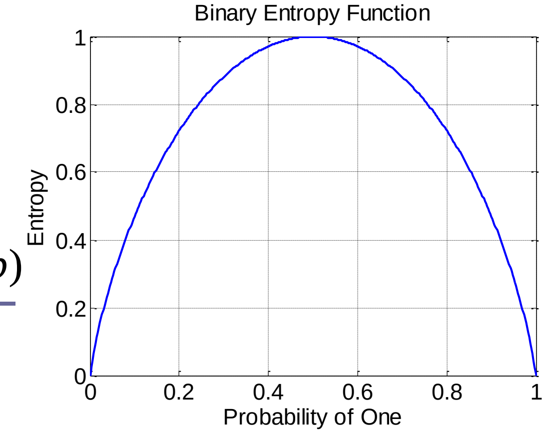
ESE 523
Information Theory

Joseph A. O’Sullivan
Samuel C. Sachs Professor
Electrical and Systems Engineering
Washington University
211 Urbauer Hall
2120E Green Hall
314-935-4173
jao@wustl.edu

Outline

$$(p,1-p) \Rightarrow$$
$$H(p) = -p\log p - (1-p)\log(1-p)$$



- Entropy $$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$$
- Joint Entropy $$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$
- Conditional Entropy $$H(X\,|\,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x\,|\,y)$$
- Relative Entropy $$D(p\,\|\,q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}$$
- Mutual Information $$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

Notation

- X: Random variable (R.V.)
- Alphabet (discrete): $\mathcal{X} = \{x_1, x_2, \ldots x_n\}$
- Probability mass function:

$$P(X = x_i) = p_i = p(i) = p(x_i)$$

$$p_i \geq 0, \quad \sum_{x \in \mathcal{X}} p_i = 1$$

- $\log = \log_2$

- Biased coin flip: $\mathcal{X}$={h,t}; p(x)=(p,1-p)
- Two dice: $\mathcal{X}$={2,3,4,5,6,7,8,9,10,11,12}; p(x)=(1,2,3,4,5,6,5,4,3,2,1)/36
- Powerball:

$$p(x) = \left[\binom{59}{5} 39 \right]^{-1} = \frac{1}{195{,}249{,}054}$$

Measure of Information: Entropy

- The **entropy** of X, $H(X)$ is:

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

- Units are "bits"
- Measure of uncertainty of a R.V.

$$H(X) = E\left[-\log p(X)\right]$$
$$= E\left[\log \frac{1}{p(X)}\right]$$

"… the eerily self-referential expectation…"
Cover and Thomas, p. 14

Entropy

- **Example 1:** Deterministic R.V.

$$p(x_i) = 1 \text{ and } p(x_j) = 0 \quad \forall j \neq i$$

$$H(X) = 0$$

- No information gained from observing the outcome

$$1 \cdot \log(1) = 1 \cdot 0 = 0$$

$$0 \log 0 \triangleq \lim_{\varepsilon \to 0^+} \varepsilon \log \varepsilon = 0$$

Proof uses l'Hôpital's rule:

$$\lim_{\varepsilon \to 0^+} \varepsilon \log \varepsilon = \lim_{\varepsilon \to 0^+} \frac{-\log(1/\varepsilon)}{1/\varepsilon} = \lim_{\varepsilon \to 0^+} \frac{1/\varepsilon}{-1/\varepsilon^2} \log e = 0$$

Entropy

- **Example 2:** Flip a "fair" coin

$$\mathcal{X} = \{\mathrm{h,t}\}; \quad p(\mathrm{h}) = p(\mathrm{t}) = \frac{1}{2}$$

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}$$

$$= 1\ \mathrm{bit}$$

Entropy

- **Example 3:** Flip a fair coin n times

$$\mathcal{X} = \{(\mathrm{h},\mathrm{h},\ldots \mathrm{h}),(\mathrm{h},\mathrm{h},\ldots \mathrm{t}),\ldots(\mathrm{t},\mathrm{t},\ldots \mathrm{t})\}$$

$$p(x_i) = \frac{1}{2^n} \quad i = 1, 2, \ldots 2^n$$

$$H(X) = -\sum_{i=1}^{2^n} \frac{1}{2^n} \log \frac{1}{2^n}$$

$$= n \text{ bits}$$

Entropy

- **Example 4:** Powerball, or any other uniform distribution.

$$\mathcal{X} = \{x_1, x_2, \ldots, x_M\}; \quad p(x_i) = \frac{1}{M}, \text{ for all } i$$

$$H(X) = \sum_{i=1}^{M} \frac{1}{M} \log M = \log M$$

$$H(\text{Powerball}) = \log(195249054) = 27.5407$$

Entropy

- **Example 5:** Flip a fair coin 2 times and add the number of heads

$$\mathcal{X} = \{0,1,2\}; \quad p(0) = p(2) = \frac{1}{4},\ p(1) = \frac{1}{2}$$

$$H(X) = -\frac{1}{4}\log\frac{1}{4} - \frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4}$$

$$= \frac{3}{2}\ \text{bits}$$

Properties and Remarks

- Entropy is the expected number of binary questions one needs to ask to determine the value of a R.V.
  - Last example: on average how many yes-no questions to determine outcome? Answer: 1.5 questions
- Entropy is nonnegative
- Base change: other units

$$H_b(X) = E\left[-\log_b p(x)\right] = \log_b a \cdot E\left[-\log_a p(x)\right]$$

