Ergodic Theory for Interested Computer Scientists

John Mount[*]

February 4, 2012

Abstract

We describe ergodic theory in modern notation accessible to interested computer scientists.

The ergodic theorem (http://en.wikipedia.org/wiki/Ergodic_theory (link)) is an important principle of recurrence and averaging in dynamical systems. However, there are some inconsistent uses of the term, much of the machinery is intended to work with deterministic dynamical systems (not probabilistic systems, as is often implied) and often the conclusion of the theory is mis-described as its premises.

By "interested computer scientists" we mean people who know math and work with probabilistic systems[1], but know not to accept mathematical definitions without some justification (actually a good attitude for mathematicians also).

There is precedent for the claim that the connection between ergodic theory and randomized algorithms has been overstated. For instance in Jacob T. Schwartz's 1962 essay "The Pernicious Influence of Mathematics on Science"[KRS92] we have:

> ... the delicious ingenuity of the Birkhoff ergodic theorem has created the general impression it must play a central role in the foundations of statistical mechanics ... this dictum is promulgated, with a characteristically straight face, in Dunford-Schwartz, *Linear Operators*, Vol I, Chap. 7. .

And there is documented discomfort with the claim that the use of the word ergodic in different fields actually means the same thing (from Rota's "A Mathematician's Gossip"[Rot97]):

> Ergodic theory is torn between two extremes: on the one hand, it pretends to be a generalization of the mechanics of flow (well measurable flows instead of continuous flows, but let it pass); on the other hand, it is the study of shift operators originating from stationary stochastic processes and their generalizations.

The first uses and etymology[2] of the neologism "ergodic" is itself in controversy.[Mat88] And the term "ergodic" was used to describe dynamical systems before the much of the axiomitization of probability was available to reliably translate the concept into probabilistic terms (will will return to this in Section 3.4).

We will state the traditional definitions, show which ideas carry over from dynamical systems to a randomized algorithm setting and show which ideas do not. The summary being: the conclusions of the ergodic theorem transfer over, but the heavy machinery needed when working with deterministic dynamical systems does not (and is not needed). We also try to call out the common

---

[*]email: mailto:jmount@win-vector.com article hosted at: http://www.win-vector.com/blog/2012/02/ergodic-theory-for-interested-computer-scientists/ (link).

[1]Called randomized algorithms in computer science and related to statistical mechanics from physics.

[2]Roughly: the term "ergodic" is derived from the Greek for "work" plus either "path" or "form".

roots of some popular randomized algorithms (allowing experience and intuition to be translated from one family of randomized algorithms to another).

Or: we will use several pages to provide examples for a few appropriate paragraphs of Breiman's "Probability"[Bre92]

Contents

1   Introduction

Ergodicity related to the property of recurrence or representativeness of dynamical systems. The consequence of ergodicity are often confused with and stated as the defining properties. Some of the confusion is because the results of the ergodic theorem are very closely linked to the defining properties (despite having been a major accomplishment to prove the ergodic theorem).

There are at least two definitions of ergodic. The deep concept *ergodic transformation* (what is used when analyzing dynamical systems) and the simpler concept of *ergodic sets* (what is actually used in randomized algorithms and statistical mechanics). These definitions are related, but not interchangeable. Stating which definition you are working with eliminates a lot of confusion.

We will not give a potted history of the principles of ergodicity, but we will give a potted derivation of some of the theory.

1.1   The Dynamical Definition

In modern usage[Bre92, BKS91] a measure preserving transformation $T()$ is an *ergodic transformation* if it has one of two additional equivalent properties:

1. If $f()$ is function such that $f(T(x)) = f(x)$ with probability 1 then $f()$ is a constant function (except on on an exceptional set of measure 0).

2. If $T^{-1}(E) = E$ then the measure of the set E is either 1 (the typical case) or 0 (the exceptional cases).

We put off formally defining all of our terms in Section 2.

The ergodic theorem is: for a transform $T()$ that is measure preserving with respect to $\mu()$ and ergodic we have:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k(x)) = \int_{x \in X} f(x) \mathrm{d}\mu(x)$$

This is the classic "time averages converge to space averages" often called the "bug on a hotplate" statement. That is a particle (or bug) being moved around by repeated applications of $T()$ sees on average the same value of $f()$ (our "temperature") as the average of $f()$ over the entire domain (average with respect to $\mu()$). This is very useful as there are many cases where we have no idea how to efficiently estimate the integral on the right, but can estimate the sum on the left.

In modern treatments of probability [Bre92] functions like $f()$ have interpretation as random variables or experiments (they map unseen state that may be truly random to observables in a deterministic fashion). Under this treatment the conclusion of the ergodic theorem is (almost surely in the choice of x):

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k(x)) = \mathrm{E}_{\mu()}[f(x)|x \text{ is in a recurrent state}].$$

However, computer scientist do not actually tend to use the actual ergodic theorem. They apply the conclusion of the ergodic theorem (time averages approach space averages) but they do not appeal to the ergodic theorem to achieve these properties (they instead assume or establish them through simpler means). Of the major concepts of the ergodic theorem (measure preserving, invariant sets, recurring sets, ergodicity and time averages approaching space averages) only two concepts (recurring sets and time averages approaching space averages) are commonly used in the construction of randomized algorithms. The heavy lifting that the ergodic notation and theorem perform is only need to show a conservative iterated *deterministic* system has the properties we would more easily design into a nice randomized system (though the analogy does give us a hint what to design for).

