cdots + \xi_n}{n}$ converges to zero almost surely. by the law of large numbers. We want to develop some methods for proving the central limit theorem, i.e. the covergence of the distribution of $\frac{\xi_1 + \cdots + \xi_n}{\sqrt{n}}$ to some Gaussian distribution with mean 0 variance σ^2. Under the best of situations, since the covariance $\rho_k = E[X_n X_{n+k}]$ may not be 0 for all $k \neq 0$, if we assume that $\sum_{-\infty<j<\infty} |\rho_j| < \infty$, we get

$$\begin{aligned}
\sigma^2 &= \lim_{n\to\infty} \frac{1}{n} E[(\xi_1 + \cdots + \xi_n)^2] \\
&= \lim_{n\to\infty} \sum_{|j|\leq n} (1 - \frac{|j|}{n})\rho_j \\
&= \sum_{-\infty<j<\infty} \rho_j \\
&= \rho_0 + 2\sum_{j=1}^{\infty} \rho_j.
\end{aligned}$$

The standard central limit theorem with $\sqrt{n}$ scaling is not likely to work if the covariances do not decay rapidly enough to be summable. When the covariances $\{\rho_k\}$ are all 0 for $k \neq 0$ the variance calculation yields $\sigma^2 = \rho_0$ just as in the independent case, but there is no guarantee that the central limit theorem is valid.

A special situation is when $\{\xi_j\}$ are square integrable martingale differences. With the usual notation for the σ-fields $\mathcal{F}_n^m$ for $m \leq n$ (remember that m can be $-\infty$ while n can be $+\infty$) we assume that

$$E\{\xi_n | \mathcal{F}_{n-1}\} = 0 \quad a.e.$$

and in this case by conditioning we see that $\rho_k = 0$ for $k \neq 0$. It is a useful and important observation that in this context the central limit theorm always holds. The distribution of $Z_n = \frac{\xi_1 + \cdots + \xi_n}{\sqrt{n}}$ converges to the normal distribution with mean 0 and variance $\sigma^2 = \rho_0$. The proof is a fairly simple modification of the usual proof of the central limit theorem. Let us define

$$\psi(n, j, t) = \exp[\frac{\sigma^2 t^2 j}{2n}] E\{ \exp[i\, t\, \frac{\xi_1 + \cdots + \xi_j}{\sqrt{n}}] \}$$

and write

$$\psi(n,n,t) - 1 = \sum_{j=1}^{n} [\psi(n,j,t) - \psi(n,j-1,t)]$$

leaving us with the estimation of

$$\Delta(n,t) = \Big| \sum_{j=1}^{n} \big[\psi(n,j,t) - \psi(n,j-1,t)\big] \Big|.$$

Theorem 6.11. *For an ergodic stationary sequence $\{\xi_j\}$ of square integrable martingale diffrences, the central limit theorem is always valid.*

Proof. We let $S_j = \xi_1 + \cdots + \xi_j$ and calculate

$$\begin{aligned}\big[\psi(n,j,t) &- \psi(n,j-1,t)\big] \\ &= \exp[\frac{\sigma^2 t^2 j}{2n}] E\bigg\{ \exp[it\frac{S_{j-1}}{\sqrt{n}}] \bigg[\exp[it\frac{\xi_j}{\sqrt{n}}] - \exp[-\frac{\sigma^2 t^2}{2n}] \bigg] \bigg\}.\end{aligned}$$

We can replace it with

$$\theta(n,j,t) = \exp[\frac{\sigma^2 t^2 j}{2n}] E\bigg\{ \exp[it\frac{S_{j-1}}{\sqrt{n}}] \bigg[\frac{(\sigma^2 - \xi_j^2)t^2}{2n} \bigg] \bigg\}$$

because the error can be controlled by Taylor's expansion. In fact if we use the martingale difference property to kill the linear term, we can bound the difference, in an arbitrary finite interval $|t| \le T$, by

$$\begin{aligned}&|\big[\psi(n,j,t) - \psi(n,j-1,t)\big] - \theta(n,j,t)| \\ &\quad \le C_T E\bigg\{ \Big| \exp[it\frac{\xi_j}{\sqrt{n}}] - 1 - it\frac{\xi_j}{\sqrt{n}} + \frac{t^2 \xi_j^2}{2n} \Big| \bigg\} + C_T |\exp[-\frac{\sigma^2 t^2}{2n}] - 1 + \frac{\sigma^2 t^2}{2n}|\end{aligned}$$

where C_T is a constant that depends only on T. The right hand side is independent of j because of stationarity. By Taylor expansions in the variable $\frac{t}{\sqrt{n}}$ of each of the two terms on the right, it is easily seen that

$$\sup_{\substack{|t| \le T \\ 1 \le j \le n}} |\big[\psi(n,j,t) - \psi(n,j-1,t)\big] - \theta(n,j,t)| = o(\frac{1}{n}).$$

