Information

Reading for this lecture:

Elements of Information Theory (EIT), Chapters 1 & Sections 2.1-2.8.

Information ...

Sources of Information:

Apparent randomness:

Uncontrolled initial conditions
Actively generated: Deterministic chaos

Hidden regularity:

Ignorance of forces
Limited capacity to model structure

Information ...

Issues:

- What is information?
- How do we measure unpredictability?
- How do we quantify structure?
- Information $\neq$ Energy

History of information:

- Boltzmann (19th Century):
  - Equilibrium, large-scale systems (indistinguishable microstates)
- Hartley-Shannon-Wiener (Early 20th):
  - Communication & Cryptography
- Current threads (late 20th century):
  - Coding, Statistics, Dynamics, and Learning

Information ...

Information as uncertainty:

Observe something unexpected:

Gain information

Bateson: "A difference that makes a difference"

Information ...

Information as uncertainty:

How to formalize?

Shannon's approach:

A measure of surprise.

Connection with Boltzmann's thermodynamic entropy

Self-information of an event $\propto -\log \Pr(\text{event})$.

Predictable: No surprise $-\log 1 = 0$

Completely unpredictable: Maximally surprised

$$-\log \frac{1}{\text{Number of Events}} = \log(\text{Number of Events})$$

Information ...

How to measure?

Information = f(Pr(events))?

What is f( )? Maps probability distribution to a number.

Random variables:

$$X, Y; \text{ events } x, y \in \{1, 2, \ldots, k\}$$

Distribution:

$$\Pr(X) = (p_1, \ldots, p_k)$$

$$\Pr(Y) = (q_1, \ldots, q_k)$$

Shorthand: $X \sim p(x)$

$$Y \sim q(x)$$

Information ...

Khinchin axioms for a measure of information:

Entropy: $H(X) = H(p_1, \ldots, p_k)$

(1) Maximum at equidistribution:

$$H(p_1, \ldots, p_k) \leq H\left(\tfrac{1}{k}, \ldots, \tfrac{1}{k}\right)$$

(2) Continuous function of distribution:

$$H(p_1, \ldots, p_k) \text{ versus } p_i$$

(3) Expansibility:

$$H(p_1, \ldots, p_k) = H(p_1, \ldots, p_k, p_{k+1} = 0)$$

(4) Additivity of independent systems:

$$X \perp Y \Rightarrow H(X, Y) = H(X) + H(Y)$$

Information ...

Khinchin axioms for a measure of information ...

Theorem:

Then get unique (up to a factor) functional form,

The Shannon entropy:

$$H(X) \propto -\sum_{i=1}^{k} p_i \log p_i$$

Information ...

Shannon axioms for a measure of information:

Entropy: $H(X) = H(p_1, \ldots, p_k)$

(1) Maximum surprise:

$$H(\tfrac{1}{2}, \tfrac{1}{2}) = 1$$

(2) Continuous function of distribution:

$$H(p_1, \ldots, p_k) \text{ versus } p_i$$

(3) Merging:

$$H(p_1, p_2, p_3, \ldots, p_k)$$

$$= H(\overbrace{p_1 + p_2, p_3, \ldots, p_k}^{k-1 \text{ events}}) + (p_1 + p_2) H(\overbrace{\tfrac{p_1}{p_1+p_2}, \tfrac{p_2}{p_1+p_2}}^{2 \text{ events}})$$

Information ...

Shannon axioms for a measure of information ...

Theorem (ditto):

Also get Shannon entropy:

$$H(X) \propto -\sum_{i=1}^{k} p_i \log p_i$$

Information ...

Shannon Entropy: $X \sim P \qquad x \in \mathcal{X} = \{1, 2, \ldots, k\}$

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

Note: $0 \log 0 = 0$

$$H(X) = \langle -\log_2 p(x) \rangle$$

Units:

Log base 2: $H(X) = [\mathrm{bits}]$

Natural log: $H(X) = [\mathrm{nats}]$

Properties:

1. Positivity: $H(X) \geq 0$
2. Predictive: $H(X) = 0 \;\Leftrightarrow\; p(x) = 1$ for one and only one x
3. Random: $H(X) = \log_2 k \;\Leftrightarrow\; p(x) = U(x) = 1/k$

Information ...

