0.1 (Title Page)

Dynamical Systems,
Stochastic Processes,
and Information Theory

---

1 Dynamical Systems

1.1 Dynamical Systems

A **dynamical system** is a mathematical object used to model physical systems which evolve *deterministically* in time.

The *state* of the physical system is represented by a point in a **state space**.

As time passes, the state of the system changes in a *deterministic* way. Thus, the point moves around the state space in a *predictable* fashion.

The movement of the point through the statespace therefore defines a *endomorphism* from the statespace to itself.

1.2 Dynamical Systems

Definition 1.1

A **dynamical system**[1] is a pair (X, T), where:

- X is a set (the **statespace**)
- $T : X \longrightarrow X$ is a function (the **transformation**) .

Usually T is a *bijection*; ie. the flow of time in the system is *reversible* .

---

[1] Here, we consider only **discrete time** dynamical systems, to keep things simple.

2 Flows and Vector Fields

2.1 Flows on Manifolds

Let M be a *smooth manifold* , and V a *vector field* on M.

V defines a **flow** on the manifold.

Starting at any point $x \in M$, imagine being dragged through M along a trajectory

$$\gamma_x : (a_x, b_x) \longrightarrow M$$

dictated by V.

Here, (a_x, b_x) is some time-interval where γ_x is well-defined.

2.2 Integral Curves

The trajectory γ_x satisfies the conditions:

- The trajectory passes through x at time 0; that is:

$$\gamma_x(0) = x.$$

- The *velocity vector* of a particle moving along the trajectory is always equal to the value of the vector field at that point. That is: For all times $t \in (a_x, b_x)$, we have

$$\dot{\gamma_x}(t) = V(\gamma_x(t))$$

2.3 The Exponential Map

For "nice" vector fields, V, the path γ_x is well-defined for *all* times; the trajectory does not "run off the edge of the manifold" in finite time.

Hence, we get a trajectory $\gamma_x : \mathbb{R} \longrightarrow M$.

Do this for all $x \in M$. This collection of trajectories defines a map $F : \mathbb{R} \times M \longrightarrow M$

$$F(x,t) \ := \ \gamma_x(t)$$

This is the **exponential map** associated with the vector field V. We denote it by $\exp(V)$, as in:

$$\exp(t \cdot V)(x) \ := \ \gamma_x(t)$$

2.4 The Exponential Map

The exponential map is a smooth *group action* of the real numbers on the manifold. That is:

- For all $t \in \mathbb{R}$, the map $\exp(t.V) : M \longrightarrow M$ is a *diffeomorphism.*
- $\exp((t+s) \cdot V) \ = \ \exp(t \cdot V) \circ \exp(s \cdot V)$.
- $\exp(0 \cdot V) \ = \ \mathbf{Id}_M$.
- $\exp(t \cdot s \cdot V) \ = \ \exp(t \cdot (s.V))$.

To create a discrete-time dynamical system, look at the **time-one map** :

$$T \ := \ \exp(1 \cdot V)$$

T is a *diffeomorphism* from M to itself, and therefore defines a dynamical system.

---

3 Flows and Differential Equations

3.1 Ordinary Differential Equations

Any system of **ordinary differential equations** on $\mathbb{R}^D$ induces a **vector field**.

The *flow* defined by the vector field provides the *solution* for the ODE.

3.2 ODEs and Vector Fields

Consider an *ordinary differential equation* concerning a function $u : \mathbb{R} \longrightarrow \mathbb{R}^D$, given by:

$$\frac{\partial^N \vec{u}}{\partial t^N} \;=\; F\left(\vec{u},\; \frac{\partial \vec{u}}{\partial t},\; \frac{\partial^2 \vec{u}}{\partial t^2},\; \ldots,\; \frac{\partial^{N-1} \vec{u}}{\partial t^{N-1}}\right).$$

Let $M := \underbrace{\mathbb{R}^D \times \ldots \times \mathbb{R}^D}_{N}$.

Define vector field V on M by:

$$\begin{aligned} & V(\vec{x}_0,\; \vec{x}_1, \ldots,\; \vec{x}_{N-1}) \\ & \quad = \quad \left(\vec{x}_1,\; \vec{x}_2, \ldots,\; \vec{x}_{N-1},\; F(\vec{x}_0, \ldots, \vec{x}_{N-1})\right) \end{aligned}$$

3.3 ODEs and Vector Fields

Thus,

$$\begin{aligned} \frac{\partial \vec{u}}{\partial t} &= \vec{x}_1, \\ \frac{\partial \vec{x}_1}{\partial t} &= \vec{x}_2, \\ \frac{\partial \vec{x}_2}{\partial t} &= \vec{x}_3, \\ &\vdots \\ \frac{\partial \vec{x}_{N-1}}{\partial t} &= \vec{x}_N, \\ \textit{and} \quad \frac{\partial \vec{x}_N}{\partial t} &= F\left(\vec{u}, \vec{x}_1, \ldots, \vec{x}_N\right), \end{aligned}$$

which is what we want.

