Lecture 03. Information theory

Xin Chen

These slides are based on slides from Mahdi Roozbahani

Outline

- Logistics
- Motivation
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

Logistics

- Create your team as soon as possible.
- Textbook and reading materials
- Homework 1 will come out by the end of this week.
- Attendance sheet will be posted.
- We start our office hour this week.

Recap

Outline

- Logistics
- Motivation ⟵
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

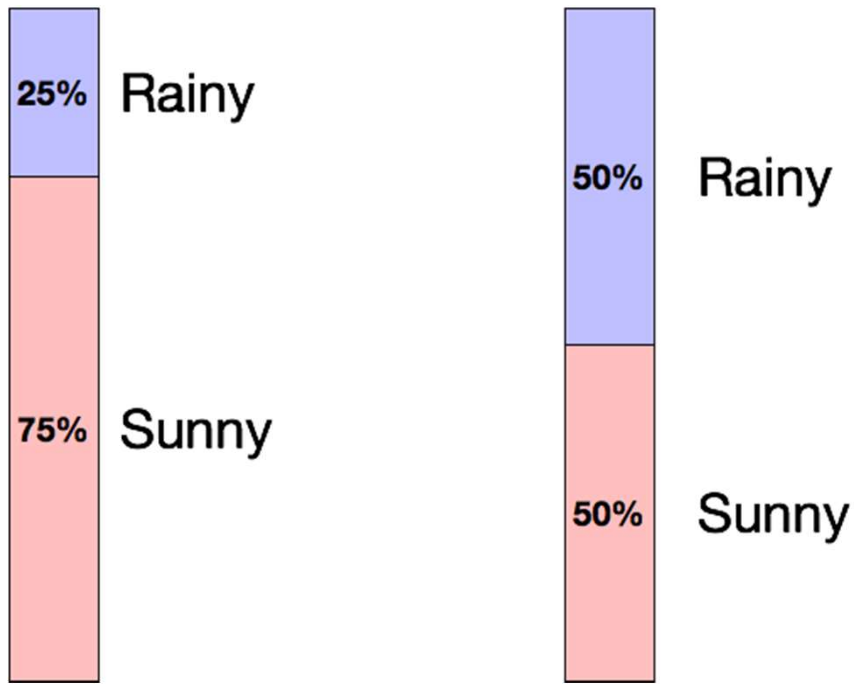
Uncertainty and Information

Information is processed data
whereas knowledge is information that is modeled to be useful.

You need **information** to be able to get **knowledge**

- information $\neq$ knowledge
  Concerned with abstract possibilities, not their meaning

Uncertainty and Information

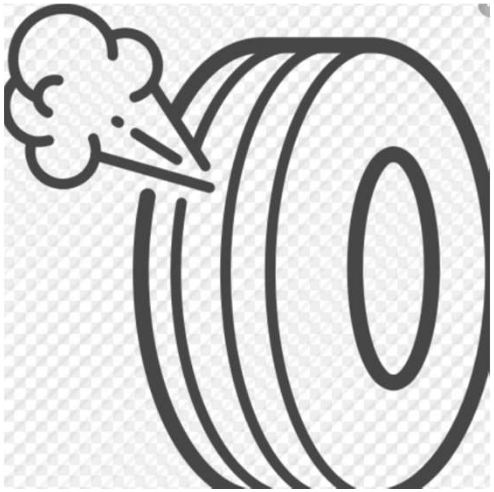


Which day is more uncertain?

How do we quantify uncertainty?

High entropy correlates to high information or the more uncertain

Physics and chemistry



How to explain these behaviors?

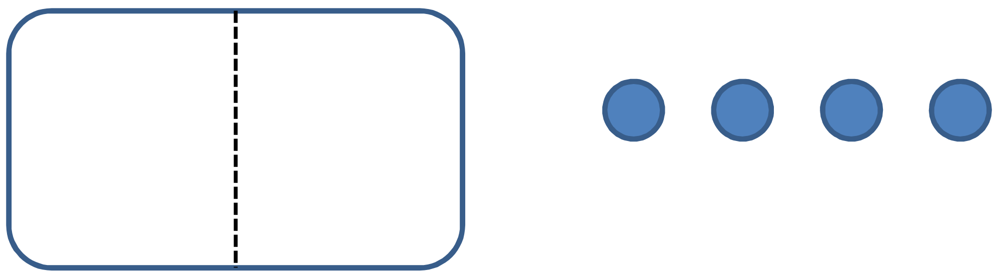
Design English Dictionary

- Each word is used in people's lives with various frequencies
  - Frequent: a, an, the
  - Infrequent: adomania, opia

- The question is how to encode these words.

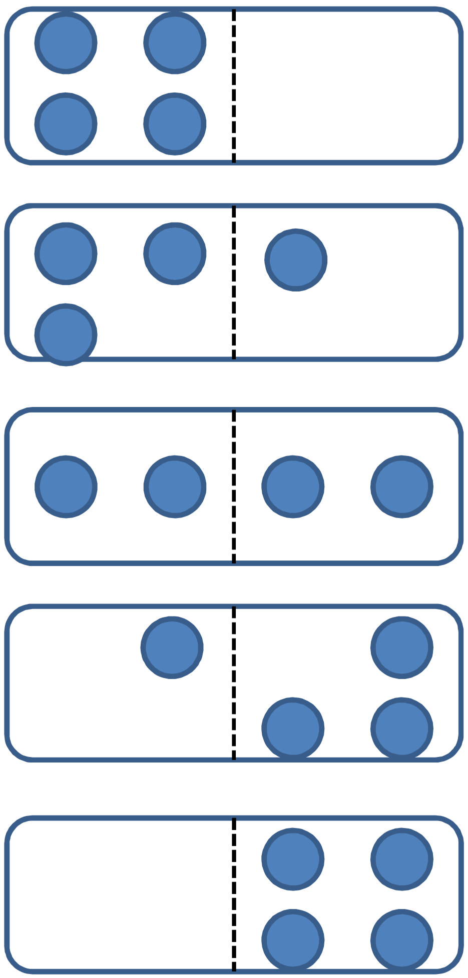
- The goal is to minimize the size of the information.
  - Intuitively, you don't want to say a long sentence for: "how are you?", "this is an apple."