If you are willing to appeal directly to random sources you can get to the important measure conclusions quickly (without a lot of machinery). If you desire to explain how a deterministic system (like classical Newtonian physics) could *even appear* random (as Boltzmann and Maxwell were attempting) they you need the full ergodic theorem (both premises and conclusions).

1.2 The Probabilistic Definition

The related concept *ergodic set E* for $T()$ is defined as set of non-transient states or a state that is visited infinitely often[3] by the iterated process $T()$ (deterministic or random)[KS76].[4]

This definition is very close to (but not quite the same as) the notion of invariant sets found in Breiman's set up of the ergodic theorem (Definition 6.2, invariant sets are measurable sets A such that $T^{-1}(A) = A$). We will now work an example to show some of the flavor and demonstrate that knowing $T(E) = E$ is not enough to show the set E is invariant.

To see the difference in definitions: take the following dynamical system on the non-negative integers ($\mathbb{Z}^{\geq 0}$) defined by the operator $T(x) = \max(0, x - 1)$ which is surjective on $\mathbb{Z}^{\geq 0}$ and measure preserving under the measure $\mu(E)$ that returns 1 if $0 \in E$ and 0 otherwise.[5] Under $T()$ the only ergodic set is $\{0\}$ and the only invariant set is $\mathbb{Z}^{\geq 0}$ itself.

---

[3]For infinite state spaces replace visited with visited arbitrarily near.

[4]This is stronger than merely asking $T()$ to be surjective on E (i.e. $T(E) = E$).

[5]Measure preserving is technical and defined later in our write-up. We are using the same definition of measure preserving as Breiman in this example.

It is the concept of ergodic sets that turns out to be immediately useful in the design of randomized algorithms.

2 Notation

All of the concepts we are discussing (and trying to distinguish between) are both technical and subtle. So will use the notation of Kolmogorov axiomitized probability theory (which itself is based on measure theory). These axioms seem very formal and introduce new concepts like σ-algebras. But all theory is trying to do is:

1. Perform detailed accounting of what you consider as your source of randomness.

   Roughly all randomness is accounted for as some unobservable that is thought of as being written down ahead of time. All observable random variables are then normal deterministic functions of this hidden state.[6] This allows us to leave randomness as axiomatic (or undefined, avoiding philosophical problems) and makes almost all operations (like observing a random variable or performing an experiment) ordinary mathematics (mere composition of deterministic functions).

   For example: suppose our random state is a sequence of real numbers that appears as if each number distributed as an independent Gaussian with mean 0 and standard deviation 1. We could then design a sequence of random variables implementing "fair coin flips" by saying the k-th coin flip is "heads" if and only if the k entry of our random state is ≥ 0.

   This separation of concerns make things much easier. We can use axiom of choice and compactness style arguments to say there are infinite sequences of real numbers where each number appears to be distributed independently as a Gaussian with mean zero and standard deviation 1. And we can then ignore all such technicalities in arguing that checking if each of these variables is at least zero behaves like we would hope an independent sequence of fair coins would.

   We will use this notion in Section 4.2 when we try to write down exactly what is meant by a randomized transform (until then we will intentionally leave the concept as an unspecified intuition).

2. Limit what sets you will even attempt to compute measure for.

   By limiting what sets you will even be asked to compute measure for (or equivalently: limiting what experiments you are expected to able to perform) you can get by with a weaker theory of measures. Your theory of measure does not have to worry about sets you know you will not be asked about. This demarcation of what sets you will be asked about is what allows the necessarily incomplete theory of integration on $\mathbb{R}^n$ (you can easily construct sets and functions such that you can not evaluate the integral of the chosen function over your chosen set) be used to implement a complete theory of measure (there are no finite compositions of measurable sets that are not themselves measurable).

   A usable theory of measure has to be complete in the following sense: composing two measurable observables into a new observable must result in a *measurable observable*. You can't claim that you can observe individual coin flips, but that observing if two coins are both heads is impossible (this would be an unacceptably incomplete theory). In fact we also want some non-finite operations (like certain sums and limits) to always work. σ-algebras are just the name for the accounting system that tracks which sets you will be asked to compute measure for.

   We will use this notion in Section 2.1 where we define measure and measurablity.

---

[6]Actually we restrict random variables to be measurable functions, or functions that have bounded norm.

These are the two things to look for when reading axiomitized probability. σ-algebras seem complicated, but they are actually designed to make everything easier.

2.1 Measure

Let X be the set of points we are going to work with (often X is $\mathbb{R}^n$, $\mathbb{Z}^n$, $\mathbb{Z}^{\geq 0}$ or $\mathbb{Z}_k = \{0, \cdots, k-1\}$). To apply measure theory we must explicitly specify Σ: our σ-algebra or set of subsets we of X we consider measurable. We will insist that Σ have the following standard properties:

- $\emptyset \in \Sigma$.
- $X \in \Sigma$.
- $E \in \Sigma$ implies $(X \setminus E) \in \Sigma$.
- $E_i \in \Sigma$ for $i \in \mathbb{Z}^{\geq 0}$ implies $\bigcup_{i=1}^{\infty} E_i \in \Sigma$.