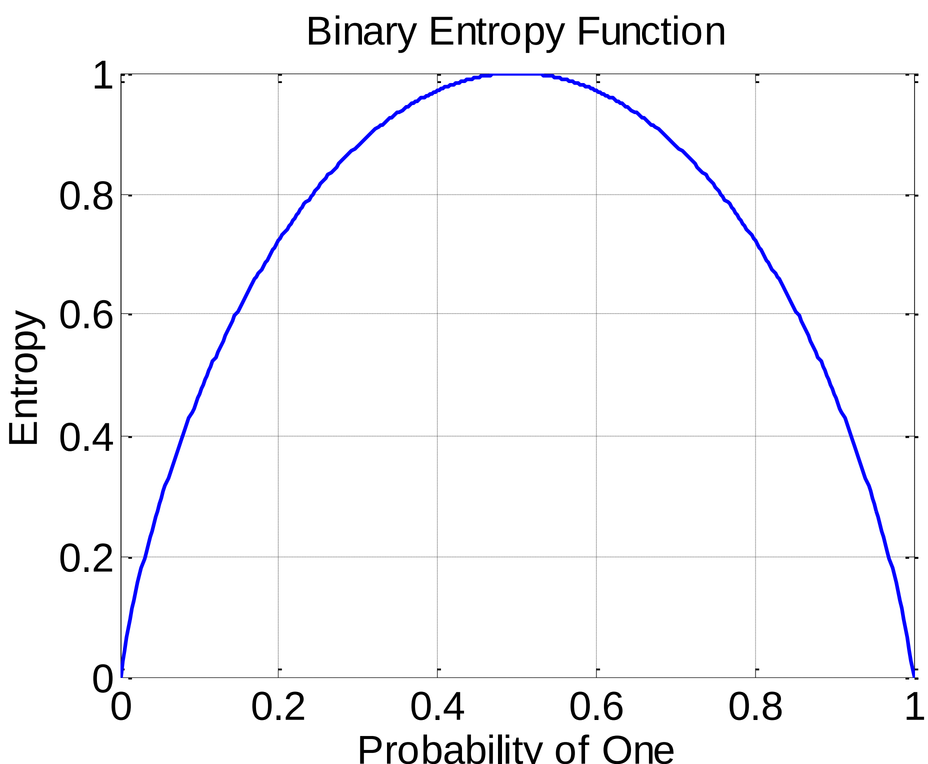
"nats" for base *e*

- 1 nat = $\log_2 e$ bits = 1.44 bits

Binary Entropy Function

$$\mathcal{X} = \{0,1\}; \quad p(1) = p$$

$$H(X) = -p\log p - (1-p)\log(1-p) = H(p)$$



Matlab Function entropy.m

```
function ent=entropy(p)
np=size(p);
if length(np)>1,
    p=reshape(p,prod(np),1);
end
ip=find(and(p>0,p<1));
pp=p(ip)/sum(p(ip));
hhp=-pp.*log2(pp);
ent=sum(hhp);
```

Matlab Function plotbinentropy.m

```
p=0.0025:0.0025:1-0.0025;
onep=1-p;
ent=-p.*log2(p)-onep.*log2(onep);
ent=[0 ent 0];
p=[0 p 1];
figure1=figure;
axes1 = axes('FontSize',16,'Parent',figure1);
title(axes1,'Binary Entropy Function');
xlabel(axes1,'Probability of One');
ylabel(axes1,'Entropy');
box(axes1,'on');
hold(axes1,'all');
plot(p,ent,'LineWidth',2)
```

Example 6: Entropy as Answer to Combinatorics Question, Lecture 1

- Assume $|\mathcal{X}| = m$.
- There are n trials.
- How many ways are there to get $k_1, k_2, \ldots k_m$ of the elements $(k_1+k_2+ \ldots +k_m=n)$?
- Operational role of entropy for a combinatorics question.

$$\binom{n}{k_1 k_2 \ldots k_m} = \frac{n!}{k_1!k_2!\ldots k_m!}$$

$$= 2^{n\left(-\frac{k_1}{n}\log\frac{k_1}{n}-\frac{k_2}{n}\log\frac{k_2}{n}\ldots-\frac{k_m}{n}\log\frac{k_m}{n}+o(n)\right)}$$

$$= 2^{nh\left(\frac{k_1}{n},\frac{k_2}{n},\ldots,\frac{k_m}{n}\right)+no(n)}$$

$\Rightarrow$ Theorem:

$$\frac{1}{n}\log\binom{n}{k_1 k_2 \ldots k_m} \xrightarrow[n\to\infty]{} h\left(p_1, p_2, \ldots, p_m\right)$$

$$\text{if } \frac{k_1}{n} \xrightarrow[n\to\infty]{} p_1, \frac{k_2}{n} \xrightarrow[n\to\infty]{} p_2, \ldots, \frac{k_m}{n} \xrightarrow[n\to\infty]{} p_m$$

Definitions

- The **joint entropy** of R.V.'s X and Y is:

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

$$= E\left[-\log p(X,Y)\right]$$

- The **conditional entropy** of Y given X is:

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

$$= E\left[-\log p(Y \mid X)\right]$$

Entropies

Theorem: $$H(X,Y) = H(X) + H(Y \mid X)$$
$$= H(Y) + H(X \mid Y)$$

Proof:

$$p(x,y) = p(x)p(y \mid x)$$

$$\Rightarrow \log p(x,y) = \log p(x) + \log p(y \mid x)$$

$$\Rightarrow \sum_{x,y} p(x,y) \log p(x,y) = \sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(y \mid x)$$

$$\Rightarrow \sum_{x,y} p(x,y) \log p(x,y) = \sum_{x} p(x) \log p(x) + \sum_{x,y} p(x,y) \log p(y \mid x)$$

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

Definition

- The **relative entropy** between probability distribution functions *p(x)* and *q(x)* is:

$$D(p \| q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p \left[\log \frac{p(X)}{q(X)} \right]$$

- Not a true distance:

$$D(p \| q) \neq D(q \| p)$$

Matlab Function relentropy.m

```
function relent=relentropy(p,q)
np=size(p);
nq=size(q);
if np~=nq,
    errormess='Matlab function relentropy error:  dim mismatch'
    return
end
if length(np)>1,
    p=reshape(p,prod(np),1);
    q=reshape(q,prod(np),1);
end
ip=find(and(p>0,q>0));
rpq=p(ip).*log2(p(ip)./q(ip));
% Gives the wrong answer if q(k)=0 and p(k)~=0
relent=sum(rpq);
```

Matlab Function plotrelentropy.m

```
function re=plotrelentropy(q);
p=0.0025:0.0025:1-0.0025;
onep=1-p;
re=p.*log2(p/q)+onep.*log2(onep/(1-q));
re=[-log2(1-q) re -log2(q)];
p=[0 p 1];
figure1=figure;
axes1 = axes('FontSize',16,'Parent',figure1);
title(axes1,strcat('Binary Relative Entropy q=',num2str(q)));
xlabel(axes1,'Probability p');
ylabel(axes1,'Relative Entropy D(p||q)');
box(axes1,'on');
hold(axes1,'all');
plot(p,re,'LineWidth',2)
grid
```