An example with Probability



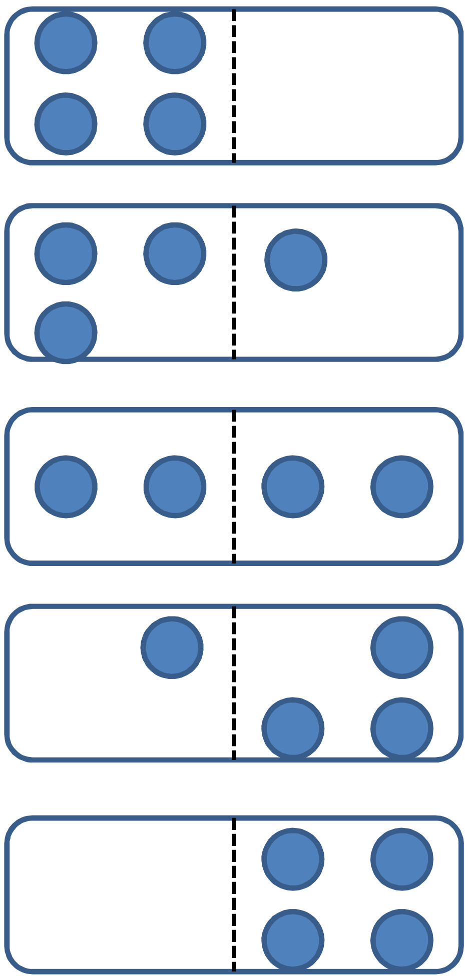
Assume that the particles are can move to anywhere in the container.

An example with Probability



$$P = \frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{2} = \frac{1}{16}$$

An example with Probability



Q1: What is the relationships among these states?

Q2: Can we have a single term to represent the information as knowledge?

MOTIVATION: COMPRESSION

- Suppose we observe a sequence of events:
  - Coin tosses
  - Words in a language
  - notes in a song
  - etc.
- We want to record the sequence of events in the smallest possible space.
- In other words we want the shortest representation which preserves all information.
- Another way to think about this: How much information does the sequence of events actually contain?

MOTIVATION: COMPRESSION

To be concrete, consider the problem of recording coin tosses in unary.

$$T, T, T, T, H$$

Approach 1:

H	T
0	00

$$00, 00, 00, 00, 0$$

We used 9 characters

MOTIVATION: COMPRESSION

To be concrete, consider the problem of recording coin tosses in unary.

$$T, T, T, T, H$$

Approach 2:

H	T
00	0

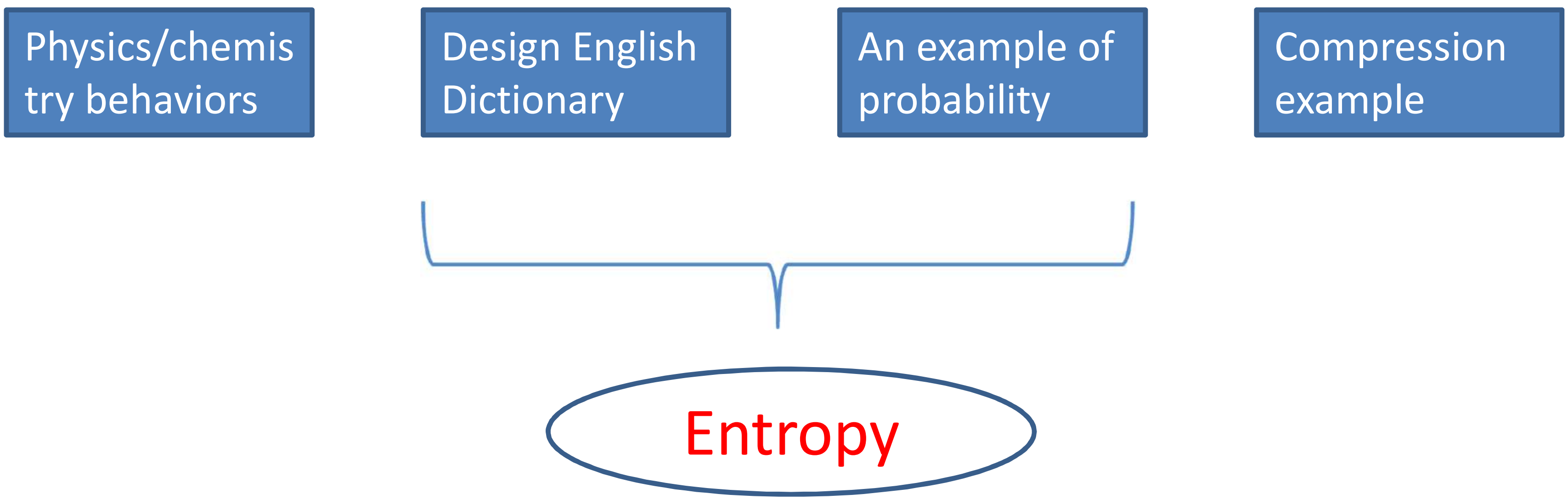
$$0, 0, 0, 0, 00$$

We used 6 characters

Motivation: Compression

- Frequently occuring events should have short encodings
- We see this in english with words such as “a”, “the”, “and”, etc.
- We want to maximise the information-per-character
- seeing common events provides little information
- seeing uncommon events provides a lot of information