---

4 Ergodic Theory

4.1 Invariant Measures

Definition 4.1 *Invariant Measure*

An **invariant measure** for (X, T) is a measure μ on X which is **invariant** under the action of the transformation.

That is, for any measurable set $U \subset X$:

$$\mu[U] \;=\; \mu\left[T^{-1}U\right].$$

Normally, μ is a *probability measure.*

4.2 Invariant Measures

Start at some point $x \in X$, and follow its **orbit** through X over time:

$$\ldots T^{-2}(x),\; T^{-1}(x),\; x,\; T(x),\; T^2(x),\; T^3(x), \ldots$$

The orbit spends "a lot of time" in some parts of X, and "very little time" in other parts of X.

An invariant measure tells you "how much time" the average point spends in different parts of the space.

4.3 Measure Preserving Dynamical System

A dynamical system (X, T), together with a T-invariant measure μ, is called a **measure-preserving dynamical system** (**MPDS**).

Formally, a **measure-preserving dynamical system** consists of a quadruple $(X, \mathcal{X}, \mu;\; T)$. Here,

- X is a set.
- $\mathcal{X}$ is a *sigma-algebra* on X.
- μ is a *measure* defined over $\mathcal{X}$.

- $T : X \longrightarrow X$ is a function which is *measurable* with respect to $\mathcal{X}$, and *preserves* the measure μ. (ie. μ is *invariant* with respect to T .)

.

4.4 Constructing Invariant Measures

Let X be a topological space, $T : X \longrightarrow X$ a homeomorphism. Let $x \in X$. Define the measure μ_x by:

For any open subset $U \subset X$,

$$\mu_x[U] \ := \ \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{1}_U \left[T^n(x)\right] \quad (1)$$

(here, $\mathbb{1}_U$ is the *characteristic function* of U.)

The measure assigned to U is the *long term average frequency* with which the orbit of x enters U.

An average like (1) is called an **ergodic average** .

4.5 Krylov-Bogolioubov Theorem

If X is a **compact topological space** and $T : X \longrightarrow X$ is a homeomorphism, then the ergodic average (1) will always produce an invariant measure. More generally...

Theorem 4.2 *Krylov and Bogolioubov*

Let ν be a probability measure, on X. Consider the sequence of measures: $\nu_1,\ \nu_2,\ \ldots$, where

$$\nu_{_N} \ := \ \frac{1}{N} \sum_{n=0}^{N-1} \left(\nu \circ T^n\right)$$

1. The sequence $\left\{\nu_n|_{n\in\mathbb{N}}\right\}$ always has **cluster points** in the **weak topology** on $\mathcal{M}(X)$.

2. Any cluster point is a T-invariant measure.

□

4.6 Constructing Invariant Measures

If μ is an **ergodic** measure, then the *Birkhoff Ergodic Theorem* says that μ is *always* defined by an ergodic average.

The *Ergodic Decomposition Theorem* says that *any* $T-$invariant measure can be decomposed as a *convex integral combination* of ergodic measures.

Thus, in a sense, *every* invariant measure can be built out of ergodic averages.

4.7 Ergodic Measures

Definition 4.3 *Ergodic*

Let (X, T) be a dynamical system. A $T-$invariant probability measure μ is called **ergodic** if all $T-$invariant subsets of X have trivial measure. In other words, if $U \subset X$ is such that

$$T(U) = U,$$

then either

- $\mu[U] = 0$, (ie. μ says U is "almost nothing"),
  or
- $\mu[U] = 1$ (ie. μ says U is "almost everything").

4.8 Ergodic Theorems

Ergodic measures have very nice behaviour with respect to *ergodic averages.*

Theorem 4.4 *Birkhoff Ergodic Theorem*

Let (X, T) be a dynamical system and μ be a $T-$invariant, ergodic measure.

- Let $U \subset X$ be measurable. Then for $\mu-$almost all $x \in X$,

$$\mu[U] \;=\; \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{1}_U \left[T^n(x)\right]$$

- More generally, let $f : U \longrightarrow \mathbf{C}$ be an $\mathbf{L}^1$ function. Then for $\mu-$almost all $x \in X$,

$$\int_X f \, d\mu \;=\; \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f\left[T^n(x)\right]$$

4.9 Ergodic Theorems

The *Birkhoff Ergodic Theorem* is the starting point of an entire branch of *ergodic theory*; the study of *ergodic averages.*

There are many other *ergodic theorems*, describing the convergence of different kinds of ergodic averages in different kinds of dynamical systems.

---

5 Stochastic Processes

5.1 Stochastic Processes

Stochastic processes are used to model the evolution of systems which do *not* behave **deterministically** in time.

Some systems appear to behave "randomly" merely because we lack complete information about their internal state or dynamics.

Others are genuinely random in their evolution.

