An Introduction to Ergodic Theory
Normal Numbers: We Can't See Them, But They're Everywhere!

Joseph Horan
Department of Mathematics and Statistics
University of Victoria

November 29, 2013

Abstract

We present an introduction to ergodic theory, using as the basic example the unit interval on the real line, together with Borel sets and Lebesgue measure, as well as the standard "multiply by $k \mod (1)$" map. After briefly developing the example, we state the Birkhoff Ergodic Theorem, and subsequently use it to prove the Borel Normal Number Theorem, thus showing the application of ergodic theory in a field seemingly far removed from analysis.

1 Introduction

One can loosely define the field of ergodic theory as the study of time and space averages, and when they are related. But that requires asking, "What do we mean by time and space averages?" Thus we must start with some definitions.

Definition 1.1. A *(measure-theoretic) dynamical system* is a tuple $(X, \mathfrak{B}, \mu, \tau)$, where $(X, \mathfrak{B}, \mu)$ is a measure space, μ is a positive measure, and τ is an *measure-preserving transformation* on X, that is, a map such that $\tau^{-1}(A) \in \mathfrak{B}$, and $\mu(\tau^{-1}(A)) = \mu(A)$, for all measurable sets $A \in \mathfrak{B}$.

It is interesting and important to note that the properties of *finite* measure spaces, ones with $\mu(X) < \infty$, are very different from *infinite* measure spaces, with $\mu(X) = \infty$. Here, we will consider the finite case, and by a rescaling, require that $\mu(X) = 1$, so that $(X, \mathfrak{B}, \mu)$ is a *probability* space. As well, there are substantial differences between *discrete*, as opposed to *continuous*, dynamical systems. In a discrete setting, we consider the iteration of the map on points in X, whereas in a continuous setting, points move continuously, usually according to some sort of flow. Here, we will consider the discrete case, and give the following definition.

Definition 1.2. For $x \in X$, the *orbit* of x under τ is the set $[x]_\tau = \{\tau^n(x)\}_{n=0}^{\infty}$.

The intuitive notion here is that X is the space of all states of the system, and $x \in X$ is a specific state. Applying τ to x yields $\tau(x)$, which is the *new* state of the system after one time step, and the orbit of x is the sequence of states the system will be in, given that it starts at state x. Furthermore, if f is a function on X, it can be considered an observable of the system; it yields some quantity dependent on the state of the system. This is the basic idea of dynamics.

Now, where does the term "ergodic" come into play?

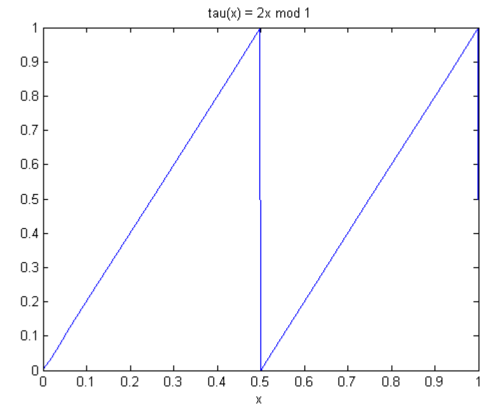
Definition 1.3. We say that a pair (μ, τ), where μ and τ are as above, is *ergodic* if for any set τ-invariant set A (ie. $\tau^{-1}(A) = A$), we have either $\mu(A) = 0$ or $\mu(X \setminus A) = 0$.

For a τ-invariant set A, if $x \in A$, then $\tau(x) \in A$, and if $x \notin A$, then $\tau(x) \notin A$. So $X \setminus A$ is τ-invariant as well, so we can partition the space as $X = A \cup (X \setminus A)$, and instead of studying τ on all of X, we can study it separately on A and $X \setminus A$. If (μ, τ) is ergodic, then this means that one of those sets has zero measure; hence the τ-invariant sets with non-trivial measure must have necessarily take up essentially the whole space, in a measure-theoretic sense. Hence the system is, in some sense, indecomposable.

It would probably be wise to give an example at this point. This is, if not *the* canonical example, certainly one of the canonical examples of an ergodic dynamical system.