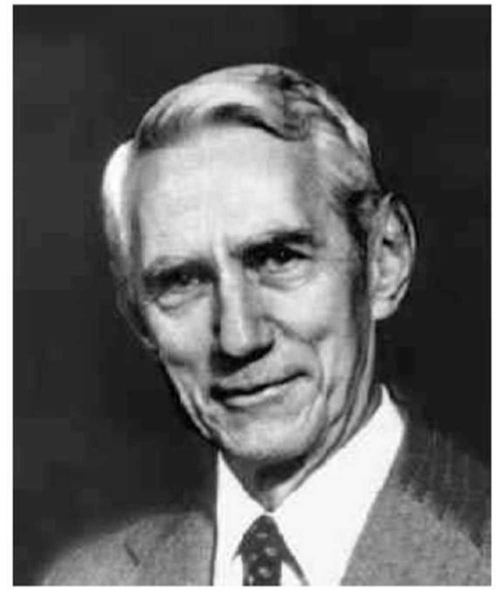
Application examples



Entropy is a direct measure of disorder.

Information Theory

- Information theory is a mathematical framework which addresses questions like:
  - How much information does a random variable carry about?
  - How efficient is a hypothetical code, given the statistics of the random variable?
  - How much better or worse would another code do?
  - Is the information carried by different random variables complementary or redundant?


Claude Shannon

Outline

- Logistics
- Motivation
- Entropy ←
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

Entropy

- Entropy $H(Y)$ of a random variable Y

$$H(Y) = -\sum_{k=1}^{K} P(y = k) \log_2 P(y = k)$$

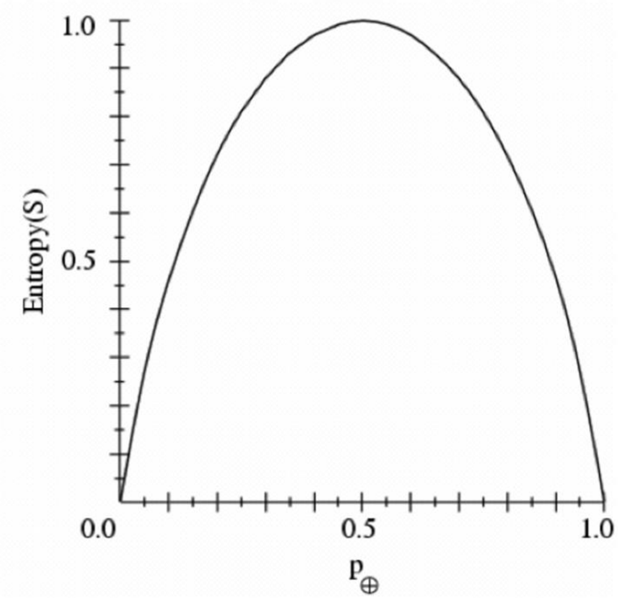
- $H(Y)$ is the expected number of bits needed to encode a randomly drawn value of Y (under most efficient code)

- Information theory:

  Most efficient code assigns $-\log_2 P(Y = k)$ bits to encode the message $Y = k$, So, expected number of bits to code one random Y is:

$$\sum_{k=1}^{K} P(y = k) \log_2 P(y = k)$$

Entropy



- S is a sample of coin flips
- p_+ is the proportion of heads in S
- p_- is the proportion of tails in S
- Entropy measure the uncertainty of S

$$H(S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_-$$

Entropy Computation: An Example

$$H(S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_-$$

head	0
tail	6

P(h) = 0/6 = 0 P(t) = 6/6 = 1

Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

head	1
tail	5

P(h) = 1/6 P(t) = 5/6

Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (1/6) = 0.65

head	2
tail	4

P(h) = 2/6 P(t) = 4/6

Entropy = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

Information

Let X be a random variable with distribution p(x)

$$I(X) = \log_2\left(\frac{1}{p(x)}\right)$$

Have you heard a picture is worth 1000 words?

Information obtained by random word from a 100,000 word vocabulary:

$$I(word) = \log\left(\frac{1}{p(x)}\right) = \log\left(\frac{1}{1/100000}\right) = 16.61\ bits$$

A 1000 word document from same source:

$$I(document) = 1000 \times I(word) = 16610$$

A 640*480 pixel, 16-greyscale video picture (each pixel has 16 bits information):

$$I(Picture) = \log\left(\frac{1}{1/16^{640*480}}\right) = 1228800$$

A picture is worth (a lot more than) 1000 words!