Therefore

$$\sup_{|t|\le T}\sum_{j=1}^{n}|\big[\psi(n,j,t)-\psi(n,j-1,t)\big]-\theta(n,j,t)|=n\,o(\frac{1}{n})\to 0.$$

We now concentrate on estimating $|\sum_{j=1}^{n}\theta(n,j,t)|$. We pick an integer k which will be large but fixed. We divide $[1,n]$ into blocks of size k with perhaps an incomplete block at the end. We will now replace $\theta(n,j,t)$ by

$$\theta_k(n,j,t)=\exp[\frac{\sigma^2t^2kr}{2n}]E\Big\{\exp[it\frac{S_{kr}}{\sqrt{n}}]\Big[\frac{(\sigma^2-\xi_j^2)t^2}{2n}\Big]\Big\}$$

for $kr+1\le j\le k(r+1)$ and $r\ge 0$.

Using stationarity it is easy to estimate for $r\le\frac{n}{k}$,

$$|\sum_{j=kr+1}^{k(r+1)}\theta_k(n,j,t)\,|\le C(t)\frac{1}{n}E\Big\{\,|\sum_{j=kr+1}^{k(r+1)}(\sigma^2-\xi_j^2)|\Big\}=C(t)\frac{k}{n}\delta(k)$$

where $\delta(k)\to 0$ as $k\to\infty$ by the L_1 ergodic theorem. After all $\{\xi_j^2\}$ is a stationary sequence with mean σ^2 and the ergodic theorem applies. Since the above estimate is uniform in r, the left over incomplete block at the end causes no problem and there are approximately $\frac{n}{k}$ blocks, we conclude that

$$|\sum_{j=1}^{n}\theta_k(n,j,t)|\le C(t)\delta(k).$$

On the other hand, by stationarity,

$$\begin{aligned}\sum_{j=1}^{n}&|\theta_k(n,j,t)-\theta(n,j,t)|\\&\le n\sup_{1\le j\le n}|\theta_k(n,j,t)-\theta(n,j,t)|\\&\le C(t)\sup_{1\le j\le k}E\Big\{\Big|\exp[\frac{\sigma^2t^2j}{2n}]\exp[it\frac{S_{j-1}}{\sqrt{n}}]-1\Big||\sigma^2-\xi_j^2|\Big\}\end{aligned}$$

and it is elementary to show by dominated convergence theorem that the right hand side tends to 0 as $n\to\infty$ for each finite k.

This concludes the proof of the theorem. □

One may think that the assumption that $\{\xi_n\}$ is a martingale difference is too restrictive to be useful. Let $\{X_n\}$ be any stationary process with zero mean. We can often succeed in writing $X_n = \xi_{n+1} + \eta_{n+1}$ where ξ_n is a martingale difference and η_n is negligible, in the sense that $E[(\sum_{j=1}^n \eta_j)^2] = o(n)$. Then the central limit theorem for $\{X_n\}$ can be deduced from that of $\{\xi_n\}$. A cheap way to prove $E[(\sum_{j=1}^n \eta_j)^2] = o(n)$ is to establish that $\eta_n = Z_n - Z_{n+1}$ for some stationary square integrable sequence $\{Z_n\}$. Then $\sum_{j=1}^n \eta_j$ telescopes and the needed estimate is obvious. Here is a way to construct Z_n from X_n so that $X_n + (Z_{n+1} - Z_n)$ is a martingale difference.

Let us define

$$Z_n = \sum_{j=0}^{\infty} E\big\{X_{n+j}|\mathcal{F}_n\big\}$$

There is no guarantee that the series converges, but we can always hope. After all, if the memory is weak, prediction j steps ahead should be futile if j is large. Therefore if X_{n+j} is becoming independent of $\mathcal{F}_n$ as j gets large one would expect $E\big\{X_{n+j}|\mathcal{F}_n\big\}$ to approach $E[X_{n+j}]$ which is assumed to be 0. By stationarity n plays no role. If Z_0 can be defined the shift operator T can be used to define $Z_n(\omega) = Z_0(T^n\omega)$. Let us assume that $\{Z_n\}$ exist and are square integrable. Then

$$Z_n = E\big\{Z_{n+1}|\mathcal{F}_n\big\} + X_n$$

or equivalently

$$\begin{aligned} X_n &= Z_n - E\big\{Z_{n+1}|\mathcal{F}_n\big\} \\ &= [Z_n - Z_{n+1}] + [Z_{n+1} - E\big\{Z_{n+1}|\mathcal{F}_n\big\}] \\ &= \eta_{n+1} + \xi_{n+1} \end{aligned}$$

where $\eta_{n+1} = Z_n - Z_{n+1}$ and $\xi_{n+1} = Z_{n+1} - E\big\{Z_{n+1}|\mathcal{F}_n\big\}$. It is easy to see that $E[\xi_{n+1}|\mathcal{F}_n] = 0$.

