Bridging Probability And Dynamic System Using Ergodic Theory

Yongli Peng, 1500010664

Abstract

In this report, we want to treat the random process as a dynamic system and show how this angle can offer quite different results in both fields. After clarifying the details in this change of view, we will give many examples to illustrate what it can bring from this viewpoint. The results in the former part of this report is mainly from probability book, which don't appear anywhere in the book of dynamic system, mainly because they focus on different things from what is cared about in dynamic system. The results in latter part is mainly from books of dynamic system. But each of them also has a corresponding version in the book about probability. It's quite interesting people in different area discovery the same results with different technique to prove them.

1 Treat Random Processes As Dynamic Systems

1.1 Basic Setting

We show first how to treat a stochastic process as a dynamic system, mainly by considering it as a symbolic system. First we shall give a formal definition for stochastic process in probability.

Definition 1.1. Given a **probability space** $(\Omega, \mathcal{B}, P)$, a **random variable** (abbreviated as r.v.) is a measurable map X from $(\Omega, \mathcal{F})$ to $(A, \mathcal{B}_A)$ (where $\mathcal{B}$ denotes the Borel sets on that space).

This random variable is often associated with the case when A is the real line and $\mathcal{B}_A$ is the Borel field on $\mathbb{R}$ in probability. Moreover, this measurability ensures that there exists a probability measure inheriting the probability measure induced by this map:

$$P_X(B) = P(X^{-1}(B)) = P(\{\omega : X(\omega) \in B\}); B \in \mathcal{B}_A$$

So $(A, \mathcal{B}_A, P_f)$ becomes a probability space and the induced measure P_f is called the **distribution** of f.

Definition 1.2. A **random process** is a family of random variables $\{X_t; t \in \mathbb{T}\}$ on a common probability space $(\Omega, \mathcal{B}, P)$, where $\mathbb{T}$ is an index set, often interpreted as time in probability. More generally, a random process can be viewed as a function $X : \Omega \times \mathbb{T} \to A$, measurable in Ω and usually continuous in $\mathbb{T}$.

This is an ordinary, elementary way to define a random process. But a random process can be constructed from a single random variable together with a dynamic system consisting of a probability space as well as a family of transformations on the space. First for a fixed $\omega \in \Omega$, $X(\omega)$ can be regarded as a function from $\mathbb{T}$ to A, denoted by $A^{\mathbb{T}}$. In our textbook [Sun12], we have actually defined the product σ-field $\mathcal{B}_{A^{\mathbb{T}}}$ on $A^{\mathbb{T}}$ for countable case, namely when $\mathbb{T} = \mathbb{Z}$ or $\mathbb{Z}_+$. In Gray's book [GG88], the product σ-field $\mathcal{B}_{A^{\mathbb{T}}}$ on $A^{\mathbb{T}}$ for general $\mathbb{T}$ is defined in a similar way using the smallest σ-field containing all the rectangles. Then we can check this map from $(\Omega, \mathcal{B})$ to $(A^{\mathbb{T}}, \mathcal{B}_{A^{\mathbb{T}}})$ is measurable since each component mapped to a single A is measurable. Furthermore from the above discussion it has induced a probability measure P_X on $(A^{\mathbb{T}}, \mathcal{B}_{A^{\mathbb{T}}})$. Finally when we require $\mathbb{T}$ to be a semigroup allowing an addition (not necessarily commutable) , we can give a group action of $\mathbb{T}$ on itself using this addition and this induce a transformation on $A^{\mathbb{T}}$: for each $(a_t)_{t\in\mathbb{T}}$ where $a_t \in A$ and s-induced transformation S, $S((a_t)_{t\in\mathbb{T}}) = (a_{t+s})_{t\in\mathbb{T}}$. This is the most general case. More typically, when $\mathbb{T} = \mathbb{Z}$ or $\mathbb{Z}_+$ the transformations are like the time shifts and the system is similar to a symbolic system. When $\mathbb{T} = \mathbb{R}$ the system acts like the flow.

1.2 Ergodic And Invaraint

In ergodic theory, we usually need the measure to be invariant (or preserved) w.r.t T and care if it's ergodic. If we try to interpret these two concepts in random process, we immediately find that the invariance indicates the process to be stationary, namely the distribution of each r.v. in the process is the same, since the volume of each rectangle in $A^{\mathbb{T}}$ is just the joint distribution in probability. As for ergodicity, it's equivalent to recurrent in probability, which means when you trace along the time, each point will be attained at some time point in the future. Note here I just consider the countable case for simplicity. As for general cases I think it can be proved similarly, but I have not considered its details in this observation. Thus set $\mathbb{T} = \mathbb{Z}$ and reminded of a proposition in our textbook saying that:

Proposition 1.3. *Let $T : (X, \mathcal{B}, m) \to (X, \mathcal{B}, m)$ be measure-preserving, then*

