Hidden Gibbs Models: Theory and Applications

– DRAFT –

Evgeny Verbitskiy

Mathematical Institute, Leiden University, The Netherlands
Johann Bernoulli Institute, Groningen University, The Netherlands

April 4, 2015
## Contents

1. **Introduction** 3
   1.1. Hidden Models / Hidden Processes 4
   1.2. Hidden Gibbs Models 7

I. **Theory** 9

2. **Gibbs measures** 10
   2.1. Gibbs-Bolztmann Ansatz 10
   2.2. Gibbs measures for Lattice Systems 10
     2.2.1. Translation Invariant Gibbs states 12
     2.2.2. Regularity of Gibbs states 15
   2.3. Gibbs measures in Dynamical Systems 16

A. **Prerequisites** 17
   A.1. Notation 17
   A.2. Measure theory 17
   A.3. Stochastic processes 20
   A.4. Ergodic theory 20
   A.5. Entropy 21
     A.5.1. Shannon’s entropy rate per symbol 22
     A.5.2. Kolmogorov-Sinai entropy of measure-preserving systems 22
1. Introduction

Very often the true dynamics of a physical system is hidden from us: we are able to observe only a certain measurement of the state of the underlying system. For example, air temperature and the speed of wind are easily observable functions of the climate dynamics, while electroencephalogram provides insight into the functioning of the brain.

Partial Observability  Noise  Coarse Graining

Observing only partial information naturally limits our ability to describe or model the underlying system, detect changes in dynamics, make predictions, etc. Nevertheless, these problems must be addressed in practical situations, and are extremely challenging from the mathematical point of view.

Mathematics has proved to be extremely useful in dealing with imperfect data. There are a number of methods developed within various mathematical disciplines such as

- Dynamical Systems
- Control Theory
- Decision Theory
- Information Theory
- Machine Learning
- Statistics

to address particular instances of this general problem – dealing with imperfect knowledge.

Remarkably, many problems have a common nature. For example, correcting signals corrupted by noisy channels during transmission, analyzing neuronal spikes, analyzing
1. Introduction

genomic sequences, and validating the so-called renormalization group methods of the-
toretical physics, all have the same underlying mathematical structure: a hidden Gibbs
model. In this model, the observable process is a function (either deterministic or ran-
dom) of a process whose underlying probabilistic law is Gibbs. Gibbs models have been
introduced in Statistical Mechanics, and have been highly successful in modeling phys-
ical systems (ferromagnets, dilute gases, polymer chains). Nowadays Gibbs models are
ubiquitous and are used by researchers from different fields, often implicitly, i.e., without
knowledge that a given model belongs to a much wider class.

A well-established Gibbs theory provides an excellent starting point to develop Hidden
Gibbs theory. A subclass of Hidden Gibbs models is formed by the well-known Hidden
Markov Models. Several examples have been considered in Statistical Mechanics and
the theory of Dynamical Systems. However, many basic questions are still unanswered.
Moreover, the relevance of Hidden Gibbs models to other areas, such as Information
Theory, Bioinformatics, and Neuronal Dynamics, has never been made explicit and
exploited.

Let us now formalise the problem, first by describing the mathematical paradigm to
model partial observability, effects of noise, coarse-graining, and then, introducing the
class of Hidden Gibbs Models.

1.1. Hidden Models / Hidden Processes
§ 1.1. Hidden Models / Hidden Processes:

- Underlying time-dependent process \( \{X_t\} \),
- observable process \( \{Y_t\} \) is a function of \( \{X_t\} \).

The process \( \{X_t\} \) is either
- a stationary random \( \text{ (stochastic) process,} \)
- a deterministic dynamical process

\[
X_{t+1} = f(X_t)
\]

We will only consider discrete time (for simplicity). However, many of what we discuss applies to continuous time processes as well.

The process \( \{Y_t\} \) is a function of \( \{X_t\} \).

The function can be
- deterministic \( Y_t = \Phi(X_t) \) for all \( t \),
- or random \( Y_t \sim P_{X_t}(\cdot) \),

i.e. \( Y_t \) is chosen according to some probability distribution, which depends on \( \{X_t\} \).

N.B. We will implicitly assume that \( Y_t \) is chosen independently for every \( t \), e.g.

\[
Y_t = X_t + Z_t, \quad \{Z_t\} \text{ independent noise}
\]
Remark: Equivalence of "random" and "deterministic" settings

There is one-to-one correspondence between stationary processes and measure-preserving dynamical systems.

Proposition: Let \((\Omega, \mathcal{A}, \mu, T)\) be a measure-preserving dynamical system.

Then for any measurable \(\varphi: \Omega \to \mathbb{R}\),

\[ X^\omega_t = \varphi(T^t \omega), \quad \forall \omega \in \Omega, \quad t \in \mathbb{Z}_+ \text{ or } \mathbb{Z}, \]

is a stationary process.

Exercise 1: Prove Proposition 1.

In the opposite direction, suppose \(\{X_t\}\) is a stationary stochastic process on some measurable space \((\Omega, \mathcal{A}, \mu)\).