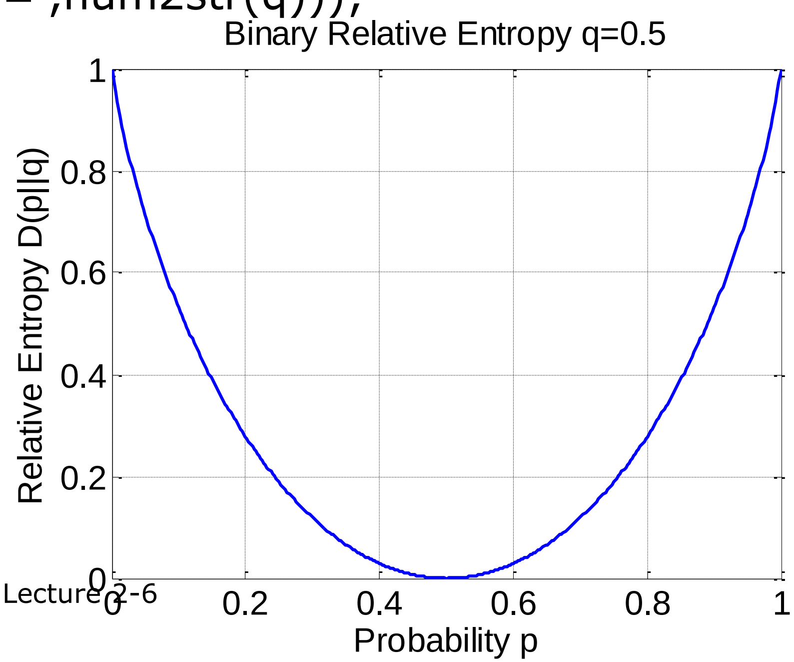
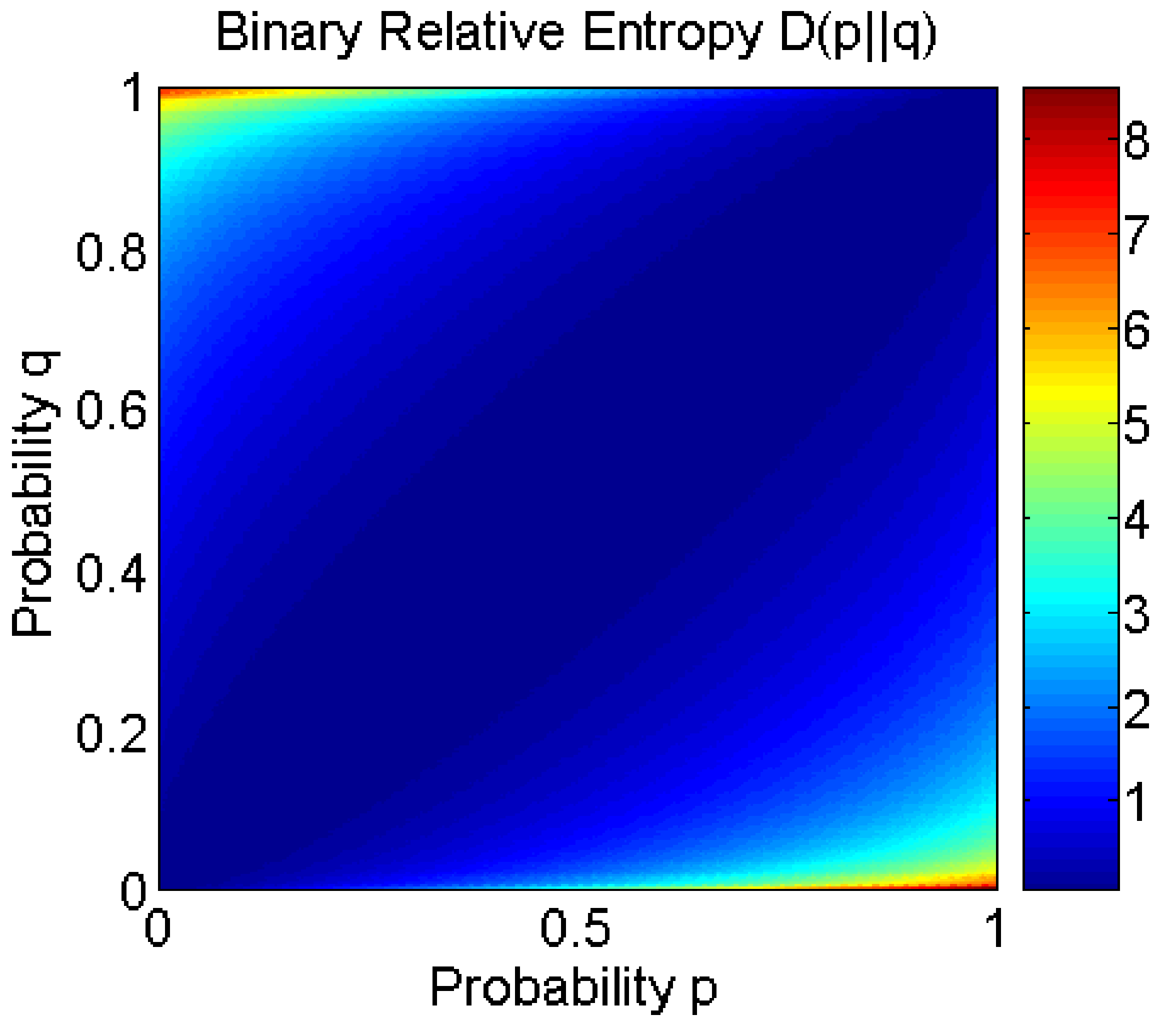


Image of Relative Entropy Function



Definition

- The **mutual information** between X and Y is:

$$I(X;Y) = D(p(x,y) \| p(x)p(y))$$

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

- Some Properties:

1) $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
2) $I(X;Y) = H(X) + H(Y) - H(X,Y)$
3) $I(X;Y) = I(Y;X)$
4) $I(X;Y) \geq 0$

Properties of Mutual Information

1) $I(X;Y) = H(X) - H(X \mid Y)$

$$= H(Y) - H(Y \mid X)$$

Proof: $I(X;Y) = E\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right]$

$$= E\left[\log \frac{1}{p(X)}\right] + E\left[\log p(X \mid Y)\right]$$

$$= E\left[\log \frac{1}{p(X)}\right] - E\left[\log \frac{1}{p(X \mid Y)}\right]$$

$$= H(X) - H(X \mid Y)$$

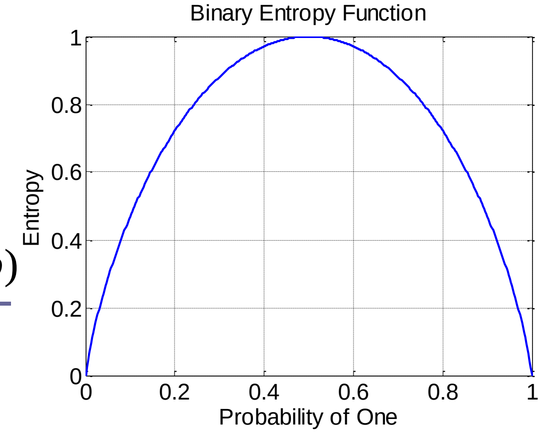
Matlab Function mutualinformation.m

```
function info=mutualinformation(p)
p=p/sum(sum(p));
px=sum(p,2);
py=sum(p,1);
info=entropy(px)+entropy(py)-entropy(p);
```

$$
\begin{aligned}
2)\ I(X;Y) &= H(X) - H(X|Y) \\
&= H(X) - \left[H(X,Y) - H(Y)\right] \\
&= H(X) + H(Y) - H(X,Y)
\end{aligned}
$$