- *T is ergodic $\iff$ $\forall A \in \mathcal{B}$ with $m(A) > 0$, we have $m(\cup_{i=1}^{\infty} T^{-i} A) = 1$.*

In product space, since the σ-field is generated by the rectangles, we only need to verify the condition for the rectangles. In particular we only need to consider those rectangles with one coordinate. For a general rectangle we can first decompose it into rectangles with continuous coordinates. Then for a rectangle with k coordinates in this form, such as $[a_0, \ldots, a_{k-1}]$, we can construct another random process from the present one. Let $\tilde{X} = \{(X_t, X_{t+1}, \ldots, X_{t+k}); t \in \mathbb{Z}\}$ and the measure $P_{\tilde{X}}$ inherit the original measure P_X, namely the process can only go from $(X_t, \ldots, X_{t+k})$ to $(X_{t+1}, \ldots, X_{t+k+1})$. Otherwise the set is of zero measure. In this way, the discussion will reduce to considering the rectangle with only one coordinates. In this case, $m(\cup_{i=1}^{\infty} T^{-i} A) = 1$ is just a formal definition of recurrent. Therefore, if we consider transitive group actions (action that can attain any $t \in \mathbb{T}$ from a specific t'), we have:

Proposition 1.4. *For any transitive action τ,*

1. *the system $(A^{\mathbb{T}}, \mathcal{B}_{A^{\mathbb{T}}}, P_X, \tau)$ is invariant $\iff$ the corresponding random process X is stationary.*
2. *It's ergodic $\iff$ the corresponding process X is recurrent.*

1.3 Examples

To end this section, we give two examples illustrating the above definition, whose property will be discussed further in the following sections.

Example 1.5. (Bernoulli Shift) Let $Y = \{0, 1, \ldots, k-1\}$ denotes k symbols. $(p_0, p_1, \ldots, p_{k-1})$ denotes a probability vector, i.e. $p_i > 0$ for each i and $\sum_{i=0}^{k-1} p_i = 1$. Then we can give a measure μ on $(Y, 2^Y)$ by setting $\mu(i) = p_i$. If $(X, \mathcal{B}, m)$ denotes the product probability space of $(Y, 2^Y, \mu)$, namely

$$(X, \mathcal{B}, m) = \prod_{-\infty}^{\infty} (Y, 2^Y, \mu)$$

Define the transformation T as the left translation: $T(x_i)_{-\infty}^{\infty} = (x_{i+1})_{-\infty}^{\infty}$ and let the value of m on a rectangle be: $m[a_0, a_1, \ldots, a_n] = p_{a_0} p_{a_1} \ldots p_{a_n}$. This system $(X, \mathcal{B}, m, T)$ is called a (two-sided) Bernoulli shift. We have proved in our textbook that m is an invariant measure of T. Moreover, it's ergodic. But if we consider the corresponding random process, we can easily find the underlying process is just a series of i.i.d (independent and identical distributed) r.v, which is commonly seen in the statistics. In probability, it's well acknowledged that this process is stationary and recurrent, which indicates it's invariant and ergodic.

Example 1.6. (Markov Shift) The Markov shift can be seen as a generalization of the Bernoulli shift. Given a probability vector $\mathbf{p} = (p_0, p_1, \ldots, p_{k-1})$ and a stochastic matrix

$$\mathbf{P} = \begin{bmatrix} p_{00} & \cdots & p_{0,k-1} \\ \vdots & & \vdots \\ p_{k-1,0} & \cdots & p_{k-1,k-1} \end{bmatrix}, \quad p_{ij} \geq 0, \quad \sum_{j=0}^{k-1} p_{ij} = 1$$

which satisfy the relation: $\mathbf{p} \cdot \mathbf{P} = \mathbf{p}$. Then we can just define the measure by setting $m[a_0, a_1, \ldots, a_n] = p_{a_0} p_{a_0 a_1} \ldots p_{a_{n-1} a_n}$. Then the system $(X, \mathcal{B}, m, T)$ with other things unchanged is called the

Markov shift. In fact, it corresponds to the well-known stationary Markov chain in probability. A bunch of theoretical results in probability can be applied to this system in this manner. In particular, we have in probability that each such Markov chain is recurrent if every point is attainable (we call such a process irreducible later), which implies in dynamic system that the Markov shift is ergodic iff $\forall i, j$, $\exists n$ such that $(P^n)_{ij} > 0$. It's also proved in our textbook.

2 Results From Probability Concerning Bernoulli Shift

2.1 Birkhoff Ergodic Theory And Law of Large Numbers

In this section, we take a look at the relatively simple system: Bernoulli shift, and see how considering the underlying random process can help us get a better understanding of the system.