All sets E we are going to work with will be measurable ($E \in \Sigma$).

A measure $\mu()$ is a function from Σ to $\mathbb{R}^{\geq 0}$ that is countably additive (that is the measure of a countable union of disjoint sets is the sum of the measures) and $\mu(X) = 1$ (we will insist $X \in \Sigma$ and we are only working with normalized measures). Think of $\mu()$ as summation, integration or as computing probabilities. For $x \in X$ we write $\mu(x) = \mu(\{x\})$ (assuming $\{x\} \in \Sigma$). When X is $\mathbb{R}^n$ we typically have $\mu(x) = 0$ for all $x \in X$ (sets of consisting of a single point have zero measure) and we thing of $\mu()$ as integration. When X is finite (like $\mathbb{Z}_k$) we think of $\mu()$ as summation and typically have $\mu(E) = \sum_{x \in E} \mu(x)$ (the only way we could not equality is if for some x we have $\{x\} \notin \Sigma$ meaning we don't know the measure of such a point).

These definitions are meant to capture the intuition of probabilities (they are non-negative and countable additive) and provide solid foundations (to avoid contradictions brought in by attempting to work with non-measurable sets).

2.2 Measure Preserving Transformations

The standard definitions of the measure theory of dynamical systems look a bit artificial, so we will explain a bit as we go.

We will work with function or transform $T() : X \to X$. For $S \subseteq X$ we will write $T(S) = \{T(x) | x \in S\}$ and $T^{-1}(S) = \{x | T(x) \in S\}$. $T() : X \to X$ is called measurable with respect to $\mu()$ if for all $E \in \Sigma$: $T^{-1}(E) \in \Sigma$. This definition is designed so that continuous functions from $\mathbb{R}^n \to \mathbb{R}$ are measurable under the usual Borel measure.[7] Just as the inverse of a continuous function may not be itself continuous we may have that a function that is an inverse of a measurable $T()$ may not be measurable.[8] Because we are insisting $X \in \Sigma$ we have for measurable $T()$: $T^{-1}(X)$ exists so we have $T()$ must in fact map X onto all of X (so measurable $T()$ are onto or surjective). Roughly think of measurable as meaning: nice, bounded norm and continuous.

A measurable $T()$ is called a measure preserving transformation with respect to $\mu()$ if for all $E \in \Sigma$:

$$\mu(T^{-1}(E)) = \mu(E). \tag{1}$$

This definition is non-intuitive but very clever.[9] We need to explain why you would consider this as a definition (instead of something like $\mu(T(E)) = \mu(E)$).

---

[7]Intuitively continuous functions are those that a small change in input produces a small change in output. In topological terms this is provably equivalent to the E being an open set implies $T^{-1}(E)$ is an open set. And this is why the definition of measurable is stated in terms of $T^{-1}()$ instead of in terms of $T()$.

[8]Take for example $T(x) = x^2$ over the complex numbers. This is measurable under the standard Borel measure. And $T^{-1}(x) = \{\pm\sqrt{x}\}$ is not.

[9]Somethings mathematical definitions are clever both in the good and bad senses of the word.

This definition does not have $T()$ preserving measure of sets upon forward application. Take for instance the "horseshoe map" or "phyllo dough function" (stretch and fold over): $T() : [0, 1] \to [0, 1]$:

$$T(x) = \begin{cases} 2x & \text{if } x \le 1/2 \\ 2 - 2x & \text{otherwise} \end{cases}.$$

This $T()$ is a measure preserving function on $[0, 1]$ (under the standard Borel measure). For instance: $T^{-1}([0, 1/2]) = [0, 1/4]) \cup [3/4, 1]$ so we can confirm $\mu(T^{-1}([0, 1/2])) = 1/2 = \mu([0, 1/2])$. Notice this $T()$ does not satisfy $\mu(T([0, 1/2])) = \mu([0, 1/2])$ (the left having measure 1 and the right measure $1/2$).

An example of non-measure preserving function on $\mathbb{R}$ is just $T(x) = 2x$. $T()$ is one to one and onto $\mathbb{R}$, but neither $T()$ or $T^{-1}()$ are measure preserving.

But why do we have this definition in terms of inverses? First there is the obvious analogy to the definitions of continuity and measurablity (both of which are phrased in the inverse form). The real issue is what interpretation is measure supposed to support in ergodic theory.

2.2.1 Interpretation of Measure Preserving

As we have seen measure preserving is not taken to mean $\mu(T(E)) = \mu(E)$ (measure preserved under application) but instead $\mu(T^{-1}(E)) = \mu(E)$. We will try to explain the reasoning behind this.

Suppose we are looking at a deterministic dynamical system $T()$. Look at the sequence of points

$$x, T(x), T^2(x), T^3(x), \cdots = x_1, x_2, x_3, x_4, \cdots .$$

In the sense of dynamical systems *any* $T()$ defined on all of X is "preserving measure under application" as we know with certainty at time k the point we are following has moved to exactly one point by $T^k()$. The problem of "preserving measure under application" is just not that interesting.