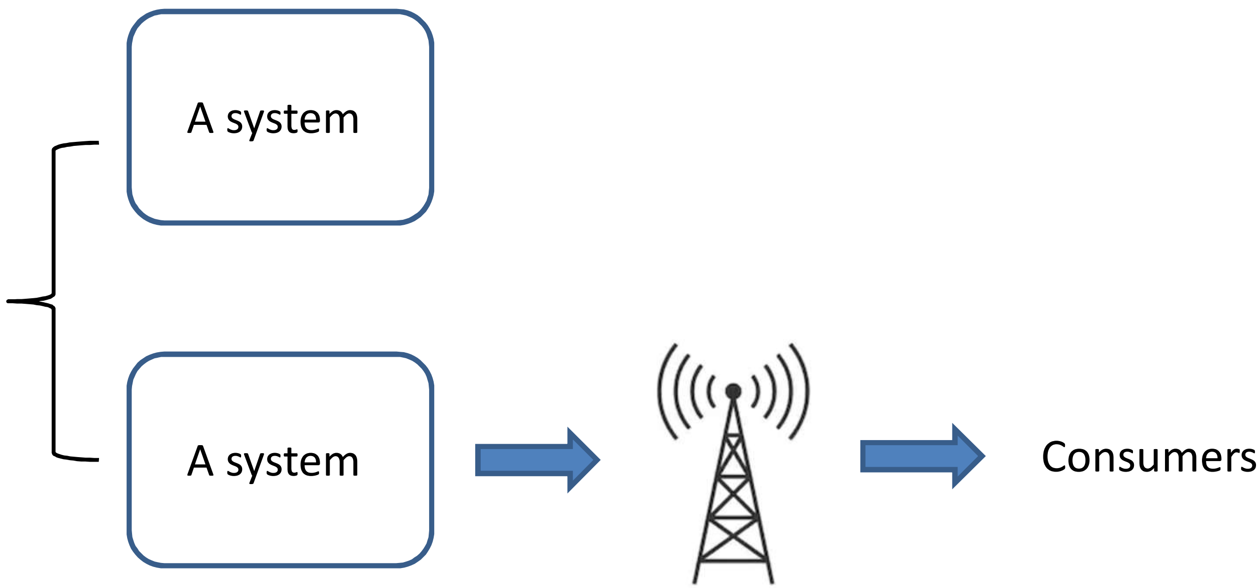
Understand entropy with the example

Physics/chemis try behaviors

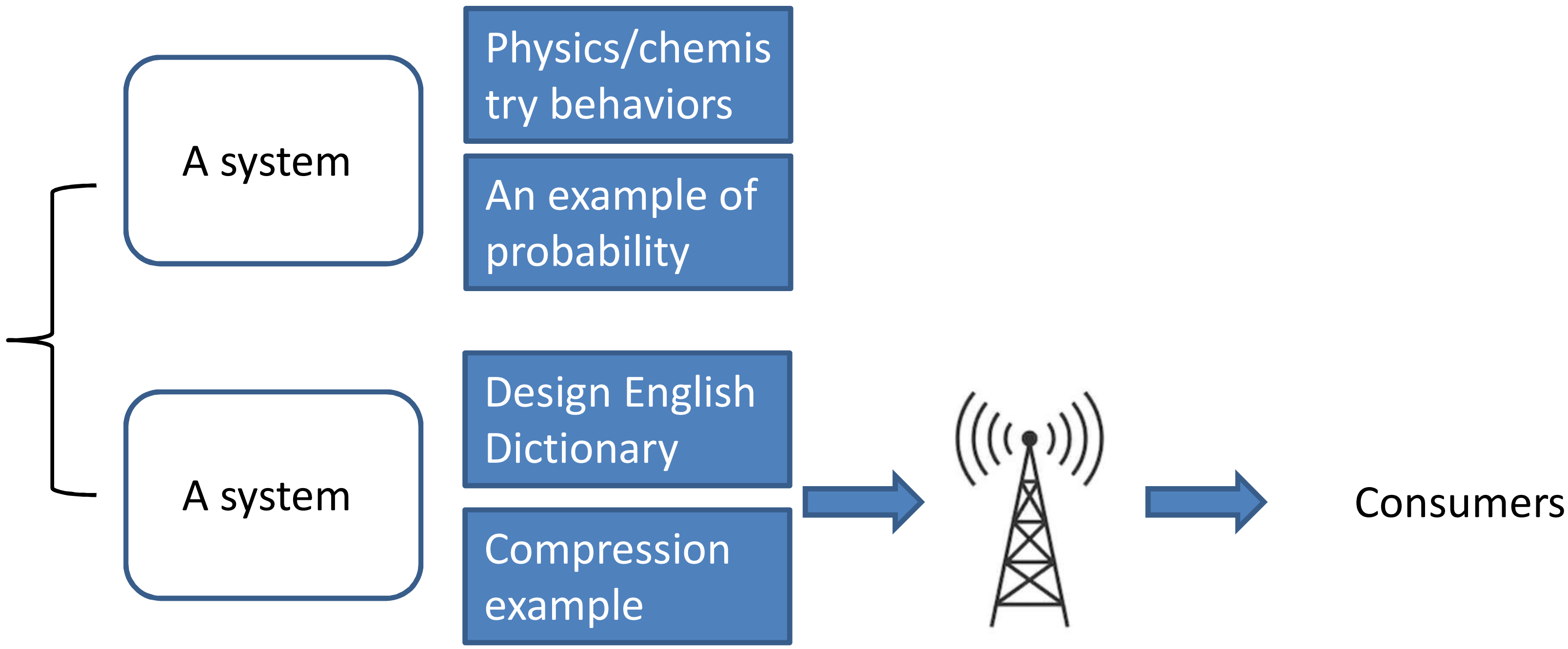
Design English Dictionary

An example of probability

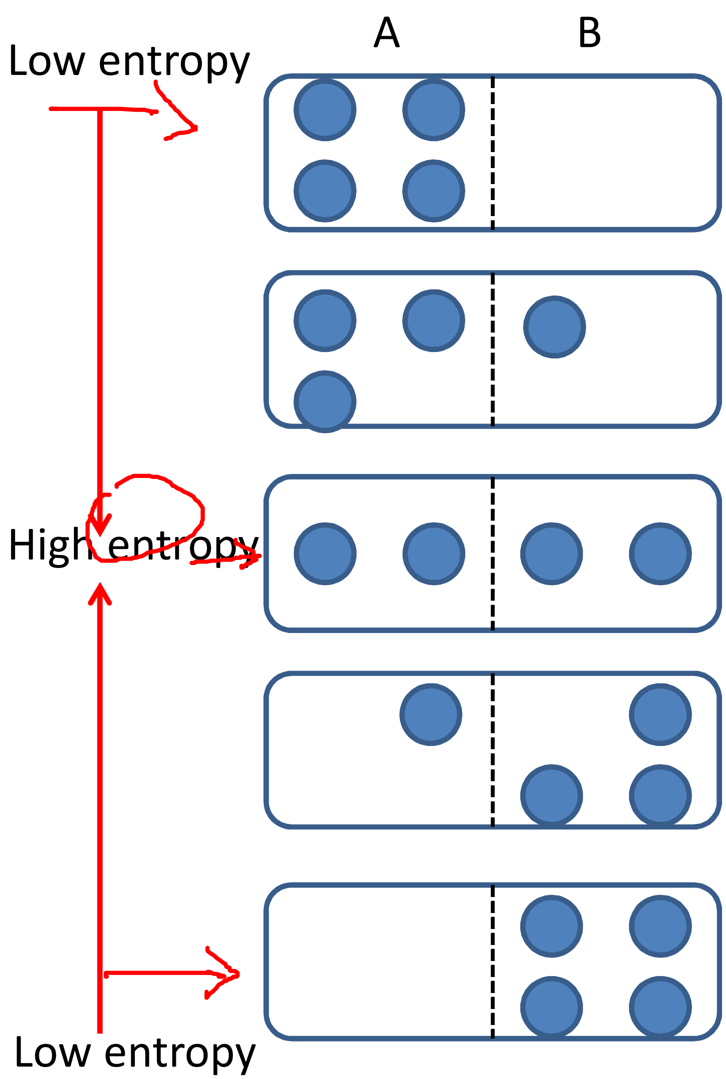
Compression example



Understand entropy with the example



The Probability Example



P(A) =4 P(B) = 0 Entropy = -0-0 = 0

P(A) =3 P(B) = 1 Entropy = $-\frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{3}{4}\log\left(\frac{3}{4}\right) = 0.81$

P(A) =2 P(B) = 2 Entropy = $-\frac{2}{4}\log\left(\frac{2}{4}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right) = 1$

