Information Theory I and II

This two-quarter course will be a 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 will include applications to machine learning, artificial intelligence, natural language processing, computer science and complexity, biological science, and communications.

Prof. Jeff Bilmes
EE514a/Fall 2019/Info. Theory I – Lecture 1 - Sep 25th, 2019
Information Theory I and II

- This two-quarter course will be a 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 will include applications to machine learning, artificial intelligence, natural language processing, computer science and complexity, biological science, and communications.
This two-quarter course will be a 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.
Information Theory I and II

This two-quarter course will be a 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 will include applications to machine learning, artificial intelligence, natural language processing, computer science and complexity, biological science, and communications.
Course Web Pages

- our canvas page (https://canvas.uw.edu/courses/1319497)

[Course page details]

[Canvas page link]

[Web page link]

[Assignment dropbox link]

[Discussion board link]

[Email instruction]

[Prof. Jeff Bilmes information]
Course Web Pages

- our canvas page (https://canvas.uw.edu/courses/1319497)
- our web page (https://class.ece.uw.edu/514/bilmes/ee514_fall_2019/)
Course Web Pages

- our canvas page (https://canvas.uw.edu/courses/1319497)
- our web page (https://class.ece.uw.edu/514/bilmes/ee514_fall_2019/)
- our assignment dropbox (https://canvas.uw.edu/courses/1319497/assignments) which is where all homework will be due. It will be due exclusively electronically, in PDF format. No paper homework accepted.
Course Web Pages

- our canvas page (https://canvas.uw.edu/courses/1319497)
- our web page (https://class.ece.uw.edu/514/bilmes/ee514_fall_2019/)
- our assignment dropbox (https://canvas.uw.edu/courses/1319497/assignments) which is where all homework will be due. It will be due exclusively electronically, in PDF format. No paper homework accepted.
- our discussion board (https://canvas.uw.edu/courses/1319497/discussion_topics) is where you can ask questions. Please use this rather than email so that all can benefit from answers to your questions.
Religious Accommodations

Washington state law requires that UW develop a policy for accommodation of student absences or significant hardship due to reasons of faith or conscience, or for organized religious activities. The UW’s policy, including more information about how to request an accommodation, is available at Religious Accommodations Policy (https://registrar.washington.edu/staffandfaculty/religious-accommodations-policy/). Accommodations must be requested within the first two weeks of this course using the Religious Accommodations Request form (https://registrar.washington.edu/students/religious-accommodations-request/).
Prerequisites

- basic probability and statistics & convex analysis
- random processes (e.g., EE505 or a Stat 5xx class).
- Knowledge of python (numpy and scipy)
- 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 python exercises, so you will need to have access to python (anaconda is recommended).
- 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: Wednesday, Nov 6th, 2019, in class.
- Final exam date/time: Tuesday, December 10, 2019, 230-420 pm, MUE 154.
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 via any online bookstore.

- Reading assignment: Read Chapters 1 and 2.
Other Relevant Texts

Still Other Relevant Texts

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

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

- Slides will (mostly) be available by the early morning before lecture.
- Slides will be posted to our web page (https://class.ece.uw.edu/514/bilmes/ee514_fall_2019/).
- 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.
- Lectures slides and audio are being recorded live and will be posted to private youtube links (see our canvas page (https://canvas.uw.edu/courses/1319497)).
Class Road Map - IT-I

L1 (9/25): Overview, Communications, Information, Entropy
L2 (9/30):
L3 (10/2):
L4 (10/7):
L5 (10/9):
L6 (10/14):
L7 (10/16):
L8 (10/21):
L9 (10/23):

L10 (10/28):
L11 (10/30):
L12 (11/4):
LXX (11/6): In class midterm exam
LXX (11/11): Veterans Day holiday
L13 (11/13):
L14 (11/18):
L15 (11/20):
L16 (11/25):
L17 (11/27):
L18 (12/2):
L19 (12/4):
LXX (12/10): Final exam

Finals Week: December 9th–13th.
We currently know nothing. Hence, nothing yet to review.
Read chapters 1 and 2 in our book (Cover & Thomas, “Information Theory”).
No homework yet, but will be posted soon, watch for announcements on our canvas page (https://canvas.uw.edu/courses/1319497) (you can set it up to receive email whenever anything is posted).
Information Theory and Coding Theory

- **Information Theory** is concerned with the theoretical limitations of and potential for 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**, (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 cover IT and its application not only to communication theory but other fields as well. 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, genetics, and neurobiology
- psychology
- And many more . . .
In 1948, Claude E. Shannon of Bell Labs published a paper “The Mathematical Theory of Communications”, and single-handedly created this field. (paper can be found 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 cause errors.
General model of communication:

source -> encoder -> channel -> decoder -> receiver

noise
Source Information Possibilities

- 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 Possibilities

- 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 the representation of information of one type signal in another form.