For a stationary ergodic Markov process $\{X_n\}$ on state space $(X, \mathcal{B})$, with transition probability $\pi(x\,, dy)$ and invariant measure μ, we can prove the central limit theorem by this method. Let $Y_j = f(X_j)$. Using the Markov property we can calculate

$$Z_0 = \sum_{j=0}^{\infty} E[f(X_j)|\mathcal{F}_0] = \sum_{j=0}^{\infty} [\Pi^j f](X_0) = [[I - \Pi]^{-1} f](X_0).$$

If the equation $[I - \Pi]U = f$ can be solved with $U \in L_2(\mu)$, then

$$\xi_{n+1} = U(X_{n+1}) - U(X_n) + f(X_n)$$

is a martingale difference and we have a central limit theorem for $\frac{\sum_{j=1}^{n} f(X_j)}{\sqrt{n}}$ with variance given by

$$\sigma^2 = E^{P_\mu}\big\{[\,\xi_0\,]^2\big\} = E^{P_\mu}\big\{[U(X_1) - U(X_0) + f(X_0)]^2\big\}.$$

Exercise 6.14. Let us consider a two state Markov Chain with states $[1, 2]$. Let the transition probabilities be given by $\pi(1,1) = \pi(2,2) = p$ and $\pi(1,2) = \pi(2,1) = q$ with $0 < p, q < 1$, $p + q = 1$. The invariant measure is given by $\mu(1) = \mu(2) = \frac{1}{2}$ for all values of p. Consider the random variable $S_n = A_n - B_n$, where A_n and B_n are respectively the number of visits to the states 1 and 2 during the first n steps. Prove a central limit theorem for $\frac{S_n}{\sqrt{n}}$ and calculate the limiting variance as a function $\sigma^2(p)$ of p. How does $\sigma^2(p)$ behave as $p \to 0$ or 1 ? Can you explain it? What is the value of $\sigma^2(\frac{1}{2})$? Could you have guessed it?

Exercise 6.15. Consider a random walk on the nonnegative integers with

$$\pi(x\,,y) = \begin{cases} \frac{1}{2} & \textit{for all } x = y \geq 0 \\ \frac{1-\delta}{4} & \textit{for } y = x+1, x \geq 1 \\ \frac{1+\delta}{4} & \textit{for } y = x-1, x \geq 1 \\ \frac{1}{2} & \textit{for } x = 0, y = 1. \end{cases}$$

Prove that the chain is positive recurrent and find the invariant measure $\mu(x)$ explicitly. If $f(x)$ is a function on $x \geq 0$ with compact support solve explicitly the equation $[I - \Pi]U = f$. Show that either U grows exponentially at infinity or is a constant for large x. Show that it is a constant if and only if $\sum_x f(x)\mu(x) = 0$. What can you say about the central limit theorem for $\sum_{j=0}^{n} f(X_j)$ for such functions f?

6.6 Stationary Gaussian Processes.

Considering the importance of Gaussian distributions in Probability theory, it is only natural to study stationary Gaussian processes, i.e. stationary

processes $\{X_n\}$ that have Gaussian distributions as their finite dimensional joint distributions. Since a joint Gaussian distribution is determined by its means and covariances we need only specify $E[X_n]$ and Cov $(X_n, X_m) = E[X_n X_m] - E[X_n]E[X_m]$. Recall that the joint density on $\mathbf{R}^n$ of N Gaussian random variables with mean $m = \{m_i\}$ and covariance $C = \{\rho_{i,j}\}$ is given by

$$p(y) = [\frac{1}{\sqrt{2\pi}}]^N \frac{1}{\sqrt{DetC}} \exp[-\frac{1}{2} < (y-m), C^{-1}(y-m) >]$$