P(A) =1 P(B) = 3 Entropy = $-\frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{3}{4}\log\left(\frac{3}{4}\right) = 0.81$

P(A) =4 P(B) = 0 Entropy = 0

Physics and chemistry

Entropy(solid water) < Entropy (liquid water)

Entropy(compressed air) < Entropy (air outside)

Definition of entropy in Compression

$$H(Y) = -\sum_{k=1}^{K} P(y=k) \log_2 P(y=k)$$

$log_2 \frac{1}{p(y=k)}$

- Frequent => High probability =>Less bits
- Infrequent => Low probability => More bits

$log_2 \frac{1}{1} = 0$

$log_2 \frac{1}{1/1024} = 10$

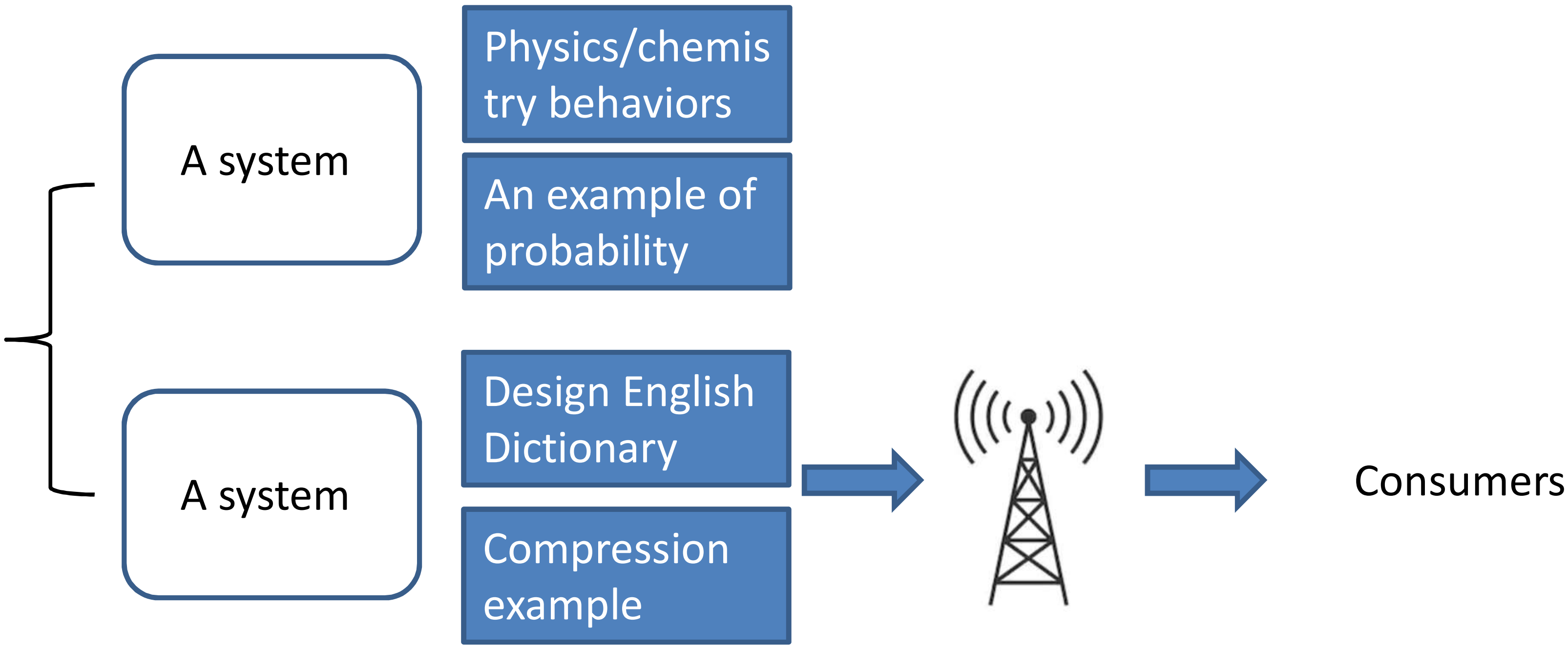
Why using log 2 as the definition???