First let's take an investigation on the well-known BET (Birkhoff Ergodic Theory):

Theorem 2.1. *(Birkhoff Ergodic Theory) Let $T : (X, \mathcal{B}, m) \to (X, \mathcal{B}, m)$ is measure-preserving, and $f \in L^1(X, \mathcal{B}, m)$, then:*

1. *$\{\frac{1}{n}\sum_{i=0}^{n-1} f(T^i x)\}_{n\geq 0}$ converges to a function $f^*(x) \in L^1(X, \mathcal{B}, m)$ for m - a.e. $x \in X$.*
2. *$f^* \circ T(x) = f^*(x)$ for m - a.e. $x \in X$ and $\int f dm = \int f^* dm$.*

More specificly, if T is even ergodic, we have for $f \in L^1(X, \mathcal{B}, m)$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) = \int f dm$$

for m - a.e. $x \in X$.

For the Bernoulli shift, we have observed it is measure-preserving and ergodic. Therefore if we let f be the projection of $(a_t)_{-\infty}^{\infty}$ onto its value when $t = 0$, i.e. let $f((a_t)_{-\infty}^{\infty}) = a_0$. We shall have (notice the measure P_X is inherited from original P)

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} a_i = \sum_{i=0}^{k-1} i p_i$$

for P - a.e. $\omega \in \Omega$. This is just the SLLN (strong law of large numbers) for the case of finite sample space in probability (the term in the right hand side is just the expectation in probability). Therefore the BET provides another approach to prove the SLLN. Whereas the use of Borel-Cantelli lemma and Markov's Inequality may provide a novel proof of BET in the case of Bernoulli shift (but it's still a little tricky to tackle general f).

Theorem 2.2. *(Strong Law of Large Numbers) Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of i.i.d r.v. with $E[X_i] = \mu$ and $E|X_i| < \infty$ ($X_i \in L^1(\Omega, \mathcal{F}, P)$), then $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu$ for a.s. (P-a.e.) $\omega \in \Omega$.*

This theorem asserts the BET is true for a product measure space $(X, \mathcal{B}, m) = \prod_{-\infty}^{\infty}(Y, 2^Y, \mu)$ formed by general probability measure μ on Y equipped with the left translation T as the transformation.

Well, the above discussion may not be interesting for dynamic systems (but at least it's quite exciting for probability since SLLN is the most basic theorem in this field). So in the following discussions we may derive some results from probability but not existing in dynamic system to show that considering the underlying random process for these systems indeed offer some insights into the system.

2.2 Central Limit Theorem

In this section, we try to interpret another basic theorem in probability, Central Limit Theorem, in the context of Bernoulli shift. In probability theory, the central limit theorem (CLT) establishes that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed. The theorem is a key concept in probability theory

because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.

There's a set of weak-convergence theorem in probability expressing the fact that a sum of i.i.d. random variables, or alternatively, random variables with specific types of dependence, will tend to be distributed according to one of a small set of attractor distributions. When the variance of the i.i.d. variables is finite and fixed, the attractor distribution is the normal distribution. In contrast, the sum of a number of i.i.d. random variables with power law tail distributions (decreasing as $|x|^{-\alpha-1}1$ where $0 < \alpha < 2$ and therefore having infinite variance) will tend to an alpha-stable distribution with parameter α as the number of variables grows. Later we will show this is just an optimization of the entropy with various constrictions.

First we may introduce the famous Gaussian distribution in the context of dynamic system (just introduce the induced measure). I found reference in Sinai's book [CFS12] calling it Gauss measure, and it also discusses Gauss dynamic systems, which we shall mention later.

Definition 2.3. The measure τ on the product measure space is said to be a **Gauss measure** if the joint distribution of any finite number of r.v. $(x(s_1), \ldots, x(s_r))$ is an r-dimensional Gauss distribution, both in continuous and discrete cases. In measure theory's word, the Radon Nikodym derivative of the measure of a rectangle with k coordinates w.r.t the k-dimensional Lebesgue measure is $\frac{1}{(2\pi)^{k/2}(det\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-a)^T\Sigma^{-1}(x-a)\right)$, where a is the expectation vector and Σ is the covariance matrix.

From the definition we shall see even if the expression of the Gauss distribution is a little bit complicated, many researchers in statistics like to use it because it's totally determined if the mean value as well as the variance is fixed. This property is also popping up in the following discussion of CLT.

