This two-quarter course will be an thorough introduction to information theory.

Information Theory I: entropy, mutual information, asymptotic equipartition properties, data compression to the entropy limit (source coding theorem), Huffman, Lempel-Ziv, convolutional codes, communication at the channel capacity limit (channel coding theorem), method of types, differential entropy, maximum entropy.

Information Theory II (EE515, Spring 2012): ECC, turbo, LDPC and other codes, Kolmogorov complexity, spectral estimation, rate-distortion theory, alternating minimization for computation of RD curve and channel capacity, more on the Gaussian channel, network information theory, information geometry, and some recent results on use of polymatroids in information theory.

Additional topics throughout will include information theory as it is applicable to pattern recognition, natural language processing, computer science and complexity, biological science, and communications.
Course Web Pages

- our web page (http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/)
- our assignment dropbox (https://canvas.uw.edu/courses/847774/assignments) which is where all homework will be due, electronically, in PDF format. No paper homework accepted.
- our discussion board (https://canvas.uw.edu/courses/847774/discussion_topics) is where you can ask questions. Please use this rather than email so that all can benefit from answers to your questions.

Prerequisites

- basic probability and statistics & convex analysis
- random processes (e.g., EE505 or a Stat 5xx class).
- Knowledge of matlab.
- The course is open to students in all UW departments.
Homework

- There will be a new problem set assignment every 1 to 2 weeks (about 6-7 problem sets for the quarter).
- You will have approximately 1 to 1.5 weeks to solve the problem set.
- Problem sets might also include matlab exercises, so you will need to have access to matlab, please let me know if you currently do not.
- The problem sets that are longer will take longer to do, so please do not wait until the night before they are due to start them.

Exams

- We will have an in-class midterm and an in-class final.
- Midterm exam date: Thursday, Oct 31st, 2013, in class.
- Final exam date/time: TBD (our current slot is Tuesday, December 10th, but we will have it one or two days later, either Wednesday Dec 11th or Thursday Dec 12th (Mark your calendars))
Grading

- Grades will be based on a combination of the final (33.3%) and midterm (33.3%) exam, and on homework (33.3%).

Our Main Text

- It should be available at the UW bookstore, or you can get it, say, via amazon.com.
- Reading assignment: Read Chapters 1 and 2.
Other Relevant Texts


- “Information Theory”, Robert Ash, Dover 1965
- “An Introduction to Information Theory”, Fazlollah M. Reza, Dover 1991
Other Relevant Texts

- “Convex Optimization”, Boyd and Vandenberghe
- “Probability & Measure” Billingsley,
- “Probability with Martingales”, Williams
- “Probability, Theory & Examples”, Durrett

On Our Lecture Slides

- Slides will (hopefully) be available by the early morning before lecture.
- Slides available on the web page, as well as a printable version.
- Updated slides with typo and bug fixes will be posted as well as (buggy) ones with hand annotations/corrections/fill-ins.
- In-class annotations are (currently) being made using GoodReader on the ipad — all PDF readers should be able to read them.
Class Road Map - IT-I

- L1 (9/26): Overview, Entropy
- L2
- L3
- L4
- L5
- L6
- L7
- L8
- L9
- L10
- L11
- L12
- L13
- L14
- L15
- L16
- L17
- L18
- L19

Finals Week: December 12th–16th.

Review

- We currently know nothing. Hence, nothing yet to review.
Cumulative Outstanding Reading

- Read chapters 1 and 2 in our book (Huang, Acero, Hon, “Spoken Language Processing”).

Homework

- We have our homework (homework 0) due next Tuesday, Oct 1st, 11:45pm electronically.
- It is not graded but it counts credit/no-credit.
Information Theory

- Information Theory is concerned with the theoretical limitations and potentials of systems that communicate. E.g., “What is the best compression or communications rate we can achieve”
- What is information? Beyond the philosophical questions that this raises, how can we mathematically quantify information in a way that is useful?

Coding Theory

- Coding Theory, (e.g., ECC) is concerned with the creation of practical encoding and decoding algorithms that can be used for communication over real-world noisy channels.
Information Theory (IT)

In this course, we stress IT and it’s application not only to communication theory, but to other fields as well. We also touch on coding theory.