$p = \frac{1}{2}$ $\quad$ $p = \frac{1}{4}$ $\quad$ … … $\quad$ $p = \frac{1}{1024}$

$\log 2 = 1$ meaning 1 bit $\quad$ $\log 4 = 2$ meaning 2 bits $\quad$ $\log 1024 = 10$ meaning 10 bits

Understand entropy with the example

Entropy is a direct measurement of disorder.



Entropy is an average number of bits needed encodes an variable..

More disordered means needing more bits for the encoding

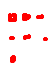
Less disordered means needing less bits for the encoding

How to explain?

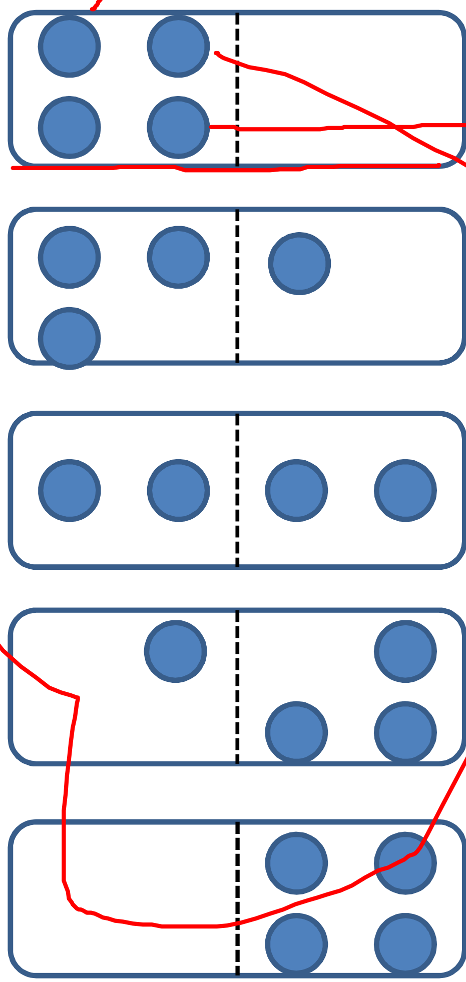
Water in solid
Low entropy

Water in liquid
Medium entropy

Water in vapor
High entropy



An example with Probability



H=0, $P = \frac{1}{16}$

H=0.81, $P = \frac{4}{16}$

H=1, $P = \frac{6}{16}$

H=0.81, $P = \frac{4}{16}$

H=0, $P = \frac{1}{16}$

Entropy = $\frac{1}{16} * log\frac{16}{1} + \frac{4}{16} * log\frac{16}{4} + \frac{6}{16} * log\frac{16}{6} + \frac{4}{16} * log\frac{16}{4} + \frac{1}{16} * log\frac{16}{1}$

$= \frac{1}{4} + \frac{1}{4} + \frac{1}{2} + \frac{1}{2} + 0.53 = 2.03$

Properties of Entropy

$$H(P) = \sum_i p_i \cdot \log \frac{1}{p_i}$$

1. Non-negative: $H(P) \geq 0$
2. Invariant wrt permutation of its inputs:
   $H(p_1, p_2, \ldots, p_k) = H(p_{\tau(1)}, p_{\tau(2)}, \ldots, p_{\tau(k)})$
3. For any *other* probability distribution $\{q_1, q_2, \ldots, q_k\}$:
   $$H(P) = \sum_i p_i \cdot \log \frac{1}{p_i} < \sum_i p_i \cdot \log \frac{1}{q_i}$$
4. $H(P) \leq \log k$, with equality iff $p_i = 1/k \ \forall i$
5. The further P is from uniform, the lower the entropy.

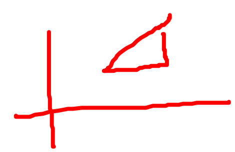
Outline

- Logistics
- Motivation
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

Joint Entropy

		Temperature		
	cold	mild	hot	
huMidity low	0.1	0.4	0.1	0.6
high	0.2	0.1	0.1	0.4
	0.3	0.5	0.2	1.0

- $H(T) = H(0.3, 0.5, 0.2) = 1.48548$
- $H(M) = H(0.6, 0.4) = 0.970951$
- $H(T) + H(M) = 2.456431$
- **Joint Entropy**: consider the space of (t, m) events $H(T, M) = \sum_{t,m} P(T = t, M = m) \cdot \log \frac{1}{P(T=t, M=m)}$
  $H(0.1, 0.4, 0.1, 0.2, 0.1, 0.1) = 2.32193$



Notice that $H(T, M) < H(T) + H(M)$!!!

Conditional Entropy

$$P(T = t|M = m)$$

	cold	mild	hot	
low	1/6	4/6	1/6	1.0
high	2/4	1/4	1/4	1.0

P(x,y)

p(x)

Conditional Entropy:

- $H(T|M = low) = H(1/6, 4/6, 1/6) = 1.25163$
- $H(T|M = high) = H(2/4, 1/4, 1/4) = 1.5$
- **Average Conditional Entropy** (aka equivocation):
  $H(T/M) = \sum_m P(M = m) \cdot H(T|M = m) =$
  $0.6 \cdot H(T|M = low) + 0.4 \cdot H(T|M = high) = 1.350978$

Conditional Entropy

$$P(M = m|T = t)$$

	cold	mild	hot
low	1/3	4/5	1/2
high	2/3	1/5	1/2
	1.0	1.0	1.0

Conditional Entropy:

- $H(M|T = cold) = H(1/3, 2/3) = 0.918296$
- $H(M|T = mild) = H(4/5, 1/5) = 0.721928$
- $H(M|T = hot) = H(1/2, 1/2) = 1.0$
- Average Conditional Entropy (aka Equivocation):
  $H(M/T) = \sum_t P(T = t) \cdot H(M|T = t) =$
  $0.3 \cdot H(M|T = cold) + 0.5 \cdot H(M|T = mild) + 0.2 \cdot H(M|T = hot) = 0.8364528$