An **encoding** = representation of information in another form using a code.
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.

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

$s$ encoder $t$ channel $r$ decoder $\hat{s}$

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 bi 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.
- Noise, random mutation.
- 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
Communication Theory

When do we know the components and what do we know about them?

- Sometimes we know the code (e.g., when it is designed by humans, e.g., Morse code)
- Other times we do not (e.g., when it is nature, speech, genetics)
- Much of machine learning (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.
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 and resource consuming.

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

\[
\log P_e \sim -C0 R
\]

Error Exponent

\[
E(R) \sim 0
\]

Prof. Jeff Bilmes
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.
Communication & Information

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
Communication & Information

What is information?

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.
Oranges are 99¢/pound.
Oranges are 99¢/pound.

It is cloudy in Seattle today.
Oranges are 99¢/pound.

It is cloudy in Seattle today.

You are taking an information theory course right now.
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.
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 2020.
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 2020.

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,”
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 2020.

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
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 2020.

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?
Need a mathematical model of a source

- Assume a source (writer, speaker, etc.) conveys one of a number of messages.
- 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 \( \Rightarrow \) greatest information gained on average.
- Ex: constant random variable, least choice, least uncertainty \( \Rightarrow \) least information about the source, on average.
Entropy

We’ll define entropy shortly, but intuitively:

- 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.
- $\text{entropy} = 0 \Rightarrow \text{no choice.}$
- $\text{entropy} = \log N \Rightarrow \text{maximum choice}$
- 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.
Are Humans Random?

- IT uses randomness to measure information. 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 (e.g., poetry, music, art).
- 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 properties of a human with a random process?
- Yes. Humans (and natural organisms and signals in general) do exhibit purposeful statistical regularity.
Original model of communication

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.
Communication Theory

Original model of communication

- Source
- Encoder
- Channel
- Decoder
- Receiver
- Noise

General model of communication expanded:

- Source
- Source encoder
- Channel encoder
- Channel decoder
- Source decoder
- Receiver
- Noise
Communication Theory

Original model of communication

Source → Encoder → Channel → Decoder → Receiver

General model of communication expanded:

Source → Source Encoder → Channel Encoder → Channel → Channel Decoder → Source Decoder → Receiver

- Can we do source coding and channel coding separately without them knowing about each other and retain optimality?
Communication Theory

Original model of communication

- source
- encoder
- channel
- decoder
- receiver

噪音

General model of communication expanded:

- source
- source coder
- channel encoder
- channel
- channel decoder
- source decoder
- receiver

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
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
On Source Coding

What makes a good code?

- Lossless codes (such as Huffman, Lempel-Ziv, bzip, bzip2, etc.), compress to the theoretical limit of entropy, and do so without error.
- Code length, we want a short code on average. What does “average” mean? We’ll see.
- Fidelity loss (e.g., JPEG, MPEG, CDMA, TDMA). Lossy compression/communication. 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.
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.
What is entropy?

- Events $E_k$ each occur with probability $p_k$. $p_k$ indicates the likelihood of the event $E_k$ happening.
- Shannon/Hartley information of event $E_k$ is $I(E_k) = \log(1/p_k)$, indicating:
  - 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 $E_k$.
  - A measure of 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.
  - 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.
  - $I(E_k) = -\log p(E_k)$ = 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.
What is entropy?

- Events $E_k$ each occur with probability $p_k$. $p_k$ indicates the likelihood of the event $E_k$ happening.
- Shannon/Hartley information of event $E_k$ is $I(E_k) = \log(1/p_k)$, indicating:
  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 $E_k$. 

All logs are base 2 (by default), so $\log \equiv \log_2$ unless otherwise stated. $\ln$ will be natural log.
What is entropy?