In fact, the philosophy under CLT (central limit theorem) is rather simple: we just want to estimate the difference between the measures appearing in SLLN. Let $\mu_{n,x} = \frac{1}{n}\sum_{i=0}^{n-1} \delta_{T^i x}$, from the BET we know if μ is an ergodic measure, then $\mu_{n,x} \xrightarrow{*} \mu$ in the sense of weak-$*$ topology for m-a.e. x. But we are now curious about how far is $\mu_{n,x}$ from μ. Like in the process of proving ergodic decomposition theorem, here we shall not consider a single x but treat it as a r.v. Then we shall have $\int f d(\mu_{n,x} - \mu) = \frac{1}{n}\sum_{i=0}^{n-1} f(T^i x) - \int f d\mu$ is not a number, but a r.v and can induce a measure. The CLT states that $\int f d(\sqrt{n}(\mu_{n,x} - \mu)) \xrightarrow{*} \tau$ where τ is just the Gauss measure mentioned above (it's of zero mean, but its variance will be determined by f):

Theorem 2.4. *(Central Limit Theorem)* $\int f d(\sqrt{n}(\mu_{n,x} - \mu)) \xrightarrow{*} \tau$ *as n tend to infinity, where* τ *denotes the Gauss measure with zero mean and variance determined by f, here* $f \in L^2(X, \mathcal{B}, m)$.

Therefore, when considering dynamic systems of a product measure space, turn to its underlying random process may sometimes be helpful to have a further insight into the system. In the case of Bernoulli shift we can actually estimates how far the time averaging is from the integral or the expectation. Although this treatment is not in the general sense and can be applied to other systems, it can provide some subtle results towards the system concerning product space, which cannot be obtained from the general methods in dynamic system. We shall give further examples illustrating this in the later discussion of Markov shift.

2.3 Large Deviations Theory

We show in this subsection that a famous result in probability can induce more subtle estimates towards the system of Bernoulli shift. It emerged in the 1930s with the Swedish mathematician Harald Cramér's study of a sequence of i.i.d. random variables $(Z_i)_{i\in\mathbb{N}}$. Namely, Cramér studied the behavior of the distribution of the average $X_n = \frac{1}{n}\sum_{i=1}^{n}$ as $n \to \infty$. He found that the tails of the distribution of X_n decay exponentially as $e^{-n\lambda(x)}$ where the factor $\lambda(x)$ in the exponent is the Legendre–Fenchel transform of the cumulant-generating function $\Psi_Z(t) = \log Ee^{tZ}$. A very incomplete list of mathematicians who have made important advances afterwards on this field would include Petrov,Sanov, S.R.S. Varadhan (who has won the Abel prize for his contribution to the theory), D. Ruelle (who has worked on dynamic systems), O.E. Lanford, Amir Dembo, and Ofer Zeitouni.

We know by SLLN that if $\bar{X} = \frac{X_1+\ldots+X_n}{n}$, then $\bar{X} \to EX$ for m-a.e. $\omega \in \Omega$, but in probability we can have a more precise estimation of $P(|\bar{X} - EX| \geq t)$. This is just the philosophy of large

deviation theory as well as the various similar style inequalities in probability. In other words, it implies how fast this deviation will tend to the Gaussian measure.

Theorem 2.5. *(Hoeffding's Inequality) Let $X_1, \ldots, X_n$ be i.i.d random variables bounded by the interval $[a, b]$ (in Bernoulli shift we just have $a = 0$ and $b = k - 1$) and $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$, then*

$$P(|\bar{X} - EX| \geq t) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

This is frequently used in statistics, but in probability people care more about the general case for general distributions. Then we shall get

Theorem 2.6. *(Cramer's Theorem) The logarithmic moment generating function of a random variable is defined as: $\Lambda(t) = \log E[\exp(tX_1)]$. Let $X_1, X_2, \ldots$ be a series of i.i.d. real random variables with finite logarithmic moment generating function, $\Lambda(t) < \infty$ for all $t \in \mathbb{R}$.*

Then the Legendre transform of Λ : $\Lambda^*(x) := \sup_{t \in \mathbb{R}}(tx - \Lambda(t))$ satisfies,

$$\lim_{n \to \infty} \frac{1}{n} \log\left(P\left(\sum_{i=1}^{n} X_i \geq nx\right)\right) = -\Lambda^*(x)$$

Note here in this section we mainly interprets those subtle results in probability into dynamic systems to have some insights in this product dynamic system. But these examples also show the things we care, the techniques we use in probability may be a little far from what researchers do in dynamic systems. The most challenging thing in combining these two fields might be: in probability we usually care the behavior of the process when it's not stationary (which means we don't expect we have an invariant measure whereas in ergodic theory we make it a prequisite) while in ergodic theory we usually care if a transformation is measure-preserving, under what conditions is it ergodic or periodic, how to calculate its entropy, etc. But in probability, people mainly focused on process with Markov property, which in most situations will be ergodic if equipped with an invariant measure. Moreover, we have calculated its entropy in the class, even the topological entropy of Bernoulli shift is well calculated. But for general cases, the entropy of a system is quite hard to obtain. Therefore people in probability always turn to consider the change of entropy, namely entropy production rate, which is inspired mainly by statistical physics.

In conclusion, although we can link these two field in some sense, there's much difference between these two areas ranging from its goal and the technique. But there're still things similar in both area, like there're theories developed in both field seperatedly, using different techniques, but saying the same thing. So we will focus on the dynamic system's viewpoint in the following section.