But what about backward measure? Suppose we ran an experiment or had a function (or random variable) $f() : X \to \{0, 1\}$ that we wished to evaluate. Going forward if we know x_1 we can evaluate $f(x_k)$ for any $k \ge 1$ as $f(T^k(x_1))$. Going backward if we only knew x_k (for some large k) we may not be able to infer x_1 (as $T^{-1}()$ can be multi-valued). So going backwards has uncertainty and here is where we need help and tools. Suppose $f()$ is itself measurable which in this case means $\chi_{f()} = \{x | f(x) = 1\} \in \Sigma$ and that instead of knowing the detailed state x_k we only know the observation $f(x_k) = 1$. $x_k \in \chi_{f()}$ implies $x_1 \in (T^{-1})^k(\chi_{f()})$ Since $T()$ is measurable we have $(T^{-1})^k(\chi_{f()}) \in \Sigma$ and when $T()$ is measure preserving we also have:

$$\mu((T^{-1})^k(\chi_{f()})) = \mu(\chi_{f()}).$$

And that is the merit of the definition: $f()$ tells us exactly as much information about the location of x_k as it does about the location of $(T^{-1})^k(x_k)$. With respect to $f()$ we know as much about the observed x_k and the unobserved x_1. This is as much as we could possibly hope for: running $T()$ backwards does not increase uncertainty of any observable experiment (measurable $f()$).

2.2.2 Justifying the name "Measure Preserving"

We have seen the concept of measure preserving transforms is useful: but why is it called "measure preserving" and not something like "inverse measure preserving?" We can explain this using the theory of adjoints: measure preserving $T()$ naturally corresponds to an operator $\widehat{T}()$ on the space of measures such that $\widehat{T}(\mu()) = \mu()$ (justifying the name). Measure preserving transformations leave the measure $\mu()$ unaltered. The natural correspondence used is the adjoint over the inner product defined by function application.

This is worth seeing in detail.

We will write the application of measure a measure $\kappa()$ to a measurable set E as the inner product:
$$< \kappa(), E >= \kappa(E).$$
Now look for the adjoint of $T()$. That is an operator $\widehat{T}()$ on the space of measures such that:
$$< \widehat{T}(\kappa()), E >=< \kappa(), T(E) > . \tag{2}$$

It is easy to see that $\widehat{T}()$ is just the operator $\widehat{T}()$ such that $\widehat{T}(\kappa()) = \widehat{\kappa}()$ where $\widehat{\kappa}()$ is the measure that for all measurable E: $\widehat{\kappa}(E) = \kappa(T(E))$. This is just function application again, or in computer scientist terminology $\widehat{T}()$ is curried function. So far we have done nothing but introduce notation.

Now plug $\mu()$ and any set of the form $T^{-1}(E)$ into Equation 2:

$$
\begin{aligned}
< \widehat{T}(\mu()), T^{-1}(E) > &= < \mu(), T(T^{-1}(E)) > & (3)\\
&= < \mu(), E > & (4)\\
&= < \mu(), T^{-1}(E) > . & (5)
\end{aligned}
$$

We move from Equation 3 to Equation 4 using the fact $T(T^{-1}(E)) = E$. We move from Equation 4 to Equation 5 using the fact that $T()$ is measure preserving with respect to $\mu()$. We now have for all E:
$$\left\langle \widehat{T}(\mu()), T^{-1}(E) \right\rangle = \left\langle \mu(), T^{-1}(E) \right\rangle .$$

So $\widehat{T}(\mu())$ and $\mu()$ agree on measure of for all sets F that can be written as $T^{-1}(E)$ for some $E \in \Sigma$. This isn't necessarily all sets in Σ; but it is enough establish that $\widehat{T}(\mu())$ and $\mu()$ must be the same function over Σ.[10]

So:
$$\widehat{T}(\mu()) = \mu()$$

(hence the name measure preserving, it sends $\mu()$ to itself as a map on the set of measures).

The summary being: the statement $\mu(T^{-1}(E)) = \mu(E)$ is exactly the extra statement needed to drive a proof that $\widehat{T}(\mu()) = \mu()$. Hence the name "measure preserving" what is being preserved under $T()$ (or more accurately its adjoint $\widehat{T}()$) is the measure function $\mu()$ itself not the volume or area calculated by it.

3 Ergodicity

3.1 The ergodic property

Ergodicity is a property of measure preserving operators that is *in addition to* preservation of measure (the two concepts are often confused).

Let $T()$ be a measure preserving transform. $T()$ is called "ergodic" if it is also "sufficiently mixing":
$$T^{-1}(E) = E \text{ implies } \mu(E) = 0 \text{ or } 1. \tag{6}$$

Once again this seems backwards and a bit more complicated than hoped (as we are going to be working with $T()$, not $T^{-1}()$). Breiman states that this property was exactly the unproven step in Gibbs' work that Birkhoff made an explicit requirement (leading to the ergodic principle used to prove the ergodic theorem).[Bre92] One should have an immediate suspicion that the arguments would be similar to what we worked through in Section 2.2.2. The definition is what it is, not because it is the most obvious definition- but because it is the one that is needed elsewhere.

