Information Theory with Applications, Math6397
Lecture Notes from August 26, 2014

taken by Bernhard G. Bodmann

0 Course Info

0.1 Syllabus

Instructor: Bernhard Bodmann, bgb@math.uh.edu

Office: PGH 604, (ph) 713 743 3581

Hours: Mo, We 1:30-2:30pm

Texts: A. I. Khinchin, Mathematical Foundations of Information Theory, Dover, 2001, reprint of 1957 edition (approx. \$10); optional texts: T. S. Han and K. Kobayashi, Mathematics of Information and Coding, Translations of mathematical monographs, v. 203, American Mathematical Society, 2002 (approx. \$80 for AMS members); I. Csiszár and J. Körner, Information Theory, 2nd edition, Cambridge University Press, Cambridge, 2011 (approx. \$100).

Homework and grade: Course notes taken by students in LaTeX, up to 4 homework sets with elementary problems, also including group projects that involve small programming tasks in Matlab.

Goal: Understand and apply principles of information theory

0.2 Background knowledge

0.2.1 Definition. A *probability space* $(\Omega, F, \mathbb{P})$ consists of a set of outcomes Ω, a σ-algebra F containing subsets of Ω which are called events, and a probability measure $\mathbb{P}$ that associates with each event of F a probability. A *random variable* X is a map $X : \Omega \to \mathbb{A} \subset \mathbb{R}$, with $\mathbb{A}$ called the *alphabet*. A stochastic process is a map $X : \Omega \times \mathbb{Z} \to \mathbb{A} \subset \mathbb{R}$, where the second argument is often thought of as a (discrete) time. A random variable or a stochastic process induce a probability measure on subsets of the alphabet or subsets of sequences from the alphabet. This induced measure is written as $\mathbb{P}_X$.

The shift operator τ applied to a sequence of outcomes acts by $\tau(\dots, \omega_{-2}, \omega_{-1}, \omega_0, \omega_1, \dots) = (\dots, \omega_{-1}, \omega_0, \omega_1, \omega_2, \dots)$ and it applies to an event $\mathcal{A}$ by $\tau(\mathcal{A}) = \{\tau(\omega) : \omega \in \mathcal{A}\}$. An event $\mathcal{A}$

is called *shift invariant* if $\tau(\mathcal{A}) = \mathcal{A}$. A process X is called *stationary* if $\mathbb{P}_X(\mathcal{A}) = \mathbb{P}_X(\tau(\mathcal{A}))$ for all events $\mathcal{A}$ in the σ-algebra. A process X is called *ergodic* if each $\mathcal{A}$ for which $\tau(\mathcal{A}) = \mathcal{A}$ has probability $\mathbb{P}_X(\mathcal{A}) \in \{0, 1\}$.

Ergodicity

Later, we will see that ergodic processes have a nice property that relates averages over the probability space to averages over all shifts of one outcome, Birkhoff's ergodic theorem.

Convergence

When considering sequences of random variables, we distinguish pointwise convergence, almost-sure convergence (with probability one), and convergence in distribution. We will recall the weak and the strong laws of large numbers, and the central limit theorem.

Inequalities

Among the inequalities used in this course are the Hölder, Minkowski, Jensen and Chebyshev inequalities.

1 Basics of Information Theory

A brief history

Information theory is the science related to the storage and transmission of data. Many outstanding researchers contributed to this field over the years.

- Shannon (1948)
  - channel coding (reliable transmissions)
  - source coding (compression)
- Huffman (1952) compression
- Kolmogorov (1965) and Chaitin (1966) complexity and algorithmic information theory
- Amari (1985) geometric formulation of information theory
- Slepian and Wolf (1973) correlated data streams
- Han and Kobayashi (1980's) multiterminal information systems (internet)
- Holevo (1973) quantum transmissions

1.1 Entropy

In Shannon's words, information is "anything previously uncertain". A quantitative measure for uncertainty, a lack of knowledge, is entropy.

1.1.1 Definition. Let $(\Omega, F, \mathbb{P})$ be a probability space. Given a random variable $X : \Omega \to \mathbb{A}$, whose alphabet $\mathbb{A}$ is at most countable, and the induced probability measure $\mathbb{P}_X$ on $\mathbb{A}$, we write

$$H(X) \equiv H(\mathbb{P}_X) \equiv -\sum_{a\in\mathbb{A}} \mathbb{P}_X(a) \log \mathbb{P}_X(a)$$

for the entropy of X, with the convention $0 \log 0 = 0$.