3 Results From Probability Concerning Markov Shift

3.1 Mixing

We will introduce the mixing property in dynamic system in this part. Then we shall show its correspondence in the context of Markov process. Interestingly, this property is proved in both fields but with quite different techniques. Eventually, we will prove it from dynamic system's viewpoint.

Definition 3.1. We say that the system $(X, \mathcal{B}, m, T)$ is **mixing** if

$$\lim_{n \to \infty} m(T^{-n}(A) \cap B) = m(A)m(B)$$

for any $A, B \in \mathcal{B}$ measurable.

Technically, we just need to verify the above identity w.r.t the sets in an algebra $\mathcal{A}$ that generates the σ-field $\mathcal{B}$. For the case of Bernoulli shift and Markov shift, we just need to verify the identity subject to the rectangles.

Example 3.2. Every Bernoulli shift is mixing. Given any two rectangles $A = [A_p, \ldots, A_q]$ and $B = [B_r, \ldots, B_s]$. We have:

$$\begin{aligned} m(T^{-n}(A) \cap B) &= m([B_r, \ldots, B_s, X, \ldots, X, A_p, \ldots, A_q])) \\ &= m([B_r, \ldots, B_s])m([A_p, \ldots, A_q]) = m(A)m(B) \end{aligned}$$

for every $n > s - p$. Hence it remains to be true when n tend to infinity. So Bernoulli shift is mixing.

Then we list the corresponding theory in Markov process as below:

Theorem 3.3. *If the Markov process is irreducible and aperiodic, π is the invariant measure, then*

$$\lim_{n\to\infty} \sum_{j\in A} |p_{ij}^{(n)} - \pi_j| = 0$$

for any $i \in A$. More specifically, $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$

Here **irreducible** is clarified at the beginning: $\forall i, j \in A$, there exists $n \geq 1$ such that $p_{ij}(n) > 0$, where $p_{ij}(n)$ is the (ij)-element of the matrix $\mathbf{P}^n$. We say a Markov process (or the stochastic matrix $\mathbf{P}$) is **aperiodic** if there exists $n \geq 1$ such that $p_{ij}(n) > 0$ for every $i, j \in A$.

First we shall show how this property is equivalent to the mixing property. In the definition of mixing, if we just set $A = \{(x)_{-\infty}^{\infty} : x_0 = i\}$ and $B = \{(x)_{-\infty}^{\infty} : x_0 = j\}$, then the left hand side is just $m(\{(x)_{-\infty}^{\infty} : x_0 = j, x_n = i\}) = \pi_j p_{ji}(n)$ while the right hand side is $\pi_j \pi_i$. Therefore it can be derived that $p_{ji}(n) \to \pi_i$. This is just what we said for the Markov process above.

More formally, we have the following theorem from the dynamic systems' side:

Theorem 3.4. *Let $T : M \to M$ be a mixing transformation relative to some invariant probability μ. Let υ be any probability measure on M, absolutely continuous w.r.t μ. Then $\tilde{T}_*^n \upsilon$ weak-* converges to μ, where $\tilde{T}_*^n \upsilon$ is inherited from υ induced by f^n.*

Proof. Let $\phi = 1_B$ and $\Psi = \frac{d\upsilon}{d\mu}$. Note that $\phi \in L^\infty(\mu)$ and $\Psi \in L^1(\mu)$. Hence,

$$\int (1_B \circ T^n \frac{d\upsilon}{d\mu} d\mu) \to \int \phi d\mu \int \Psi d\mu = \int 1_B d\mu \int \frac{d\upsilon}{d\mu} d\mu$$

where the convergence is the result of the mixing property. The sequence on the left-hand side coincides with $\int (1_B \circ T^n d\upsilon = \upsilon(T^{-n}(B))$ while the right-hand side is just $\mu(B) \int 1 d\upsilon = \mu(B)$. So $\tilde{T}_*^n \upsilon \xrightarrow{*} \mu$. □

Remark 1. In this theorem we need the initial measure υ to be absolutely continuous w.r.t μ, which is also required in Markov process where we need the process to be irreducible. It's natural to have this irreducible behavior since you can never imagine having a path with positive measure and its reversed path of zero measure (in dynamic system a path is just a rectangle with continuous coordinates). So starting from any measure given the mixing property, you will always converge to the invariant measure.

Theorem 3.5. *If the stochastic matrix $\mathbf{P}$ is irreducible, then the following are equivalent:*

1. *The Markov shift $(X, \mathcal{B}, m, T)$ is mixing*
2. *The stochastic matrix $\mathbf{P}$ is aperiodic*
3. $\lim_{n\to\infty} p_{ij}(n) = \pi_j$

Proof. (2) $\Leftarrow$ (3): Since we assume the matrix is irreducible, then $\pi_j > 0$ for every j, then $\lim_{n\to\infty} p_{ij}(n) = \pi_j$ implies $p_{ij}(n) > 0$ for every i,j and every n sufficiently large.