IT involves or is related to many fields:

- communications theory
- cryptography
- computer science
- physics (statistical mechanics)
- mathematics (in particular, probability and statistics)
- philosophy of science
- linguistics and natural language processing
- speech recognition
- Pattern recognition and machine learning
- economics
- Biology and genetics
- psychology
- And many more

Communications Theory

- In 1948, Claude E. Shannon of Bell Labs published a paper “The Mathematical Theory of Communications”, and single-handedly created this field. (paper is on the web)
- Shannon’s work grew out of solving problems of electrical communication during WWII, but IT applies to many other fields as well. Would IT exist if WWII didn’t happen?
- Many of the results of this course were published back in that original paper. But the field has become very large, with influence on many other fields (e.g., IEEE trans. information theory, 6 times/year).
- Key idea: Communication is sending information from one place and/or time to another place and/or time over a medium that might have errors.
Communication Theory

General model of communication:

- The “decoder” is the inverse system of the “encoder” and it attempts to recover the original source signal or some “subpart” of the original source signal.
Source Information

- voice
- words
- pictures
- music, art
- Galileo space probe orbiting Jupiter
- human cells about to reproduce
- human parents about to reproduce
- sensory input of biological organism
- or any signal at all (any binary data).

Channel

- telephone line
- high frequency radio link
- space communication link
- storage (disk, tape, internet, TCP/IP, facebook, twitter), transmission through time rather than space, could be degradation due to decay
- biological organism (send message from brain to foot, or from ear to brain)
The destination of the information transmitted
- Person,
- Computer
- Disk
- Analog Radio or TV
- internet streaming audio system

some signal with time-varying frequency response, cross-talk, thermal noise, impulsive switch noise, etc.

Represents our imperfect understanding of the universe. Thus, we treat it as random, often however obeying some rules, such as that of a probability distribution..
- processing done before placing info into channel
- First stage: data reduction (keep only important bits or remove source redundancy),
- followed by redundancy insertion catered to channel.
- A code = a mechanism for representation of information of one signal by another.
- An encoding = representation of information in another form.

- exploit and then remove redundancy
- remove and fix any transmission errors
- restore the information in original form
Ex: Transmitting an Image

From: David J.C. MacKay “Information Theory, Inference, and Learning Algorithms”, 2003. Transmitting 10,000 source bits over a BSC with $f = 10\%$ using a repetition code and the majority vote algorithm. The probability of decoded bit error has fallen to about 3%; the rate has fallen to 1/3.

Ex: DNA Code

- DNA or the chromosomes within each cell encode all the info about each body
- Source = two parents
- Encoder = your imagination 😊
- Channel = biological combination, creation of haploid gametes, meiosis, mutation, and so on.
- Decoder = further mitosis creating the new child.
Ex: Morse Code

- Morse code, series of dots and dashes to represent letters
- most frequent letter sent with the shortest code, 1 dot
- Note: codewords might be prefixes of each other (e.g., “E” and “F”).
- uses only binary data (single current telegraph, size two “alphabet”), could use more (three, double current telegraph), but this is more susceptible to noise (binary in computer rather than ternary).

Ex: Human Speech

- Source = human thought, speakers brain
- Encoder = Human Vocal Tract
- Channel = air, sound pressure waves
- Noise = background noise (cocktail party effect)
- Decoder = human auditory system
- Receiver = human thought, listeners brain
When do we know the components and what do we know about them?

- Sometimes we know the code (e.g., when it is human-made)
- Other times we do not (e.g., when it is nature, speech, genetics)
- Much of pattern recognition (e.g., object recognition in a sound source, an image source, etc.) can be seen as the decoder aspects of the model (we don’t know for certain the code).
- Nor do we know how the brain does it.