- Events $E_k$ each occur with probability $p_k$. $p_k$ indicates the likelihood of the event $E_k$ happening.
- Shannon/Hartley information of event $E_k$ is $I(E_k) = \log(1/p_k)$, indicating:
  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.
What is entropy?

- Events $E_k$ each occur with probability $p_k$. $p_k$ indicates the likelihood of the event $E_k$ happening.
- Shannon/Hartley information of event $E_k$ is $I(E_k) = \log(1/p_k)$, indicating:
  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 of 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.
What is entropy?

- Events $E_k$ each occur with probability $p_k$. $p_k$ indicates the likelihood of the event $E_k$ happening.
- Shannon/Hartley information of event $E_k$ is $I(E_k) = \log(1/p_k)$, indicating:
  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.
What is entropy?

- Events $E_k$ each occur with probability $p_k$. $p_k$ indicates the likelihood of the event $E_k$ happening.
- Shannon/Hartley information of event $E_k$ is $I(E_k) = \log(1/p_k)$, indicating:
  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.
What is entropy?

- Events $E_k$ each occur with probability $p_k$. $p_k$ indicates the likelihood of the event $E_k$ happening.
- Shannon/Hartley information of event $E_k$ is $I(E_k) = \log(1/p_k)$, indicating:
  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” or “unexpectedness” are better words, self information is a measure of this. I.e., $1/p_k$ is how surprised we are that $k$ happened when it does.

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.*
Uses of entropy in IT

Entropy uses:

- measure information in the communication theory model.
- Surprise of an event \( \{X = x\} \) is measured as \( \log \frac{1}{p(x)} \), and there are reasons for using \( \log \). Entropy is the average surprise.
- The lower bound on min number of guesses (on average) to guess the value of a random variable.
- The minimum number of bits to compress a source.
- The optimal coding “length”, of a random source.
- The minimum description length (MDL) of a random source that can be achieved without probability of error.
(discrete) Entropy Definition

- 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 \( g : X \rightarrow \mathbb{R} \), 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 \( \sum_x p(x) \log \frac{1}{p(x)} \). That is

\[
\sum_x p(x)g(x) = \sum_x p(x) \log \frac{1}{p(x)} \quad (1.1)
\]

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

$p(x)$

$x$
Entropy Of Distributions

\[ p(x) \]

\[ x \]

Low Entropy
Entropy Of Distributions

\[ p(x) \]

Low Entropy

\[ x \]

\[ p(x) \]

High Entropy

\[ x \]
Entropy Of Distributions

$p(x)$

\[ x \]

Low Entropy

$p(x)$

\[ x \]

High Entropy
Entropy Of Distributions

Low Entropy

High Entropy

In Between
Entropy Of Distributions

Low Entropy

High Entropy

In Between

$p(x)$

$x$

$p(x)$

$x$

$p(x)$

$x$
Entropy

- A measure of the true **average uncertainty**, or **average surprise** 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.

\[
\begin{align*}
    p(X=1) &= 1 - p(X=0) \\
    H(X) &= -p \log p - (1-p) \log (1-p) = H(p) \\
    \text{Note, greatest uncertainty (value 1) when } p = 0.5 \text{ and least uncertainty (value 0) when } p = 0 \text{ or } p = 1.
\end{align*}
\]
Binary Entropy

- Binary alphabet, $X \in \{0, 1\}$ say.
- $p(X = 1) = p = 1 - p(X = 0)$.
Binary Entropy

- Binary alphabet, $X \in \{0, 1\}$ say.
- $p(X = 1) = p = 1 - p(X = 0)$.
- $H(X) = -p \log p - (1 - p) \log(1 - p) = H(p)$. 

Note, greatest uncertainty (value 1) when $p = 0.5$ and least uncertainty (value 0) when $p = 0$ or $p = 1$.

Note also: concave in $p$. 

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:

\[
H(p) = \begin{cases} 
1 & \text{if } p = 0 \\
0 & \text{if } p = 1 \\
-2p \log p - 2(1 - p) \log (1 - p) & \text{otherwise}
\end{cases}
\]
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:

\[
H(p) = \begin{cases} 
1 & p = 0.5 \\
-\log(2p) & 0 < p < 0.5 \\
-\log(2(1-p)) & 0.5 < p < 1 \\
0 & p = 0 \text{ or } p = 1
\end{cases}
\]

Note, greatest uncertainty (value 1) when \( p = 0.5 \) and least uncertainty (value 0) when \( p = 0 \) or \( p = 1 \).
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 showing the binary entropy function $H(p)$]

- Note, greatest uncertainty (value 1) when $p = 0.5$ and least uncertainty (value 0) when $p = 0$ or $p = 1$.
- Note also: concave in $p$. 
Two random variables $X$ and $Y$ have joint entropy.