(2) $\Rightarrow$ (3): Now suppose $\mathbf{P}$ is aperiodic. Then we may apply the theorem of Perron-Forbenius to the matrix $\mathbf{P}$. since π is an eigenvalue of $\mathbf{P}$ with positive coefficients, we get the maximal eigenvalue to be 1 and all the other eigenvalues are smaller than 1 in the absolute value. It's easy to see the hyperplane $H = \{(h_1, \ldots, h_d) : h_1 + \ldots + h_d = 0\}$ is invariant is invariant under $\mathbf{P}$ if treating it as a linear map. Then the decomposition $\mathbf{R}^d = \mathbf{R}\pi \oplus H$ is invariant under $\mathbf{P}$ and the spectral radius of H is less than 1. Therefore $\mathbf{P}^n$ will converges to the projection of the first coordinates in the decomposition. So $\lim_{n\to\infty} p_{ij}(n) = \pi_j$.

(1) $\Rightarrow$ (3): Suppose the measure m is mixing. From the informal argument, let $A = \{(x)_{-\infty}^{\infty} : x_0 = i\}$ and $B = \{(x)_{-\infty}^{\infty} : x_0 = j\}$, then we can derive (3).

(2) $\Rightarrow$ (1): Now suppose that the matrix $\mathbf{P}$ is aperiodic. We want to conclude μ is mixing. Then we just need to prove the identity in the definition w.r.t the rectangles. Set $A = [a_m, \ldots, a_q]$ and $B = [b_r, \ldots, b_s]$ (the subscript denotes its coordinates). Then we can calculate:

$$m(A \cap T^{-l}(B)) = m(A)m(B)\frac{1}{\pi_{b_r}} p_{a_q, b_r}(r - q + l)$$

for every $l > q - r$. Then using (3) and let $l \to \infty$ we can verify the identify in the definition of mixing in this case. This completes the proof. □

3.2 Kac Theorem

In this part we proof a certain generalization of the Poincare recurrence theorem, the Kac theorem, and in turn present its version in probability, showing there's some similarity in the development of probability and dynamic systems.

Let $T : X \to X$ be a measurable transformation and m be a finite measure invariant under f. Let $E \subset X$ be any measurable set with $m(E) > 0$. Consider the *first-return* time function $\rho_E : E \to \mathbb{N} \cup \{\infty\}$, defined by:

$$\rho_E(x) = \min\{n \geq 1 : T^n(x) \in E\}$$

if the set on the right-hand side is non-empty and $\rho_E(x) = 0$ if, on the contrary, x has no iteration in E. According to the Poincare recurrence theorem, the second case occurs only on a set with zero measure.

The Kac theorem shows this function is integrable and even provides the value of integral. For the statement we need the following notation:

$$E_0 = \{x \in E : T^n(x) \notin E,\ \forall n \geq 1\}$$

$$E_0^* = \{x \in X : T^n(x) \notin E,\ \forall n \geq 0\}$$

In other words, E_0 is the set of points in E that never return to E. E_0^* is the set of points in M that never enter E. We have show $m(E_0) = 0$.

Theorem 3.6. *(Kac) With the above setting, the function $\rho_E(x)$ is integrable and*

$$\int_E \rho_E d\mu = m(X) - m(E_0^*)$$

Proof. For each $n \geq 1$, define

$$E_n = \{x \in E : T(x) \notin E, \ldots, T^{n-1}(x) \notin E, T^n(x) \in E\}$$

$$E_n^* = \{x \in X : x \notin E, T(x) \notin E, \ldots, T^{n-1}(x) \notin E, T^n(x) \in E\}$$

that is, E_n is the set of points return to E for the first time exactly at time n, $E_n = \{x \in E : \rho_E(x) = n\}$, and E_n^* is the set points that are not in E and enter E for the first time exactly at time n. It's clear these sets are measurable and hence $\rho_E(x)$ is a measurable function. Moreover, these sets for $n \geq 0$ constitute a partition: they are pairwise disjoint and their union is the whole of X. So

$$m(X) = \sum_{n=0}^{\infty}(m(E_n) + m(E_N^*)) = m(E_0^*) + \sum_{n=1}^{\infty}(m(E_n) + m(E_n^*))$$