Entropy is said to measure the uncertainty inherent in $\mathbb{P}_X$. Usually, we will choose the natural logarithm, which corresponds to measuring information in nats, as opposed to the binary logarithm, which measures information in bits.

We compile elementary properties of H.

- $H(X) \geq 0$, because for each $a \in \mathbb{A}$, $0 \leq \mathbb{P}_X(a) \leq 1$ and $-t \ln t \geq 0$ for all $t \in [0, 1]$.
- $H(X) = 0$ is equivalent to the existence of some $a \in \mathbb{A}$ such that $\mathbb{P}_X(a) = 1$, because $t \ln t = 0$ if and only if $t \in \{0, 1\}$, but then one and only one outcome can have probability one, because $\sum_{a\in\mathbb{A}} \mathbb{P}_X(a) = 1$.
- For all $\lambda \in [0, 1]$, X, Y $\mathbb{A}$-valued random variables,

  $$H(\lambda\mathbb{P}_X + (1 - \lambda)\mathbb{P}_Y) \geq \lambda H(\mathbb{P}_X) + (1 - \lambda)H(\mathbb{P}_Y) .$$

  This is because $f : t \mapsto -t \ln t$ is concave on $[0, \infty)$, so $f(\lambda p_1 + (1 - \lambda)p_2) \geq \lambda f(p_1) + (1 - \lambda)f(p_2)$. Inserting this for each term in the expression for $H(\lambda\mathbb{P}_X + (1 - \lambda)\mathbb{P}_Y)$ gives

  $$-\sum_{a\in\mathbb{A}}(\lambda\mathbb{P}_X + (1 - \lambda)\mathbb{P}_Y) \ln(\lambda\mathbb{P}_X + (1 - \lambda)\mathbb{P}_Y) \geq -\sum_{a\in\mathbb{A}}(\lambda\mathbb{P}_X \ln \mathbb{P}_X + (1 - \lambda)\mathbb{P}_Y \ln \mathbb{P}_Y)$$

  and after splitting the sum and factoring out λ or $1 - \lambda$, the desired inequality emerges.
- $H(X)$ can be infinite for some X, e.g. for $\mathbb{P}_X(a) = \frac{c}{a(\ln a)^2}$ where c is chosen so that $\sum_{a\in\mathbb{A}} \mathbb{P}_X(a) = 1$. Showing this is an exercise with the integral comparison criterion.

The third property means that if we randomly select among two sources of information, with probabilites λ and $1 - \lambda$, then the entropy of the resulting distribution is at least as big as the weighted average of the individual entropies: Mixing can only create entropy.

1.1.1 Binary entropy

In the simplest case, $X : \Omega \to \{0, 1\}$, $\mathbb{P}_X(0) = p$, $0 \leq p \leq 1$, and then

$$H(X) = -p \ln p - (1 - p) \ln(1 - p) .$$

We see that this is symmetric with respect to reflections about $p = 1/2$ and has its maximum at $p = 1/2$.

1.1.2 Entropy of joint distributions

1.1.2 Definition. Let $(\Omega, F, \mathbb{P})$ be a probability space. Given two random variables $X : \Omega \to \mathbb{A}$ and $Y : \Omega \to \mathbb{B}$, we denote by $\mathbb{P}_{X,Y}$ the probability measure for their joint distribution, $\mathbb{P}_{X,Y}(a,b) = \mathbb{P}(X = a$ and $Y = b)$. Later, we use a similar notation for more than two random variables. We write

$$H(X,Y) = -\sum_{a\in\mathbb{A}, b\in\mathbb{B}} \mathbb{P}_{X,Y}(a,b) \log \mathbb{P}_{X,Y}(a,b) \,.$$

The conditional probability of Y given $X = a$ is

$$\mathbb{W}(b|a) = \begin{cases} P_{X,Y}(a,b)/\mathbb{P}(X=a), & \text{if } \mathbb{P}(X=a) \neq 0 \\ 0, & \text{else} \end{cases}$$

which relates to the conditional probability by $\mathbb{P}_{X,Y}(a,b) = \mathbb{P}(X = a)\mathbb{W}(b|a)\,, a \in \mathbb{A}, b \in \mathbb{B}$.

1.1.3 Definition. For $a \in \mathbb{A}$, the entropy of $\mathbb{W}(\cdot|A)$ is written as

$$H(Y|a) = -\sum_{b\in\mathbb{B}} \mathbb{W}(b|a) \ln \mathbb{W}(b|a) \,.$$

1.1.4 Question. What do we expect from a notion of conditional entropy? Could $H(Y|a)$ serve this role?