---

[10]This is a classic linear algebra argument, we only need a enough check sets so that we have a full rank set of checks to establish equality. We do not need to check every set.

3.2 The consequence of ergodicity

Under the above conditions the result of the ergodic theorem is (almost certainly with the choice of x):

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k(x)) = \int_{x \in X} f(x) \mathrm{d}\mu(x) = \mathrm{E}_{\mu()}[f()] \tag{7}$$

(where $T^k()$ is $T()$ applied k-times and f() is an arbitrary measurable function). The amazing thing (and what makes the ergodic theorem useful) is establishing connectedness (with respect to $T()$) is enough to show enough mass (mass measured according to $\mu()$) is moving around to get averages right.

This time we can demonstrate a nice forward implication: for ergodic $T()$ and measurable E with $\mu(E) \notin \{0, 1\}$ then $T(E) \setminus E \neq \emptyset$ (we escape all sets except sets of measure 1 or 0, this is very related to the what computer scientist call "expansion"). This is a consequence of the conclusion of the ergodic theorem. Define $f_E(x) = 1$ if $x \in E$ (and 0 otherwise) and pick $x \in E$ according to the distribution $\mu()$ restricted to E. Apply the ergodic theorem to get $\mathrm{E}_{\mu()}[f_E()] = \mu(E)$ almost certainly with our choice of x. Since $\mu(E) < 1$ this means $T^k(x)$ must have escaped E some fraction of the time (else the limit on the left-hand side of Equation 7 would converge to 1, not to $\mu(E)$).

Notice we didn't even need $T()$ to be random in any sense (and this is in fact why things like the linear congruential pseudo random generator appear attractive- they are ergodic on some subset of the integers, also the ergodic properties of dynamical systems are why real world coin flips appear random). We emphasize: there is no randomness needed in the definition of $T()$, the ergodic property ensures that all measurable test-functions are starting to approach their expectations (under $\mu()$) as we average over time. So we will say almost surely (with choice of starting x): a sample drawn from the $T()$ process *at a random time index* is distributed nearly by $\mu()$ over our probability space (because pretty much only a random variable can simultaneously match expectations of all test functions including various "are you in this region?" indicator functions).

This result gets applied again and again: evaluate a large sum on the left side of Equation 7 and you have a good estimate of the integral on the right hand side.

3.3 Some examples

A trivial example ergodic transformation: $T(x) = (x + 1)$ modulo 997 (x an integer from 0 to 996). This $T()$ is measure preserving under the uniform measure $\mu(x) = 1/997$ and it is ergodic as 997 is prime (so there is no S such that $T^{-1}(S) = S$ other than the set of all integers from 0 to 996 which has the required measure of 1). The ergodic theorem in this case says if you were to run $T()$ many times and observe at random time you would be see x in a uniform random position (not trapped in some subset). Notice we instead only inspect at times k*997 we would always see x with the exact same value (as this is period of the operator $T()$), a decidedly non-random behavior. This is very weak application of the ergodic theorem, but it is captures some of the spirit of the theorem.

To properly understand the principle of a theorem you need to also demonstrate a concrete counterexample. For our counterexample let our space of points be pairs (x,y) where x and y are both integers and $T((x, y)) = (x + 1 \text{ modulo } 997, y + 1 \text{ modulo } 997)$. This $T()$ is not ergodic with respect to the uniform distribution $\mu()$ because the set $S = \{(x, y) | x = y\}$ has $\mu()$-measure 1/997 (so not equal to 0 or 1) yet $T^{-1}(S) = S$ (violating the ergodic condition). And failing to meet the conditions we happen to also fail to achieve outcomes. Consider the test-function $f((x, y)) = (x - y)^2$. For any starting point (x,y) $f((T^k)((x, y)) = f((x, y))$ for all k, so no average of f() over time is going to approach the global average value of f() over all possible states (as f() is not a constant function). $T()$ looks very non random and the sets that $T()$ stays trapped in are themselves are the most obvious evidence of this failure.

3.4 Some Historic Context

According to "On the origin of the notion of "Ergodic Theory""[Mat88] ergodic theory formalizes some notions of recurrence that were used in the study of dynamical systems by Boltzmann, Maxwell and others in the 1860s through 1870s. P. and U. Ehrenfest comment on the property in 1911 and von Neuman and Birkhoff have published on the topic by 1932. All just before Kolmogorov publishes his foundational work of measure-theoretic probability in 1933.

This is to emphasize the deterministic roots of ergodic theory (despite the current measure theoretic clothes).

4 Randomized Transforms

Typically the computer scientist wants to work with a randomized transform over a finite set of points (or objects). To do this requires a slight adaptation of the dynamical system terminology.

Suppose we have a set of finite set of points X and a target measure $\mu()$ we wish to sample from. In this case the ergodic theorem is restated as:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k(x)) = \frac{1}{|X|} \sum_{x \in X} f(x)\mu(x) = \mathrm{E}_{\mu()}[f()]. \tag{8}$$

This remains useful because the sum on the right could be very difficult to estimate when when $|X|$ is very large. And often even $|X|$ itself is hard to compute (this is called the partition function in statistical physics).