In either case, the appropriate tool to model the system is a stochastic process.

5.2 Stochastic Processes

In a *dynamical system* (X, T), each point x had a unique *orbit* through X :

$$\ldots T^{-2}(x),\ T^{-1}(x),\ x,\ T(x),\ T^2(x),\ T^3(x), \ldots$$

Now, however, starting at x, there are many *possible* future paths the system could travel through, and many possible past histories it could conceivably have had.

Hence, in a stochastic process, we want to put a *probability distribution* on the space of all possible orbits through x.

5.3 Stochastic Processes

So, we want a probability distribution on the space of all **orbits** in the space X.

Every orbit corresponds to some particular "history" for the system. Intuitively, we are saying that some "histories" are more likely than others.

5.4 Stochastic Processes

An orbit in X is a $\mathbf{Z}-$indexed sequence of points

$$\ldots x_{-3},\ x_{-2},\ x_{-1},\ x_0,\ x_1,\ x_2,\ x_3, \ldots$$

Thus, the space of all such possible orbits is $X^{\mathbf{Z}}$. This is called **sequence space** .

Definition 5.1 *Stochastic Process*

Let X be a set. A **stochastic process** with *statespace* X is a probability measure on $X^{\mathbf{Z}}$.

5.5 Events as Subsets of Statespace

Intuitively, a subset of state space corresponds to an "event".

For example, if the X is the **weather** statespace, then the event

It is raining.

corresponds to the set

$U := \{x \in X \ ; \ \text{It is raining in state } x \ \}.$

5.6 Causality and Physical Law

Science consists of trying to find *physical laws* which describe relationships of *cause* and *effect* between events occuring at different times.

In other words, science is about discovering relationships of *causality* .

In a *stochastic process*, causality manifests through *correlations in probability* between events occuring at different times.

5.7 Causality and Physical Law

Example:

Consider a physical law of the form:

If event A happens today, then event B will happen tommorrow.

Mathematically:

The conditional probability *of event B happening at time* 1, given *that event A happened at time* 0, *is 1.*

5.8 Causality and Physical Law

As an equation:

$$\mathbf{P}_{\text{rob}}\left[B \text{ at time } 1 \;\middle|\; A \text{ at time } 0 \right] \;=\; 1.$$

If A and B are treated as *subsets* of the statespace X, then we can rewrite this:

$$\mathbf{P}_{rob}\left[T(x) \in B \;\middle|\; x \in A\right] \;=\; 1.$$

5.9 Nondeterministic Causality

In a *deterministic* setting, causal relationships are always of this simple, "all-or-nothing" variety.

In real science, however, things are rarely so simple. Normally, scientific "laws" take a more probabilistic form:

If event A happens today, then event B is very likely *to happen tommorrow.*

Mathematically:

$$\mathbf{P}_{rob}\left[T(x) \in B \;\middle|\; x \in A\right] \;>\; 1 - \epsilon.$$

(where ϵ is "small") .

5.10 Nondeterministic Causality

Sometimes we may even have a weaker statement:

If event A happens today, then event B is more likely *to happen tommorrow.*

which can be written mathematically:

$$\mathbf{P}_{rob}\left[T(x) \in B \;\middle|\; x \in A\right] \;>\; \mathbf{P}_{rob}\left[T(x) \in B\right].$$

Stochastic processes thus define a kind of *nondeterministic causality*, via *causal* relationships which are not completely *deterministic* in nature.

5.11 Nondeterministic Causality and the Philosophy of Science

In the *Philosophy of Science* , there has long been dispute over how the scientific desideratum of *causality* can be meaningful in a possibly *nondeterministic* universe.

The emergence of *quantum mechanics* as a fundamentally nondeterministic theory of nature has made this issue particularly pressing.

The *nondeterministic causality* found in stochastic processes can help resolve this controversy.

5.12 The Shift Map

Intuitively, the "flow of time" in a stochastic process can be simulated in *sequence space* by the map which shifts each sequence "forward" by one time unit. In other words, we define the map:

$$\mathcal{S}_{hift} : X^{\mathbf{Z}} \longrightarrow X^{\mathbf{Z}}$$

so that, if

$$\vec{x} := \left(\dots x_{-3},\ x_{-2},\ x_{-1},\ \boxed{x_0},\ x_1,\ x_2,\ x_3, \dots \right)$$

(where the "zeroth element" is in the box)
Then:

$$\mathcal{S}_{hift}\ (\vec{x}) := \left(\dots x_{-2},\ x_{-1},\ x_0,\ \boxed{x_1},\ x_2,\ x_3, x_4 \dots \right)$$

This is called the **shift map** .

5.13 Examples of Shifts

The sequences illustrated here takes its values on the Roman Alphabet, $\mathcal{A} := \{\mathsf{A}, \mathsf{B}, \mathsf{C}, \dots, \mathsf{Z}\}$

The sequence $\vec{a}$										
. . . . .	-4	-3	-2	-1	0	1	2	3	4	.....
. . . . .	G	V	J	Q	E	H	K	X	Z	.....