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

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

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

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

\[
H(X_1, \ldots, X_N) = \sum_{x_1, x_2, \ldots, x_N} p(x_1, \ldots, x_N) \log \frac{1}{p(x_1, \ldots, x_N)}
= E \log \frac{1}{p(x_1, \ldots, x_N)} \tag{1.4, 1.5}
\]
Conditional Entropy

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

\[
H(Y | X = x) = \mathbb{E} \log \frac{1}{p(Y | X = x)} 
\]

(1.6)
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 $H(Y|X = x)$

\[
H(Y|X = x) = E \log \frac{1}{p(Y|X = x)} \\
= - \sum_y p(y|x) \log p(y|x)
\]
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 $H(Y|X = x)$

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

  $$= - \sum_y p(y|x) \log p(y|x)$$  \hspace{1cm} (1.7)

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

  $$H(Y|X) = \sum_x p(x) H(Y|X = x)$$  \hspace{1cm} (1.8)

  $$= - \sum_x p(x) \sum_y p(y|x) \log p(y|x)$$  \hspace{1cm} (1.9)

  $$= - \sum_{x,y} p(x,y) \log p(y|x) = E \log \frac{1}{p(Y|X)}$$  \hspace{1cm} (1.10)
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 \( H(X, Y) = H(X) + H(Y) \).
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)
Proposition 1.5.4 (Chain Rule for Entropy)

\[ H(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} H(X_i|X_1, X_2, \ldots, X_{i-1}) \]  \hspace{1cm} (1.13)

Proof.

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

\[ p(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} p(x_i|x_1, \ldots, x_{i-1}) \]  \hspace{1cm} (1.14)
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

\[ \ln x \leq x - 1 \quad (1.18) \]
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 $i$. 
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 \( H(X) - \log n \leq 0 \).
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 $H(X) - \log n \leq 0$.

$$H(X) - \log n$$
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 \( H(X) - \log n \leq 0 \).

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

(1.19)
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 \( H(X) - \log n \leq 0 \).

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

\[
= \log_2 e \sum_x p(x) \ln \frac{1}{p(x)n}
\]
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 $H(X) - \log n \leq 0$.

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

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

\[
\leq \log e \sum_x p(x) \left[ \frac{1}{p(x)n} - 1 \right]
\]
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 \( H(X) - \log n \leq 0 \).

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

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

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

\[
= \log e \left[ \sum_x \frac{1}{n} - \sum_x p(x) \right]
\]
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 $H(X) - \log n \leq 0$.

$$H(X) - \log n = - \sum_x p(x) \log p(x) - \sum_x p(x) \log n$$

$$= \log_2 e \sum_x p(x) \ln \frac{1}{p(x)n}$$

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

$$= \log e \left[ \sum_x \frac{1}{n} - \sum_x p(x) \right] = 0$$
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.
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 
  \[
  - \sum_i p_i \log p_i = - \sum_i \frac{1}{n} \log \frac{1}{n} = - \log \frac{1}{n} = \log n.
  \]
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 $-\sum_i p_i \log p_i = -\sum_i \frac{1}{n} \log \frac{1}{n} = -\log \frac{1}{n} = \log n$.

- Implications: entropy increases when the distribution becomes more uniform.
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

$$-\sum_i p_i \log p_i = -\sum_i \frac{1}{n} \log \frac{1}{n} = -\log \frac{1}{n} = \log n.$$

- 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.
Permutations

- What if we permute the probabilities themselves?
Permutations

- 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.
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.
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 $H(p_\sigma)$?
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

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

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

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

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

and

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

- We defined the information in event \( \{X = x\} \) as
  
  \[
  I(\{X = x\}) = I(x) = \log \frac{1}{p(x)},
  \]

  but why log?
Why log?

- We defined the information in event $\{X = x\}$ as $I(\{X = x\}) = I(x) = \log \frac{1}{p(x)}$, but why log?