Let \(\hat{\Omega} = \Omega^{\mathbb{Z}_+}\) be the space of all possible realizations \(\{X_t\}_{t \geq 0}\).

\[ \mathcal{A} = \text{product } \sigma\text{-algebra} = \mathcal{A}_\Omega \times \mathcal{Z}_+ \]

is the minimal \(\sigma\text{-algebra} \text{ s.t. all projections are measurable}.

\[ \hat{\mu} = \text{the corresponding measure/Law of } \{X_t\} \]

\[ T: \hat{\Omega} \to \hat{\Omega} \text{ left shift } (T \tilde{\omega})_n = \tilde{\omega}_{n+1}. \]

Observable \(\varphi: \hat{\Omega} \to \mathbb{R}, \varphi(\tilde{\omega}) = \tilde{\omega}_0 \) (first coordinate).
1.2. Hidden Gibbs Models

The key assumption of the model is that the observable process is a function, either deterministic or random, of the underlying process, whose governing probability distribution belongs to the class of Gibbs distributions.

(i) The most famous example covered by the proposed dynamic probabilistic models are the so-called Hidden Markov Models (HMM). Suppose \( \{X_n\} \) is a Markov process, assuming values in a finite state space \( A \). In HMM’s, the observable output process \( Y_n \) with values in \( B \), is dependent only on \( X_n \), and the value \( Y_n \in B \) is chosen according to the so-called emission distributions:

\[
\Pi_{ij} = \mathbb{P}(Y_n = j \mid X_n = i), \quad i \in A, \ j \in B,
\]

for each \( n \) independently. The Hidden Markov models can be found in a broad range of applications: from speech and handwriting recognition, to bioinformatics and signal processing.

(ii) A Binary Symmetric Channel (BSC) is a widely used model of a communication channel in coding theory and information theory. In this model, a transmitter sends a bit (a zero or a one), and the receiver receives a bit. Typically, the bit is transmitted correctly, but, with a small probability \( \epsilon \), the bit will be inverted or flipped. We can represent the action of the channel as

\[
Y_n = X_n \oplus Z_n,
\]

where \( \oplus \) is the exclusive OR operation, and \( Z_n \) is the state of the channel: 0 – transmit ”as is”, 1 – flip. The channel can be memoryless: in this case, \( \{Z_n\} \) is a sequence of independent Bernoulli random variables with \( \mathbb{P}(Z_n = 1) = \epsilon \), or the channel can have memory, e.g., if \( \{Z_n\} \) is Markov (Gilbert-Elliott channel, suitable to model burst errors). If \( \{X_n\} \) is a binary Markov process, then this model is a particular example of a Hidden Markov model.

(iii) Neuronal spikes are sharp pronounced deviations from the baseline in recordings of neuronal activity. Often, the electrical neuronal activity is transformed into 0/1 processes, indicating the presence of absence of spikes. These temporal binary processes – spike trains, are clearly functions of the underlying neuronal dynamics. In the field of Neural Coding/Decoding, the relationship between the stimulus and the resulting spike responses is investigated. Another, particularly interesting type of spike processes
1. Introduction

Figure 1.1.: (a) The spike trains emitted by 30 neurons in the visual area of a monkey brain are shown for a period of 4 seconds. The spike train emitted by each of the neurons is recorded in a separate row. (source W.Maas, http://www.igi.tugraz.at/maass/). (b) EEG recordings of a patient with Rolandic epilepsy provided by the Epilepsy Centre Kempenhaeghe. Epileptic interictal spikes are clearly visible in the last 3 epochs.

originates from the so-called interictal (between the seizures) spikes. The presence of interictal spikes is used diagnostically as a sign of epilepsy.

(iv) The \textbf{decimation} is a classical example of renormalization transformation in Statistical Mechanics. Consider a $d$-dimensional lattice system $\Omega = A^{\mathbb{Z}^d}$, with the single-spin space $A$. The image of a spin configuration $x \in \Omega$, under the decimation transformation with spacing $b$, is again a spin configuration $y$ in $\Omega$, $y = T_b(x)$, given by

$$y_n = x_{bn} \quad \text{for all} \quad n \in \mathbb{Z}^d.$$ 

Under the decimation transformation, only every $b^d$-th spin survives. Figure on the right depicts the action of the decimation transformation $T_2$ on the two-dimensional integer lattice $\mathbb{Z}^2$. 
Part I.

Theory
2. Gibbs measures

Last modified on April 3, 2015

Existing literature on rigorous mathematical theory of Gibbs states is quite extensive. The classical references are the books of Ruelle [9] and Georgii [5, 6]. In the context of dynamical systems, the book of Bowen [7, 8] provides a nice introduction to the subject.

2.1. Gibbs-Bolzmann Ansatz

The Gibbs distribution prescribes the probability of the event that the system with the finite state space \( S \) is in state \( x \in S \), to be equal to

\[
P(x) = \frac{1}{Z} \exp\left(-\frac{1}{kT} H(x)\right),
\]

(2.1.1)

where \( H(x) \) is the energy of the state \( x \), \( k \) is the Boltzmann constant, \( T \) is the temperature of the system, and

\[Z = \sum_{y \in S} \exp\left(-\frac{1}{kT} H(y)\right)\]

is a normalising factor, known as the partition function.

Exercise. Show that the probability measure \( P \) on \( S \) given by (2.1.1) is precisely the unique probability measure on \( S \) which maximizes the following functional, defined on probability measures on \( S \) by

\[
P \mapsto H(P) - \beta \mathbb{E}H = H(P) = \text{Entropy} - \beta \cdot \text{Energy},
\]

where \( \beta = 1/kT \) is the inverse temperature.