Last Class Outline

$$(p,1-p) \Rightarrow$$
$$H(p) = -p\log p - (1-p)\log(1-p)$$



- Entropy $$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$$
- Joint Entropy $$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$
- Conditional Entropy $$H(X\,|\,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x\,|\,y)$$
- Relative Entropy $$D(p\,\|\,q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}$$
- Mutual Information $$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

Example: Entropy and Mutual Information

$$p(x,y) = \begin{bmatrix} \frac{1}{8} & \frac{1}{4} & \frac{1}{8} \\ 0 & \frac{1}{4} & \frac{1}{8} \\ 0 & 0 & \frac{1}{8} \end{bmatrix}; \text{ values of } x \text{ in columns, } y \text{ in rows}$$

$$p(x) = \begin{bmatrix} \frac{1}{8} & \frac{1}{2} & \frac{3}{8} \end{bmatrix}; \quad p(y) = \begin{bmatrix} \frac{1}{2} & \frac{3}{8} & \frac{1}{8} \end{bmatrix}$$

$$H(X) = H(Y) = -\frac{1}{8}\log\frac{1}{8} - \frac{1}{2}\log\frac{1}{2} - \frac{3}{8}\log\frac{3}{8} = 2 - \frac{3}{8}\log 3$$

$$H(X,Y) = -4\left(\frac{1}{8}\log\frac{1}{8}\right) - 2\left(\frac{1}{4}\log\frac{1}{4}\right) = 2.5$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = 1.5 - 0.75\log 3 = 0.311278124$$

Telescoping Sums: Entropy

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

$$-E[\log p(X_1, X_2, \ldots X_n)]$$
$$= -E[\log(p(X_1)p(X_2 | X_1)p(X_3 | X_2, X_1)\ldots p(X_n | X_{n-1}, \ldots, X_1))]$$
$$= -\sum_{i=1}^{n} E[\log p(X_i | X_{i-1}, \ldots, X_1)]$$

- Comments:
  - Generalization of two variable case $H(X_1, X_2) = H(X_1) + H(X_2 | X_1)$
  - Example for three variables

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2 | X_1) + H(X_3 | X_2, X_1)$$

Telescoping Sums: Mutual Information

- Definition: $I(X;Y\mid Z)=H(X\mid Z)-H(X\mid Y,Z)$

- Theorem: $I(X_1,X_2,\ldots X_n;Y)=\sum_{i=1}^{n} I(X_i;Y\mid X_{i-1},\ldots X_1)$

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

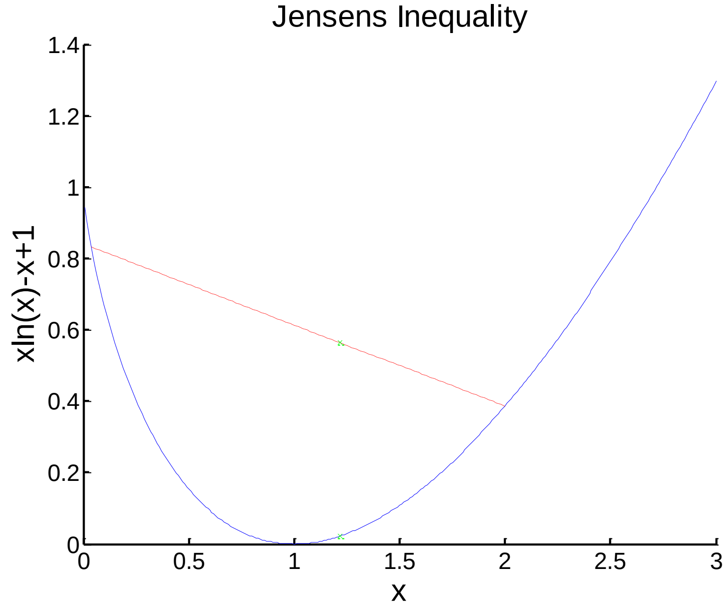
$$H(X_1,X_2,\ldots,X_n\mid Y)=\sum_{i=1}^{n} H(X_i\mid X_{i-1},\ldots X_1,Y)$$

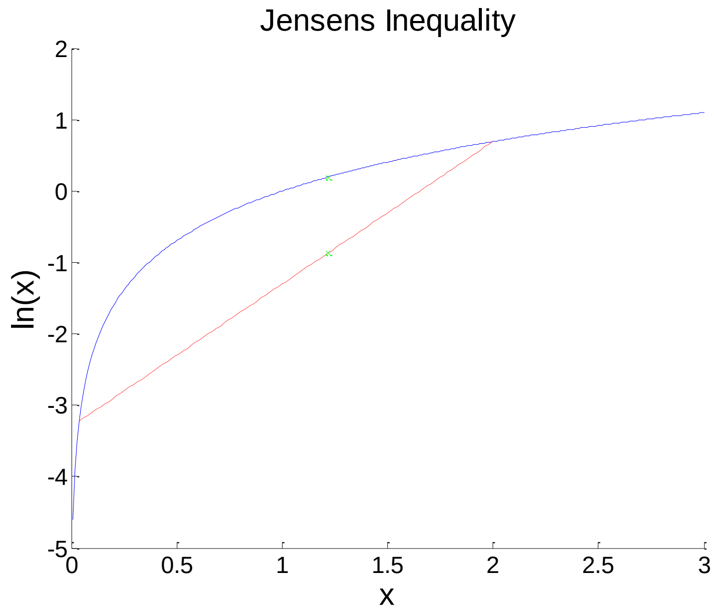
Towards Jensen's Inequality:
Convex and Concave Functions

- Definition: A function f is *convex* over (a,b) if for any x_1,x_2 ∈ (a,b) and λ ∈ [0,1],

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$$

- A function f is concave if $-f$ is convex. f is strictly convex or concave if the inequalities are strict for λ ≠ 0 or 1.