Here m is the vector of means and C^{-1} is the inverse of the positive definite covariance matrix C. If C only positive semidefinite the Gaussian distribution lives on a lower dimensional hyperplane and is singular. By stationarity $E[X_n] = c$ is independent of n and Cov $(X_n, X_m) = \rho_{n-m}$ can depend only on the difference $n-m$. By symmetry $\rho_k = \rho_{-k}$. Because the covariance matrix is always positive semidefinite the sequence ρ_k has the positive definiteness property

$$\sum_{k,j=1}^{n} \rho_{j-k} z_j \bar{z_k} \geq 0$$

for all choices of n and complex numbers $z_1 \cdots, z_n$. By Bochner's theorem (see Theorem 2.2) there exists a nonnegative measure μ on the circle that is thought of as $S = [0, 2\pi]$ with end points identified such that

$$\rho_k = \int_0^{2\pi} \exp[ik\theta] d\mu(\theta)$$

and because of the symmetry of ρ_k, μ is symmetric as well with respect to $\theta \to 2\pi - \theta$. It is convenient to assume that $c = 0$. One can always add it back. Given a Gaussian process it is natural to carry out linear operations that will leave the Gaussian character unchanged. Rather than working with the σ-fields $\mathcal{F}_n^m$ we will work with the linear subspaces $\mathcal{H}_n^m$ spanned by $\{X_j : m \leq j \leq n\}$ and the infinite spans $\mathcal{H}_n = \vee_{m \leq n} \mathcal{H}_n^m$ and $\mathcal{H}^m = \vee_{n \geq m} \mathcal{H}_n^m$, that are considered as linear subspaces of the Hilbertspace $\mathcal{H} = \vee_{\substack{m,n \\ n \geq m}} \mathcal{H}_n^m$ which lies inside $L_2(P)$. But $\mathcal{H}$ is a small part of $L_2(P)$, consisting only of linear functions of $\{X_j\}$. The analog of Kolmogorov's tail σ-field are the subspaces $\wedge_m \mathcal{H}^m$ and $\wedge_n \mathcal{H}_n$ that are denoted by $\mathcal{H}^\infty$ and $\mathcal{H}_{-\infty}$. The analog of Kolmogorov's zero-one law would be that these subspaces are trivial having in them only the zero function. The symmetry in ρ_k implies that the processes $\{X_n\}$ and $\{X_{-n}\}$ have the same underlying distributions

so that both tails behave identically. A stationary Gaussian process $\{X_n\}$ with mean 0 is said to be purely non determinstic if the tail subspaces are trivial.

In finite dimensional theory a Covarance matrix can be diagonalized or better still written in special form T^*T, which gives a linear representation of the Gaussian random variables in terms of canonical or independent standard Gaussian random variables. The point to note is that if X is standard Gaussian with mean zero and covariance $I = \{\delta_{i,j}$, then for any linear trnasformation T, $Y = TX$ iagain Gaussian with mean zero and covariance $C = TT^*$. In other words if

$$Y_i = \sum_j t_{i,k} X_k$$

then

$$C_{i,j} = \sum_k t_{i,k} t_{j,k}$$

In fact for any C we can find a T which is upper or lower diagonal i.e. $t_{i,k} = 0$ for $i > k$ or $i < k$. If the indices correspond to time, this can be interpreted as a causal representation interms of current and future or past variables only.

The following questions have simple answers.

Q1. When does a Gaussian process have a moving average representation in terms of independent Gaussians i.e a representation of the form

$$X_n = \sum_{m=-\infty}^{\infty} a_{n-m} \xi_m$$

with

$$\sum_{n=-\infty}^{\infty} a_n^2 < \infty$$

in terms of i.i.d. Gaussians $\{\xi_k\}$ with mean 0 and variance 1 ?

If we have such a representation then the Covariance ρ_k is easily calculated as the convolution

$$\rho_k = \sum_j a_j a_{j+k} = [a * \bar{a}](k)$$

and that will make $\{\rho_k\}$ the Fourier coefficients of the function

$$f = |\sum_j a_j e^{i\,j\,\theta}|^2$$

which is the square of a function in $L_2(S)$. In other words the spectral measure μ will be absolutely continuous with a density f with repect to the normalized Lebesgue measure $\frac{d\theta}{2\pi}$. Conversely if we have a μ with a density f its square root will be a function in L_2 and will therefore have Fourier coefficients a_n in l_2 and a moving average representation holds in terms of i.i.d. random variables with these weights..