- Intuition says we want the information about an event to be inversely related to the probability, but there are many such relationships that might be useful.
Why log?

- We defined the information in event \( \{ X = x \} \) as 
  \[ I(\{ X = x \}) = I(x) = \log \frac{1}{p(x)}, \] 
  but why log?

- Intuition says we want the information about an event to be inversely related to the probability, but there are many such relationships that might be useful.

- Other possible functions include 
  \[ I(x) = p(x)^{-1/n} \] 
  for some \( n > 0 \).
Why log?

- We defined the information in event \( \{X = x\} \) as
  \[
  I(\{X = x\}) = I(x) = \log \frac{1}{p(x)},
  \]
  but why log?

- Intuition says we want the information about an event to be inversely related to the probability, but there are many such relationships that might be useful.

- Other possible functions include \( I(x) = p(x)^{-1/n} \) for some \( n > 0 \).

- Or any monotone non-decreasing non-negative concave function?
Why log?

- We defined the information in event \( \{ X = x \} \) as 
  \[ I(\{ X = x \}) = I(x) = \log \frac{1}{p(x)}, \] but why log?
- Intuition says we want the information about an event to be inversely related to the probability, but there are many such relationships that might be useful.
- Other possible functions include 
  \[ I(x) = p(x)^{-1/n} \] for some \( n > 0 \).
- Or any monotone non-decreasing non-negative concave function?
- Another example. \( I(x) = \) number of prime factors in \( \lceil 1/p(x) \rceil \)
Why log?

- We defined the information in event \(\{X = x\}\) as
  \[ I(\{X = x\}) = I(x) = \log \frac{1}{p(x)}, \]
  but why log?

- Intuition says we want the information about an event to be inversely related to the probability, but there are many such relationships that might be useful.

- Other possible functions include \(I(x) = p(x)^{-1/n}\) for some \(n > 0\).
- Or any monotone non-decreasing non-negative concave function?
- Another example. \(I(x) = \text{number of prime factors in } \lceil 1/p(x) \rceil\)
- But log, as well will see, has a number of attractions.
Why log?

- We defined the information in event \( \{X = x\} \) as 
  \[ I(\{X = x\}) = I(x) = \log \frac{1}{p(x)} \], but why log?
- Intuition says we want the information about an event to be inversely related to the probability, but there are many such relationships that might be useful.
- Other possible functions include \( I(x) = p(x)^{-1/n} \) for some \( n > 0 \).
- Or any monotone non-decreasing non-negative concave function?
- Another example. \( I(x) = \text{number of prime factors in} \left\lceil \frac{1}{p(x)} \right\rceil \)
- But log, as well will see, has a number of attractions.
- Khinchin’s uniqueness theorem: Assume a measure of information, over distributions, satisfies: (1) maximum value at uniform distribution, (2) Joint entropy of two random variables is sum of marginal plus conditional (chain rule), and (3) augmenting distribution by a zero-probability event makes no change. Only satisfying measure is entropy, up to multiplicative positive constant.
Why \(\log\)?: Khinchin’s uniqueness theorem

- For a distribution on \(n\) symbols with probabilities \(p = (p_1, p_2, \ldots, p_n)\), let \(H(p) = H(p_1, p_2, \ldots, p_n)\) be the entropy of that distribution.
Why $\log$?: Khinchin’s uniqueness theorem

- For a distribution on $n$ symbols with probabilities $p = (p_1, p_2, \ldots, p_n)$, let $H(p) = H(p_1, p_2, \ldots, p_n)$ be the entropy of that distribution.

- Consider any information measure, say $\mathcal{H}(p)$ on $p$, and consider the following three natural and desirable properties.
Why \( \log \): Khinchin’s uniqueness theorem

- For a distribution on \( n \) symbols with probabilities \( p = (p_1, p_2, \ldots, p_n) \), let \( H(p) = H(p_1, p_2, \ldots, p_n) \) be the entropy of that distribution.

- Consider any information measure, say \( \mathcal{H}(p) \) on \( p \), and consider the following three natural and desirable properties.
  1. \( \mathcal{H}(p) \) takes its largest value when \( p_i = 1/n \) for all \( i \).
Why \( \log \): Khinchin’s uniqueness theorem

- For a distribution on \( n \) symbols with probabilities \( p = (p_1, p_2, \ldots, p_n) \), let \( H(p) = H(p_1, p_2, \ldots, p_n) \) be the entropy of that distribution.