4.1 Finite Deterministic Transforms

The first issue is: if $T()$ is a deterministic measure preserving ergodic transform over onto (or surjective over) a finite set of points then $T()$ becomes one to one (or injective on) on this set.[11] Therefore the only distributions that are going to support measure preserving maps (which in this finite case connect a single subset of points with non-zero measure) are uniform (that is $\mu(x) = 0$ or c for all $x \in X$).

In this situation you get the conclusions of ergodicity for trivial reasons. The time based expectation is exactly the sum over all reachable states because $T()$ necessarily cycles through all of the states with non-zero measure in one large sequence.

But this supplies an intuition about ergodic theory: it is trying to characterize when you can state strong relations between surjectiveness and injectiveness in the non-finite case. And we can show some flavor of applying ergodicity if we prove the above statements using ergodic style arguments.

First: if X is finite then since $T()$ is surjective (onto) it must also be injective (one to one). So in this case $T^{-1}()$ is also a function from X onto X and is measure preserving. Thus $\mu(\{T(x)\}) = \mu(x)$ for all $x \in X$. Now pick an $x \in X$ with $\mu(x) > 0$. Look at the sequence of points $T^k(x)$ ($k \geq 0$). At some point this sequence must start to repeat and cycle. Let $E \subseteq \{T^k(x)|k \geq 0\}$ be the set of points that recur. We have $\mu(E) > 0$ and $T(E) = E = T^{-1}(E)$ which implies (by the ergodicity of $T()$) $\mu(E) = 1$. So all of the points x with $\mu(x) > 0$ are in a single cycle under $T()$ and all have the same value for $\mu(x)$.

To work with distributions other than uniform we must work with infinite collections of points. To get back some of the power of the original theory (and be able to work with non-uniform distributions) we augment our finite state space X into a specific larger (infinite) state space.

---

[11]This is practically the definition of finiteness: injective and surjective are always equivalent.

4.2 Random Tape as an Infinite Object

The obvious construction to get a deterministic system to simulate a random process is to add an external source of randomness. This is likely the construction most computer scientists have in mind when then think of the ergodic theorem being applied to a random process.

Starting with a finite state space X and our distribution $\mu()$ we design a measure preserving $T()$ on that operates on tuples of form $X \times \mathbb{Z}$. We assume we have access to an auxiliary function $\Omega() : \mathbb{Z} \to [0,1]$. The computer scientist intuition is that $\Omega()$ is our "randomness tape".

$T()$ is defined as

$$T((x,z)) = (\tau(x, \Omega(z)), z + 1)$$

where $\tau()$ is a deterministic function from $X \times [0,1]$ to X we will define shortly. This is a bit awkward- but should be very familiar to a computer scientist as the left part of our tuple is finite state and the right part is infinite tape. From the symbolic dynamics point of view we are trying to encode $T()$ as a shift operator.

We define $\tau()$ as follows. Label the points in X as $1, \cdots, n$. Let P be an n by n non-negative real matrix with all columns summing to 1. Then

$$\tau(x,y) = w \in X \text{ where } \sum_{i < \text{label}(w)} P(i, \text{label}(x)) < y \text{ and } \sum_{i \le \text{label}(w)} P(i, \text{label}(x)) \ge y. \tag{9}$$

Or in words: we $\tau()$ assigns a $P(w,x)$ fraction of the range of the tape input y for moving from x to w. The intended intuition is: when y is drawn uniformly from the interval $[0,1]$ then $P(w,x)$ represents the probability of $\tau(x,y)$ moving from x to w (this is the transpose of the traditional notation, but it lets us treat P as a standard left-operator). To avoid confusion we will write $P(w,x)$ as $P(x \to w)$.

Frankly this construction buys us nothing. It is better to a get rid of the deterministic $T()$ or reliance on shift operators (the idea of getting all of our randomness at once). It is better to move on to true random processes based more directly on P (move to Markov chains).

From Breiman:

> The basic property characterizing Markov chains is a probabilistic analogue of the familiar property of dynamical systems.

Which we take to mean- in a Markov chain the state is not a deterministic function of the previous state (as it was in a dynamical system) but shares the property that no state other than the previous state has any influence on the next state (in probability terms we are conditionally independent of the past given the most recent state). We will expand on this in our next section.

4.3 Markov Chains Directly

A random process (or sequence of observations) $x_1, x_2, \cdots$ is called a Markov chain if:

$$P(x_k = v | x_1, \cdots, x_{k-1}) = P(x_k = v | x_{k-1}) \text{ (for all choices of } x, v, k).$$

For example a sequence of independent coin-flips is forms a Markov chain. And (in the other extreme) the deterministic sequence $H, T, H, T, \cdots$ (formed by deterministicly turning a coin over each step) also forms a Markov chain.