Q2. When does a Gaussian Process have a representation that is causal i.e. of the form

$$X_n = \sum_{j\geq 0} a_j \xi_{n-j}$$

with

$$\sum_{j\geq 0} a_j^2 < \infty?$$

If we do have a causal representation then the remote past of the $\{X_k\}$ process is clearly part of the remote past of the $\{\xi_k\}$ process. By Kolmogorov's zero-one law, the remote past for independent Gaussians is trivial and a causal representation is therefore possible for $\{X_k\}$ only if its remote past is trivial. The converse is true as well. The subspace $\mathcal{H}_n$ is spanned by $\mathcal{H}_{n-1}$ and X_n. Therefore either $\mathcal{H}_n = \mathcal{H}_{n-1}$, or $\mathcal{H}_{n-1}$ has codimension 1 in $\mathcal{H}_n$. In the former case by stationarity $\mathcal{H}_n = \mathcal{H}_{n-1}$ for every n. This inturn implies $\mathcal{H}_{-\infty} = \mathcal{H} = \mathcal{H}^{\infty}$. Assuming that the process is not identically zero i.e. $\rho_0 = \mu(S) > 0$ this makes the remote past or future the whole thing and definitely nontrivial. So we may assume that $\mathcal{H}_n = \mathcal{H}_{n-1} \oplus \mathbf{e_n}$ where $\mathbf{e_n}$ is a one dimensional subspace spanned by a unit vector ξ_n. Since all our random variables are linear combinations of a Gaussian collection they all have Gaussian distributions. We have the shift operator U satsfying $UX_n = X_{n+1}$ and we can assume with out loss of generality that $U\xi_n = \xi_{n+1}$ for every n. If we start with X_0 in our Hilbert space

$$X_0 = a_0\xi_0 + R_{-1}$$

with $R_1 \in \mathcal{H}_{n-1}$. We can continue and write

$$R_{-1} = a_1\xi_{-1} + R_{-2}$$

and so on. We will then have for every n

$$X_0 = a_0\xi_0 + a_1\xi_{-1} + \cdots + a_n\xi_{-n} + R_{-(n+1)}$$

with $R_{-(n+1)} \in \mathcal{H}_{-(n+1)}$. Since $\wedge_n \mathcal{H}_{-n} = \{0\}$ we conclude that the the expansion

$$X_0 = \sum_{j=0}^{\infty} a_j \xi_{-j}$$

is valid.

Q3. What are the conditions on the spectral density f in order that the process may admit a causal representation. From our answer to Q1. we know that we have to solve the following analytical problem. Given the spectral measure μ with a non negative density $f \in L_1(S)$, when can we write $f = |g|^2$ for some $g \in L_2(S)$, that admits a Fourier representation $g = \sum_{j\geq 0} a_j e^{i\,j\,\theta}$ involving only positive frequencies. This has the following neat solution which is far from obvious.

Theorem 6.12. *The process determined by the spectral density f admits a causal representation if and only if $f(\theta)$ satisfies*

$$\int_S \log f(\theta) d\theta > -\infty$$

Remark 6.8. Notice that the condition basically prevents f from vanishing on a set of positive measure or having very flat zeros.

The proof will use methods from the theory of functions of a complex variable. Define

Proof.

$$g(\theta) = \sum_{n\geq 0} c_n \exp[i\,n\,\theta]$$

as the Fourier series of some $g \in L_2(S)$. Assume $c_n \neq 0$ for some $n > 0$. In fact we can assume without loss of generality that $c_0 \neq 0$ by removing a suitable factor of $e^{i\,k\,\theta}$ which will not affect $|g(\theta)|$. Then we will show that

$$\frac{1}{2\pi}\int_S \log|g(\theta)| d\theta \geq \log|c_0|.$$

Consider the function

$$G(z) = \sum_{n \geq 0} c_n z^n$$

as an analytic function in the disc $|z| < 1$. It has boundary values

$$\lim_{r \to 1} G(re^{i\theta}) = g(\theta)$$

in $L_2(S)$. Since G is an analytic function we know, from the theory of functions of a complex variable, thatlog $|G(re^{i\theta})|$ is subharmonic and has the mean value property

$$\int_S \log |G(re^{i\theta})| d\theta \geq \log |G(0)| = \log |c_0|$$

Since $G(re^{i\theta})$ has a limit in $L_2(S)$, the positive part of $\log |G|$ which is dominated by $|G|$ is uniformly integrable. For the negative part we apply Fatou's lemma and derive our estimate.