- Consider any information measure, say \( \mathcal{H}(p) \) on \( p \), and consider the following three natural and desirable properties.
  1. \( \mathcal{H}(p) \) takes its largest value when \( p_i = 1/n \) for all \( i \).
  2. If we define the conditional information as

\[
\mathcal{H}(Y|X) \triangleq \sum_x p(x) \mathcal{H}(p(y_1|x), p(y_2|x), \ldots, p(y_n|x)) \tag{1.28}
\]

\[
= \sum_x p(x) \mathcal{H}(Y|X = x), \tag{1.29}
\]

then we wish to have additivity in the following way

\[
\mathcal{H}(X, Y) = \mathcal{H}(X) + \mathcal{H}(Y|X) \tag{1.30}
\]
Why $\log$?: Khinchin’s uniqueness theorem

For a distribution on $n$ symbols with probabilities $p = (p_1, p_2, \ldots, p_n)$, let $H(p) = H(p_1, p_2, \ldots, p_n)$ be the entropy of that distribution.

Consider any information measure, say $\mathcal{H}(p)$ on $p$, and consider the following three natural and desirable properties.

1. $\mathcal{H}(p)$ takes its largest value when $p_i = 1/n$ for all $i$.

2. If we define the conditional information as

$$
\mathcal{H}(Y|X) \triangleq \sum_x p(x)\mathcal{H}(p(y_1|x), p(y_2|x), \ldots, p(y_n|x))
$$

then we wish to have additivity in the following way

$$
\mathcal{H}(X, Y) = \mathcal{H}(X) + \mathcal{H}(Y|X)
$$

3. For a distribution on $n + 1$ symbols, then if the probability of one is zero, we wish for $\mathcal{H}(p_1, p_2, \ldots, p_n, 0) = \mathcal{H}(p_1, p_2, \ldots, p_n)$
Why $\log$?: Khinchin’s uniqueness theorem

**Theorem 1.5.6 (Khinchin’s Theorem)**

If $\mathcal{H}(p_1, \ldots, p_n)$ satisfies the above 3 properties for all $n$ and for all $p$ such that $p_i \geq 0$, $\forall i$ and $\sum_i p_i = 1$ (i.e., all probability distributions), then

$$\mathcal{H}(p_1, \ldots, p_n) = -\lambda \sum_i p_i \log p_i$$

(1.31)

for $\lambda$ a positive constant.

- Thus, we get entropy for some logarithmic base.
What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).
Entropy and The Guessing Game

- What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

- Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

So we ask 5 questions 30% of the time, 4 questions 20% of the time, etc. Average number of questions asked is $(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (5, 4, 3, 2, 1)^T = 3.35$.

If we first ask about $x_1$, then $x_2$ it is better, i.e., $(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (1, 2, 3, 4, 5)^T = 2.65$.

Is this the best we can do?
Entropy and The Guessing Game

What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Consider the following fixed successive guessing strategy:
Entropy and The Guessing Game

- What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

- Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

  \[
  \begin{array}{c|ccccc}
  x & x_1 & x_2 & x_3 & x_4 & x_5 \\
  p(x) & 0.3 & 0.2 & 0.2 & 0.15 & 0.15
  \end{array}
  \]

- Consider the following fixed successive guessing strategy:  
  a. Is $X = x_5$?
Entropy and The Guessing Game

What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Consider the following fixed successive guessing strategy:  

a. Is $X = x_5$?  
b. Is $X = x_4$?
Entropy and The Guessing Game

- What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).
- Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

- Consider the following fixed successive guessing strategy: a. Is $X = x_5$? b. Is $X = x_4$? c. Is $X = x_3$?
Entropy and The Guessing Game

- What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

- Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

- Consider the following fixed successive guessing strategy: a. Is $X = x_5$? b. Is $X = x_4$? c. Is $X = x_3$? d. Is $X = x_2$?


Entropy and The Guessing Game

- What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

- Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

- Consider the following fixed successive guessing strategy:  
  a. Is $X = x_5$?  
  b. Is $X = x_4$?  
  c. Is $X = x_3$?  
  d. Is $X = x_2$?  
  e. Is $X = x_1$?

So we ask 5 questions 30% of the time, 4 questions 20% of the time, etc.