The sequence $\mathcal{S}_{\mathit{shift}}(\vec{a})$										
. . . . .	-4	-3	-2	-1	0	1	2	3	4	.....
. . . . .	V	J	Q	E	H	K	X	Z	D	.....

The sequence $\mathcal{S}_{\mathit{shift}}^{-3}(\vec{a})$										
. . . . .	-4	-3	-2	-1	0	1	2	3	4	.....
. . . . .	F	P	I	G	V	J	Q	E	H	.....

5.14 Stationary Stochastic Processes

A basic tenet of modern science is that physical laws are *unchanging over time.*

In a stochastic process, physical laws are represented by *probability correlations* between events occuring at different times.

Thus, we would like these correlations to be unchanging over time.

We want the *probability* of a certain event, or sequence of events, to be *invariant* under the action of the *shift map* .

Definition 5.2 *Stationary Process*

Let μ be a probability measure on $X^{\mathbf{Z}}$, determining a stochastic process. The stochastic process is called **stationary** if μ is $\mathcal{S}_{\mathit{shift}}-$invariant. In other words, for all measurable $U \subset X^{\mathbf{Z}}$,

$$\mu[U] \ = \mu\left[\mathcal{S}_{\mathit{shift}}\, U\right].$$

5.15 Stationary Stochastic Processes as Measure Preserving Dynamical Systems

Any *stationary stochastic process* is thus a *measure-preserving dynamical system* .

Suppose we have a stationary stochastic process on the statespace X. This is a *shift-invariant probability measure* , μ, on the *sequence space* $X^{\mathbf{Z}}$.

If we think of $X^{\mathbf{Z}}$ as a *new* statespace, then $\mathcal{S}_{\mathit{shift}} : X^{\mathbf{Z}} \longrightarrow X^{\mathbf{Z}}$ is a bijection, and μ is *invariant* with respect to $\mathcal{S}_{\mathit{shift}}$.

Thus, $\left(X^{\mathbf{Z}}, \mu;\ \mathcal{S}_{\mathit{shift}}\right)$ is a *measure-preserving dynamical system.*

5.16 Measure Preserving Dynamical Systems as Stationary Stochastic Processes

We can also go in the reverse direction.

Any *measure preserving dynamical system* can be represented as a *stationary stochastic process* .

This is best understood through the concept of a **time series** .

---

6 Time Series

6.1 Time Series

Suppose we continuously observe some physical system, like the sun. Over time we collect data. This data does not provide *complete information* about the sun, but *does* provide us with some *predictive capability* .

For example, we might formulate a *correlation* like:

> *A fluctuation in the solar magnetic field will be followed soon after by a burst of solar radiation.*

6.2 Time Series: Measurements as Functions

Mathematically, a measurement is like a *function*:

$$f : X \longrightarrow Y,$$

where:

- X is the *statespace* of the sun.
- Y is the *statespace* of our measurement apparatus.

Example: Measuring the *luminosity* of the sun is a function $f : X \longrightarrow [0, \infty)$.

If the measurement provides *full information* about the sun, then f is an *injective* map.

Normally our measurements provide only *partial* information, so f is *many-to-one* .

6.3 Time Series

If we make a record of this data over time, we get an infinite sequence of data points. This is called a **time series** .

Let (X, T) be a **dynamical system**.

Let $f : X \longrightarrow Y$ be a **measurement**.

If the system is in state $x \in X$ at time *zero*, its **orbit** is

$$\begin{aligned} &\ldots,\ T^{-3}(x),\ T^{-2}(x),\ T^{-1}(x), \\ &\qquad x,\ T(x),\ T^2(x),\ T^3(x), \ldots \end{aligned}$$

Thus, we would get the time series:

$$\begin{aligned} &\ldots f\left(T^{-3}(x)\right),\ f\left(T^{-2}(x)\right),\ f\left(T^{-1}(x)\right), \\ &\qquad f(x),\ f\left(T(x)\right),\ f\left(T^2(x)\right),\ f\left(T^3(x)\right), \ldots \end{aligned}$$

6.4 Time Series

Thus, any initial state $x \in X$ induces a $\mathbf{Z}-$indexed **sequence** of measurement values in Y.

The measurement function f induces a map

$$F : X \longrightarrow Y^{\mathbf{Z}}$$

where

$$\begin{aligned} F(x) \quad := \quad &\left[\ldots, f\left(T^{-2}(x)\right),\ f\left(T^{-1}(x)\right),\right. \\ &\left. f(x),\ f\left(T(x)\right),\ f\left(T^2(x)\right), \ldots\right] \end{aligned}$$

is the **time series** of data generated by the state x.

6.5 Time Series as Stochastic Processes

Past measurements can help predict future measurements.

