Information Theory
Lecture 1: Introduction & Overview

Mark Reid

Research School of Computer Science
The Australian National University

28th November, 2014
1 Course Overview
   - What is Information?
   - Motivating Examples

2 Basic Concepts
   - Probability
   - Information and Entropy
   - Joint Entropy, Conditional Entropy and Chain Rule
   - Mutual Information, Divergence
This short course is based on my COMP2610/COMP6261 course at ANU — a 26 hour, 2nd year undergraduate/Masters level course co-developed with Aditya Menon (NICTA) & Edwin Bonilla (NICTA).

The ANU version of the course studies the fundamental limits and potential of the *representation* and *transmission* of information.

- Mathematical Foundations
- Coding and Compression
- Communication
- Probabilistic Inference
- Kolmogorov Complexity
Mackay (ITILA, 2006) available online:
http://www.inference.phy.cam.ac.uk/mackay/itila

David MacKay’s Lectures:
http://www.inference.phy.cam.ac.uk/itprnn_lectures/
The History of Information Theory

James Gleick
THE INFORMATION
A History, a Theory, a Flood

&

Information Theory and the Digital Age
by Aftab, Cheung, Kim, Thakkar, and Yeddanapudi.
Uses of Information Theory

- Statistical physics (thermodynamics, quantum information theory);
- Computer science (machine learning, algorithmic complexity, resolvability);
- Probability theory (large deviations, limit theorems);
- Statistics (hypothesis testing, multi-user detection, Fisher information, estimation);
- Economics (gambling theory, investment theory);
- Biology (biological information theory);
- Cryptography (data security, watermarking);
- Networks (self-similarity, traffic regulation theory).
What Is Information? (1)

According to a dictionary definition, information can mean:

1. Facts provided or learned about something or someone: 
   *a vital piece of information.*

2. What is conveyed or represented by a particular arrangement or sequence of things: 
   *genetically transmitted information.*

In this course: information in the context of *communication*:

- Explicitly include uncertainty, modelled probabilistically
- Shannon (1948): “Amount of unexpected data a message contains”
  - A theory of information transmission
  - Source, destination, transmitter, receiver
What is Information? (2)

Fig. 1 — Schematic diagram of a general communication system.

From Shannon (1948)
Information is a message that is *uncertain* to receivers:
- If we receive something that we already knew with absolute certainty then it is non-informative.
- Uncertainty is crucial in measuring information content
- We will deal with uncertainty using probability theory

**Information Theory**

Information theory is the study of the fundamental *limits* and *potential* of the *representation* and transmission of information.
Examples
Example 1: What Number Am I Thinking of?

- I have in mind a number that is between 1 and 20
- You are allowed to ask me one question at a time
- I can only answer yes/no
- Your goal is to figure out the number as quickly as possible
- What strategy would you follow?
Example 1: What Number Am I Thinking of?

- I have in mind a number that is between 1 and 20
- You are allowed to ask me one question at a time
- I can only answer yes/no
- Your goal is to figure out the number as quickly as possible
- What strategy would you follow?

Your strategy + my answers = a code for each number

Some variants:
- What if you knew I was twice as likely to pick numbers more than 10?
- What if you knew I never chose prime numbers?
- What if you knew I only ever chose one of 7 or 13?

What is the optimal strategy/coding?
Can you read this sentence without any vowels?
Can you read this sentence without any vowels?

Written English (and other languages) has much redundancy:

- Approximately 1 bit of information per letter
- Naively there should be almost 5 bits per letter

(For the moment think of “bit” as “number of yes/no questions”)

How much redundancy can we safely remove?
(Note: “rd” could be “read”, “red”, “road”, etc.)
Example 3: Error Correction

Hmauns hvae the aitliby to cerroct for eorrrs in txet and iegmas.