Communication Theory: On Error

- How do we decrease errors in a communications system?
- Physical: use more reliable components in circuitry, broaden spectral bandwidth, use more precise and expensive electronics, increase signal power.
- All of this is more expensive.
- Question: Given a fixed imperfect analog channel and transmission equipment, can we achieve perfect communication over an imperfect communication line?
- Yes: Key is to add redundancy to signal. Ex. Speech.
- Encoder adds redundancy appropriate for channel. Decoder exploits and then removes redundancy.
Communication Theory: On Error

- Question: If you transmit information at a higher rate, does the error necessarily go up?
- Answer: Surprisingly, not always.
- Surprisingly, for a given noisy channel (where the channel has exceedingly small probability of transmitting without error) one can achieve perfect communication at a given rate.
- If that rate has not exceed a critical value, then one can increase the rate without increasing error.
- Let $R =$ rate of code (bits per channel use), and $P_e$ be probability of error. Then

\[
\text{Error Exponent } E(R) = \lim_{R \to C_0} \frac{\log P_e}{R}
\]

Communication & Information

What is information?

OED says:

1. facts provided or learned about something or someone.
2. what is conveyed or represented by a particular arrangement or sequence of things.
What is information?

Websters says:

1. the communication or reception of knowledge or intelligence
2a. knowledge obtained from investigation, study, or instruction.
2b. the attribute inherent in and communicated by one of two or more alternative sequences or arrangements of something that produce specific effects
2c. a signal or character representing data
2d. something (as a message, experimental data, or a picture) which justifies change in a construct (as a plan or theory) that represents physical or mental experience or another construct
2e. a quantitative measure of the content of information; specifically: a numerical quantity that measures the uncertainty in the outcome of an experiment to be performed

Wikipedia says:

Information in its most restricted technical sense is a message (utterance or expression) or collection of messages in an ordered sequence that consists of symbols, or it is the meaning that can be interpreted from such a message or collection of messages. Information can be recorded or transmitted. It can be recorded as signs, or conveyed as signals. Information is any kind of event that affects the state of a dynamic system. The concept has numerous other meanings in different contexts. Moreover, the concept of information is closely related to notions of constraint, communication, control, data, form, instruction, knowledge, meaning, mental stimulus, pattern, perception, representation, and especially entropy.
Information

- Oranges are 99¢/pound.
- It is cloudy in Seattle today.
- You are taking an information theory course right now.
- It is a balmy tropical climate in Seattle. As in other places in the Pacific North-West, warm, sunny days are the norm.
- Richard Dawkins will win the U.S. Presidential Election in 2016.

Poetry: “I heard an echo in a hollow place. No sound of blowing wind or drifting sand, some ancient voice was this, a captive trace of gone-by speech, of argument, demand.”

A Painting

Music Click

Such information has semantic meaning, but how do we quantify it?

IT is a formal mathematical theory, uses probability and statistics to make things mathematically precise. We don’t mean semantics, we want a “quantifiable” meaning.

Key is communication, relaying a message from someone to someone else.

I’m communicating to you now, hopefully. How much am I communicating to you? Can we quantify it?
### Mathematical model of source

- Assume a source (writer, speaker, etc.) conveys one of a number of messages.
- The message source randomly chooses one among many possible messages.
- Information conveyed by a message corresponds to how unlikely that message is. Information should be inversely related to probability in some way.
- That which is predictable conveys little or no information.
- The probability distribution of those messages determines the inherent information contained in the source, on average.
- Ex: uniform distribution, greatest choice, or uncertainty about source ⇒ greatest information gained on average.
- Ex: constant random variable, least choice, least uncertainty ⇒ least information about the source, on average.

### Entropy

- Entropy $H$ measures uncertainty or information:
- Entropy is a measure of choice, the choice that the source exercises in selecting the messages that are transmitted.
- $\text{entropy} = 0 \Rightarrow$ no choice.
- $\text{entropy} = \log N \Rightarrow$ maximum choice.
- Entropy is the uncertainty of the receiver, how much uncertainty does the receiver seeing a source have about the source.
- Entropy measures amount of information (complexity) in a source.
- The more random you are, the more choice you have.
- We'll define entropy shortly.
Are Humans Random?