Example 6.1

The data points:

$$\dots f\left(T^{-3}(x)\right),\ f\left(T^{-2}(x)\right),\ f\left(T^{-1}(x)\right)$$

should help predict the value of $f(x)$.
It seems that a time series is like a *stochastic process.*
Indeed it is.

6.6 Time Series as Stochastic Processes

To interpret a time series as a stochastic process, we need a *probability measure* on $Y^{\mathbf{Z}}$.

Let μ be a *invariant probability measure* on the statespace X.

Use the function F to *push forward*[2] the measure μ to a probability measure ν on $Y^{\mathbf{Z}}$.

For any subset $U \subset Y^{\mathbf{Z}}$, define:

$$\nu[U] \ := \ \mu\left[F^{-1}(U)\right].$$

---

[2]Of course, U and F must be *measurable* with respect to some suitably chosen sigma-algebra on $Y^{\mathbf{Z}}$.

6.7 Time Series as Dynamical Systems

Because μ is $T-$invariant, the measure ν is $\mathcal{S}_{hift}-$invariant.

Thus, the time series induces a *stationary stochastic process.*

But every stationary stochastic process is a *measure preserving dynamical system.*

Thus, a time series is a measure preserving dynamical system.

6.8 Time Series as Representations

The function $F : X \longrightarrow Y^{\mathbf{Z}}$ is actually a **homomorphism** of measure-preserving dynamical systems:

$$F \circ T \;=\; \mathcal{S}_{hift} \circ F$$

In other words, we have a *commuting diagram*:

$$\begin{array}{ccc} X & \xrightarrow{T} & X \\ {\scriptstyle F}\downarrow & & \downarrow{\scriptstyle F} \\ Y^{\mathbf{Z}} & \xrightarrow[\mathcal{S}_{hift}]{} & Y^{\mathbf{Z}} \end{array}$$

6.9 Time Series as Representations

The time series thus acts as a *representation* of the MPDS $(X, \mu;\ T)$.

When is this representation "faithful"? When does it contain all information about $(X, \mu;\ T)$?

In general, f is not injective; it does not pass "full information" about the state of X.

F passes much *more* information: *all past measurements* ever made by f, and *all future measurements* that ever *will* be made.

Is it possible that F provides complete information about X?

6.10 Generating Time Series

Suppose that the chronological record of measurements provides total information about the current state of the system.

Thus, the function $F : X \longrightarrow Y^{\mathbf{Z}}$ is *injective*[3].

Thus, F is an **isomorphism** from the MPDS $(X, \mu;\ T)$ to the MPDS $(Y^{\mathbf{Z}}, \nu;\ \mathcal{S}_{hift})$.

We can *completely reconstruct* the dynamical system (X, T) from the probability distribution ν on $Y^{\mathbf{Z}}$.

We say that the time series **generates** the dynamical system.

6.11 Partitions

If the space Y is finite, then a function $f : X \longrightarrow Y$ is called a **partition**.

f cuts X up into finitely many "pieces", and "labels" them with the different elements of Y.

Think of Y as an "alphabet" of "letters". Sequences in Y are thus "words" written with these letters.

If $x \in X$, then the sequence of letters:

$$\Big[\ \ldots,\quad f\left(T^{-2}(x)\right),\ f\left(T^{-1}(x)\right),\ f(x),$$

$$f\left(T(x)\right),\ f\left(T^{2}(x)\right),\ f\left(T^{3}(x)\right),\ldots\ \Big]$$

is called the **name** of x.

6.12 Generating partitions

Let $f : X \longrightarrow Y$ be a *partition.*

If the *time series* of f *generates* $(X, \mu; T)$, then we say it is a **generating partition**.

---

[3]Actually, F need only be "almost" injective, in the sense that, for any measurable sets $U, V \subset X$, if $F(U) = F(V)$, then $U \triangle V$ has measure zero.

Theorem 6.2 *Krieger Generator Theorem*

If $(X, \mu;\ T)$ is a MPDS with *finite entropy*, then it has a *generating partition.*
□

Thus, any "nice" dynamical system is isomorphic to a *stationary stochastic process* on a *finite alphabet.*

---

7 Information Theory

7.1 Information Theory

Information Theory was invented by Claude Shannon in 1948, to address the problem of efficiently *transmitting* or *encoding* large amounts of data.

The idea of information theory is to take advantage of *patterns*, *redundancies*, and *statistical regularities* in the data to find more efficient ways to encode it.

Modern **compression algorithms**, such as *lharc*, *gzip*, *pkzip*, are the products of information theory.

7.2 Information Theory

Let $\mathcal{A}$ be some **alphabet** of symbols. Suppose the **cardinality** of $\mathcal{A}$ is A.

A **message** of length N, written in the alphabet $\mathcal{A}$, is a **sequence** of elements of $\mathcal{A}$. In other words, it is an element of $\mathcal{A}^N$.