How much noise is it possible to correct for and how?
Overview of ANU Course

- How can we quantify information?
  - Probability, Basic Properties
  - Entropy & Information, Results & Inequalities

- How can we make good guesses?
  - Probabilistic Inference
  - Bayes Theorem and Applications

- How much redundancy can we safely remove?
  - Compression
  - Source Coding Theorems, Kraft Inequality
  - Block, Huffman, and Lempel-Ziv Coding

- How much noise can we correct and how?
  - Noisy-Channel Coding
  - Repetition Codes, Hamming Codes

- What is randomness?
  - Kolmogorov Complexity & Algorithmic Information Theory
Overview of ANU Course

- How can we quantify information?
  - Probability, Basic Properties
  - Entropy & Information, Results & Inequalities

- How can we make good guesses?
  - Probabilistic Inference
  - Bayes Theorem and Applications

- How much redundancy can we safely remove?
  - Compression
  - Source Coding Theorems, Kraft Inequality
  - Block, Huffman, and Lempev-Ziv Coding

- How much noise can we correct and how?
  - Noisy-Channel Coding
  - Repetition Codes, Hamming Codes

- What is randomness?
  - Kolmogorov Complexity & Algorithmic Information Theory

- Applications to Machine Learning
  - Max. entropy, online learning, & more
**Overview of Short Course**

- **Day 1**: Overview & Basic Concepts
  - Definitions: Probability, Entropy, Information, Divergence
  - Basic Properties & Relationships

- **Day 2**: Inequalities & Key Results
  - Probabilistic Inequalities
  - Information Theoretic Inequalities
  - Source Coding Theorems
  - Noisy-Channel Coding Theorem

- **Day 3**: Information Theory & Machine Learning
  - Online Learning
  - Exponential Families
  - Clustering
Outline

1. Course Overview

2. Basic Concepts
   - Probability
   - Information and Entropy
   - Joint Entropy, Conditional Entropy and Chain Rule
   - Mutual Information, Divergence
Probability

Let $X, Y$ be random variables taking values in $\{x_i\}_{i=1}^N$ and $\{y_j\}_{j=1}^M$ (resp.)

Sum Rule / Marginalization:

$$p(X = x_i) = \sum_j p(X = x_i, Y = y_j)$$

Product Rule:

$$p(X = x_i, Y = y_j) = p(Y = y_j | X = x_i) \cdot p(X = x_i)$$

Bayes Rule:

$$p(Y = y | X = x) = \frac{p(X = x | Y = y) \cdot p(Y = y)}{p(X = x)}$$
An Illustration of a Distribution over Two Variables

joint

marginal

conditional
Definition: Independent Variables

Two variables $X$ and $Y$ are statistically independent, denoted $X \perp \perp Y$, if and only if their joint distribution factorizes into the product of their marginals:

$$X \perp \perp Y \iff p(X, Y) = p(X)p(Y)$$

We may also consider random variables that are conditionally independent given some other variable.

Definition: Conditionally Independent Variables

Two variables $X$ and $Y$ are conditionally independent given $Z$, denoted $X \perp \perp Y|Z$, if and only if

$$p(X, Y|Z) = p(X|Z)p(Y|Z)$$

Intuitively, $Z$ is a common cause for $X$ and $Y$. 
Say that a message comprises an answer to a single, yes/no question — e.g., Will rain tomorrow or not?

Informally, the amount of information in such a message is how unexpected or “surprising” it is.

- If you are 90% sure it will not rain tomorrow, learning that it is raining is more surprising than if you learnt it was not raining.

Information

For $X$ a random variable with outcomes in $\mathcal{X}$ and distribution $p(X)$ the information in learning $X = x$ is $h(x) = \log_2 \frac{1}{p(x)} = - \log_2 p(x)$.

The information in observing $x$ is large when $p(x)$ is small and vice versa. Rare events are more informative.
The entropy of a random variable $X$ is the average information content of its outcomes.