Average number of questions asked is

$$(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (5, 4, 3, 2, 1) = 3.35.$$

If we first ask about $x_1$, then $x_2$ it is better, i.e.,

$$(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (1, 2, 3, 4, 5) = 2.65.$$
Entropy and The Guessing Game

- What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).
- Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability
  \[
  \begin{array}{c|ccccc}
  x & x_1 & x_2 & x_3 & x_4 & x_5 \\
  p(x) & 0.3 & 0.2 & 0.2 & 0.15 & 0.15 \\
  \end{array}
  \]

- Consider the following fixed successive guessing strategy:  
  a. Is $X = x_5$?  
  b. Is $X = x_4$?  
  c. Is $X = x_3$?  
  d. Is $X = x_2$?  
  e. Is $X = x_1$?

- So we ask 5 questions 30% of the time, 4 questions 20% of the time, etc.
Entropy and The Guessing Game

What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Consider the following fixed successive guessing strategy:  

- a. Is $X = x_5$?  
- b. Is $X = x_4$?  
- c. Is $X = x_3$?  
- d. Is $X = x_2$?  
- e. Is $X = x_1$?

So we ask 5 questions 30% of the time, 4 questions 20% of the time, etc.

Average number of questions asked is

$$(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (5, 4, 3, 2, 1)^T = 3.35.$$
Entropy and The Guessing Game

What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq X$ (the domain of the r.v. $X$).

Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$p(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.3 0.2 0.2 0.15 0.15</td>
</tr>
</tbody>
</table>

Consider the following fixed successive guessing strategy:

a. Is $X = x_5$?
b. Is $X = x_4$?
c. Is $X = x_3$?
d. Is $X = x_2$?
e. Is $X = x_1$?

So we ask 5 questions 30% of the time, 4 questions 20% of the time, etc.

Average number of questions asked is

$$(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (5, 4, 3, 2, 1)^T = 3.35.$$ 

If we first ask about $x_1$, then $x_2$ it is better, i.e.,

$$(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (1, 2, 3, 4, 5)^T = 2.65.$$
Entropy and The Guessing Game

- What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

- Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

- Consider the following fixed successive guessing strategy: a. Is $X = x_5$? b. Is $X = x_4$? c. Is $X = x_3$? d. Is $X = x_2$? e. Is $X = x_1$?

- So we ask 5 questions 30% of the time, 4 questions 20% of the time, etc.

- Average number of questions asked is

$$(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (5, 4, 3, 2, 1)^T = 3.35.$$  

- If we first ask about $x_1$, then $x_2$ it is better, i.e.,

$$(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (1, 2, 3, 4, 5)^T = 2.65.$$  

- Is this the best we can do?
Consider the following strategy.

- $X \in \{x_2, x_3\}$
- $X \in \{x_1\}$
- $X \in \{x_4\}$
- $X \in \{x_5\}$

Average number of questions:

$2(0.2 + 0.2 + 0.3) + 3(0.15 + 0.15) = 2.3$

Note, $H(X) = 2.71$.

In general, the average number of questions is always $\geq H(X)$.
Consider the following strategy.

![Decision Tree Diagram]

- **Entropy and The Guessing Game**

- Average number of questions
  
  \[
  2(0.2 + 0.2 + 0.3) + 3(0.15 + 0.15) = 2.3
  \]
Entropy and The Guessing Game

Consider the following strategy.

Average number of questions
\[ 2(0.2 + 0.2 + 0.3) + 3(0.15 + 0.15) = 2.3 \]

Note, \( H(X) = 2.271 \)
Consider the following strategy.

Average number of questions
\[ 2(0.2 + 0.2 + 0.3) + 3(0.15 + 0.15) = 2.3 \]

Note, \( H(X) = 2.271 \)

In general, the average number of questions is always \( \geq H(X) \).
Entropy and The Guessing Game

- Difference between the best & worst first question in terms of how it splits the distribution.

- Worst: $X = x_5$? $p(X = x_5) = 0.15$, $p(X \neq x_5) = 0.85$, and the entropy $H(0.15, 0.85) = 0.6098$.

- Best: $X \in \{x_2, x_3\}$? $p(X \in \{x_2, x_3\}) = 0.4$, $p(X \notin \{x_2, x_3\}) = 0.6$, and the entropy $H(0.4, 0.6) = 0.971$.