$$\Pr(1) = p \ \& \ \Pr(0) = 1 - p$$

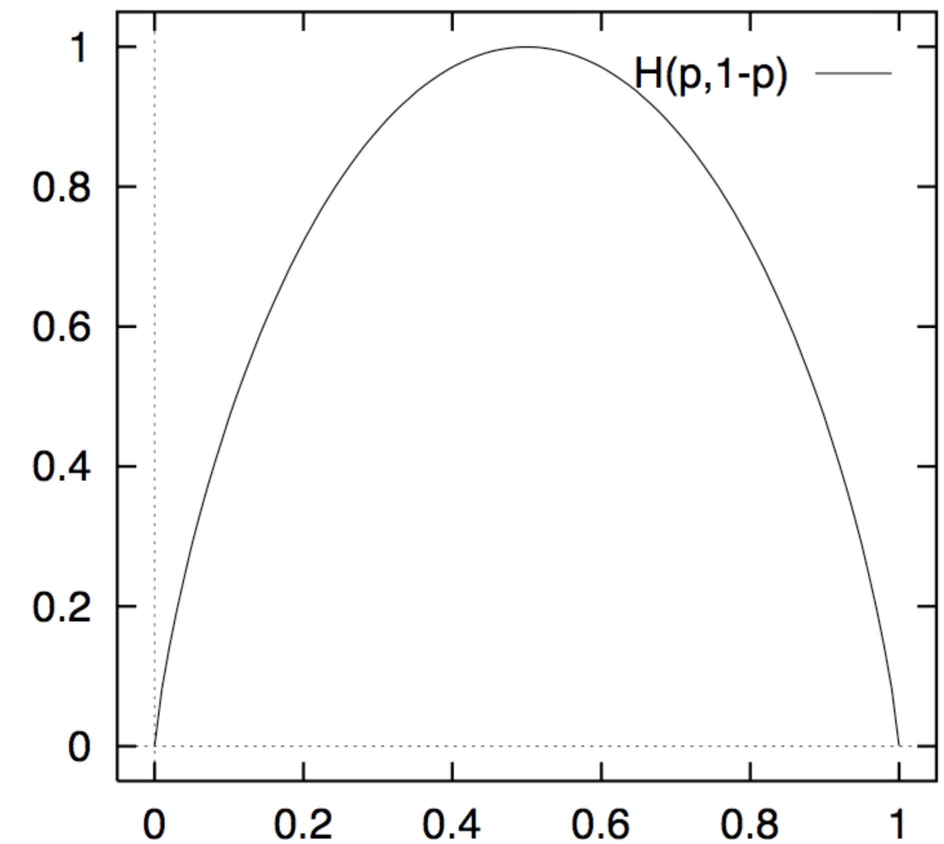
Examples: Binary random variable X

$$\mathcal{X} = \{\text{Heads}, \text{Tails}\} \quad \Pr(\text{Heads}) = p, \quad \Pr(\text{Tails}) = 1 - \text{p}$$

$H(X)$?

Binary entropy function:

$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$



Fair coin: $p = \frac{1}{2}$

$$H(p) = 1 \text{ bit}$$

Completely biased coin: $p = 0 \ (\text{or } 1)$

$$H(p) = 0 \text{ bits}$$

Recall: $0 \cdot \log 0 = 0$

Information ...

Example: IID Process over four events

$$\mathcal{X} = \{a, b, c, d\} \qquad \Pr(X) = (\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8})$$

Entropy: $H(X) = \frac{7}{4}$ bits

Number of questions to identify the event?

x = a? (must always ask at least one question)
x = b? (this is necessary only half the time)
x = c? (only get this far a quarter of the time)