Let $X$ be a discrete r.v. with possible outcomes $\mathcal{X}$ and distribution $p(X)$. The entropy of $X$ — or, equivalently, $p(X)$ — is

$$H(X) = \mathbb{E}_X [h(X)] = - \sum_x p(x) \log_2 p(x)$$

where we define $0 \log 0 \equiv 0$, as $\lim_{p \to 0} p \log p = 0$.

**Example 1:** $\mathcal{X} = \{a, b, c, d\}$; $p(a) = p(b) = \frac{1}{8}$, $p(c) = \frac{1}{4}$, $p(d) = \frac{1}{2}$. Entropy $H(X) = 2 \frac{1}{8} \log_2 8 + \frac{1}{4} \log_2 4 + \frac{1}{2} \log_2 2 = 2 \frac{3}{8} + \frac{2}{4} + \frac{1}{2} = 1.75$.

**Example 2:** $\mathcal{X} = \{a, b, c, d\}$; $p(a) = p(b) = p(c) = p(d) = \frac{1}{4}$. Entropy $H(X) = 4 \frac{1}{4} \log_2 4 = 2$. 
Example 3 — Bernoulli Distribution

Let $X \in \{0, 1\}$ with $X \sim \text{Bern}(X|\theta)$: $p(X = 0) = 1 - \theta$ and $p(X = 1) = \theta$. Entropy of $X$ is $H(X) = H_2(\theta) := -\theta \log \theta - (1 - \theta) \log(1 - \theta)$.

- Minimum entropy $\rightarrow$ no uncertainty about $X$, i.e. $\theta = 1$ or $\theta = 0$
- Maximum when $\rightarrow$ complete uncertainty about $X$, i.e. $\theta = 0.5$
- For $\theta = 0.5$ (e.g. a fair coin) $H_2(X) = 1$ bit.
Property: Concavity

**Proposition**

Let \( \mathbf{p} = (p_1, \ldots, p_N) \). The function \( H(\mathbf{p}) := -\sum_{i=1}^{N} p_i \ln p_i \) is concave.

First derivative is \( \nabla H(\mathbf{p}) = - (\ln p_1 + 1, \ldots, \ln p_N + 1)^\top \) and so second derivative is \( \nabla^2 H(\mathbf{p}) = \text{diag} \left(-p_1^{-1}, \ldots, -p_N^{-1}\right) \), which is negative semi-definite so \( H(\mathbf{p}) \) is concave.

We can switch between \( \log_2 \) and \( \ln \) since for \( x > 0 \)
\[
\log_2 x = \log_2 e^{\ln x} = \ln x \cdot \log_2 e.
\]

When entropy is defined using \( \log_2 \) its base is 2 and units are bits. When entropy is defined using \( \ln \) it has base \( e \) and units of nats.
Categorical distributions with 30 different states:

- The more sharply peaked the lower the entropy
- The more evenly spread the higher the entropy
- Maximum for *uniform* distribution: \( H(X) = -\log \frac{1}{30} \approx 3.40 \text{ nats} \)

When will the entropy be minimum?
Property: Maximised by Uniform Distribution

Proposition
Let $X$ take values from $\mathcal{X} = \{1, \ldots, N\}$ with distribution $p = (p_1, \ldots, p_N)$ where $p_i = p(X = i)$. Then $H(X) \leq \log_2 N$ with equality iff $p_i = \frac{1}{N} \ \forall i$.

Sketch Proof:
Objective: $\max_p H(X) = - \sum_{i=1}^{N} p_i \log p_i \text{ s.t. } \sum_{i=1}^{N} p_i = 1$. Lagrangian:

$$\mathcal{L}(p) = - \sum_{i} p_i \log p_i + \lambda \left( \sum_{i} p_i - 1 \right). \quad (1)$$

$\nabla \mathcal{L}(p) = 0$ gives $\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_i p_i - 1 = 0$ and $\frac{\partial \mathcal{L}}{\partial p_i} = -(\log p_i + 1) + \lambda = 0$ so $\log p_i = \lambda - 1 \implies p_i = 2^{\lambda-1}$. Summing $p_i$ gives $1 = \sum_i 2^{\lambda-1} = N.2^{\lambda-1}$. Taking logs: $0 = \log_2 N + \lambda - 1$ so $p_i = 2^{- \log_2 N} = \frac{1}{N}$.