Example 1.1. Let $X = [0, 1]$, $\mathfrak{B}$ be the Borel sets on $[0, 1]$, λ be the Lebesgue measure on $[0, 1]$, and $\tau : [0, 1] \to [0, 1]$, $\tau(x) = kx \mod 1$. Clearly, $(X, \mathfrak{B}, \lambda)$ is a probability space. It is relatively straight forward to show that τ is a measure-preserving transformation; it is simply the observation that the pullback of an interval (a, b) is k disjoint intervals each with measure $\frac{b-a}{k}$. Furthermore, τ is also ergodic, the proof of which we shall omit.

The following diagram shows the map for the case $k = 2$, as drawn by Matlab. The lines at $x = 0.5, 1$ are discontinuities.



2 The Birkhoff Ergodic Theorem

We now discuss the idea of time and space averages.

Definition 2.1. Let $f \in L^1(X)$. For $x \in X$, the *time average* of f at x is

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

when the limit exists. If $\mu(X) < \infty$, then the *space average* of f on X is

$$\frac{1}{\mu(X)} \int_X f \; d\mu.$$

Intuitively, this should make sense; the time average of f is the average value of the function on the orbit of x, and the space average of f is the average value of the function on the entire space. A good question to ask is "How are these related?" A reasonably good answer is the following theorem.

Theorem 2.1 (Birkhoff Ergodic Theorem)**.** *Let $(X, \mathfrak{B}, \mu, \tau)$ be a dynamical system, and $f \in L^1(X)$. Then we have:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} (f(\tau^i(x))) = \tilde{f}(x), \; \mu - a.e.$$

Moreover, $\tilde{f}$ is τ-invariant, ie. $\tilde{f}(x) = \tilde{f}(\tau(x))$, and $\int_X \tilde{f} = \int_X f$, so $\tilde{f} \in L^1(X)$. If (μ, τ) is ergodic, then $\tilde{f}$ is constant, with

$$\tilde{f} = \frac{1}{\mu(X)} \int_X f \; d\mu, \; \mu - a.e.$$

This theorem is one of the foundational results of ergodic theory. It states that the time average of a function on an orbit is itself an almost everywhere well-defined function, which is constant on the orbit, and has the same integral over the whole space. Moreover, if we have an ergodic system, we see that *the time average of the function is equal to the space average of the function*, for all but a set of measure zero. This matches what we should expect from the definition; if the orbits reach most of the space, then the average of f over them should be the space average. It's at least plausible that this is the case. More practically, we get information about the function in the long-term by looking at the whole space at the present time; quite useful for many applications.

For a more in-depth treatment of ergodic theory, one such source is [2]; the proof of Birkhoff's theorem given there is elementary, though perhaps not as intuitive as one might hope.

3 The Borel Normal Number Theorem

Now that we have some background knowledge, let us see how we can apply the theory. It turns out that a relatively straightforward application of Birkhoff's theorem yields a rather astounding result in number theory. First, we state some definitions and notation, to set up the theorem. Recall that when using base b expansions, we always choose a representation ending with all zeros, instead of ending with infinitely many instances of $b-1$.

Definition 3.1. Fix a positive integer $b > 1$. Let $x \in [0,1]$, with base b expansion $0.x_1x_2\ldots$. Let $w \in W_b = \{0, 1 \ldots, b-1\}^*$ be a finite word, of length $|w|$ (the $*$ operator is the Kleen star, signifying to take all finite strings over those symbols). Then

$$N^n_{b,w}(x) = \#\{1 \le i \le n - |w| + 1 \ : \ x_i \ldots x_{i+|w|-1} = w\}$$

is the number of times the word w appears in the first n digits of the base b expansion of x.

x is said to be *simply normal with respect to the base* b if when $w = i,\ 0 \le i \le b-1$, we have

$$\lim_{n\to\infty} \frac{N^n_{b,w}(x)}{n} = \frac{1}{b},$$

and is *normal with respect to the base* b if we have, for any word w,

$$\lim_{n\to\infty} \frac{N^n_{b,w}(x)}{n} = \frac{1}{b^{|w|}}.$$

x is said to be *simply normal* if it simply normal with respect to *any* base b, and is called *(absolutely) normal* if it is normal with respect to any base b.