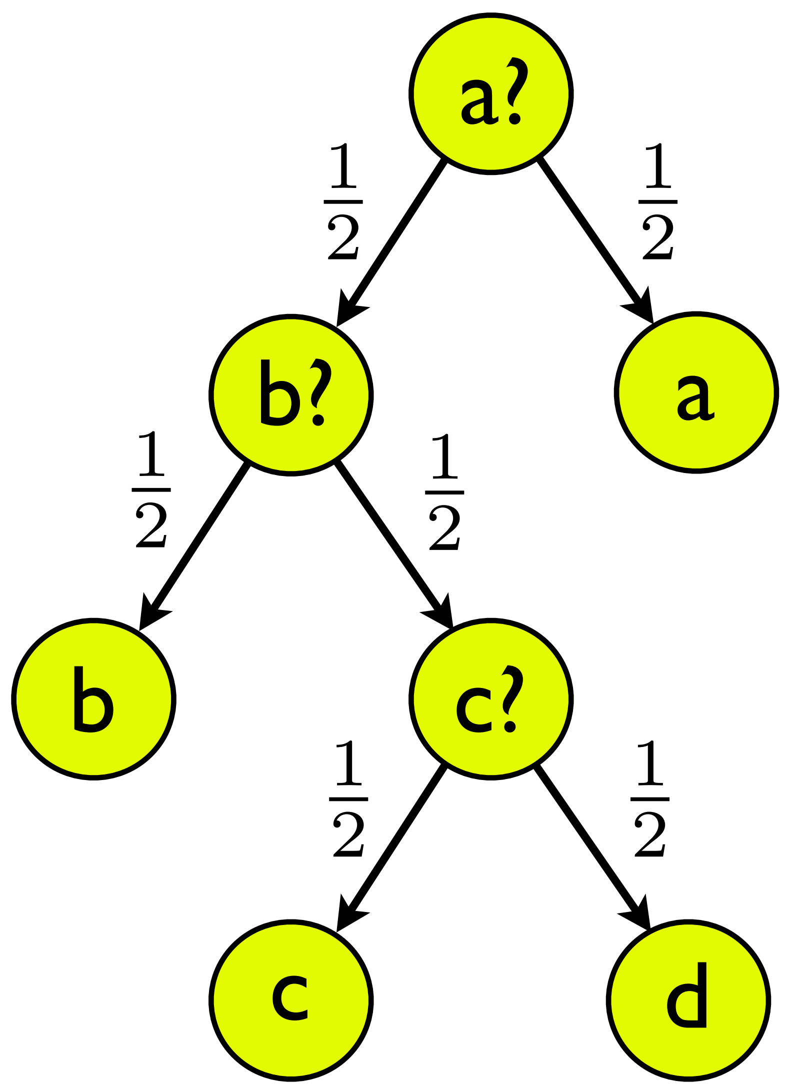
Average number: $1 \cdot 1 + 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} = 1.75$ questions

Interpretation? Optimal way to ask questions.

Information ...

Example: IID Process over four events ...

Average number: $1 \cdot 1 + 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} = 1.75$ questions



$$\Pr(X) = (\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8})$$

Information ...

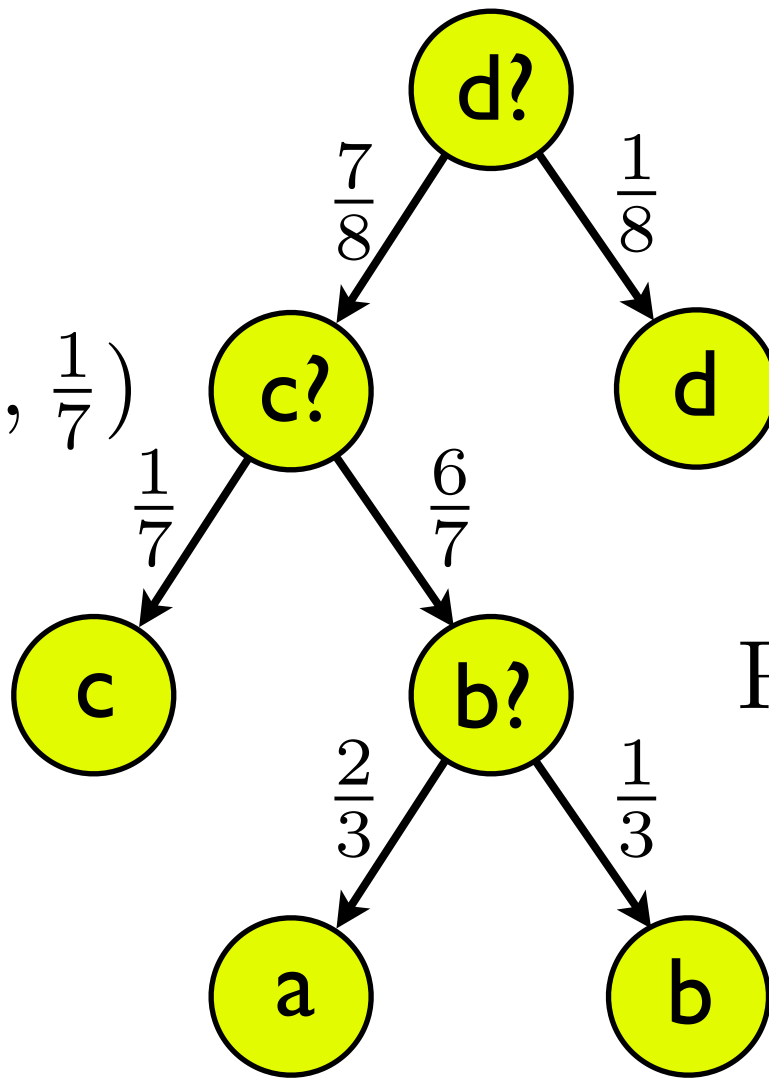
Example: IID Process over four events ...