Note that $\log_2 N$ is number of bits needed to describe an outcome of $X$. 
Property: Decomposability

For a r.v. $X$ on $\mathcal{X} = \{x_1, \ldots, x_N\}$ with probability distribution $p = (p_1, \ldots, p_N)$:

$$H(X) = H(X^{(1)}) + (1 - p_1)H(X^{(2:N)})$$

$X^{(1)} \in \{0, 1\}$ indicates if $X = x_1$ or not, so:

$$p(X^{(1)} = 1) = p(X = x_1) = p_1 \text{ and } p(X^{(1)} = 0) = p(X \neq x_1) = 1 - p_1$$

$X^{(2:N)} \in \{x_2, \ldots, x_N\}$ is r.v. over outcomes except $x_1$ and

$$p(X^{(2:N)} = x) = p(X = x|X \neq x_1) = \left(\frac{p_2}{1-p_1}, \ldots, \frac{p_{|\mathcal{X}|}}{1-p_1}\right)$$
The joint entropy $H(X, Y)$ of a pair of discrete random variables with joint distribution $p(X, Y)$ is given by:

$$H(X, Y) = \mathbb{E}_{X, Y} \left[ \log \frac{1}{p(X, Y)} \right]$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x, y)}$$

*Easy to remember:* This is just the entropy $H(Z)$ for a random variable $Z = (X, Y)$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with distribution $p(Z) = p(X, Y)$. 
Joint Entropy:
Independent Random Variables

If $X$ and $Y$ are statistically independent we have that:

$$H(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x, y)}$$

$$= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x)p(y) \left[ \log p(x) + \log p(y) \right]$$

$$= - \sum_{x \in \mathcal{X}} p(x) \log p(x) \sum_{y \in \mathcal{Y}} p(y) - \sum_{y \in \mathcal{Y}} p(y) \log p(y) \sum_{x \in \mathcal{X}} p(x)$$

$$= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} + \sum_{y \in \mathcal{Y}} p(y) \log \frac{1}{p(y)}$$

$$= H(X) + H(Y)$$

Entropy is additive for independent random variables.
Also, $H(X, Y) = H(X) + H(Y)$ implies $p(X, Y) = p(X)p(Y)$. 
An Axiomatic Characterisation

Why that definition of entropy? Why not another function?

Suppose we want a measure $H(X)$ of “information” in r.v. $X$ so that

1. $H$ depends on the distribution of $X$, and not the outcomes themselves
2. The $H$ for the combination of two variables $X, Y$ is at most the sum of the corresponding $H$ values
3. The $H$ for the combination of two independent variables $X, Y$ is the sum of the corresponding $H$ values
4. Adding outcomes with probability zero does not affect $H$
5. The $H$ for an unbiased Bernoulli is 1
6. The $H$ for a Bernoulli with parameter $p$ tends to 0 as $p \to 0$

Then, the only possible choice for $H$ is

$$H(X) = - \sum_x p(x) \log_2 p(x)$$
Conditional Entropy

The conditional entropy of $Y$ given $X = x$ is the entropy of the probability distribution $p(Y|X = x)$:

$$H(Y|X = x) = \sum_{y \in Y} p(y|X = x) \log \frac{1}{p(y|X = x)}$$

The conditional entropy of $Y$ given $X$, is the average over $X$ of the conditional entropy of $Y$ given $X = x$:

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x)$$

$$= \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log \frac{1}{p(y|x)}$$

$$= \mathbb{E}_{X,Y} \left[ \frac{1}{p(Y|X)} \right]$$

Average uncertainty that remains about $Y$ when $X$ is known.
The joint entropy can be written as:

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

\[ = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \left[ \log p(x) + \log p(y|x) \right] \]

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

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

The joint uncertainty of $X$ and $Y$ is the uncertainty of $X$ plus the uncertainty of $Y$ given $X$. 
Relative Entropy

Definition

The relative entropy or Kullback-Leibler (KL) divergence between two probability distributions $p(X)$ and $q(X)$ is defined as:

$$D_{KL}(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_{p(X)} \left[ \log \frac{p(X)}{q(X)} \right].$$

- Note:
  - Both $p(X)$ and $q(X)$ are defined over the same alphabet $\mathcal{X}$

- Conventions:

  $$0 \log \frac{0}{0} \overset{\text{def}}{=} 0 \quad 0 \log \frac{0}{q} \overset{\text{def}}{=} 0 \quad p \log \frac{p}{0} \overset{\text{def}}{=} \infty$$
Relative Entropy

Properties:

- \( D_{KL}(p\|q) \geq 0 \)
- \( D_{KL}(p\|q) = 0 \iff p = q \)
- \( D_{KL}(p\|q) \neq D_{KL}(q\|p) \)

Observations:

- Not a true distance since is not symmetric and does not satisfy the triangle inequality
- Hence, “KL divergence” rather than “KL distance”
- Very important in machine learning and information theory. The “right” distance for distributions.
Mutual Information

Let $X, Y$ be two r.v. with joint $p(X, Y)$ and marginals $p(X)$ and $p(Y)$:

**Definition**

The *mutual information* $I(X; Y)$ is the relative entropy between the joint distribution $p(X, Y)$ and the product distribution $p(X)p(Y)$:

$$I(X; Y) = D_{KL}(p(X, Y) \parallel p(X)p(Y))$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

Measures “how far away” the joint distribution is from independent.

Intuitively, *how much information, on average, does $X$ convey about $Y$.*
We can re-write the definition of mutual information as:

\[
I(X; Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}
\]

\[
= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x|y)}{p(x)}
\]

\[
= - \sum_{x \in \mathcal{X}} \log p(x) \sum_{y \in \mathcal{Y}} p(x, y) - \left( - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x|y) \right)
\]

\[
= H(X) - H(X|Y)
\]

The average reduction in uncertainty of $X$ due to the knowledge of $Y$. 
Mutual Information:

Properties

- Mutual Information is non-negative:

\[ I(X; Y) \geq 0 \]

- Since \( H(X, Y) = H(X) + H(Y|X) \) we have that:

\[ I(X; Y) = H(X) + H(Y) - H(X, Y) \]

- Above is symmetric in \( X \) and \( Y \) so

\[ I(X; Y) = I(Y; X) \]

- Finally:

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

Sometimes the entropy is referred to as *self-information*
Breakdown of Joint Entropy

\[ H(X) \quad H(Y) \]

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

\[ H(X,Y) \]
The conditional mutual information between $X$ and $Y$ given $Z = z_k$:

$$I(X; Y|Z = z_k) = H(X|Z = z_k) - H(X|Y, Z = z_k).$$

Averaging over $Z$ we obtain:

The conditional mutual information between $X$ and $Y$ given $Z$:

$$I(X; Y|Z) = H(X|Z) - H(X|Y, Z)$$

$$= \mathbb{E}_{p(X,Y,Z)} \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}$$
Summary:
- Probability (Joint, Marginal, Conditional, Dependence)
- Information, Entropy (Joint, Conditional) & Properties
- Relative Entropy & (Conditional) Mutual Information

Next Time:
- Probabilistic Inequalities (Markov, Chebyshev)
- Information Theoretic Inequalities (Gibbs, Kraft, Data Processing)
- Source Coding Theorems
- Noisy-Channel Coding Theorem

Questions?