2.2. Gibbs measures for Lattice Systems

Equation (2.1.1) – known as the Gibbs-Boltzmann Ansatz – makes perfect sense for systems where the number of possible states is finite. Typically for systems with infinitely many states, each state occurs with probability zero, and a new interpretation of (2.1.1) is required. This problem was successfully solved in late 1960’s by R.L. Dobrushin [1], O. Lanford and D. Ruelle [2]. The key is to identify Gibbs states as those probability
measures on the \textbf{infinite} state space which have very specific \textbf{finite volume} conditional probabilities, parametrized by boundary conditions, and given by expressions similar to (2.1.1).

The definition of Gibbs measures is applicable to rather general lattice systems $\Omega = \mathbb{A}^\mathbb{L}$, where $\mathbb{A}$ is a finite or a compact spin-space, and $\mathbb{L}$ is a lattice. For simplicity, let us assume that $\mathbb{A}$ is finite, and $\mathbb{L}$ is an integer lattice $\mathbb{Z}^d$, $d \geq 1$. Elements $\omega \in \Omega$ are functions on $\mathbb{Z}^d$ with values in $\mathbb{A}$, equivalently, at each site $n \in \mathbb{Z}^d$, “sits” a spin $\omega_n$ with a value in $\mathbb{A}$. If $\omega \in \Omega$, and $\Lambda$ is a subset of $\mathbb{Z}^d$, $\omega_\Lambda$ will denote the restriction of $\omega$ to $\Lambda$, i.e., $\omega_\Lambda \in \mathbb{A}^\Lambda$. Fix a finite volume $\Lambda$ is $\mathbb{Z}^d$, and some configuration $\omega_\Lambda^c$ outside of $\Lambda$, we will refer to $\omega_\Lambda^c$ as the \textit{boundary conditions}. Gibbs measures are defined by prescribing the conditional probability of observing configuration $\sigma_\Lambda$ in the finite volume $\Lambda$, given the boundary conditions $\omega_\Lambda^c$,

$$P(\sigma_\Lambda | \omega_\Lambda^c) = \frac{1}{Z(\omega_\Lambda^c)} e^{-\beta H_\Lambda(\sigma_\Lambda \omega_\Lambda^c)} =: \gamma_\Lambda(\sigma_\Lambda | \omega_\Lambda^c),$$ \hspace{1cm} (2.2.1)

where $\beta = \frac{1}{kT}$ is the inverse temperature, $H_\Lambda$ is the corresponding energy, and the normalising factor – partition function $Z$, now depends on the boundary conditions $\omega_\Lambda^c$. The analogy with (2.1.1) is clear.

Naturally, the family of conditional probabilities $\Gamma = \{\gamma_\Lambda(\cdot | \sigma_\Lambda^c)\}$, called \textbf{specification}, indexed by finite sets $\Lambda$ and boundary conditions $\sigma_\Lambda^c$, cannot be arbitrary, and must satisfy some consistency conditions.

In order to ensure the consistency of a the family of probability kernels $\{\gamma_\Lambda\}$, the energies $H_\Lambda$ must be of a very special form, namely:

$$H_\Lambda(\sigma_\Lambda \omega_\Lambda^c) = \sum_{V \cap \Lambda \neq \emptyset} \Psi(V, \sigma_\Lambda \omega_\Lambda^c),$$ \hspace{1cm} (2.2.2)

where the sum is taken over all finite subsets $V$ of $\mathbb{Z}^d$, and the collection of functions

\[\Psi = \{\Psi(V, \cdot) : \Omega \to \mathbb{R} : V \subset \mathbb{Z}^d, |V| < \infty\}\]

called the interaction, has the following properties:

\begin{itemize}
  \item \textbf{(locality):} for all $\eta \in \Omega$, $\Psi(V, \eta) = \Psi(V, \eta_V)$, meaning that the contribution of volume $V$ to the total energy only depends on the spins in $V$.
  \item \textbf{(uniform absolute summability):}
\end{itemize}

\[\|\|\| \Psi \|\| = \sup_{n \in \mathbb{Z}^d} \sum_{V \ni n} \sup_{\eta \in \Omega} |\Psi(V, \eta_V)| < +\infty.\]

\textbf{Definition.} A probability measure $P$ on $\Omega = \mathbb{A}^{\mathbb{Z}^d}$ is called a Gibbs state for the interaction $\Psi = \{\Psi(\Lambda, \cdot) | \Lambda \in \mathbb{Z}^d\}$ (denoted by $P \in \mathcal{G}(\Psi)$ if (2.2.1) holds $P$-a.s., equivalently, if the so-called Dobrushin-Lanford-Ruelle equations hold for all $f \in L^1(\Omega, P)$.

11
2. Gibbs measures

\[(f \in C(\Omega, \mathbb{R})) \text{ and all } \Lambda \in \mathbb{Z}^d\]

\[\int_{\Omega} f(\omega) \mathbb{P}(d\omega) = \int_{\Omega} (\gamma_\Lambda f)(\omega) \mathbb{P}(d\omega),\]

where

\[(\gamma_\Lambda f)(\omega) = \sum_{\sigma_\Lambda \in A^\Lambda} \gamma_\Lambda(\sigma_\Lambda|\omega_{\Lambda^c}) f(\sigma_\Lambda \omega_{\Lambda^c}).\]

This is equivalent to saying that

\[\mathbb{E}_\mathbb{P}(f|\mathcal{F}_{\Lambda^c}) = \gamma_\Lambda f,\]

where \(\mathcal{F}_{\Lambda^c}\) is the \(\sigma\)-algebra generated by spins in \(\Lambda^c\).