- This is a random theory. But are humans random?
- Humans utilize “semantics” (whatever that is), and may convey “meaning” or “information” in a source beyond how improbable or unpredictable it is.
- Example: death. After a long bout with cancer, it is predictable, but it has extraordinary meaning.
- Information theory ignores such semantics.
- On the other hand, can we model certain human properties of a human with a random process?
- Yes. Humans (and natural organisms and signals in general) do exhibit purposeful statistical regularity.

Communication Theory

Original model of communication

General model of communication expanded:

- Can we do source coding and channel coding separately without them knowing about each other and retain optimality?
- Source Coding: shrinks source down to ultimate limit, data compression, $H$, the entropy of the source
- Channel coding: achieves ultimate transmission rate $C$, channel capacity, pushes as many bits through the channel pipe as possible without error.
On Source Coding

What makes a good code?

- Lossless codes (such as Huffman, Lempel-Ziv, bzip, bzip2, etc.), the compress to the theoretical limit of entropy, and do so without error.
- Code length (want a short code on average) What does “average” mean? We’ll see.
- Fidelity lossy (e.g., JPEG, MPEG, CDMA, TDMA). You can compress more if you accept error.
- Rate-distortion - underlying tradeoff between rate of code and the underlying distortion, but there are limits of rate for any given distortion.

5 basic questions in information theory for communication

1. how to measure information & define a unit of measure
2. how to define an info source & measure rate of info supplied by source
3. how to define a channel & rate of info trans. through a channel
4. how to study joint rate of trans. from source through channel to receiver? how to maximize rate of transfer?
5. How to study noise, & how noise limits rate of info trans. without limiting reliability.
What is entropy?

- Events $E_k$ each occur with probability $p_k$. $p_k$ indicates the likelihood of the event $E_k$ happening.
- Shannon information of event $E_k$ is $I(E_k) = \log(1/p_k)$. It indicates:
  1. a measure of surprise of finding out $E_k$. If $p_k = 1 \Rightarrow$ no surprise in finding out that $E_k$ occurred, while $p_k = 0 \Rightarrow$ infinite surprise in finding out $E_k$.
  2. A measure if information gained in finding out $E_k$ (information gained is equal to surprise). $p_k = 1 \Rightarrow$ No information is gained, while $p_k = 0 \Rightarrow$ infinite information is gained.
  3. A measure of the “uncertainty” of $E_k$ (but really unexpectedness). Unexpectedness is the thing that determines interest, or information. Also, information required to resolve this particular unexpectedness.
  4. $I(E_k) = -\log p(E_k) ==$ the self information of that event, or that message. Why is it called self-information? We’ll soon see.
- All logs are base 2 (by default), so $\log \equiv \log_2$ unless otherwise stated. $\ln$ will be natural log.

The word “Uncertainty” doesn’t really apply to the individual event or information thereof (as it is described in some texts).

- If $p(x) = 0$, we are as certain about it not happening as we are certain about $x$ happening when $p(x) = 1$.
- That’s why “surprise, unexpectedness, etc.” are better words, self information is a measure of this.
- Unexpectedness is the thing that determines interest, or information. We care about that which is unexpected.
Entropy - What’s in a name?

- In Shannon’s 1948 paper, he used the term “entropy” which came from the “disorder” in a thermodynamical system.
- The term “Entropy” came from Rudolph Clausius, 1865:

  > Since I think it is better to take the names of such quantities as these, which are important for science, from the ancient languages, so that they can be introduced without change into all the modern languages, I propose to name the magnitude $S$ the entropy of the body, from the Greek word “trope” for “transformation.” I have intentionally formed the word “entropy” so as to be as similar as possible to the word “energy” since both these quantities which are to be known by these names are so nearly related to each other in their physical significance that a certain similarity in their names seemed to me advantageous.