Query in a different order:

Average number: $1 \cdot 1 + 1 \cdot \frac{7}{8} + 1 \cdot \frac{6}{7} \approx 2.7$ questions

$$\Pr(X) = (\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8})$$

$$\Pr(a, b, c) = (\tfrac{4}{7}, \tfrac{2}{7}, \tfrac{1}{7})$$



$$\Pr(a, b) = (\tfrac{2}{3}, \tfrac{1}{3})$$

Information ...

Example: IID Process over four events

Entropy: $H(X) = \frac{7}{4}$ bits

At each stage, ask questions that are most informative.

Choose partitions of event space that give "most random" measurements.

Theorem:
Entropy gives the smallest number of questions to identify an event, on average.

Information ...

Interpretations of Shannon Entropy:

Observer's *degree of surprise* in outcome of a random variable

Uncertainty *in* random variable

Information required to *describe* random variable

A measure of *flatness* of a distribution

Information ...

Two random variables: $(X, Y) \sim p(x, y)$

Joint Entropy: Average uncertainty in X and Y occurring

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x, y)$$

Independent:

$$X \perp Y \Rightarrow H(X, Y) = H(X) + H(Y)$$

Conditional Entropy: Average uncertainty in X, knowing Y

$$H(X|Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log_2 p(x|y)$$

$$H(X|Y) = H(X,Y) - H(Y)$$

Not symmetric: $H(X|Y) \neq H(Y|X)$

Information ...

Example: Dining on campus

Food served at cafeteria is a random process:

Random variables:

Dinner one night: $D \in \{\mathrm{Pizza}, \mathrm{Meat\ w/Vegetable}\} = \{P, M\}$

Lunch the next day: $L \in \{\mathrm{Casserole}, \mathrm{Hot\ Dog}\} = \{C, H\}$

After many meals, estimate:

$$\Pr(P) = \tfrac{1}{2} \text{ \& } \Pr(M) = \tfrac{1}{2}$$

$$\Pr(C) = \tfrac{3}{4} \text{ \& } \Pr(H) = \tfrac{1}{4}$$

Entropies:

$$H(D) = 1 \text{ bit}$$

$$H(L) = H(\tfrac{3}{4}) \approx 0.81 \text{ bits}$$

Information ...

Example: Dining on campus ...

Also, after many meals, estimate the joint probabilities:

$$\Pr(P, C) = \frac{1}{4} \;\&\; \Pr(P, H) = \frac{1}{4}$$

$$\Pr(M, C) = \frac{1}{2} \;\&\; \Pr(M, H) = 0$$

Joint Entropy: $H(D, L) = 1.5 \text{ bits}$

Dinner and Lunch are not independent:

$$H(D, L) = 1.5 \text{ bits} \neq H(D) + H(L) = 1.81 \text{ bits}$$

Suspect something's correlated: What?

Information ...

Example: Dining on campus ...