**Theorem 2.1.** If \(\Psi\) is a uniformly absolutely summable interaction, then the set of Gibbs measures for \(\Psi\) is not empty.

Thus the Dobrushin-Lanford-Ruelle (DLR) theory guarantees existence of at least one Gibbs state for any interaction \(\Psi\) as above, i.e., a Gibbs state consistent with the corresponding specification. It is possible that multiple Gibbs states are consistent with the same specification; in this case, one speaks of a phase-transition.

### 2.2.1. Translation Invariant Gibbs states
Translation Invariant Gibbs states

Lattice (group) \( \mathbb{Z}^d \) acts on \( \mathbb{A}^{\mathbb{Z}^d} \) by translations:

\[ \mathbb{Z}^d \ni k \rightarrow S^k, \]

where \( S^k(w) = \hat{w} \) with

\[ \hat{w}_n = w_{n+k} \quad \text{for all} \quad n \in \mathbb{Z}^d, \]

i.e. configuration \( w \) is shifted by \( k \).

Measure \( \mathbb{P} \) on \( \mathcal{L} = \mathbb{A}^{\mathbb{Z}^d} \) is called translation invariant if

\[ \mathbb{P}(S^{-k}A) = \mathbb{P}(A) \]

for every Borel set \( A \).

When are Gibbs states translation invariant?

What are sufficient conditions for existence of translation invariant Gibbs states?
Definition. An interaction \( \Phi = \{ \Phi(\Lambda, \cdot) \mid \Lambda \in \mathbb{Z}^d \} \) is called translation invariant if
\[
\Phi_{\Lambda + k}(S^k w) = \Phi(\Lambda)(w) \quad (\Phi_{\Lambda + k} \circ S^k = \Phi_{\Lambda})
\]
for all \( \Lambda \in \mathbb{Z}^d, k \in \mathbb{Z}^d, w \in \mathcal{S} = \mathcal{A} \mathbb{Z}^d \).

Example: Ising model of interacting spins, \( A = \{-1, 1\}^2 \)
\[
\Phi_{\Lambda}(w) = \begin{cases} 
-\theta w_i, & \text{if } \Lambda = \{i, i\} \\
-\theta w_i w_j, & \text{if } \Lambda = \{i, j\} \text{ and } i \neq j \\
0, & \text{otherwise}
\end{cases}
\]

\( \theta > 0 \): the interaction is ferromagnetic
\( \theta < 0 \): the interaction is anti-ferromagnetic,
\( \theta \): is magnetic field (external)

If \( \Phi \) is translation invariant, then so is the specification
\[
\Phi_{\Lambda + k}(S^k \sigma | S^k w) = \Phi_{\Lambda}(\sigma | w)
\]

Theorem: If \( \mathcal{F} = \{ \Phi_{\Lambda}(\cdot, \cdot) \} \) is a translation invariant specification, then the set of translation invariant Gibbs measures for \( \mathcal{F} \) is not empty.
2.2.2. Regularity of Gibbs states

If the interaction $\Psi$ is as above, then the corresponding specification $\Gamma = \{\gamma_\Lambda(\cdot|\cdot)\}$ has the following important regularity properties:

(i) $\Gamma$ is uniformly non-null or finite energy: the probability kernels $\gamma_\Lambda(\cdot|\cdot)$ are uniformly bounded away from 0 and 1.

(ii) $\Gamma$ is quasi-local: for any $\Lambda$, the kernels are continuous in product topology: i.e., as $V \nearrow \mathbb{Z}^d$

$$
\sup_{\omega, \eta, \zeta, \chi} \left| \gamma_\Lambda(\omega|\eta \setminus \Lambda \zeta V^c) - \gamma_\Lambda(\omega|\eta \setminus \Lambda \chi V^c) \right| \to 0,
$$

in other words, changing remote spins (spins in $V^c$), has diminishing effect of the distribution of the spins in $\Lambda$.

Gibbs states and regularity

The Gibbs measures have regular (uniformly non-null & quasi-local) conditional probabilities (specifications). By the Kozlov-Sullivan theorem [3, 4] the opposite is also true: any measure consistent with a regular specification is Gibbs for some interaction $\Psi$.

Gibbs probability distributions have been highly useful in describing physical systems (ferromagnets, dilute gases, polymers). However, the intrinsically multi-dimensional theory of Gibbs random fields is of course applicable to one-dimensional random processes as well. To define a Gibbs random process $\{X_n\}$ we have to prescribe two-sided conditional distributions

$$
P(X_0 = a_0|X_{-\infty}^1, X_{-\infty}^+ = a_{-\infty}^+, X_1^{+\infty} = a_{1}^{+\infty}), \quad a_i \in A,
$$

i.e., conditional distribution of the present given the complete past and the complete future. Defining processes via the one-sided conditional probabilities $\mathbb{P}(X_0 = a_0|X_{-\infty}^1 = a_{-\infty}^1)$ – using only past values, is probably more familiar (Markov processes). However, two-sided descriptions have definite advantages in some applications. Moreover, all “classical” processes such as Bernoulli, Markov are Gibbs, and the class of Gibbs processes is extremely large.