Now observe that

$$T^{-1}(E_n^*) = E_{n+1}^* \cup E_{n+1}$$

So given m is invariant, $m(E_n^*) = m(E_{n+1}^*) + m(E_{n+1})$. Applying the relation successively, we find that $m(E_n^*) = m(E_k^*) + \sum_{i=n+1}^{k} m(E_i)$ for every $k > n$. The convergence of the summation series in the first formula implies that $m(E_k^*) \to 0$ when $k \to \infty$. So after taking the limit, we shall have

$$m(E_n^*) = \sum_{i=n+1}^{\infty} m(E_i)$$

Replace this identity in the first formula, we find that

$$m(X) - m(E_0^*) = \sum_{n=1}^{\infty}(\sum_{i=n}^{\infty} m(E_i)) = \sum_{n=1}^{\infty} nm(E_n) = \int \rho_E d\mu$$

as we want to prove. □

In some cases, for example when the system is ergodic, $m(E_0^*)$ has zero measure. Then the conclusion of the Kac theorem means that

$$\frac{1}{m(E)}\int_E \rho_E d\mu = \frac{m(X)}{m(E)}$$

for every measurable set E with positive measure. The left-hand side is the *mean return time* to E. This is just what we assert in probability in the context of Markov process.

Theorem 3.7. *If the Markov process is irreducible, the followings are equivalent:*

1. *There's a state $i \in A$ such that $\int_{\{i\}} \rho_{\{i\}} dP_i < \infty$, where P_i is the induced Markov measure starting from a Dirac measure supported at $\{i\}$ with matrix $\mathbf{P}$.*
2. *Every states $i \in A$ satisfies (1).*
3. *There exist an invariant measure π.*

Moreover, when the above conditions is satisfied, we also have:

$$\pi_i = \frac{1}{E_i \rho_i}, \quad \forall i \in A$$

where $E_i\rho_i = \int_{\{i\}} \rho_{\{i\}} dP_i$.

Remark 2. Since $E_i\rho_i = \int_{\{i\}} \rho_{\{i\}} dP_i$ is just $\frac{1}{m(E)}\int_E \rho_E d\mu$ in the Kac theorem. This theory in probability possess the same result as the Kac theorem. Therefore if we just constrict ourselves to the path starting at a single point or a specific set E, these two theorems are equivalent, regardless of whether the measure is invariant (since the later behavior is totally determined by the matrix $\mathbf{P}$).

Remark 3. In probability, ρ_E defined here is quite important. It can introduce one of the most important concepts in advanced probability, martingale. So this again displays how much the topics being discussed in these two areas different from the other since I only found Kac theorem in [VO16] for a generalization of Poincare recurrence theorem but with no further discussion.

3.3 Entropy

In the last part, we discuss a little about the entropy. As a matter of fact, in probability we have the concept of entropy but we don't discuss its property further. However, recently inspired by the statistical physics, there are a lot of discussions about the concept entropy production rate, which is focused on the change of entropy in the time evolution. As for the entropy we discussed in the class, though we can lift these process listed above to dynamic systems and calculate their entropy, all the process is covered in class, which means there's little left in this topic if we only discuss the regular Markov process or i.i.d. process as usual in probability. So we may try to find process with properties other than Markovian to consider this topic.

Proposition 3.8. *A Bernoulli shift with the probability vector $(p_0, \dots, p_{k-1})$ possess the entropy being $h_m(T) = -\sum_{i=0}^{k-1} p_i \ln p_i$*

Proposition 3.9. *A Markov shift with the probability vector as $\mathbf{p} = (p_0, \dots, p_{k-1})$ and the stochastic matrix as $\mathbf{P} = (p_{ij})_{k\times k}$ has the entropy being $h_m(T) = -\sum_{i,j} p_i p_{ij} \ln p_{ij}$*

Moreover, using the variational principle, we can just calculate the maximal of these entropy to have a guessing about its topological entropy. Because we have $\sum_{j=0}^{k-1} p_{ij} = 1$ for every $0 \le i \le k-1$ and $f(x) = x\ln x$ is convex, it's easy to derive $\sum_{j=0}^{k-1} p_{ij}\ln p_{ij} \ge k \cdot \frac{1}{k}\ln\frac{1}{k} = -\ln k$. Therefore for Markov shift we shall get $h_m(T) \le \sum_{i=0}^{k-1} \ln k = \ln k$ and so does the Bernoulli shift. Moreover, we have proved in the class the topological entropy for Bernoulli shift is just $\ln k$. Since Bernoulli shift is a special form of Markov shift, the topological entropy of Markov shift is also this value and the equality in variational principle can be attained when the measure is distributed uniformly on $\{0, \dots, k-1\}$.

Therefore, for finite and discrete case, the discussion of entropy is quite trival. As for infinite cases, we can just let $k \to \infty$, then both the entropy and the topological entropy will tend to infinity and there's nothing can be done.

But there's something interesting in the continuous cases. If we just define the entropy of a measure on continuous case to be $-\int p(x) \ln p(x) dx$, then after adding some constrictions we can get something interesting. First if we have no restrictions, this entropy can also tend to infinity like the discrete case. But if we fix its mean $\int xp(x)dx = 0$ and variance $\int x^2 p(x)dx = \sigma^2$, using calculus of variation we can derive it attains its maximal when it's a Gauss distribution. Moreover, if we require $x > 0$ and fix the mean $\int xp(x)dx = \lambda$, we shall get another famous distribution, exponential distribution, in probability which maximize its entropy. Therefore, considering the entropy in continuous case can provide an interpretation why these two distributions appear frequently in probability and why we mainly discuss Gauss process and Poisson process in continuous case.