Towards Jensen's Inequality: Sufficient Condition for Convexity

- Theorem: Suppose that f is twice continuously differentiable. If d^2f/dx^2 is nonnegative (positive) everywhere, then f is convex (strictly convex).
- Proof: Using a Taylor series approximation,

$$f(x) = f(x_0) + \frac{df}{dx}(x_0)(x - x_o) + \frac{1}{2}\frac{d^2 f}{dx^2}(x^*)(x - x_o)^2$$

where x^* is some value between x and x_0. Take

$$x_0 = \lambda x_1 + (1-\lambda)x_2$$

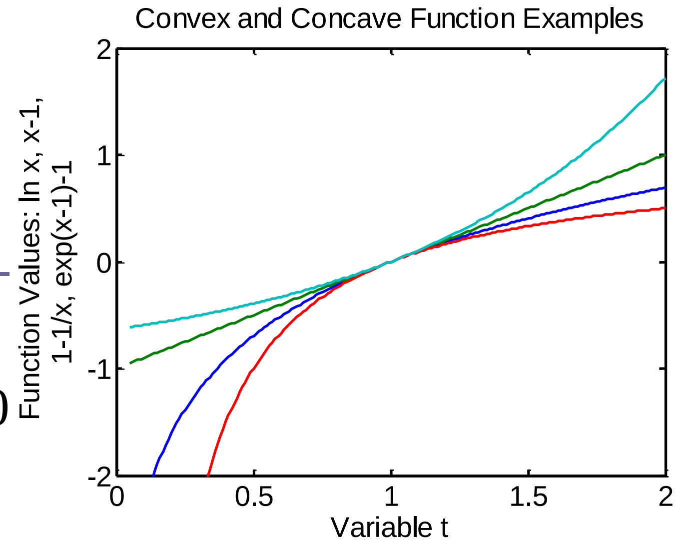
$$f(x_1) \geq f(x_0) + \frac{df}{dx}(x_0)(x_1 - \lambda x_1 - (1-\lambda)x_2)$$

$$f(x_2) \geq f(x_0) + \frac{df}{dx}(x_0)(-\lambda x_1 + \lambda x_2)$$

$$x_2 - x_0 = \lambda(x_2 - x_1)$$

$$\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(x_0) = f\left(\lambda x_1 + (1-\lambda)x_2\right)$$

Convex and Concave Function Examples



- $f(x) = \log x$ is strictly concave.
  - Proof: $df/dx = \log e/x$; $d^2f/dx^2 = -\log e/x^2 < 0$
- $f(x) = -x \log x$ is strictly concave.
  - Proof: $df/dx = -\log x - \log e$; $d^2f/dx^2 = -\log e/x < 0$
  - Comment: This implies concavity of entropy $H(X) = -\Sigma\, p(x) \log p(x)$
- x^m for $m \geq 1$ is convex for $x > 0$
- e^x is strictly convex
- Comment: several information inequalities can be derived from $x - 1 \geq \ln x \geq 1 - \frac{1}{x}$
- Relative entropy is nonnegative $$\boxed{D(p \,\|\, q) = E_p\left[\log \frac{p(X)}{q(X)}\right] \geq \log e\, E_p\left[1 - \frac{q(X)}{p(X)}\right] = 0}$$

Jensen's Inequality

- **Theorem (Jensen's Inequality)**: If f is convex over (a,b) and X is a random variable taking values in (a,b), then

  $$E[f(X)] \geq f(E[X])$$

  If f is strictly convex, then equality implies that $X=E[X]$ with probability one.

Proof of Jensen's Inequality

Proof by induction. Let $|\mathcal{X}| = 2$, $\mathcal{X} = \{x_1, x_2\}$. Then

$pf(x_1) + (1-p)f(x_2) \geq f(px_1 + (1-p)x_2)$, by definition.

Assume $|\mathcal{X}| = k$ and that for any set of cardinality k -1 the theorem holds. Then

$$\sum_{i=1}^{k} p_i f(x_i) = p_k f(x_k) + (1-p_k)\sum_{i=1}^{k-1} \frac{p_i}{1-p_k} f(x_i)$$

$$\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} \frac{p_i}{1-p_k} x_i\right), \text{ by induction hypothesis}$$

$$\geq f\left(p_k x_k + (1-p_k)\left(\sum_{i=1}^{k-1} \frac{p_i}{1-p_k} x_i\right)\right) = f(E[X]), \text{ by definition.} \quad \square$$

Relative Entropy is Nonnegative

- **Theorem**: $D(p||q) \geq 0$ with equality if and only if (iff) $p = q$.
- Proof uses Jensen's inequality. The function $\log x$ is strictly concave so $-\log x$ is strictly convex.

$$D(p\,||\,q) = E_p\left[\log\frac{p(X)}{q(X)}\right] = E_p\left[-\log\frac{q(X)}{p(X)}\right]$$

$$\geq -\log\left[E_p\left(\frac{q(X)}{p(X)}\right)\right] = -\log(1) = 0$$

Refinement: Need to restrict sums to $A = \{x \in \mathcal{X} \mid p(x) > 0\}$

Relative Entropy is Nonnegative

- **Theorem**: $D(p\|q) \geq 0$ with equality iff $p = q$.
- Proof uses Jensen's inequality. $-\log x$ is strictly convex.

$$D(p\,\|\,q) = E_p\left[\log\frac{p(X)}{q(X)}\right] = \sum_{x\in\mathcal{A}} p(x)\left[-\log\frac{q(x)}{p(x)}\right]$$

$$\underset{(a)}{\geq} -\log\left[\sum_{x\in\mathcal{A}} p(x)\left(\frac{q(x)}{p(x)}\right)\right] = -\log\left[\sum_{x\in\mathcal{A}} q(x)\right]$$

$$\underset{(b)}{\geq} -\log(1) = 0$$

where $\mathcal{A} = \{x \in \mathcal{X} \mid p(x) > 0\}$,

(a) follows from Jensen's Inequality, and

(b) follows from $\sum_{x\in\mathcal{A}} q(x) \leq 1.$

Start Here Sept. 8, 2011
Outline

- Concavity of entropy
- Log-sum inequality
- Convexity of relative entropy
- Conditioning reduces entropy
- Convexity and concavity of mutual information (toward optimization)
- Data processing inequality
- Chapter 3: Asymptotic equipartition property

Entropy is Concave and Bounded

- **Theorem**: Entropy is concave and bounded above by the log of the cardinality of the set, with equality iff the random variable is uniformly distributed.
- Proof: Concavity follows from concavity of $-x\log x$.

$$H(X) = E_p\left[\log\frac{1}{p(X)}\right] \underset{\uparrow}{\leq} \log\left[E_p\frac{1}{p(X)}\right] = \log\left[\sum_{x\in\mathcal{X}}\frac{p(x)}{p(x)}\right] = \log|\mathcal{X}|$$

$$\overbrace{\text{equality iff } \frac{1}{p(x)} = \text{constant} = |\mathcal{X}|}$$

Log Sum Inequality

Theorem: For any nonnegative numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$, with $\sum_{i=1}^{n} b_i > 0$. Assume that if $b_i = 0$ then $a_i = 0$ $\left(0 \log \frac{0}{0} = 0\right)$. Then

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\left(\sum_{i=1}^{n} a_i\right)}{\left(\sum_{i=1}^{n} b_i\right)}$$