The theory of Gibbs states is extremely rich [5, RuelleBook]. In particular, there is a dual possibility to characterize Gibbs states locally (via specifications) or globally (via variational principles). A variety of probabilistic techniques is available for Gibbs distributions: limit theorems, coupling techniques, cluster expansions, large deviations.
2.3. Gibbs measures in Dynamical Systems

Among all Gibbs measures on a lattice system $\Omega = \mathbb{A}^{\mathbb{Z}^d}$, a special subclass is formed by translation invariant measures. The lattice group $\mathbb{Z}^d$ acts on $\Omega$ by shifts: $(S^n\omega)_k = \omega_{k+n}$ for all $\omega, k, n \in \mathbb{Z}^d$. The Gibbs measure $P$ is invariant under this action, if $P(S^{-n}C) = P(C)$ for any event $C$ and all $n$.

Shortly after the founding work of Dobrushin, Lanford, and Ruelle, Sinai extended the notion of invariant Gibbs measures to more general phase-spaces like manifolds, i.e., spaces lacking the product structure of lattice systems, thus making the notion of Gibbs states applicable to invariant measures of Dynamical Systems. In the subsequent works, it was shown that the important class of the natural invariant measures (now known as the Sinai-Ruelle-Bowen measures) of a large class of chaotic dynamical systems are Gibbs. That sparked a great interest in these objects in the Dynamical Systems and Ergodic Theory. In Dynamical Systems literature one can find several definitions (not always equivalent) of Gibbs states. For example, the following definition due to Bowen is very popular. Suppose $f$ is a continuous homeomorphism of a compact metric space $(X, d)$, and $P$ is an $f$-invariant Borel probability measure on $X$. Then $P$ is called Gibbs for a continuous potential $\psi : X \to \mathbb{R}$, if for some $P \in \mathbb{R}$ and every sufficiently small $\epsilon > 0$, there exists $C = C(\epsilon) > 1$ such that the inequalities

$$\frac{1}{C} \leq \frac{P(\{y \in X : d(f^j(x), f^j(y)) < \epsilon \ \forall j = 0, \ldots, n-1\})}{\exp(-nP + \sum_{j=0}^{n-1} \psi(f^j(x)))} \leq C$$

hold for all $x \in X$ and every $n \in \mathbb{N}$. In other words, the set of points staying $\epsilon$-close to $x$ for $n$-units of time, has measure roughly $e^{-H_n(x)}$ with $H_n = nP - \sum_{j=0}^{n-1} \psi(f^j(x))$. Capocaccia gave the most general definition of Gibbs invariant measures of multidimensional actions (dynamical systems) on arbitrary spaces, hence extending the Dobrushin-Lanford-Ruelle (DLR) construction substantially.
A. Prerequisites in Measure Theory, Probability, and Ergodic Theory

Last modified on April 3, 2015

We start by recalling briefly basic notions and facts which will be used in subsequent chapters.

A.1. Notation

The sets of natural numbers, integers, non-negative integers, and real numbers will be denoted by $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Z}_+$, $\mathbb{R}$, respectively. The set of extended real numbers $\mathbb{R}$ is $\mathbb{R} \cup \{\pm \infty\}$. If $\Lambda$ and $V$ are some non-empty sets, by $V^\Lambda$ we will denote the space of $V$-valued functions defined on $\Lambda$:

$$V^\Lambda = \{ f : \Lambda \to V \}.$$

If $\Pi \subseteq \Lambda$ and $f \in V^\Lambda$, by $f|_\Pi$ we will denote the restriction of $f$ to $\Pi$, i.e., $f|_\Pi \in V^\Pi$. If $\Pi_1$ and $\Pi_2$ are disjoint, $f \in V^{\Pi_1}$, and $g \in V^{\Pi_2}$, by $f|_{\Pi_1} g|_{\Pi_2}$ we will denote the unique element $h \in V^{\Pi_1 \cup \Pi_2}$ such that $h|_{\Pi_1} = f$ and $h|_{\Pi_2} = g$.

A.2. Measure theory

A.2.1. Suppose $\Omega$ is some non-empty set, a collection $\mathcal{A}$ of subsets of $\Omega$ is called a $\sigma$-algebra if it satisfies the following properties:

(i) $\Omega \in \mathcal{A}$;

(ii) if $A \in \mathcal{A}$, then $A^c = \Omega \setminus A \in \mathcal{A}$;

(iii) for any sequence $\{A_n\}_{n \in \mathbb{N}}$ where $A_n \in \mathcal{A}$ for every $n \in \mathbb{N}$, then $\bigcup_n A_n \in \mathcal{A}$.

Equivalently, $\sigma$-algebra $\mathcal{A}$ is a collection of subsets of $\Omega$, closed under countable operations like intersection, union, complement, symmetric difference. An element of $\mathcal{A}$ is called a measurable subset of $\Omega$. An intersection of two $\sigma$-algebras is again a $\sigma$-algebra. This property allows us, for a given collection of sets $\mathcal{C}$ of $\Omega$, to define uniquely the minimal $\sigma$-algebra $\sigma(\mathcal{C})$ containing $\mathcal{C}$, as an intersection of all $\sigma$-algebras containing $\mathcal{C}$.
A. Prerequisites

A pair \((\Omega, \mathcal{A})\), where \(\mathcal{A}\) is some \(\sigma\)-algebra of subsets of \(\Omega\), is called a measurable space. If the space \(\Omega\) is metrizable (e.g., \(\Omega = [0, 1]\) or \(\mathbb{R}\)), one can define the Borel \(\sigma\)-algebra of \(\Omega\), denoted by \(\mathcal{B}(\Omega)\), as the minimal \(\sigma\)-algebra containing all open subsets of \(\Omega\).