Conditional Entropy

- Conditional entropy $H(Y|X)$ of a random variable Y given X_i

Discrete random variables:

$$H(Y|X_i) = \sum_{x \in X} p(x_i) H(Y|X = x_i) = \sum_{x \in X, y \in Y} p(x_i, y_i) log \frac{p(x_i)}{p(x_i, y_i)}$$

Continuous:

$$H(Y|X_i) = -\int \left(\sum_{k=1}^{K} P(y = k|x_i) \log_2 P(y = k) \right) p(x_i) dx_i$$

- Quantify the uncertainty in Y after seeing feature X_i
- $H(Y)$ is the expected number of bits needed to encode a randomly drawn value of Y
  - given X_i, and
  - average over the likelihood of seeing particular value of x_i

Mutual Information

- Mutual information: quantify the reduction in uncertainty in Y after seeing feature X_i

$$I(X_i, Y) = H(Y) - H(Y|X_i)$$

- The more the reduction in entropy, the more informative a feature.

- Mutual information is symmetric
  - $I(X_i, Y) = I(Y, X_i) = H(X_i) - H(X_i|Y)$
  - $I(Y, X_i) = \int \sum_k^K p(x_i, y = k) \log_2 \frac{p(x_i, y=k)}{p(x_i)p(y=k)} \; dx_i$
  - $= \int \sum_k^K p(x_i|y = k)p(y = k) \log_2 \frac{p(x_i|y = k)}{p(x_i)} \; dx_i$

Properties of Mutual Information

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X/Y) \\
&= \sum_x P(x) \cdot \log \frac{1}{P(x)} - \sum_{x,y} P(x,y) \cdot \log \frac{1}{P(x|y)} \\
&= \sum_{x,y} P(x,y) \cdot \log \frac{P(x|y)}{P(x)} \\
&= \sum_{x,y} P(x,y) \cdot \log \frac{P(x,y)}{P(x)P(y)}
\end{aligned}
$$

Properties of Average Mutual Information:

- Symmetric (but $H(X) \neq H(Y)$ and $H(X/Y) \neq H(Y/X)$)
- Non-negative (but $H(X) - H(X/y)$ may be negative!)
- Zero iff X, Y independent

CE and MI: Visual Illustration

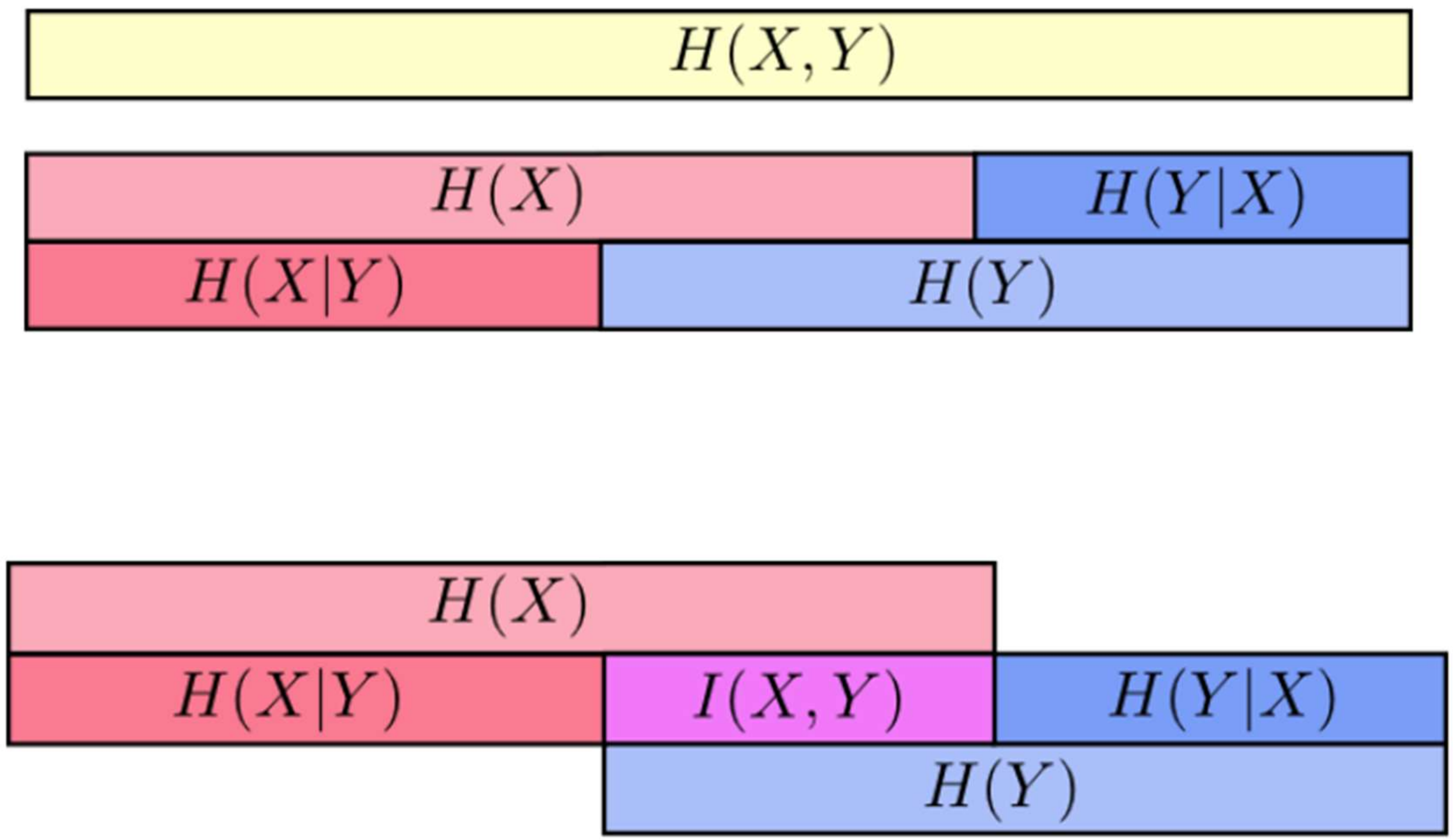
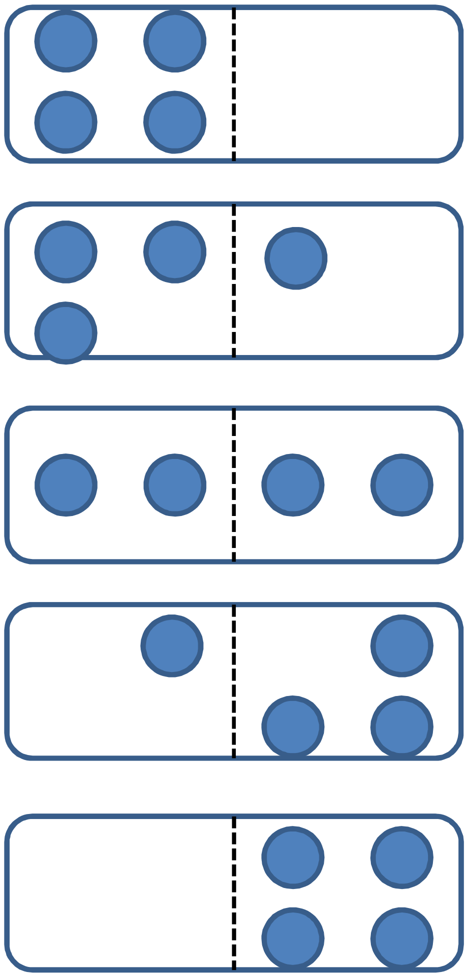