We will translate Markov chains into the notation we have worked out above (using P to record the transition dependencies). Let us directly consider a randomized transform process $T()$ that is a randomized function from finite set X onto X such that the probability that $T(x) = y$ is $P(x \to y)$. In this case different applications of the process $T()$ to the same x can have different outcomes. This is essentially the random process of the previous Section 4.2, but without worrying so much about the bookkeeping around the randomness or trying to encode the whole system as shift-operator.

However I would note the precise accounting of where randomness is coming from and when it is being used (encoding the system as a great big shift operator) has led to some amazing results (such as coupling from the past [WP96, PW97, Fil97]).

A little algebra shows for $T()$ to be measure preserving (with respect to $\mu()$) we must have P such that:

$$\text{for all } x \in X : \sum_{y \in X} \mu(y) P(y \to x) = \mu(x) \tag{10}$$

or in vector notation:

$$P\mu = \mu. \tag{11}$$

It is natural to call this conservation of flow with respect to $\mu()$ (instead of continuing to say "measure preserving").

And this is immediately equivalent to Kirchhoff's conservation of flow laws of the form:

$$\text{for all } x \in X : \sum_{y \in X} \mu(y) P(y \to x) = \sum_{z \in X} \mu(x) P(x \to z) \tag{12}$$

(in-flow equals out-flow).

The detailed conservation of flow with respect to $\mu()$ is easy to show. The natural idea is that we should be working with formal sums (the 18th century way of thinking) or vector spaces over our set of states (the modern notation). Under this set-up the matrix P from Section 4.2 has a natural interpretation as an operator over these vectors (and the vectors entries can in term be interpreted as probabilities of being in each state).[12] Since all columns of P sum to 1 it has the all ones vector as a left-eigenvector with eigenvalue 1. So a right-eigenvector corresponding to this eigenvalue would have have a conservation of flow. We will design P to have $\mu()$ as the unique such right-eigenvector. Uniqueness is simple: we just need the system to be connected and stationary (such as there being a k such that P^k has only positive entries which is simple to ensure as long as our state space is connected and there is at least one x such that $P(x \to x) > 0$, see for example[KS76]).

To implement this conservation with respect to $\mu()$ (or have $\mu()$ be the right-eigenvector with eigenvalue 1) we need to add some more conditions to the design of P. One method do design this to use the "detailed balance" conditions of the so-called Metropolis-Hastings sampler:[MRR+53] pick P so that

$$\mu(x) P(x \to y) = \mu(y) P(y \to x).$$

This form is also called "time reversibility" and the idea is: under distribution $\mu()$ each edge is equally likely to be seen taken in the forward direction as in the reverse. This is a very strong and useful condition. In terms of linear algebra it means that if D is the diagonal matrix such that $D_{x,x} = \sqrt{\mu(x)}$ then $D^{-1}PD$ is a symmetric matrix and therefore P is itself similar to a diagonal matrix (so P is a simple type of operator and easily characterized by its eigenvalues, as we have just seen). This treatment using linear algebra to directly characterize probabilistic processes over states is incredibly powerful.[Big94, Chu96]

We can design for detailed balance in a number of ways. Let $\delta(x, y) : X \times X \to \{0, 1\}$ be any symmetric function such that the graph on the vertices X with edges (x, y) such that $x \neq y$ and $\delta(x, y) = 1$ is connected. Then for $x \neq y$ any of the following definitions of P give us the desired conservation with respect to $\mu()$:

- $P(x \to y) = c\delta(x, y)\mu(y)$.
- $P(x \to y) = c\delta(x, y) \min(1, \mu(y)/\mu(x))$.

---

[12]So instead of intentionally working in the dual/adjoint space of measures we instead extend our primal space from states or sets of states to vectors indexed by states. This extra record keeping ability gives us enough power to represent and prove everything we need to know in the extended primal space. Also the bounded measures and measurable distributions are isomorphic- so it really is the same record keeping in disguise.

- $P(x \to y) = c\delta(x, y)\sqrt{\mu(y)/\mu(x)}$.

(where $c > 0$ is chosen so $P(x \to y) < 1$ and $P(x \to x) = 1 - \sum_{y \neq x} P(x \to y) > 0$ for all x, y). We can have $\delta()$ be the constant function 1 (all states can try to move to each other). But it is more more typical to have $\delta()$ be near neighborhoods so $\sum_y \delta(x, y)$ is much smaller than $|X|$ (so it is much less work to simulate the Markov chain than to directly sample from X). Notice several of these schemes require only ratios of the density which is very helpful in cases where we have access to an un-normalized measure (for example using Markov chain methods to estimate the value of the partition function itself!).

Proof techniques to show Markov chain achieve the desired distribution quickly (or are "rapidly mixing") using isoperimetric arguments and Cheegar bounds revolutionized theoretical computer science.[AM84, JVV86, SJ89] And this is why we work to show various random processes are in fact Markov chains, once we know they are we get access to a lot of tools and results.

Our claim is that most of the influence of ergodic theory on Markov chains is the consideration of conservation of flow as paramount (called preservation of measure in ergodic theory) and that once you make conservation of flow central you get quickly derive a very powerful theory.