A.2.2. Suppose \((\Omega_1, \mathcal{A}_1)\) and \((\Omega_2, \mathcal{A}_2)\) are two measurable spaces. The map \(T : \Omega_1 \to \Omega_2\) is called \((\mathcal{A}_1, \mathcal{A}_2)\)-measurable if for any \(A_2 \in \mathcal{A}_2\), the full preimage of \(A_2\) is a measurable subset of \(\Omega_1\), i.e.,

\[
T^{-1}A_2 = \{\omega \in \Omega_1 : T(\omega) \in A_2\} \in \mathcal{A}_1.
\]

By definition, a random variable is a measurable map from \((\Omega_1, \mathcal{A}_1)\) into \((\mathbb{R}, \mathcal{B}(\mathbb{R}))\). If \((\Omega, \mathcal{A})\) is a measurable space, and \(T : \Omega \to \Omega\) is a measurable map, then \(T\) is called a measurable transformation of \((\Omega, \mathcal{A})\) or an endomorphism of \((\Omega, \mathcal{A})\). If, furthermore, \(T\) is invertible and \(T^{-1} : \Omega \to \Omega\) is also measurable, then \(T\) is called a measurable isomorphism of \((\Omega, \mathcal{A})\).

A.2.3. Suppose \((\Omega, \mathcal{A})\) is a measurable space. The function \(\mu : \mathcal{A} \to [0, +\infty]\) is called a measure if

- \(\mu(\emptyset) = 0\);
- for any countable collection of pairwise disjoint sets \(\{A_n\}_n, A_n \in \mathcal{A}\), one has

\[
\mu\left(\bigcup_n A_n\right) = \sum_n \mu(A_n).
\]

The triple \((\Omega, \mathcal{A}, \mu)\) is called a measure space. If \(\mu(\Omega) = 1\), then \(\mu\) is called a probability measure, and \((\Omega, \mathcal{A}, \mu)\) a probability (measure) space.

A.2.4. Suppose \((\Omega, \mathcal{A}, \mu)\) is called a measure space. Then for any \(A \in \mathcal{A}\), the indicator function of \(A\) is

\[
\mathbb{I}_A(\omega) = \begin{cases} 
1, & \omega \in A, \\
0, & \omega \notin A.
\end{cases}
\]

The Lebesgue integral of \(f = \mathbb{I}_A\) is defined as

\[
\int \mathbb{I}_A \, d\mu = \mu(A).
\]

A function \(f\) is called simple if there exist \(K \in \mathbb{N}\), measurable sets \(\{A_k\}_{k=1}^K\) with \(A_k \in \mathcal{A}\), and non-negative numbers \(\{\alpha_k\}_{k=1}^K\) such that

\[
f(\omega) = \sum_{k=1}^K \alpha_k \mathbb{I}_{A_k}(\omega).
\]

The Lebesgue integral of the simple function \(f\) is defined as

\[
\int f \, d\mu = \sum_{k=1}^K \alpha_k \mu(A_k).
\]
A.2. Measure theory

Suppose \( \{f_n\} \) is a monotonically increasing sequence of simple functions and \( f = \lim_n f_n \). Then the Lebesgue integral of \( f \) is defined as

\[
\int f \, d\mu = \lim_n \int f_n \, d\mu.
\]

It turns out that every non-negative measurable function \( f : \Omega \to \mathbb{R}_+ \) can be represented as a limit of increasing sequence of simple functions.

For a measurable function \( f : \Omega \to \mathbb{R} \), let \( f_+ \) and \( f_- \) be the positive and the negative part of \( f \), respectively, i.e. \( f = f_+ - f_- \). The function \( f \) is called Lebesgue integrable if

\[
\int f_+ \, d\mu, \quad \int f_- \, d\mu < +\infty
\]

and the Lebegsue integral of \( f \) is then defined as

\[
\int f \, d\mu = \int f_+ \, d\mu - \int f_- \, d\mu.
\]

The set of all Lebesgue integrable function of \((\Omega, \mathcal{A}, \mu)\), will be denoted by \( L^1(\Omega, \mathcal{A}, \mu) \), or \( L^1(\Omega) \) and \( L^1(\mu) \), if the latter cases do not lead to confusion.

If \((\Omega, \mathcal{A}, \mu)\) is a probability space, we will sometimes denote the Lebesgue integral \( \int f \, d\mu \) by \( \mathbb{E}_\mu f \) or \( \mathbb{E} f \).

A.2.5. Conditional expectation with respect to a sub-\(\sigma\)-algebra. Suppose \((\Omega, \mathcal{A}, \mu)\) is a probability space, and \(\mathcal{F}\) is some sub-\(\sigma\)-algebra of \(\mathcal{A}\). Suppose \( f \in L^1(\Omega, \mathcal{A}, \mu) \) is a \(\mathcal{A}\)-measurable Lebesgue integrable function. The conditional expectation of \( f \) given \(\mathcal{F}\) will be denoted by \( \mathbb{E}(f|\mathcal{F}) \), and is by definition an \(\mathcal{F}\)-measurable function on \(\Omega\) such that

\[
\int_C \mathbb{E}(f|\mathcal{F}) \, d\mu = \int_C f \, d\mu
\]

for every \( C \in \mathcal{F} \).