Entropy

- Notation: $p(x) = P_X(X = x)$. The event is \{X = x\}.
- Given random variable $X$, expected value $EX = \sum_x xp(x)$.
- Given function $g : \mathcal{X} \to \mathbb{R}$, expected value of random variable $g(X)$ is $Eg(X) = \sum_x g(x)p(x)$.
- Now take $g(x) = \log \frac{1}{p(x)}$, thus $g(x)$ is the unexpectedness of finding out event $X = x$.
- Then, take expected value of this $g$ (which is self-referential but well-defined) giving $\sum_x p(x)g(x) = \sum_x p(x)\log \frac{1}{p(x)}$. That is

  $$\sum_x p(x)g(x) = \sum_x p(x)\log \frac{1}{p(x)}$$

  (1.1)

- This is the average or expected surprise, or expected unexpectedness in a random variable $X$ and is the definition of entropy.
Definition 1.5.1 (Entropy)

Given a discrete random variable $X$ over a finite sized alphabet, the entropy of the random variable is:

$$H(X) \triangleq E \log \frac{1}{p(X)} = \sum_x p(x) \log \frac{1}{p(x)} = -\sum_x p(x) \log p(x) \quad (1.2)$$

- Entropy is in units of “bits” since logs are base 2 (units of “nats” if base $e$ logs).
- Measures the degree of uncertainty in a distribution.
- Measures the disorder or spread of a distribution.
- Measures the “choice” that a source has in choosing symbols according to the density (higher entropy means more choice).

Entropy Of Distributions

$p(x)$

Low Entropy

$p(x)$

High Entropy

$p(x)$

In Between
Entropy

- A measure of the true average “uncertainty”, which is a measure over the entire distribution.
- Remember this, entropy measures average or expected degree of uncertainty of the outcome a probability distribution.
- Measures of disorder, or spread. High entropy distributions, should be flat, more uniform, while low entropy should be few modal.
- A measure of choice that the source has in choosing elements of $E_k$.

Binary Entropy

- Binary alphabet, $X \in \{0, 1\}$ say.
- $p(X = 1) = p = 1 - p(X = 0)$.
- $H(X) = -p \log p - (1 - p) \log(1 - p) = H(p)$.
- As a function of $p$, we get:

![Graph of H(p) vs. p](image)

- Note, greatest uncertainty (value 1) when $p = 0.5$ and least uncertainty (value 0) when $p = 0$ or $p = 1$.
- Note also: concave in $p$. 
Joint Entropy

- Two random variables $X$ and $Y$ have joint entropy.

$$H(X, Y) = - \sum_x \sum_y p(x, y) \log p(x, y) = E \log \frac{1}{p(x, y)} \quad (1.3)$$

- Obvious generalizations to vectors $X_{1:N} = (X_1, X_2, \ldots, X_N)$.

$$H(X_1, \ldots, X_N) = \sum_{x_1, x_2, \ldots, x_N} p(x_1, \ldots, x_N) \log \frac{1}{p(x_1, \ldots, x_N)} \quad (1.4)$$

$$= E \log \frac{1}{p(x_1, \ldots, x_N)} \quad (1.5)$$

Conditional Entropy

- For two random variables $X, Y$ related via $p(x, y)$, knowing the event $X = x$ can change the entropy of $Y$.

- Event conditional entropy $H(Y|X = x)$

$$H(Y|X = x) = E \log \frac{1}{p(Y|X = x)} \quad (1.6)$$

$$= - \sum_y p(y|x) \log p(y|x) \quad (1.7)$$

- Averaging over all $x$, we get the conditional entropy $H(Y|X)$.

$$H(Y|X) = \sum_x p(x) H(Y|X = x) \quad (1.8)$$

$$= - \sum_x p(x) \sum_y p(y|x) \log p(y|x) \quad (1.9)$$

$$= - \sum_{x,y} p(x, y) \log p(y|x) = E \log \frac{1}{p(Y|X)} \quad (1.10)$$
Chain rule for Entropy

Proposition 1.5.2 (Chain Rule for Entropy)

\[ H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \]  

(1.11)

Proof.

\[-\log p(x, y) = -\log p(x) - \log p(y|x) \]  

(1.12)

then take expected value of both sides to get result.

Corollary 1.5.3

If \( X \perp \perp Y \) then \( H(X, Y) = H(X) + H(Y) \).

General Chain rule for Entropy

Proposition 1.5.4 (Chain Rule for Entropy)

\[ H(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} H(X_i|X_1, X_2, \ldots, X_{i-1}) \]  

(1.13)

Proof.