- In general, it is better to first ask questions who, when seen as a random variable, have higher entropy.

- This will quickly reduce the remaining uncertainty.

- Note the relationship to $H(Y|X) + H(X) = H(X,Y)$. If we ask a question with large $H(X)$, the residual uncertainty $H(Y|X)$ is made smaller.
Given two random variables $X$ and $Y$, how much information do they have about each other?

If we know $X$ how much do we learn about $Y$? If we know $Y$, how much do we learn about $X$?

If they are independent, $X \perp \!\!\!\!\!\!\perp Y$, then knowing $X$ should tell us nothing about $Y$ and vice verse.

Since we now have a measure of information in a random source, $H(X)$, we can quantify how much information random variables have about each other, this is mutual information.
Given event \( \{X = x, Y = y\} \), we can ask for the information provided about event \( x \) from the fact that event \( y \) occurred.

This can be quantified as:

\[
I(x; y) = \log \frac{p(x|y)}{p(x)} = \log \frac{1}{p(x)} - \log \frac{1}{p(x|y)}
\] (1.32)

- First term: surprise that \( x \) occurred
- Second term: surprise that \( x \) occurred given that \( y \) occurred.
- Difference: difference of surprise, how much the surprise has changed between not knowing \( y \) and knowing \( y \).
- Note: \( p(x|x) = 1 \), so
  \[
  I(x; x) = \log \frac{1}{p(x)} - \log 1 = \log \frac{1}{p(x)} = I(x),
  \]
  so that \( I(x) \) can be seen as a form of “self-information”.
The mutual information (MI) is the average amount of information that r.v. $X$ has about $Y$, and vice verse.

**Definition 1.6.1 (mutual information)**

\[ I(X; Y) = E_{p(x,y)} \log \frac{p(x|y)}{p(x)} = E_{p(x,y)} \log \frac{p(x|y)p(y)}{p(x)p(y)} = E_{p(x,y)} \log \frac{p(x, y)}{p(x)p(y)} = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \]
Proposition 1.6.2

\[ I(X; Y) = H(X) - H(X|Y) \]  \hspace{1cm} (1.35)

Proof.

\[ E \log \frac{p(x|y)}{p(x)} = E \log \frac{1}{p(x)} - E \log \frac{1}{p(x|y)} = H(X) - H(X|Y) \]  \hspace{1cm} (1.36)

and the other

- By symmetry, we also have \( I(X; Y) = H(Y) - H(Y|X) \).
Proposition 1.6.2

\[ I(X; Y) = H(X) - H(X|Y) \] (1.35)

Proof.

\[ E \log \frac{p(x|y)}{p(x)} = E \log \frac{1}{p(x)} - E \log \frac{1}{p(x|y)} = H(X) - H(X|Y) \] (1.36)

and the other

- By symmetry, we also have \( I(X; Y) = H(Y) - H(Y|X) \).
- Also, since \( H(X) \geq 0 \) and \( H(X|Y) \geq 0 \), we have \( I(X; Y) \leq \min(H(X), H(Y)) \).
Recall, chain rule of entropy: \( H(X,Y) = H(X) + H(Y|X) \)

MI: \( I(X; Y) = H(X) - H(X|Y) \)

Don’t get confused with the commas vs. semicolons!!

Given the above, we have

\[
I(X; Y) = H(X) + H(Y) - H(X, Y) \tag{1.37}
\]

which is a simple instance of the inclusion-exclusion principle in combinatorics.

It is sometimes useful to visualize this relationship with pictures.
Mutual Information and Entropy - Venn Diagram

Given random variable $X$, the entropy of (uncertainty in, average surprise in, information contained within, etc.) a random variable can be displayed using a 2D area, as given above.

$H(X)$
Given random variable $X$, the entropy of (uncertainty in, average surprise in, information contained within, etc.) a random variable can be displayed using a 2D area, as given above.
Given random variable $X$, the entropy of (uncertainty in, average surprise in, information contained within, etc.) a random variable can be displayed using a 2D area, as given above.

The area of the set conveys the “degree of information”
Mutual Information and Entropy - Venn Diagram

A way of looking at the relationships.

\[ H(X, Y) \]
\[ H(X) \]
\[ H(Y) \]
\[ H(X|Y) \]
\[ H(Y|X) \]
\[ I(X;Y) \]