Now for the converse. Let $f \in L_1(S)$. Assume $\int_S \log f(\theta) d\theta > -\infty$ or equivalently $\log f \in L_1(S)$. Define the Fourier coefficients

$$a_n = \frac{1}{4\pi} \int_S \log f(\theta) \exp[i\, n\, \theta]\, d\theta.$$

Because $\log f$ is integrable $\{a_n\}$ are uniformly bounded and the power series

$$A(z) = \sum a_n z^n$$

which is well defined for $|z| < 1$. We define

$$G(z) = \exp[A(z)].$$

We will show that

$$\lim_{r \to 1} G(re^{i\theta}) = g(\theta)$$

exists in $L_2(S)$ and $f = |g|^2$, g being the boundary value of an analytic function in the disc. The integral condition on $\log f$ is then the necessary and sufficient condition for writing $f = |g|^2$ with g involving only nonnegative frequencies.

$$
\begin{aligned}
|G(re^{i\theta})|^2 &= \exp\left[2\, Real\ Part\, A(re^{i\theta})\right] \\
&= \exp\left[2\sum_{j=0}^{\infty} a_j r^j \cos j\theta\right] \\
&= \exp\left[2\sum_{j=0}^{\infty} r^j \cos j\theta [\frac{1}{4\pi}\int_S \log f(\varphi)\cos j\varphi d\varphi]\right] \\
&= \exp\left[\frac{1}{2\pi}\int_S \log f(\varphi)[\sum_{j=0}^{\infty} r^j \cos j\theta \cos j\varphi d\varphi]\right] \\
&= \exp\left[\int_S \log f(\varphi) K(r,\theta,\varphi) d\varphi\right] \\
&\leq \int_S f(\varphi) K(r,\theta,\varphi) d\varphi
\end{aligned}
$$

Here K is the Poisson Kernel for the disc

$$K(r,\, \theta,\, \varphi) = \frac{1}{2\pi}\sum_{j=0}^{\infty} r^j \cos\theta \cos\varphi$$

is nonnegative and $\int_S K(r,\, \theta,\, \varphi)d\varphi = 1$. The last step is a consequence of Jensen's inequality. The function

$$f_r(\theta) = \int_S f(\varphi) K(r,\theta,\varphi) d\varphi$$

converges to f as $r \to 1$ in $L_1(S)$ by the properties of the Poisson Kernel. It is therefore uniformly integrable. Since $|G(re^{i\theta})|^2$ is dominated by f_r we get uniform integrability for $|G|^2$ as $r \to 1$. It is seen now that G has a limit g in $L_2(S)$ as $r \to 1$ and $f = |g|^2$. □

One of the issues in the theory of time series is that of prediction. We have a stochastic process $\{X_n\}$ that we have observed for times $n \leq -1$ and we want to predict X_0. The best predictor is $E^P[X_0|\mathcal{F}_{-1}]$ or in the Gaussian linear context it is the compuation of the projection of X_0 into $\mathcal{H}_{-1}$. If we have a moving average representation, even a causal one, while it is true that

X_j is spanned by $\{\xi_k : k \leq j\}$ the converse may not be true. If the two spans were the same, then the best predictor for X_0 is just

$$\hat{X}_0 = \sum_{j \geq 1} a_j \xi_{-j}$$

obtained by dropping one term in the original representation. In fact in answering **Q2** the construction yielded a representation with this property. The quantity $|a_0|^2$ is then the prediction error. In any case it is a lower bound.

Q4. What is the value of prediction error and how do we actually find the predictor ?

The situation is some what muddled. Let us assume that we have a purely nondeterministic process i.e. a process with a spectral density satisfying $\int_S \log f(\theta) d\theta > -\infty$. Then f can be represented as

$$f = |g|^2$$

with $g \in H_2$, where by H_2 we denote the subspace of $L_2(S)$ that are boundary values of analytic functions in the disc $|z| < 1$, or equivalently functions $g \in L_2(S)$ with only nonnegative frequencies. For any such g, we have an analytic function

$$G(z) = G(re^{i\theta}) = \sum_{n \geq 0} a_n r^n e^{in\theta}.$$

For any choice of $g \in H_2$ with $f = |g|^2$, we have

$$|G(0)|^2 = |a_0|^2 \leq \exp\big[\frac{1}{2\pi} \int_S \log f(\theta) d\theta\big]. \tag{6.1}$$

There is a choice of g contstructed in the proof of the thorem for which

$$|G(0)|^2 = \exp\big[\frac{1}{2\pi} \int_S \log f(\theta) d\theta\big] \tag{6.2}$$

The prediction error $\sigma^2(f)$, that depends only on f and not on the choice of g, also satisfies

$$\sigma^2(f) \geq |G(0)|^2 \tag{6.3}$$

for every choice of $g \in H_2$ with $f = |g|^2$. There is a choice of g such that

$$\sigma^2(f) = |G(0)|^2 \tag{6.4}$$

Therefore from (6.1) and (6.4)

$$\sigma^2(f) \leq \exp\big[\frac{1}{2\pi}\int_S \log f(\theta)d\theta\big] \tag{6.5}$$

On the other hand from (6.2) and (6.3)

$$\sigma^2(f) \geq \exp\big[\frac{1}{2\pi}\int_S \log f(\theta)d\theta\big] \tag{6.6}$$

We do now have an exact formula

$$\sigma^2(f) = \exp\big[\frac{1}{2\pi}\int_S \log f(\theta)d\theta\big] \tag{6.7}$$

for the prediction error.