Use chain rule of conditional probability, i.e., that

\[ p(x_1, x_2, \ldots , x_N) = \prod_{i=1}^{N} p(x_i|x_1, \ldots , x_{i-1}) \]  

(1.14)

then

\[-\log p(x_1, x_2, \ldots, x_N) = -\sum_{i=1}^{N} \log p(x_i|x_1, x_2, \ldots, x_{i-1}) \]  

(1.15)

then take expected value of both sides to get result.
Aside: Variational Bound for Log

- Convex analysis gives variational representation

\[ \ln x = \min_{\lambda} \{ \lambda x - \ln \lambda - 1 \} \quad (1.16) \]

so for any \( \lambda \), we have

\[ \ln x \leq \lambda x - \ln \lambda - 1 \quad (1.17) \]

and with \( \lambda = 1 \), we thus get

\[ \ln x \leq x - 1 \quad (1.18) \]

Max value of (discrete) Entropy

**Proposition 1.5.5**

Let \( X \in \{ x_1, x_2, \ldots, x_n \} \). Then \( H(X) \leq \log n \), and equality is achieved iff \( p(X = x_i) = \frac{1}{n} \) for all \( i \).

**Proof.**

- Approach: show that \( H(X) - \log n \leq 0 \).

\[
H(X) - \log n = - \sum_x p(x) \log p(x) - \sum_x p(x) \log n \quad (1.19)
\]

\[
= \log_2 e \sum_x p(x) \ln \frac{1}{p(x)n} \quad (1.20)
\]

\[
\leq \log e \sum_x p(x) \left[ \frac{1}{p(x)n} - 1 \right] \quad (1.21)
\]

\[
= \log e \left[ \sum_x \frac{1}{n} - \sum_x p(x) \right] = 0 \quad (1.22)
\]
Max value of (discrete) Entropy

- Since \( \ln z = z - 1 \) when \( z = 1 \), the above becomes an equality at stationary point, i.e., when \( \frac{1}{p(x)n} = 1 \), or \( p(x) = 1/n \) the uniform distribution.
- Another way to see this is if \( p_i = 1/n \), then
\[
- \sum_i p_i \log p_i = - \sum_i \frac{1}{n} \log \frac{1}{n} = - \log \frac{1}{n} = \log n.
\]
- Can be shown that this is the only set of values with this property.
- Implications: entropy increases when the distribution becomes more uniform.
- E.g., mixing. \( \lambda p_1 + (1 - \lambda)p_2 \), we have \( H(\lambda p_1 + (1 - \lambda)p_2) \geq \lambda H(p_1) + (1 - \lambda)H(p_2) \), entropy is concave.

Shuffles

- Shuffles. \( X \) is a random variable indicating positions of cards (i.e., \( X = x \) is one set of positions).
- Let \( T \) be an independent random shuffle operation (permutation), i.e., \( T \perp \perp X \).
- Then \( H(TX) \geq H(X) \).
- Follows since
\[
H(TX) \geq H(TX|T) = H(T^{-1}TX|T) = H(X|T) = H(X)
\]
(conditioning reduces entropy, we'll learn this in next few lectures).
What if we permute the probabilities themselves?

I.e., let distribution \( p = (p_1, p_2, \ldots, p_n) \) be a discrete probability distribution and \( \sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \) be a random permutation.

Let \( p_{\sigma} = (p_{\sigma_1}, p_{\sigma_2}, \ldots, p_{\sigma_n}) \) be a permutation of the distribution.

How does \( H(p) = -\sum_i p_i \log p_i \) compare with \( H(p_{\sigma}) \)?

Same, since \( H(p) = H(p_{\sigma}) = -\sum_i p_{\sigma_i} \log p_{\sigma_i} \).

Summary so far

\[
H(X) = EI(x) = -\sum_x p(x) \log p(x) \tag{1.23}
\]

\[
H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y) \tag{1.24}
\]

\[
H(Y|X) = -\sum_{x,y} p(x, y) \log p(y|x) \tag{1.25}
\]

\[
H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{1.26}
\]

\[
0 \leq H(X) \leq \log n, \quad \text{where } n \text{ is } X\text{'s alphabet size.} \tag{1.27}
\]