Equality iff a_i/b_i = constant

Proof: By Jensen's inequality

$$\sum_{i=1}^{n} \underbrace{\frac{b_i}{\left(\sum_{l=1}^{n} b_l\right)}}_{\text{Expected value}} \underbrace{\left(\frac{a_i}{b_i} \log \frac{a_i}{b_i}\right)}_{t \log t \text{ is convex}} \geq \left(\sum_{i=1}^{n} \frac{b_i}{\sum_{l=1}^{n} b_l} \frac{a_i}{b_i}\right) \log \left(\sum_{i=1}^{n} \frac{b_i}{\sum_{l=1}^{n} b_l} \frac{a_i}{b_i}\right) . \square$$

Convexity of Relative Entropy

- **Theorem**: $D(p||q)$ is convex in the pair (p,q).
- Proof: By the log sum inequality

$$D(\lambda p_1 + (1-\lambda) p_2 \,||\, \lambda q_1 + (1-\lambda) q_2) =$$

$$\sum_{x \in \mathcal{X}} [\lambda p_1(x) + (1-\lambda) p_2(x)] \log \frac{\lambda p_1(x) + (1-\lambda) p_2(x)}{\lambda q_1(x) + (1-\lambda) q_2(x)} \leq$$

Log-sum inequality

$$\lambda \sum_{x \in \mathcal{X}} p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1-\lambda) \sum_{x \in \mathcal{X}} p_2(x) \log \frac{p_2(x)}{q_2(x)} =$$

$$\lambda D(p_1 \,||\, q_1) + (1-\lambda) D(p_2 \,||\, q_2) \qquad \square$$

Concavity of Entropy Revisited

- Let $u(x)$ be a uniform distribution. Then

  $$H(p) = \log|\mathcal{X}| - D(p\|u)$$

  Proof:

  $$D(p\|u) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{1/|\mathcal{X}|}$$

  $$= \log|\mathcal{X}| + \sum_{x\in\mathcal{X}} p(x)\log p(x)$$

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

- Convexity of relative entropy implies concavity of entropy.

Jensen's Inequality Summary

- **Theorem (Jensen's Inequality)**: If f is convex over (a,b) and X is a random variable taking values in (a,b), then
$$E[f(X)] \geq f(E[X])$$
If f is strictly convex, then equality implies that $X = \mathrm{E}[X]$ with probability one.
- **Corollary**: $\mathrm{D}(p\|q) \geq 0$ with equality iff $p = q$.
- **Corollary**: $\mathrm{I}(X;Y) \geq 0$, with equality iff $p(x,y) = p(x)p(y)$; that is, iff X and Y are independent.
- **Corollary**: Conditioning reduces entropy.
$\mathrm{I}(X;Y) \geq 0 \Rightarrow \mathrm{H}(X) \geq \mathrm{H}(X|Y)$.
- *Comment*: We often use this corollary in proofs.

Mutual Information Concavity and Convexity Motivation

- Channel capacity and its computation
  - Maximize mutual information over input probability distribution
  - Maximization problems are better-behaved for concave functions
  - To show: mutual information is concave in the input probability distribution
- Rate-distortion functions and their computation
  - Minimize mutual information over channel transition probabilities
  - Minimization problems are better-behaved for convex functions
  - To show: mutual information is convex in the channel probabilities
- Computations and properties of mutual information in multiterminal information theory
  - Current research problems

Mutual Information

- View mutual information as a function of *p(x)* and of *p(y|x)*. Then mutual information is
  - a concave function of *p(x)* (for *p(y|x)* fixed) and
  - a convex function of *p(y|x)* (for *p(x)* fixed).
- Proofs follows from concavity of entropy and convexity of relative entropy.

$$I(X;Y) = H(Y) - \sum_{x \in X} p(x) H(Y \mid X = x)$$

Concavity of *H(Y)* → concavity wrt *p(x)*

$$I(X;Y) = D(p(x,y) \| p(x)p(y))$$

Consider $p(y \mid x) = \lambda p_1(y \mid x) + (1-\lambda) p_2(y \mid x)$

$$p(x,y) = p(x)\left[\lambda p_1(y \mid x) + (1-\lambda) p_2(y \mid x)\right]$$

$$= \lambda p_1(x,y) + (1-\lambda) p_2(x,y)$$

$$p(y) = \lambda p_1(y) + (1-\lambda) p_2(y)$$

Convexity of relative entropy (log-sum inequality)

$$D(p(x,y) \| p(x)p(y)) \le \lambda D(p_1(y \mid x)p(x) \| p(x)p_1(y)) + (1-\lambda) D(p_2(x,y) \| p(x)p_2(y))$$

Data Processing Inequality

- Definition: The random variables X, Y, and Z form a Markov chain in that order if $p(z|x,y) = p(z|y)$.
- Then $p(x,y,z) = p(x)p(y|x)p(z|y)$. Also, X and Z are conditionally independent given Y.

$$p(x,z\,|\,y) = \frac{p(x,y)\,p(z\,|\,y)}{p(y)} = p(x\,|\,y)\,p(z\,|\,y)$$

- Write $X \rightarrow Y \rightarrow Z$.

Data Processing Inequality

- **Theorem**: If $X \rightarrow Y \rightarrow Z$, then $\mathrm{I}(X;Y) \geq \mathrm{I}(X;Z)$.
- Note that this says Y gives more information about X than Z does.
- Proof: $$\begin{aligned} I(X;Y,Z) &= I(X;Y) + I(X;Z\,|\,Y) \\ &= I(X;Z) + I(X;Y\,|\,Z) \end{aligned}$$
- But $\mathrm{I}(X;Z|Y) = 0$, so $\mathrm{I}(X;Y) \geq \mathrm{I}(X;Z)$.
- Comment: $$H(X) - H(X\,|\,Y) \geq H(X) - H(X\,|\,Z)$$ $$H(X\,|\,Z) \geq H(X\,|\,Y)$$

Chapter 3: Asymptotic Equipartition Property

- Strong law of large numbers → weak law
- Asymptotic equipartition property (AEP)
  - All highly likely sequences are equally likely
  - The set of highly likely sequences is the typical set
  - The cardinality of the typical set is determined by entropy
- Data compression result:
  - Number of bits required to represent sequences on average equals entropy times the length of the sequence
  - Number of bits per symbol, on average, equals entropy

Theorem (Strong Law of Large Numbers)

- Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. RVs. Let $f: \mathcal{X} \rightarrow \mathcal{R}$ be an arbitrary function such that $E[|f(X)|]$ is finite. Then
$$\lim_{n\to\infty} \frac{1}{n}\left[\sum_{i=1}^{n} f(X_i)\right] = E\left[f(X)\right]$$
with probability one. If the variance of $f(X)$ is finite, this convergence is in the mean also.
- Comment: In either event, we get convergence in probability
$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - E\left[f(X)\right]\right| > \varepsilon\right) \rightarrow 0 \text{ as } n \rightarrow \infty$$

In fact (3.1) $$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - E\left[f(X)\right]\right| > \varepsilon\right) \le \frac{\sigma^2}{n\varepsilon^2} \text{ where } \sigma^2 = \text{var } f(X)$$

Theorem

- If $X_1, X_2, \ldots, X_n$ are i.i.d. with distribution $p(x)$, then

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \rightarrow H(X) \text{ in probability.}$$

- Proof: $p(X_1, X_2, \ldots, X_n) = p(X_1)\, p(X_2) \ldots p(X_n)$

Set $f(x) = -\log p(x)$ in the previous theorem.

$$E\left[f(X)\right] = E\left[\log \frac{1}{p(X)}\right] = H(X).$$

Comment: Again $-\log p(x)$ is a function of the realization.