As for the predictor, it is not quite that simple. In principle it is a limit of linear combinations of $\{X_j : j \leq 0\}$ and may not always have a simple concrete representation. But we can understand it a little better. Let us consider the spaces $\mathcal{H}$ and $L_2(S;\mu)$ of square integrable functions on S with respect to the spectral measure μ. There is a natural isomorphism between the two Hilbert spaces, if we map

$$\sum a_j X_j \longleftrightarrow \sum a_j e^{ij\theta}$$

The problem then is the question of approximating $e^{i\theta}$ in $L_2(S;\mu)$ by linear combinations of $\{e^{ij\theta} : j \leq 0\}$. We have already established that the error, which is nonzero in the purely nondeterministic case, i.e when $d\mu = \frac{1}{2\pi}f(\theta)d\theta$ for some $f \in L_1(S)$ satisfying

$$\int_S \log f(\theta)d\theta > -\infty,$$

is given by

$$\sigma^2(f) = \exp\big[\frac{1}{2\pi}\int_S \log f(\theta)d\theta\big]$$

We now want to find the best approximation.

In order to get at the predictor we have to make a very special choice of the representation $f = |g|^2$. Simply demanding $g \in L_2(S)$ will not even give causal representations. Demanding $g \in H_2$ will always give us causal representation, but there are too many of these. If we mutiply $G(z)$ by an analytic function $V(z)$ that has boundary values $v(\theta)$ satisfying $|v(\theta)| = |V(e^{i\theta})| \equiv 1$ on S, then gv is another choice. If we demand that

$$|G(0)|^2 = \exp\big[\frac{1}{2\pi}\int_S \log f(\theta)d\theta\big] \tag{6.8}$$

there is atleast one choice that will satisfy it. There is still ambiguity, albeit a trivial one among these, for we can always multiply g by a complex number of modulus 1 and that will not change anything of consequence. We have the following theorem.

Theorem 6.13. *The representation $f = |g|^2$ with $g \in H_2$, and satisfying (6.8), is unique to within a multiplicative constant of modulus 1. In other words if $f = |g_1|^2 = |g_2|^2$ with both g_1 and g_2 satisfying (8), then $g_1 = \alpha g_2$ on S, where α is a complex number of modulus 1.*

Proof. Let $F(re^{i\theta}) = \log|G(re^{i\theta})|$. It is a subharmonic function and

$$\lim_{r\to 1} F(re^{i\theta}) = \frac{1}{2}\log f(\theta)$$

Because

$$\lim_{r\to 1} G(re^{i\theta}) = g(\theta)$$

in $L_2(S)$, the functions are uniformly integrable in r. The positive part of the logarithm F is well controlled and therefore uniformly uniformly integrable. Fatou's lemma is applicable and we should always have

$$\limsup_{r\to 1}\frac{1}{2\pi}\int_S F(re^{i\theta})d\theta \le \frac{1}{4\pi}\int_S \log f(\theta)d\theta$$

But because F is subharmonic its average value on a circle of radius r around 0 is nondecreasing in r, and the lim sup is the same as the sup. Therefore

$$F(0) \le \sup_{0\le r<1}\frac{1}{2\pi}\int_S F(re^{i\theta})d\theta = \limsup_{r\to 1}\frac{1}{2\pi}\int_S F(re^{i\theta})d\theta \le \frac{1}{4\pi}\int_S \log f(\theta)d\theta$$

Since we have equality at both ends that implies a lot of things. In particular F is harmonic and and is represented via the Poisson integral interms of its boundary value $\frac{1}{2}\log f$. In particular G has no zeros in the disc. Obviuosly F is uniquely determined by $\log f$, and by the Cauchy-Riemann equations the imaginary part of $\log G$ is determined upto an additive constant. Therefore the only ambiguity in G is a multiplicative constant of modulus 1.