There are A^N possible such messages. Not all of them are equally likely to be sent. Some may never be sent. So why use N symbols to encode each message?

7.3 Encoding Bitstrings

Suppose $\mathcal{A} := \{0,\ 1\}$. Thus, our messages are **bit strings** .

There are $2^{10} = 1024$ bitstrings of length 10. But maybe only 400 of them actually correspond to "real" messages.

Since $400 < 512 = 2^9$, we only need *nine* bits to encode each message.

7.4 English Words as strings of Letters

Suppose we encode English words using bit-strings.

The Roman Alphabet, including capital letters and punctuation, uses around 78 distinct symbols:

- 26 small letters
- 26 capital letters
- 10 digits
- space , . ; : “ ’ ! @ # $ % & * ( ) - = + / { } [ ] ¡ ¿

and $78 < 128 = 2^7$, so we need **7 bits** to encode each symbol.

7.5 English Sentences as strings of Words

The average length of an English word is 10 letters (include trailing spaces and punctuation) . Thus, a message containing 100 words will probably be about 1000 letters long. It will thus require 7000 bits.

But the English language only has a vocabulary of around $60,000$ words, and $60,000 < 65,356 = 2^{16}$.

Hence, we only *really* need 16 bits to encode each English word.

Imagine we assign a unique 16 bit **code** to each English word. We can **encode** a message 100 words in length using only 1600 bits; a savings of almost 70%!

7.6 English Sentences with Syntax

First we broke our message into individual *words*. Now let's break it into **sentences**: strings of words ending in a period.

Most strings of words are *impossible* as sentences. For example, you would *never* see:

Travel computer green walking two hundred, friendly goes tomorrow.

as a sentence in English. Sentences must follow certain basic **syntax rules**.

7.7 English Sentences with Syntax

A simple English sentence is of the form:

[object phrase] [verb phrase] [subject phrase].

For example:

[John] [walked slowly to] [the store].

[Brad] [loves] [Janice].

[A manifold] [is] [a topological space].

7.8 English sentences with Syntax

English sentences can get much more complicated than this, with **prepositional clauses**, **conditional clauses**, etc.

However, all English sentences follow certain rules of **syntax** governing how the words can be assembled into sentences.

This is the subject of the field of **Syntax** in the science of **Linguistics**.

7.9 English Sentence with Syntax

Suppose the **average word-length** of an English sentence is 20 words.

With our previous encoding scheme, the average sentence will require

$$20 \times 16 \; = \; 320$$

bits to encode. There are then 2^{320} "possible sentences".

However, the vast majority of these are *impossible* , since they are *syntactically nonsensical.*

7.10 English Sentences with Syntax

If we assume[4] that only one in a million random sequences of words forms a sentence. Then there are really only

$$\frac{2^{320}}{2^{20}} = 2^{300}$$

actual English sentences of 20 words' length.

If we assigned a unique bit-string to each one, then we only need 300 bits to encode each one.

7.11 Correlation and Compression

These examples illustrate the key ideas of **information theory**:

1. By looking at **statistical correlations** between symbols over *longer* lengths of time, we can achieve more *efficient* compression.
2. If we look at very long strings of symbols, we will find that most are either **impossible** or "**very, very, very unlikely**".

   We can thus assign **short** codes to *likely* strings, **long** codes to *very unlikely* strings, and **no code** to *impossible* strings.
3. However, a **Law of Diminishing Returns** applies to this strategy.

7.12 Stochastic Signal Sources

Claude Shannon mathematically modelled a message sender as a **stochastic signal source**.

Imagine the signal source sending a continuous sequence of symbols in the alphabet $\mathcal{A}$. For example:

- An undending sequence of English sentences.

[4] Very generously!

- An undending computer file.
- A **time series** from some measuring apparatus.

Of course, in "real life", all signals end eventually. But we can model an *extremely long* signal as being "approximately" endless.

7.13 Stochastic Signal Sources

The future signals sent by the signal source are at least partially **predictable** from the past. For example

- If the first 200 pages of a book have been about Winston Churchill, the next 200 pages will "probably" be about Winston Churchill.
- If a time series has been roughly periodic for the last 2 years, it will "probably" be roughly periodic for the next two years.

A stochastic signal source is thus a **stochastic process** ; it is a probability distribution on the space $\mathcal{A}^{\mathbf{Z}}$ of *all possible messages* in the alphabet $\mathcal{A}$.

7.14 Stationary Stochastic Signal Sources

Also, assume that the *probability correlations* between symbols don't change over time. For example:

- English vocabulary and syntax are unchanging over time.
- The subject matter of a book is unchanging over the length of the book.
- The physical laws underlying some time series are unchanging over the duration of the experiment.

Hence, we can model a signal source as a **stationary stochastic process**.

7.15 Entropy

Some signals are more **compressible** than others.

- A signal containing *English prose* is *very* compressible.
- A signal consisting of a sequence of successive random coin tosses is extremely *uncompressible.*

How can we measure how "compressible" a signal is, on average?