Suppose \((\Omega, \mathcal{A}, \mu)\) is a measure space, and \( \{\mathcal{A}_n\} \) is a sequence of sub-\(\sigma\)-algebras of \(\mathcal{A}\) such that \( \mathcal{A}_n \subseteq \mathcal{A}_{n+1} \) for all \( n \). Denote by \( \mathcal{A}_\infty \) the minimal \(\sigma\)-algebra containing all \(\mathcal{A}_n\). Then for any \( f \in L^1(\Omega, \mathcal{A}, \mu) \)

\[
\mathbb{E}_\mu(f|\mathcal{A}_n) \to \mathbb{E}_\mu(f|\mathcal{A}_\infty)
\]

\(\mu\)-almost surely and in \( L^1(\Omega) \).

A.2.7. Absolutely continuous measures and the Radon-Nikodym Theorem. Suppose \( \nu \) and \( \mu \) are two measures on the measurable space \((\Omega, \mathcal{A})\). The measure \( \nu \) is absolutely continuous with respect to \( \mu \) (denoted by \( \nu \ll \mu \)) if

\[
\nu(A) = 0 \quad \text{for all } A \in \mathcal{A} \text{ such that } \mu(A) = 0.
\]
A. Prerequisites

The Radon-Nikodym theorem states that if $\mu$ and $\nu$ are two $\sigma$-finite measures on $(\Omega, \mathcal{A})$, and $\nu \ll \mu$, then there exists a non-negative measurable function $f$, called the (Radon-Nikodym) density of $\nu$ with respect to $\mu$, such that

$$\nu(A) = \int_A f d\mu, \quad \text{for all } A \in \mathcal{A}.$$  

A.3. Stochastic processes

Suppose $(\Omega, \mathcal{A}, \mu)$ a probability space. A stochastic process $\{X_n\}$ is a collection of random variables $X_n: \Omega \to \mathbb{R}$, indexed by $n \in \mathcal{T}$, where the time is $\mathcal{T} = \mathbb{Z}_+ \text{ or } \mathbb{Z}$.

Stochastic process can be described by the finite-dimensional distributions (marginals): for every $(n_1, \ldots, n_k) \in \mathcal{T}^k$ and Borel sets $A_{n_1}, \ldots, A_{n_k} \subseteq \mathbb{R}$, put

$$\mu_{n_1, \ldots, n_k}(A_{n_1}, \ldots, A_{n_k}) := \mu(\{ \omega \in \Omega : X_{n_1}(\omega) \in A_{n_1}, \ldots, X_{n_k}(\omega) \in A_{n_k}\}).$$

In the opposite direction, a consistent family of finite-dimensional distributions can be used to define a stochastic process.

Process $\{X_n\}$ is called stationary if for all $(n_1, \ldots, n_k) \in \mathcal{T}^k$, $t \in \mathbb{Z}_+$, and Borel sets $A_{n_1}, \ldots, A_{n_k} \subseteq \mathbb{R}$ one has

$$\mu_{n_1+t, \ldots, n_k+t}(A_{n_1}, \ldots, A_{n_k}) = \mu_{n_1, \ldots, n_k}(A_{n_1}, \ldots, A_{n_k})$$

i.e., the time-shift does not affect the finite-dimensional marginal distributions.

A.4. Ergodic theory

Ergodic Theory originates from the Boltzmann-Maxwell ergodic hypothesis and is the study of measure preserving dynamical systems

$$(\Omega, \mathcal{A}, \mu, T)$$

where

- $(\Omega, \mathcal{A}, \mu)$ is a (Lebesgue) probability space
- $T: \Omega \to \Omega$ is measure preserving: for all $A \in \mathcal{A}$

$$\mu(T^{-1}A) = \mu(\{x \in X : T(x) \in A\}) = \mu(A).$$

Example 1. Let $\Omega = \mathbb{T}^1 = \mathbb{R}/\mathbb{Z} \cong [0, 1)$, $\mu$ = Lebesgue measure,
A.5. Entropy

- Circle rotation: \( T_\alpha(x) = x + \alpha \mod 1 \),
- Doubling map: \( T(x) = 2x \mod 1 \).

**Example 2.** Consider a finite set (alphabet) \( A = \{1, \ldots, N\} \), and put
\[
\Omega = A^{\mathbb{Z}^+} = \left\{ \omega = (\omega_n)_{n \geq 0} : \omega_n \in A \right\}.
\]
The corresponding Borel \( \sigma \)-algebra \( \mathcal{A} \) is generated by cylinder sets:
\[
[a_{i_1}, \ldots, a_{i_n}] = \left\{ \omega : \omega_{i_1} = a_{i_1}, \ldots, \omega_{i_n} = a_{i_n} \right\}, \quad a_{i_k} \in A.
\]
A measure \( p = (p(1), \ldots, p(N)) \) on \( A \) can be extended to a measure \( \mu = p^{\mathbb{Z}^+} \) on \( (\Omega, \mathcal{A}) \)
\[
\mu \left( \left\{ \omega : \omega_{i_1} = a_{i_1}, \ldots, \omega_{i_n} = a_{i_n} \right\} \right) = p(a_{i_1}) \cdots p(a_{i_n}).
\]
The measure \( \mu \) is clearly preserved by the **left shift** \( \sigma : \Omega \to \Omega \)
\[
(\sigma \omega)_n = \omega_{n+1} \quad \forall n.
\]
Quadruple \( (\Omega, \mathcal{A}, \mu, \sigma) \) is called the **p-Bernoulli shift**.