Given the process $\{X_n\}$ with trivial tail subspaces, we saw earlier that it has a representation

$$X_n = \sum_{j=0}^{\infty} a_j \xi_{n-j}$$

in terms of standard i.i.d Gaussians and from the construction we also know that $\xi_n \in \mathcal{H}_n$ for each n. In particular $\xi_0 \in \mathcal{H}_0$ and can be approximated by linear combinations of $\{X_j : j \leq 0\}$. Let us suppose that $h(\theta)$ represents ξ_0 in $L_2(S; f)$. We know that $h(\theta)$ is in the linear span of $\{e^{i\,j\,\theta} : j \leq 0\}$. We want to find the function h. If $\xi_0 \longleftrightarrow h$, then by the nature of the isomorphism $\xi_n \longleftrightarrow e^{i\,n\,\theta}h$ and

$$1 = \sum_{j=0}^{\infty} a_j e^{-i\,j\,\theta} h(\theta)$$

is an orthonormal expansion in $L_2(S; f)$. Also if we denote by

$$G(z) = \sum_{j=0}^{\infty} a_j z^j$$

then the boundary function $g(\theta) = \lim_{r\to 1} G(re^{i\,\theta})$ has the property

$$g(-\theta)h(\theta) = 1$$

and so

$$h(\theta) = \frac{1}{g(-\theta)}$$

Since the function G that we constructed has the property

$$|G(0)|^2 = |a_0|^2 = \sigma^2(f) = \exp\big[\frac{1}{2\pi}\int_S \log f(\theta)\, d\theta\big]$$

it is the canonical choice determined earlier, to with in a multiplicative constant of modulus 1. The predictor then is clearly represented by the function

$$\hat{\mathbf{1}}(\theta) = 1 - a_0 h(\theta) = 1 - \frac{g(0)}{g(-\theta)}$$

□

Example 6.2. A wide class of examples are given by densities $f(\theta)$ that are rational trigonometric polynomials of the form

$$f(\theta) = \frac{|\sum A_j e^{i\,j\,\theta}|^2}{|\sum B_j e^{i\,j\,\theta}|^2}$$

We can always mutiply by $e^{i\,k\,\theta}$ inside the absolute value and assume that

$$f(\theta) = \frac{|P(e^{i\,\theta})|^2}{|Q(e^{i\,\theta})|^2}$$

where $P(z)$ and $Q(z)$ are polynomials in the complex variable z. The symmetry of f under $\theta \to -\theta$ means that the coefficients in the polynomial have to be real. The integrability of f will force the polynomial Q not to have any zeros on the circle $|z| = 1$. Given any two complex numbers c and z, such that $|z| = 1$ and $c \neq 0$

$$|z - c| = |\bar{z} - \bar{c}| = |\frac{1}{z} - \bar{c}| = |1 - \bar{c}z| = |c||z - \frac{1}{\bar{c}}|$$

This means in our representation for f, first we can omit terms that involve powers of z that have only modulus 1 on S. Next, any term $(z - c)$ that contributes a nonzero root c with $|c| < 1$ can be replaced by $c(z - \frac{1}{\bar{c}})$ and thus move the root outside the disc without changing the value of f. We can therefore rewrite

$$f(\theta) = |g(\theta)|^2$$

with

$$G(z) = \frac{P(z)}{Q(z)}$$

with new polynomials P and Q that have no roots inside the unit disc and with perhaps P alone having roots on S . Clearly

$$h(\theta) = \frac{Q(e^{i\,\theta})}{P(e^{i\,\theta})}$$

If P has no roots on S, we have a nice convergent power series for $\frac{Q}{P}$ with a radius of convergence larger than 1, and we are in a very good situation. If $P = 1$, we are in an even better situation with the predictor expressed as a finite sum. If P has a root on S, then it could be a little bit of a mess as the next exercise shows.

Exercise 6.16. Assume that we have a representation of the form

$$X_n = \xi_n - \xi_{n-1}$$

in terms of standard i.i.d. Gaussians. How will you predict X_1 based on $\{X_j : j \leq 0\}$?

Exercise 6.17. An autoregressive scheme is a representation of the form

$$X_n = \sum_{j=1}^{k} a_j X_{n-j} + \sigma \xi_n$$

where ξ_n is a standard Gaussian indepenedent of $\{(X_j\,, \xi_j) : j \leq (n-1)\}$. In other words the predictor

$$\hat{X}_n = \sum_{j=1}^{k} a_j X_{n-j}$$

and the prediction error σ^2 are specified for the model. Can you always find a stationary Gaussian process $\{X_n\}$ with spectral density $f(\theta)$, that is consistent with the model?