Conditional entropy of lunch given dinner:

$$\Pr(C|P) = \Pr(P,C)/\Pr(P) = \tfrac{1}{2}$$

$$\Pr(H|P) = \Pr(P,H)/\Pr(P) = \tfrac{1}{2}$$

$$\Pr(C|M) = \Pr(M,C)/\Pr(M) = 1$$

$$\Pr(H|M) = \Pr(M,H)/\Pr(M) = 0$$

$H(L|P) = 1 \text{ bit}$ Lunch unpredictable, if dinner was Pizza

$H(L|M) = 0 \text{ bits}$ Lunch predictable, if dinner was Meat w/Veg

Average uncertainty about lunch, given dinner:

$$H(L|D) = \tfrac{1}{2} \text{ bit}$$

Information ...

Example: Dining on campus ...

Other way around?

Conditional entropy of dinner given lunch:

$$\Pr(P|C) = \Pr(P,C)/\Pr(C) = \tfrac{1}{3}$$

$$\Pr(M|C) = \Pr(M,C)/\Pr(C) = \tfrac{2}{3}$$

$$\Pr(P|H) = \Pr(P,H)/\Pr(H) = 1$$

$$\Pr(M|H) = \Pr(M,H)/\Pr(H) = 0$$

$$H(D|C) = H(\tfrac{2}{3}) \approx 0.92 \text{ bits}$$

$$H(D|H) = 0 \text{ bits}$$

Average uncertainty about dinner, given lunch:

$$H(D|L) = \tfrac{3}{4} H(\tfrac{2}{3}) \approx 0.69 \text{ bits}$$

Note: $H(D|L) \neq H(L|D)$. In fact, $H(D|L) > H(L|D)$.

Information ...

Relative Entropy of Two Distributions:

$$X \sim P \ \& \ Y \sim Q, \ \text{over common } x \in \mathcal{X}$$

Relative Entropy:

$$\mathcal{D}(X||Y) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)}$$

Note: $0 \log \frac{0}{q} = 0$

$$p \log \frac{p}{0} = \infty$$

Typically applied to: $Q : \ q(x) > 0 \ , \ \forall \ x \in \mathcal{X}$

Alternate use (notation):

$$\mathcal{D}(P||Q)$$

Information ...

Relative Entropy of Two Distributions ...

Properties:

$$
\begin{aligned}
(1)\ \mathcal{D}(X||Y) &\geq 0 \\
(2)\ \mathcal{D}(X||Y) &= 0 \Leftrightarrow P = Q \\
(3)\ \mathcal{D}(X||Y) &\neq \mathcal{D}(Y||X)
\end{aligned}
$$

Also called:

Kullback-Leibler Divergence

Information Gain: Number of bits of describing X as Y

Discrimination between X & Y

Not a distance: not symmetric, no triangle inequality

Common Information Between Two Random Variables:

$$X \sim p(x) \ \& \ Y \sim p(y)$$

$$(X, Y) \sim p(x, y)$$

Mutual Information:

$$I(X;Y) = \mathcal{D}(P(x,y)||P(x)P(y))$$

$$I(X;Y) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y) \log_2 \tfrac{p(x,y)}{p(x)p(y)}$$

Information ...

Mutual Information ...

Properties:

(1) $I(X;Y) \geq 0$

(2) $I(X;Y) = I(Y;X)$

(3) $I(X;Y) = H(X) - H(X|Y)$

(4) $I(X;Y) = H(X) + H(Y) - H(X,Y)$

(5) $I(X;X) = H(X)$

(6) $X \perp Y \Rightarrow I(X;Y) = 0$

Interpretations:

Information one variable has about another
Information shared between two variables
Measure of dependence between two variables

Information ...

Example: Dining on campus ...

Mutual information:

Reduction in uncertainty about lunch, given dinner:

$$I(D;L) = H(L) - H(L|D)$$
$$= H(\tfrac{3}{4}) - \tfrac{1}{2} \approx 0.31 \text{ bits}$$

Reduction in uncertainty about dinner, given lunch:

$$I(D;L) = H(D) - H(D|L)$$
$$= 1 - H(\tfrac{2}{3}) \approx 1 - 0.69 = 0.31 \text{ bits}$$

Shared information between what's served for dinner & lunch.

Information ...

Example: Dining on campus ...

Mutual information ...

What is the shared information?

Further inquiry:

Vegetable served with dinner (Meat + Veg)

appears in lunch's casserole!

Information ...

Example: Dining on campus ...

How different are dinner and lunch?

Information Gain?