7.16 Entropy

Intuitively, for a fixed stochastic signal source X, we want to find a **compression ratio** ; a number $h \in [0,1]$, so that, we can say

> *The average uncompressed message of length N, from the source X, can be compressed into a code of length $h.N$.*

Does such a constant ratio exist?
It does, and it is called the **entropy** of the stochastic process.

7.17 Shannon-MacMillan-Briemann Theorem

Let $\mathcal{A}$ be an **alphabet** of A letters, and let $\mathcal{X}$ be an **ergodic stationary stochastic process** on $\mathcal{A}$.

The **Shannon-MacMillan-Briemann** theorem (also called the **Asymptotic Equipartition Theorem**) characterises the **entropy** of the process $\mathcal{X}$.

7.18 Shannon-MacMillan-Briemann Theorem

Let N be some "large" natural number, and let $\mathcal{X}_N := \mathcal{A}^N$ be the set of all **words** of length N in the alphabet. Thus $\mathcal{C}_{ard}[\mathcal{X}_N] = A^N$.

The SMB theorem says that, if N is large enough, we can identify a very small subset, $\mathcal{Y}$, of $\mathcal{X}$, as words which are "quite likely" to appear.

All other words are "very unlikely".

Thus, when making our code, we only need to provide for a small population of words.

7.19 Shannon-MacMillan-Briemann Theorem

Theorem 7.1

Let $\mathcal{X}$ have **entropy** h.

We can divide $\mathcal{X}_N$ into two subsets: $\mathcal{X}_N = \mathcal{Y}_N \sqcup \mathcal{Z}_N$, so that:

1. $\mathcal{C}_{ard}[\mathcal{Y}_N] = A^{h.N}$
2. All elements of $\mathcal{Y}_N$ are roughly *equally probable* to appear; each having probability approximately $\frac{1}{A^{(h \pm \epsilon).N}}$
3. The combined probability of *all* elements in $\mathcal{Z}_N$ is ϵ.

Here, ϵ is a number which can be made **arbitrarily small** as $N \to \infty$.

7.20 Shannon-MacMillan-Briemann Theorem

This theorem says that we can approach a **compression ratio** of h by using **block encoding** with sufficiently long blocks.

If we code *input blocks* of length N, then we only need to use *output blocks* of length $h.N$; for the "vast majority" of messages, this will be enough.

Very rarely, we will encounter an "exception" block (from $\mathcal{Z}_N$) . So allocate one code, Q, of length $h.N$ to mean "exception". If Z is in $\mathcal{Z}_N$, then code Z with the codeblock "Q, Z".

7.21 Entropy

There are other ways to define **entropy**, but the SMB theorem is sufficient to characterise it, so we will use it as the defining property.

Definition 7.2 *Entropy*

Let $\mathcal{X}$ be an **ergodic stochastic process** on a **finite alphabet** $\mathcal{A}$. Then there is a unique number $h \in [0,1]$ satisfying the statement of the SMB theorem. This number is called the **entropy** of the process. We denote it by:

$$h(\mathcal{X})$$

---

8 Entropy

8.1 The Entropy of a MPDS

We can also talk about the **entropy** of a **measure preserving dynamical system** .

Let $(X, \mu;\ T)$ be a **MPDS**. Let $\mathcal{P}$ be some **partition** of X. That is,

$$\mathcal{P} : X \longrightarrow \mathcal{A}$$

where $\mathcal{A}$ is some **finite set** .

Consider the **time series** induced by $\mathcal{P}$ and $(X, \mu;\ T)$. This is a **stationary stochastic process** on $\mathcal{A}$. Denote this process by $(\mathcal{P}, T)$.

8.2 The Entropy of an MPDS

The entropy of $(\mathcal{P}, T)$ tells us how "complex" the dynamics of $(X, \mu;\ T)$ are.

If the behaviour of $(X, \mu;\ T)$ is simple, then the past record of $\mathcal{P}-$data should be an excellent **predictor** of future behaviour. The stochastic process will be *predictable*; its entropy will be *low*.

If $(X, \mu;\ T)$ is complex, then the past record of $\mathcal{P}-$data is a *poor* predictor of future behaviour. The stochastic process will be *unpredictable*; its entropy will be *high*.

8.3 The Entropy of an MPDS

The entropy of $(\mathcal{P}, T)$ depends on the choice of partition $\mathcal{P}$. Different partitions will may give us very different levels of predictability.

The **supremum** of the entropy given by any partitions should reveal the "total complexity" of the MPDS $(X, \mu;\ T)$.

Definition 8.1 *Entropy*

Let $(X, \mu;\ T)$ be an MPDS. The **entropy** of $(X, \mu;\ T)$ is defined:

$$h(X, \mu;\ T)\ := \sup\{h(\mathcal{P}, T)\ ;\ \mathcal{P}\ a\ partition\ of\ X\}$$

8.4 Kolmogorov-Sinai Theorem

The Kolmogorov-Sinai theorem lets us actually *compute* the entropy of a MPDS.