Image Credit: Christopher Olah.

Outline

- Logistics
- Motivation
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence ⬅

An example that motivates Cross Entropy



	Predicted distribution		True distribution
	$q = \frac{1}{16}$	$log\frac{16}{1} = 4$	$P = \frac{2}{16}$
	$q = \frac{4}{16}$	$log\frac{16}{4} = 2$	$P = \frac{4}{16}$
	$q = \frac{6}{16}$	$log\frac{16}{6} = 1.42$	$P = \frac{4}{16}$
	$q = \frac{4}{16}$	$log\frac{16}{4} = 2$	$P = \frac{4}{16}$
	$q = \frac{1}{16}$	$log\frac{16}{1} = 4$	$P = \frac{2}{16}$

This determines how we encode

Entropy = 2.03

$\text{Entropy} = 4 * \frac{2}{16} * 2 + 2 * \frac{4}{16} * 2 + 1.42 * \frac{4}{16} = 2.35$

Cross Entropy

Cross Entropy: The expected number of bits when a wrong distribution Q is assumed while the data actually follows a distribution P

$$H(p,q) = -\sum_{x\in\mathcal{X}} p(x)\log q(x)$$

This is because:

$$H(p,q) = \mathrm{E}_p[l_i] = \mathrm{E}_p\left[\log \frac{1}{q(x_i)}\right]$$

$$H(p,q) = \sum_{x_i} p(x_i)\ \log \frac{1}{q(x_i)}$$

$$H(p,q) = -\sum_{x} p(x)\ \log q(x).$$

Kullback-Leibler Divergence

Another useful information theoretic quantity measures the difference between two distributions.

$$\mathbf{KL}[P(S)\|Q(S)] = \sum_s P(s) \log \frac{P(s)}{Q(s)}$$

$$= \underbrace{\sum_s P(s) \log \frac{1}{Q(s)}}_{\text{cross entropy}} - \mathbf{H}[P]$$

Cross Entropy = Entropy + KL Divergence

Excess cost in bits paid by encoding according to Q instead of P.

KL Divergence is a distance measurement

$$-\mathbf{KL}[P\|Q] = \sum_s P(s) \log \frac{Q(s)}{P(s)}$$

$$\sum_s P(s) \log \frac{Q(s)}{P(s)} \leq \log \sum_s P(s) \frac{Q(s)}{P(s)} \quad \text{by Jensen}$$

$$= \log \sum_s Q(s) = \log 1 = 0$$

So $\mathbf{KL}[P\|Q] \geq 0$. Equality iff $P = Q$

When $P = Q$, $KL[P||Q] = 0$

Entropy and KL Divergence in Machine learning

- Construct a model with high entropy or low entropy?
- How a modle is related to cross entropy and KL Divergence?

Take-Home Messages

- Entropy
  - ▸A measure for disorder
  - ▸Why it is defined in this way (optimal coding)
  - ▸Its properties
- Joint Entropy, Conditional Entropy, Mutual Information
  - ▸The physical intuitions behind their definitions
  - ▸The relationships between them
- Cross Entropy, KL Divergence
  - ▸The physical intuitions behind them
  - ▸The relationships between entropy, cross-entropy, and KL divergence

Lagrange Multipliers

- Min/Max a function $f(x, y, z)$, where x, y, z are subject to the constraint g(x, y, z)=c

- Lagrange Multipliers
  - Define $F(x, y, z, \lambda) = f(x, y, z) + \lambda g(x, y, z)$
  - Take partial derivative with regarding to each parameter
  - Solve all the associated equations as the potential min/max value.

- Example
  - Max $f(x, y) = x^2 y$, s.t. x^2+y^2=1
  - Max $f(x, y, z) = 8xyz,\ s.t. \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$