But they don't share event space: $D \in \{P, M\}$ & $L \in \{C, H\}$

Turns out the Pizza was vegetarian

The events are common:

Pizza and Casserole: Vegetarian

Meat w/Veg and Hot Dog: Not

$$V \in \{\mathrm{Veg}, \mathrm{Non}\}$$

$$\mathcal{D}(D||L) = \sum_{v \in V} \Pr(D = v) \log_2 \frac{\Pr(D = v)}{\Pr(L = v)}$$

$$\mathcal{D}(D||L) \approx 0.21 \text{ bits}$$

$$\mathcal{D}(L||D) \approx 0.19 \text{ bits}$$

Distance Between Two Random Variables:

$$X \sim P(x)$$

$$Y \sim Q(y)$$

$$(X, Y) \sim P(x, y)$$

Information Distance:

$$d(X, Y) = H(X|Y) + H(Y|X)$$

Or

$$d(X, Y) = H(X, Y) - I(X; Y)$$

Distance Between Two Random Variables ...

Information Distance Properties:

$$Z \sim R(z)$$

(1) Positivity: $d(X,Y) \geq 0$

(2) Equality: $d(X,Y) = 0 \iff P = Q$

(3) Symmetric: $d(X,Y) = d(Y,X)$

(4) Triangle inequality: $d(X,Y) \leq d(X,Z) + d(Z,Y)$

(5) Independence: $d(X,Y) \leq H(X) + H(Y)$

It is a distance!

Information ...

Example: Dining on campus ...

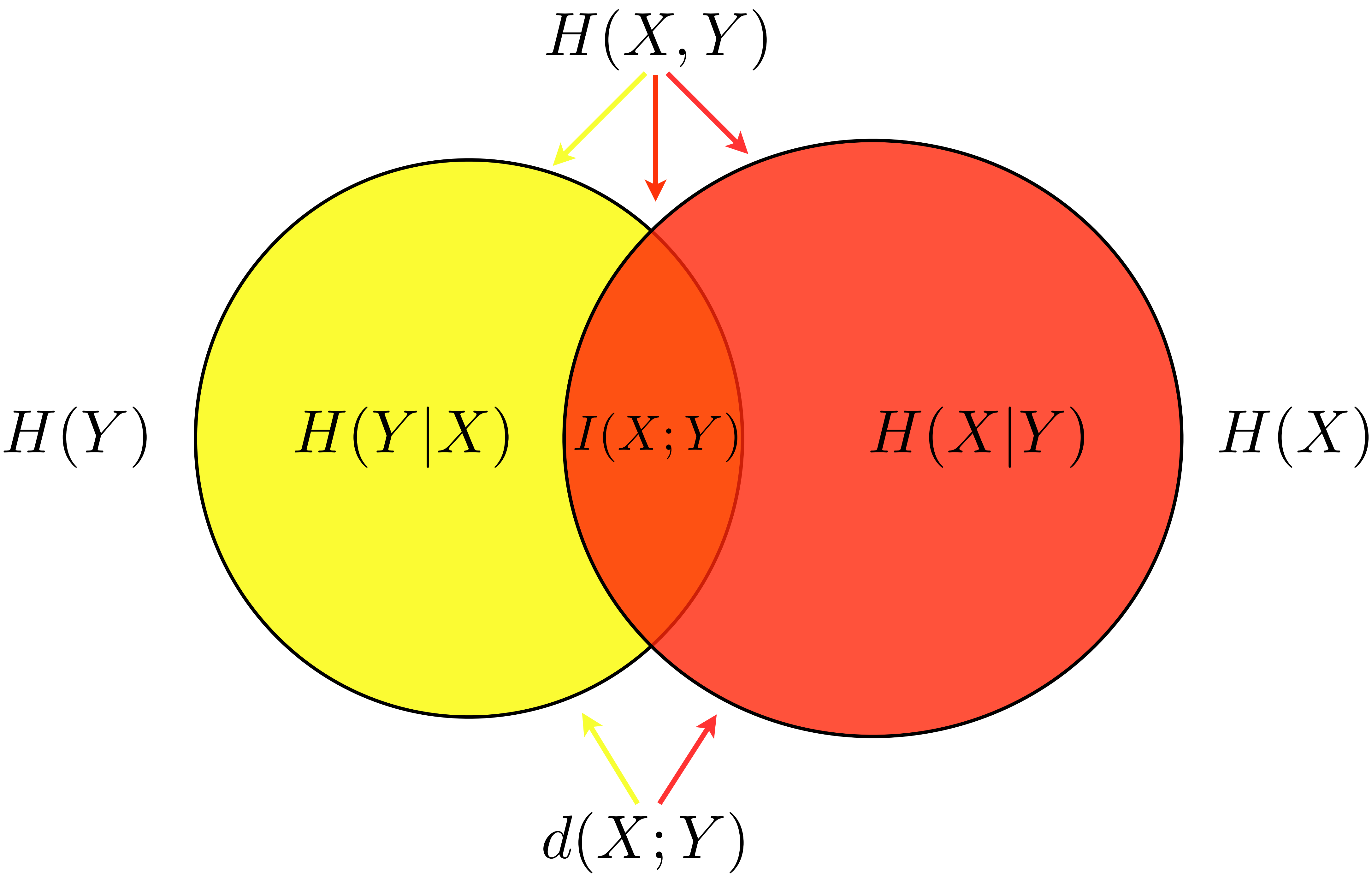
Informational distance between dinner and lunch?