Finally, if $x \in \mathbb{R}$, we can divide by b multiple times (and if necessary, multiply by -1; we assume $x \ge 0$) until we have $b^{-m}x \in [0,1]$. If the base b expansion of x is $x_{m-1}x_{m-2}\ldots x_1x_0.x_{-1}\ldots$, then the base b expansion of $b^{-m}x$ is $0.x_{m-1}x_{m-2}\ldots$. Then we say that x has any of the above properties if $b^{-m}x$ has the property, since the expansions have the same digits, just shifted.

These definitions are due to Borel, around the turn of the 20th century [4]. They are intended to describe the density of strings and randomness of the digits in a number's expansions.

Now that we've stated the definitions, let us state the theorem.

Theorem 3.1 (Borel Normal Number Theorem)**.** *Almost every real number, with respect to the Lebesgue measure, is absolutely normal.*

As an interesting historical remark which juxtaposes this theorem and Birkhoff's theorem, Borel essentially proved the theorem (up to an error which was quickly corrected) using the Strong Law of Large Numbers. Birkhoff's theorem above is exactly that law in a much, much more general setting.

This result may seem boring; if we call numbers normal, surely most numbers must be normal! However, this is very much counter-intuitive, and a terrible choice of naming; the closest we've come to writing down an absolutely normal number is defining an algorithm to compute one (see the 1917 article by Sierpinski [5], and the 2002 paper by Becher and Figueira [1]). We have demonstrated examples of normal numbers with respect to a given base; eg. Champernowne's constants $C_b,\ n \ge 2$, given by concatenating the base b expansions of the natural numbers in order, are normal with respect to base b [3], but it is unknown if they are normal in any other bases. So if we can't even write such a number down, how can there be so many of them?

Proof. Fix a base $b \geq 2$, and a word w as above. Let $x \in [0,1], x = 0.x_1x_2\dots$ in base b. The key observation is that we can divide the unit interval into $b^{|w|}$ intervals, disjoint except at the endpoints, where for each number in a subinterval, the first $|w|$ digits of its base b expansion are fixed, and all elements in the same interval have the same first $|w|$ digits. More precisely, for each integer $0 \leq i \leq b^{|w|} - 1$, we let $A^i_{b,|w|} = [\frac{i}{b^{|w|}}, \frac{i+1}{b^{|w|}}]$, and so if w_i is the length $|w|$ word corresponding to i, we get:

$$\begin{aligned} x \in A^i_{b,|w|} &\iff i \leq b^{|w|}x = x_1x_2\dots x_{|w|}.x_{|w|+1}\dots \leq i+1 \\ &\iff (x_1x_2\dots x_{|w|})_b = i \\ &\iff x_1x_2\dots x_{|w|} = w_i \\ &\iff N^{|w|}_{b,w_i}(x) = 1. \end{aligned}$$

In the converse case, $N^{|w|}_{b,w_i}(x) = 0$, so we see that $N^{|w|}_{b,w_i} = \chi_{A^i_{b,|w|}}$.

Now, if $n < |w|$, clearly $N^n_{b,w}(x) = 0$, since there aren't enough digits in which to find w. If $n > |w|$, then if we want to count the number of instances of w, we check digits starting with the first one, then shift and start checking digits from the second digit on, then from the third, etc. This is the following:

$$\begin{aligned} N^n_{b,w}(x) &= N^{|w|}_{b,w}(x) + N^{|w|}_{b,w}(bx) + \dots + N^{|w|}_{b,w}(b^{n-|w|}x) \\ &= \sum_{i=0}^{n-|w|} N^{|w|}_{b,w}(b^i x) \\ &= \sum_{i=0}^{n-|w|} \chi_{A^k_{b,|w|}}(b^i x), \end{aligned}$$

where we consider multiplication as multiplication $\mod b$, and k corresponds to the word w in base b.

So far, we haven't done anything really explicitly related to ergodic theory; we've just developed some notation and a partition (up to a finite set) of $[0,1]$. Let us define $\tau : [0,1] \to [0,1],\ \tau(x) = bx \mod 1$. Then with λ being the Lebesgue measure on $[0,1]$, we see that we're in the circumstance of the example setting back in the introduction. Notably, (λ, τ) is ergodic.

Now, if we consider our previous statements with our "new" context, $b^i x = \tau^i(x)$, so now when we're counting instances of strings of digits, we're really just counting the number of times the orbit of x under τ lies in a given interval. The dynamics of the system are "shift the base b expansion to the left."