Typical Sets

- **Definition**: The *typical set* is

$$\mathcal{A}_\varepsilon^{(n)} = \left\{ \left(x_1, x_2, \ldots, x_n\right) \in \mathcal{X}^n : \; \left| -\frac{1}{n} \log p\left(x_1, x_2, \ldots, x_n\right) - H(X) \right| \leq \varepsilon \right\}$$

- **Comment**: This is the set of sequences whose normalized log-probability is close to entropy.
- **Theorem**

$$2^{-n\left(H(X)+\varepsilon\right)} \leq p(x) \leq 2^{-n\left(H(X)-\varepsilon\right)} \quad \text{for } x \in \mathcal{A}_\varepsilon^{(n)}$$

$$P\left\{ X \in \mathcal{A}_\varepsilon^{(n)} \right\} > 1-\varepsilon \text{ for } n \text{ sufficiently large}$$

$$\left| \mathcal{A}_\varepsilon^{(n)} \right| \leq 2^{n\left(H(X)+\varepsilon\right)}$$

$$\left| \mathcal{A}_\varepsilon^{(n)} \right| \geq \left(1-\varepsilon\right) 2^{n\left(H(X)-\varepsilon\right)} \text{ for } n \text{ sufficiently large}$$

Typical Sets

- **Definition**: The *typical set* is

$$\mathcal{A}_\varepsilon^{(n)} = \left\{ \left(x_1, x_2, \ldots, x_n\right) \in \mathcal{X}^n : \ \left| -\frac{1}{n} \log p\left(x_1, x_2, \ldots, x_n\right) - H(X) \right| \le \varepsilon \right\}$$

- **Comment**: This is the set of sequences whose normalized log-probability is close to entropy.
- **Theorem**

$$2^{-n\left(H(X)+\varepsilon\right)} \le p(x) \le 2^{-n\left(H(X)-\varepsilon\right)} \quad \text{for } x \in \mathcal{A}_\varepsilon^{(n)}$$

$$P\left\{ X \in \mathcal{A}_\varepsilon^{(n)} \right\} > 1-\delta \text{ for } n \text{ sufficiently large}$$

$$\left| \mathcal{A}_\varepsilon^{(n)} \right| \le 2^{n\left(H(X)+\varepsilon\right)}$$

$$\left| \mathcal{A}_\varepsilon^{(n)} \right| \ge \left(1-\delta\right) 2^{n\left(H(X)-\varepsilon\right)} \text{ for } n \text{ sufficiently large}$$

Proof

- First line is definition of typical set.
- Second line follows from previous theorem.
- Third and fourth lines:

$$1 = \sum_{x \in \mathcal{X}^n} p(x) \geq \sum_{x \in \mathcal{A}_\varepsilon^{(n)}} p(x)$$

$$\geq \sum_{x \in \mathcal{A}_\varepsilon^{(n)}} 2^{-n\left(H(X)+\varepsilon\right)} = \left|A_\varepsilon^{(n)}\right| 2^{-n\left(H(X)+\varepsilon\right)}$$

$$\Rightarrow \left|\mathcal{A}_\varepsilon^{(n)}\right| \leq 2^{n\left(H(X)+\varepsilon\right)}$$

For the fourth line,

$P\left\{X \in \mathcal{A}_\varepsilon^{(n)}\right\} > 1 - \delta$ for n sufficiently large $\Rightarrow$

$$1 - \delta < \sum_{x \in \mathcal{A}_\varepsilon^{(n)}} p(x) \leq \left|\mathcal{A}_\varepsilon^{(n)}\right| 2^{-n\left(H(X)-\varepsilon\right)}$$

Typical Sets and the AEP

- **Theorem** $2^{-n(H(X)+\varepsilon)} \le p(x) \le 2^{-n(H(X)-\varepsilon)}$ for $x \in \mathcal{A}_\varepsilon^{(n)}$

  $P\left\{X \in \mathcal{A}_\varepsilon^{(n)}\right\} > 1-\delta$ for n sufficiently large

  $\left|\mathcal{A}_\varepsilon^{(n)}\right| \le 2^{n(H(X)+\varepsilon)}$

  $\left|\mathcal{A}_\varepsilon^{(n)}\right| \ge (1-\delta)2^{n(H(X)-\varepsilon)}$ for n sufficiently large

- Comments:
  - The typical set has probability arbitrarily close to 1.
  - The log-cardinality of the typical set is upper bounded by entropy plus ε
  - The log-cardinality is lower bounded by entropy minus ε (for n large enough)

$$H(X)+\varepsilon \ge \frac{1}{n}\log\left|\mathcal{A}_\varepsilon^{(n)}\right| \ge H(X)-\varepsilon+\frac{1}{n}\log(1-\delta) = H(X)-\varepsilon'$$

Data Compression

- **Idea**: Partition all outcomes $\mathcal{X}^n$ into the typical and nontypical sets for some ε. Design a reasonable code for the typical set and do anything else for the rest.
- **Definition**: A binary code is a mapping from $\mathcal{X}^n$ to binary sequences.
- **Theorem**: Let X_i be i.i.d. with probability distribution $p(x)$ and let $\varepsilon>0$. Then there exists a binary code that is one-to-one and

$$E\left[\frac{1}{n} l(X^n)\right] \le H(X) + \varepsilon \text{ for } n \text{ sufficiently large,}$$

where $l(\mathbf{x})$ is the length of a binary codeword assigned to $\mathbf{x}$.