$$d(D,L) = H(D|L) + H(L|D)$$

$$\begin{aligned} d(D,L) &= \tfrac{3}{4} H(\tfrac{2}{3}) + \tfrac{1}{2} H(1) \\ &\approx 0.69 + 0.5 = 1.19 \text{ bits} \end{aligned}$$

Information ...

Event Space Relationships of Information Quantifiers:



Information ...

Chain Rules:

Entropy Chain Rule:

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1)$$

$$= H(X_1) + H(X_2|X_1) + H(X_3|X_2 X_1) + \cdots$$

Information ...

Chain Rules ...

Conditional Mutual Information:

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

Mutual Information Chain Rule:

$$I(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1)$$

$$= I(X_1;Y) + I(X_2;Y|X_1) + I(X_3;Y|X_2X_1) + \cdots$$

Chain Rules ...

Conditional Relative Entropy:

$$\mathcal{D}(P(X|Y)||Q(X|Y)) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y)\log_2 \frac{p(x|y)}{q(x|y)}$$

Chain Rule:

$$\mathcal{D}(P(X,Y)||Q(X,Y))$$
$$= \mathcal{D}(P(X)||Q(X)) + \mathcal{D}(P(X|Y)||Q(X|Y))$$

Information ...

Bounds:

Uniform Distribution:

$$X \sim U(x) = 1/k$$

$$H(X) = \log |\mathcal{X}|$$

Generally: $H(X) \leq \log |\mathcal{X}|$

In fact: $H(X) = \log |\mathcal{X}| - \mathcal{D}(P(x)||U(x))$

Information ...

Bounds ...

Conditioning Reduces Entropy:

$$H(X|Y) \leq H(X)$$

Independence:

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

Information ...

Three random variables: $(X, Y, Z) \sim p(x, y, z)$

Markov Chain: $X \rightarrow Y \rightarrow Z$

$$p(x, z|y) = p(x|y)p(z|y) \qquad \text{or} \qquad I(X; Z|Y) = 0$$

Y shields X and Z from each other: $X \perp_Y Z$

Properties:

(1) $X \rightarrow Y \rightarrow Z \Rightarrow Z \rightarrow Y \rightarrow X$

(2) $Z = f(Y) \Rightarrow X \rightarrow Y \rightarrow Z$

Information ...

Data Processing Inequality:

$$X \to Y \to Z \Rightarrow I(X;Y) \geq I(X;Z)$$

Corollary:

$$Z = g(Y) \Rightarrow I(X;Y) \geq I(X;g(Y))$$

Manipulation *cannot* increase information about X.

Information ...

Dining example:

Hidden variable was "leftovers".

Knowing this, lunch and dinner are independent:

$$\text{Dinner} \perp_{\text{leftovers}} \text{Lunch}$$

Markov chain:

$$\text{Dinner} \rightarrow \text{leftovers} \rightarrow \text{Lunch}$$

Information ...

Reading for next lecture:

EIT, Secs. 5.1-5.6 and 7.1-7.7 and Chapter 4.