**Definition A.1.** Measure-preserving dynamical system \( (\Omega, \mathcal{A}, \mu, T) \) is **ergodic** if every invariant set is trivial:
\[
A = T^{-1}A \quad \Rightarrow \quad \mu(A) = 0 \text{ or } 1,
\]
equivalently, if every invariant function is constant:
\[
f(\omega) = f(T(\omega)) \quad (\mu - \text{a.e.}) \quad \Rightarrow \quad f(\omega) = \text{const} \quad (\mu - \text{a.e.}).
\]

**Theorem A.2 (Birkhoff’s Pointwise Ergodic Theorem).** Suppose \( (\Omega, \mathcal{A}, \mu, T) \) is an ergodic measure-preserving dynamical system. Then for all \( f \in L^1(\Omega, \mu) \)
\[
\frac{1}{n} \sum_{t=0}^{n-1} f(T^t(x)) \to \int f(x) \mu(dx) \quad \text{as } n \to \infty
\]
\( \mu \)-almost surely and in \( L^1 \).

### A.5. Entropy

Suppose \( p = (p_1, \ldots, p_N) \) is a probability vector, i.e.,
\[
p_i \geq 0, \quad \sum_{i=1}^{N} p_i = 1.
\]
The entropy of \( p \) is
\[
H(p) = -\sum_{i=1}^{N} p_i \log_2 p_i.
\]
A. Prerequisites

A.5.1. Shannon’s entropy rate per symbol

**Definition A.3.** Suppose $Y = \{Y_k\}$ is a stationary process with values in a finite alphabet $A$, and $\mu$ is the corresponding translation invariant measure. Fix $n \in \mathbb{N}$, and consider the distribution of $n$-tuples $(Y_0, \ldots, Y_{n-1}) \in A^n$. Denote by $H_n$ entropy of this distribution:

$$H_n = H(Y_{n-1}^0) = -\sum_{(a_0, \ldots, a_{n-1}) \in A^n} \mathbb{P}[Y_{n-1}^0 = a_{n-1}] \log_2 \mathbb{P}[Y_{n-1}^0 = a_{n-1}].$$

The entropy (rate) of the process $Y = \{Y_k\}$, equivalently, of the measure $\mathbb{P}$, denoted by $h(Y)$, $h(\mathbb{P})$, or $h_\sigma(\mathbb{P})$, is defined as

$$h(Y) = h(\mathbb{P}) = h_\sigma(\mathbb{P}) = \lim_{n \to \infty} \frac{1}{n} H_n$$

(the limit exists!).

Similarly, the entropy can be defined for stationary random fields $\{Y_n\}_{n \in \mathbb{Z}^d}$, $Y_n \in A$. Let $\Lambda_n = [0, n-1]^d \cap \mathbb{Z}^d$. Then

$$h(Y) = \lim_{n \to \infty} \frac{1}{|\Lambda_n|} \sum_{a_{\Lambda_n} \in A_{\Lambda_n}} \mathbb{P}[Y_{\Lambda_n} = a_{\Lambda_n}] \log_2 \mathbb{P}[Y_{\Lambda_n} = a_{\Lambda_n}],$$

again one can easily show that the limit exists.

A.5.2. Kolmogorov-Sinai entropy of measure-preserving systems

Suppose $(\Omega, \mathcal{A}, \mu, T)$ is a measure-preserving dynamical system. Suppose

$$\mathcal{C} = \{C_1, \ldots, C_N\}$$

is a finite measurable partition of $\Omega$, i.e., a partition of $\Omega$ into measurable sets $C_k \in \mathcal{A}$. For every $\omega \in \Omega$ and $n \in \mathbb{Z}_+$ (or $\mathbb{Z}$), put

$$Y_n(x) = Y_n^\mathcal{C}(\omega) = j \in \{1, \ldots, N\} \iff T^n(\omega) \in C_j.$$

**Proposition:** For any $\mathcal{C}$, the corresponding process $\mathcal{Y}^\mathcal{C} = \{Y_n\}$, $Y_n : \Omega \to \{1, \ldots, N\}$, is a stationary process with

$$\mathbb{P}[Y_0 = j_0, \ldots, Y_n = j_n] = \mu \left( \{ \omega \in \Omega : \omega \in C_{j_0}, \ldots, T^n(x) \in C_{j_n} \} \right).$$

**Definition A.4.** If $(\Omega, \mathcal{A}, \mu, T)$ is a measure preserving dynamical system, and $\mathcal{C} = \{C_1, \ldots, C_N\}$ is a finite measurable partition of $\Omega$, then the entropy of $(\Omega, \mathcal{A}, \mu, T)$ with
respect to $\mathcal{C}$ is defined as the Shannon entropy of the corresponding symbolic process $Y^C$:

$$h_\mu(T, \mathcal{C}) = h(Y^C).$$

Finally, the measure-theoretic or the **Kolmogorov-Sinai entropy** of $(\Omega, \mathcal{A}, \mu, T)$ is defined as

$$h_\mu(T) = \sup_{\mathcal{C} \text{ is finite}} h_\mu(T, \mathcal{C}).$$

The following theorem of Sinai eliminates the need to consider all finite partitions.

**Definition A.5.** A partition $\mathcal{C}$ is called **generating** (or, a generator) for the dynamical system $(\Omega, \mathcal{A}, \mu, T)$ if the smallest $\sigma$-algebra containing sets $T^n(C_j), j = 1, \ldots, N, n \in \mathbb{Z}$, is $\mathcal{A}$.

**Theorem A.6 (Ya. Sinai).** If $\mathcal{C}$ is a generating partition then

$$h_\mu(T) = h_\mu(T, \mathcal{C}).$$
Bibliography