We recall that every characteristic function of a measurable set in a probability space is in $L^1(X)$. In our case, we have:

$$\int_X \chi_{A^k_{b,|w|}}\ d\mu = \lambda(A^k_{b,|w|}) = \frac{1}{b^{|w|}}.$$

Thus, by the Birkhoff Ergodic Theorem, we see that for $n \geq |w|$:

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} N^n_{b,w_k}(x) &= \lim_{n\to\infty} \frac{n-|w|+1}{n} \frac{1}{n-|w|+1} \sum_{i=0}^{n-|w|} N^{|w|}_{b,w_k}(b^i x) \\ &= \lim_{n\to\infty} \frac{n-|w|+1}{n} \frac{1}{n-|w|+1} \sum_{i=o}^{n-|w|} \chi_{A^k_{b,|w|}}(\tau^i(x)) \\ &= (1)\frac{1}{\lambda([0,1])} \int_X \chi_{A^k_{b,|w|}}\ d\mu \\ &= \frac{1}{b^{|w|}} \end{aligned}$$

for almost every $x \in [0,1]$. This was independent of our choice of w, thus almost every real number in the unit interval is normal with respect to the base b.

To finish the proof, let $B_{b,w}$ be the set of $x \in [0,1]$ such that the above limit does *not* hold, for a choice of b and w. By Birkhoff's theorem, this set has measure zero. Observe that $W_b = \{0, 1, \ldots, b-1\}^*$ is a countable set, since a countable union of finite sets is clearly countable. Thus, if B is the set of all non-normal numbers, by σ-sub-additivity of measures, we obtain:

$$\lambda(B) = \lambda(\bigcup_{b=2}^{\infty} \bigcup_{w \in W_b} B_{b,w}) \leq \sum_{b=2}^{\infty} \sum_{w \in W_b} \lambda(B_{b,w}) = \sum_{b=2}^{\infty} \sum_{w \in W_b} 0 = 0.$$

Hence the set of real numbers in the unit interval which are *not* absolutely normal is of measure zero. Now, notice that the map $\tau^* : \mathbb{R} \to \mathbb{R},\ \tau(x) = bx$, maps $[x, y]$ to $[bx, by]$, with $\lambda(\tau^*(A)) = b\lambda(A)$ for any measurable set A, so that a set of measure zero in $[0, 1]$ is mapped to a set of measure zero. Negating the positive real line doesn't change measure, so that we have:

$$\lambda(\bigcup_{i=0}^{\infty} (\tau^*)^i (B \cup -B)) \leq \sum_{i=0}^{\infty} \lambda((\tau^*)^i (B \cup -B)) = \sum_{i=0}^{\infty} 2b^i \lambda(B) = \sum_{i=0}^{\infty} 0 = 0.$$

Hence, we see that the set of non-normal real numbers has measure zero, which is the theorem.

□

4 Conclusion

This is the beauty of non-constructive mathematics; we can prove results without being able to write anything concrete down. This is also the drawback of non-constructive mathematics; we have almost no idea what these numbers look like, even though they're everywhere.

Observe that ergodic theory gave us a powerful result, which did the majority of the heavy lifting in the proof. We reformulated the problem in terms of a dynamics situation, and obtained a number theoretic result essentially out of nowehere. This is but one application of ergodic theory to areas of mathematics to which you wouldn't necessarily expect it to apply. This is why ergodic theory can be extraordinarily valuable, both as a field in its own right, and as a crossover tool to apply in a wide-ranging set of topics.

References

[1] Verónica Becher and Santiago Figueira. An example of a computable absolutely normal number. *Theoretical Computer Science*, 270(1):947–958, 2002.

[2] Abraham Boyarski and Paweł Góra. *Laws of Chaos: Invariant Measures and Dynamical Systems in One Dimension.* Birkhäuser, 1997.

[3] D. G. Champernowne. The Construction of Decimals Normal in the Scale of Ten. *J. London Math. Soc.*, S1-8(4):254, 1933.

[4] M. Émile Borel. Les probabilités dénombrables et leurs applications arithmétiques. *Rendiconti del Circolo Matematico di Palermo*, 27(1):247–271, 1909.

[5] W. Sierpinski. Démonstration élémentaire du théorème de M. Borel sur les nombres absolument normaux et détermination effective d'une tel nombre. *Bull. Soc. Math. France*, 45:125–132, 1917.