4.4   The Gibbs sampler

Another important sampling/simulation technique is the Gibbs sampler. Suppose for X we have a set of partitions of state state in X: $P_1, \cdots, P_n$ such that

- For each $P_i = \{S_{i,1}, \cdots S_{i,j(i)}\}$ we have X is the disjoint union of the $S_{i,j}$.
- For each $x, y \in X$ there is a finite sequence $x_1, \cdots x_k \in X$ such that $x_1 = x$, $x_k = y$ and for each $i < k$ exists u such $\{x_i, x_{i+1}\} \subseteq P_u$.
- For each i and $x \in X$ we can let u be the index such that $x \in S_{i,u}$ and compute the value of:

$$\widehat{\mu}_i(x) = c \sum_{\{x,y\} \in S_{i,u}} \mu(y)$$

for some positive constant c.

The Gibbs sampler (for generation as opposed to optimization) runs repeats the following on a starting point x_1:

- Generate i from 1 to n (either deterministicly or randomly such that each value from 1 to n is visited with positive density).
- Choose $x_{j+1} \in X$ with probability $\mu(x_{j+1})/\widehat{\mu}_i(x_j)$.

What the Gibbs sampler does is pick a partition, then perfectly sample on the element of the given partition our current point is in. This is repeated until we have a good distribution. If i is picked uniformly at random then the Gibbs sampler is in fact a Markov chain. Also, if we make the selection of the partition deterministic with period-n (say by picking $i = 1 + ($ stepnumber modulo n)) then the random process of observing the Gibbs sampler at n step intervals is again a Markov chain.

Typically the Gibbs partition is along coordinate axes. In this case suppose $X \subseteq \mathbb{Z}^n$ and P_i is the partition of X where only x_i varies in each partition element.

$$\widehat{\mu}_i(x) = c \sum_{\substack{y \in \mathbb{Z} \\ (x_1, \cdots, x_{i-1}, y, x_{i+1}, \cdots x_n) \in X}} \mu((x_1, \cdots, x_{i-1}, y, x_{i+1}, \cdots x_n))$$

The use of disjoint partitions is essential to the technique (uniformly mixing on overlapping subsets does not itself always ensure we are uniformly mixing on the whole space).

The point to be retained is: the Gibbs sampler is not an exotic system, it is very similar to a standard Markov chain designed with the usual detailed balance conditions (so the tools available to analyze Markov chains can be used on Gibbs samplers).

5 Conclusion

It is my hope you now have a basic understanding of and respect for ergodic theory.

My opinion is: ergodic theory is most used in computer science to derive the importance of conservation of flow in designing Markov chains (the core of many randomized algorithms). The *properties* you look for in ergodic theory (conservation of flow and recurring sets) are very important to the design of randomized algorithms (in particular Markov chains). However, most of these properties can be quickly established by deliberate design of state transition probabilities (so computer scientist tend not to use the machinery behind the Ergodic theorem).

In ergodic theory itself the technical problem being addressed is that the original state space may not be rich enough to express and prove useful theorems. Dynamical system style ergodic theory solves this by moving the the dual/adjoint space of measures. Markov chain theory solves this directly by introducing a vector space (or weighted sums) over states. These different solutions are isomorphic, but have different styles and are appropriate for different applications.

References

[AM84]   N. Alon and V.D. Milman, *Eigenvalues, expanders and superconcentrators*, Foundations of Computer Science **25** (1984).

[Big94]   Norman Biggs, *Algebraic graph theory*, 2nd ed., Cambridge Mathematical Library, 1994.

[BKS91]   Tim Bedford, Michael Keane, and Caroline Series, *Ergodic theory, symbolic dynamics and hyperbolic spaces*, Oxford Science, 1991.

[Bre92]   Leo Breiman, *Probability*, SIAM, 1992.

[Chu96]   Fan Chung, *Spectral graph theory*, AMS, 1996.

[Fil97]   James Allen Fill, *An Interruptible Algorithm for Perfect Sampling via Markov Chains*, Annals of Probability (1997).

[JVV86]   M.R. Jerrum, L. G. Valiant, and V.V. Vazirani, *Random generation of combinatorial structures from a uniform distribution*, Theoretical Computer Science **43** (1986), 169–188.

[KRS92]   Mark Kac, Gian-Carlo Rota, and Jacob T. Schwartz, *Discrete thoughts*, Birkhauser, 1992.

[KS76]   John G. Kemeny and J. Laurie Snell, *Finite markov chains*, Springer, 1976.

[Mat88]   Martin Mathieu, *On the origin of the notion of "ergodic theory"*, Expo. Math **6** (1988), 373–377.

[MRR+53]   N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, *Equations of state calculations by fast computing machines*, Journal of Chemical Physics **21** (1953), no. 6, 10871092.

[PW97]   James Gary Propp and David Bruce Wilson, *Coupling from the Past: a User's Guilde*, Contemporary Mathematics (1997).

[Rot97]   Gian-Carlo Rota, *Iniscrete thoughts*, Birkhauser, 1997.

[SJ89]   Alistair Sinclair and Mark Jerrum, *Approximate counting, uniform generation and rapidly mixing markov chains*, Information and Computation **83** (1989), 93–133.

[WP96]   David Bruce Wilson and James Gary Propp, *Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics*, Random Structures and Algorithms **9** (1996), 194–203.