4 Gauss Dynamic Systems

In the last section, we provide with some results concerning with Gauss Dynamic Systems, which is just the system on a similar product space whose underlying process is a Gauss process. That is we specify the measure on that product space is a Gauss measure mentioned at the beginning. Moreover, we set $a(s_i) = \int X(s_i) dP_X$ and $b(s_i, s_j) = \int X(s_i)X(s_j) dP_X$ being its mean function and covariance function, then the system is totally determined by these two functions.

What's more, the Gauss measure is invariant w.r.t T (which is left translation in the discrete case) if

$$m(s) = m \equiv const, \qquad b(s_1, s_2) = b(s_1 + t, s_2 + t)$$

for any integer t, then we can define

$$b(s_1, s_2) = b(0, s_2 - s_1) \stackrel{\triangle}{=} b(s_2 - s_1)$$

. Further we assume $m = 0$ after a translation $\tilde{X}(s) = X(s) - m$. Then the function $b(s)$ is said to be the correlation function of the Gauss measure. Moreover it's positive definite, so by the Bohner-Khinchin theorem it may be presented in the form:

$$b(s) = \int_{-\pi}^{\pi} e^{i\lambda s} d\sigma(\lambda)$$

where σ is a finite measure on the circle S^1. The measure σ is known as the spectral measure of the Gauss measure μ since it's very similar to the Fourier transform w.r.t b in the measure sense.

Then we have the following proposition:

Proposition 4.1. *The Gauss dynamic system is ergodic if and only the spectral measure σ is absolutely continuous with respect to the Lebesgue measure on S^1.*

Proposition 4.2. *If the correlation function satisfies $b(s) \to 0$ for $s \to \infty$, then the Gauss system is mixing.*

We shall not list the proof here since they are tedious and using things concerning with the spectral in functional analysis. Interested readers may refer to Sinai's book [CFS12] at PP.188 and 356, where the author discussed a lot concentrating on the spectral analysis of Gauss dynamic systems.

5 Conclusion

In conclusion, when treating the random process as a dynamic system on the product measure space, you can find the correspondence is quite beautiful and you can interprete the results from one field into the theory in another. This viewpoint is novel and beautiful. But you'll find there's not a very solid theorem from this viewpoint (at least I haven't found one), which means I haven't found a theorem which can only be proved in one field and is hard to prove in the other field. The theorems coincides in these two fields are always quite simple. As for more difficult and advanced theorems, they always seem important in one field but is not cared about in the other (or even can be trival in the other).

This is caused mainly by the difference of focused point in these two fields. On the one side, in probability people always care how the process behaves when it's not invariant, and how it behaves when time goes too infinity. How the process becomes or approaches the invariant measure attracts people in this field. But on the other side, people study dynamic systems (mainly people study ergodic theory) always suppose we are on a space with invariant measure. We are already on the spot, but the underlying space can be too complex to do anything on it. The philosophy is quite different. Probability mainly focus on the time evolution of the measure in the long run while people study dynamic systems mainly focus on the spatial complexity (not only the measure) and how this complexity evolves along time.

But the above divergence in the philosophy is made up in the recent study. At least I know more and more people try to study process with complex space structure. It has been quite hot in probability to consider the random walk on various complex geometric space, like manifold. I have also read recently a paper concentrating on the topological recurrence in probability, namely the recurrence property concerning with the topological structure on its neighborhood. All I write above is just the results from elementary probability and elementary results in ergodic theory (since I have just taken the course and begun to learn ergodic theory). All the above perspective may not be comprehensive, even results may not be right because I'm at the beginning to combining these two areas: ergodic theory and probability. But I believe there are a lot to do in this combination of two total different fields. There might be many differences at the beginning, but I'm convinced these differences can as well boost new theorems to pop up and blossom.

For the probability part, [Str13, KS12] are what we mainly referred to and cited.

References

[CFS12] Isaac P Cornfeld, Sergej V Fomin, and Yakov Grigor'evĭc Sinai. *Ergodic theory*, volume 245. Springer Science & Business Media, 2012.

[GG88] Robert M Gray and RM Gray. *Probability, random processes, and ergodic properties*. Springer, 1988.

[KS12] Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 2012.

[Str13] Daniel W Stroock. *An introduction to Markov processes*, volume 230. Springer Science & Business Media, 2013.

[Sun12] W Sun. *Ergodic Theory*. Peking University Press, 2012.

[VO16] Marcelo Viana and Krerley Oliveira. *Foundations of ergodic theory*, volume 151. Cambridge University Press, 2016.