Theorem 8.2 *Kolmogorov-Sinai*

Let $(X, \mu;\ T)$ be an MPDS, and let $\mathcal{P}$ be a partition of X.
If $\mathcal{P}$ is a **generating partition** , then

$$h(X, \mu;\ T)\ = h(\mathcal{P}, T).$$

8.5 Chaos

In **smooth dynamical systems** , one often observes a phenomenon called **chaos** .

Loosely, a system is **chaotic** if very small perturbations to the state of the system very rapidly grow into huge deviations in its behaviour.

Thus, points in statespace very close together can have orbits which rapidly diverge.

8.6 Chaos and Prediction

Chaos makes a dynamical system's behaviour very hard to *predict* .

A very small measurement error can propagate into a wild innacuracy in long-term predictions.

Very "crude" measurements, like those obtained through **partitions** , should be especially sensitive to this unpredictability.

Is there a connection between **chaos** and **entropy** ?

8.7 Lyapunov Exponents

Let $(\mathcal{M}, T)$ be a **smooth dynamical system**

One way to measure the "chaos" of $(\mathcal{M}, T)$ is via **Lyapunov Exponents**.

Positive Lyapunov exponents correspond to dimensions in which T "spreads things apart".

Negative exponents represent dimensions where T "mashes things together".

Thus, *any* positive Lyapunov exponent suggests some degree of **chaos** .

8.8 Lyapunov Eigenspaces

For any $x \in \mathcal{M}$, let $\mathbf{T}^{n}_{x}\mathcal{M}$ be the **tangent space** of M at x, and let

$$\mathcal{D}_x\left[T\right] : \mathbf{T}^{n}_{x}\mathcal{M} \longrightarrow \mathbf{T}^{n}_{T(x)}\mathcal{M}$$

be the **derivative** of the map T at x.

For almost x, the space $\mathbf{T}^{n}_{x}\mathcal{M}$ is a direct sum of **Lyapunov eigenspaces**, which are *expanded* or *contracted* at different rates by the iterated action of T. The rates of *expansion* and *contraction* are the **Lyapunov exponents**.

8.9 Regular Points

x is a **regular point** if we have a decomposition:

$$\mathbf{T}^{n}_{x}\mathcal{M} \;=\; E^{1}_{x} \oplus E^{2}_{x} \oplus \ldots \oplus E^{M}_{x}$$

into **Lyapunov eigenspaces** , and a collection of **Lyapunov exponents**

$$\lambda_1(x) > \lambda_2(x) > \ldots > \lambda_M(x)$$

so that, for all $m \in [1..M]$, and any vector $\vec{u} \in E_x^m$,

$$\lim_{n\to\infty} \frac{1}{n} \log \left\| \mathcal{D}_x \left[T^n \right] (\vec{u}) \right\| \;=\; \lambda_m(x).$$

The set of all regular points will be denoted by $\Lambda(\mathcal{M};T)$

8.10 Oseledec's Theorem

Topologically speaking, the set $\Lambda(\mathcal{M};T)$ is usually **meager** . However, Osoledec's theorem says, from a measure-theory point of view, $\Lambda(\mathcal{M};T)$ is more than big enough.

Theorem 8.3 *(Osoledec)*

Let $(\mathcal{M},T)$ be **smooth dynamical system**, with $\mathcal{M}$ **compact**. Then $\Lambda(\mathcal{M};T)$ has **total measure** . That is, for *any* $T-$invariant Radon probability measure μ on $\mathcal{M}$,

$$\mu \left[\Lambda(\mathcal{M};T) \right] \;=\; 1.$$

□

8.11 The Ruelle-Pesin Theorem

Ruelle and Pesin succeeded in directly relating the geometric **Lyapunov exponents** to the measure-theoretic **entropy** of a smooth dynamical system.

For any regular point $x \in \mathcal{M}$, define

$$\chi(x) \;:=\; \sum_{\lambda_i(x)>0} \lambda_i(x) \cdot \mathbf{D}^{im} \left[E_x^i \right]$$

Intuitively, this measures the total *rate of expansion* at the point x.

8.12 The Ruelle-Pesin Theorem

Theorem 8.4

Let $(\mathcal{M}, T)$ be a **compact smooth dynamical system**, and let μ be any $T-$invariant Radon measure.

1. (Ruelle)

$$h(\mathcal{M}, \mu;\ T) \ \leq \ \int_{\mathcal{M}} \chi \ d\mu.$$

2. (Pesin) Further, if T is Hölder $\mathbf{C}^1$, and μ is **absolutely continuous** with respect to the **Lebesgue measure**, then

$$h(\mathcal{M}, \mu;\ T) \ = \ \int_{\mathcal{M}} \chi \ d\mu.$$

□