Proof

- To every sequence in the typical set, assign a codeword of length less than or equal to $n(H(X)+\varepsilon)+1$.
- To every sequence not in the typical set, assign a codeword of length less than or equal to $n\log|\mathcal{X}|+1$
- Then the expected length satisfies

$$E\left[\frac{1}{n}l(X^n)\right] \le H(X)+\varepsilon+\frac{1}{n}+\delta\log|\mathcal{X}|+\frac{1}{n}$$
$$=H(X)+\varepsilon'$$
$$\varepsilon'=\varepsilon+\frac{2}{n}+\delta\log|\mathcal{X}|$$

Chapter 4 Outline

- Entropy Rates of Stochastic Processes
- Two expressions: equal for stationary processes
- Markov chains
  - entropy rates
- Next Class: Markov chains
  - decreasing conditional entropy
  - second law of thermodynamics

Entropy Rates of a Stochastic Process

- Entropy rates in bits per symbol
- Stochastic process: $X_1, X_2, \ldots, X_n, \ldots$ a random sequence

  X_i is a RV; $x_i \in \mathcal{X}$; possibly confusing notation $P(X_i = x_i)$
- Structure of the random sequence must be assumed to make progress
- Definition: A stochastic process $X_1, X_2, \ldots, X_n, \ldots$ is stationary if the joint distribution is invariant to shifts; for all $l \geq 0$,

$$P\{X_1 = \alpha, X_2 = \beta, \ldots, X_n = \gamma\} = P\{X_{1+l} = \alpha, X_{2+l} = \beta, \ldots, X_{n+l} = \gamma\}$$

Entropy Rates

- **Definition**: The entropy rate of a stochastic process $\{X_i\}$ is

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$

when the limit exists.

- **Proposition**: If X_i are i.i.d., then $H(\mathcal{X}) = H(X_1)$
- **Proof**: $H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i) = nH(X_1)$
- **Comments**:
  - If X_i are independent, but not identically distributed, the first equality holds. However the limit

    $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i)$ may or may not exist

  - A second possible definition for entropy rate is

    $H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1)$, when the limit exists.

Entropy Rates

- **Theorem**: For a stationary stochastic process, $H(\mathcal{X})$ and $H'(\mathcal{X})$ exist and are equal.
- **Proof**: There are three parts: $H'(\mathcal{X})$ exists; a technical result (Cesáro mean); and $H(\mathcal{X})$ exists and equals $H'(\mathcal{X})$.
- $H'(\mathcal{X})$ exists:

$$
\begin{aligned}
0 &\le H(X_n \mid X_{n-1}, X_{n-2}, \ldots X_1) \\
&\le H(X_n \mid X_{n-1}, X_{n-2}, \ldots X_2) \text{ conditioning reduces entropy} \\
&= H(X_{n-1} \mid X_{n-2}, X_{n-3}, \ldots X_1) \text{ by stationarity}
\end{aligned}
$$

- Thus $H(X_n \mid X_{n-1}, X_{n-2}, \ldots X_1)$ is a nonincreasing sequence of nonnegative numbers. Thus it has a limit.

Proof continued

- Cesáro mean: If $a_n \to a$ and $b_n=(a_1+a_2+\ldots a_n)/n$, then $b_n \to a$.

- Completion:

$$\frac{1}{n}H(X_1,X_2,\ldots,X_n)=\frac{1}{n}\sum_{i=1}^{n}H(X_i\,|\,X_{i-1},X_{i-2},\ldots,X_1)$$

$$b_n=\frac{1}{n}\sum_{i=1}^{n}a_i$$

$$\text{Thus, } H(\mathcal{X})=\lim_{n\to\infty}\frac{1}{n}H(X_1,X_2,\ldots,X_n)$$

$$=\lim_{i\to\infty}H(X_i\,|\,X_{i-1},X_{i-2},\ldots,X_1)=H'(\mathcal{X})$$

Applications

- All results from Chapter 3 hold in this context, including definitions of typical sets, the AEP, and the data compression.
- Also

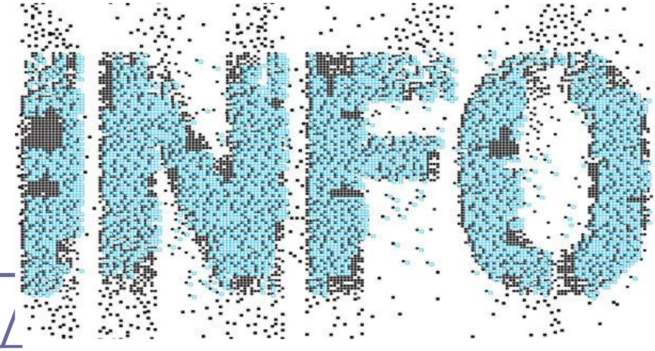
$$\frac{1}{n} H(X_1, X_2, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} H(X_i \mid X_{i-1}, X_{i-2}, \ldots, X_1)$$
$$\geq H(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1)$$

$$\frac{1}{n} H(X_1, X_2, \ldots, X_n) \leq \frac{1}{n-1} H(X_1, X_2, \ldots, X_{n-1})$$

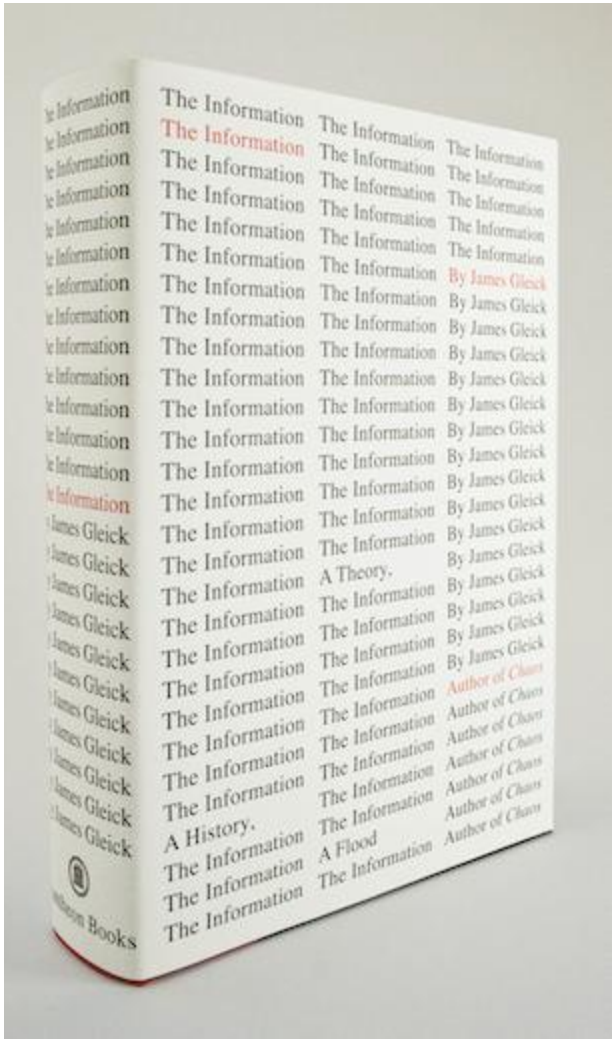
Outline September 15, 2011

- Markov Chain Properties, Classification
- Entropy rate of Markov chains
- Markov chains
  - Decreasing conditional entropy
  - Second law of thermodynamics

Information Diversion of the Day
James Gleick: *The Information: A History, a Theory, a Flood*

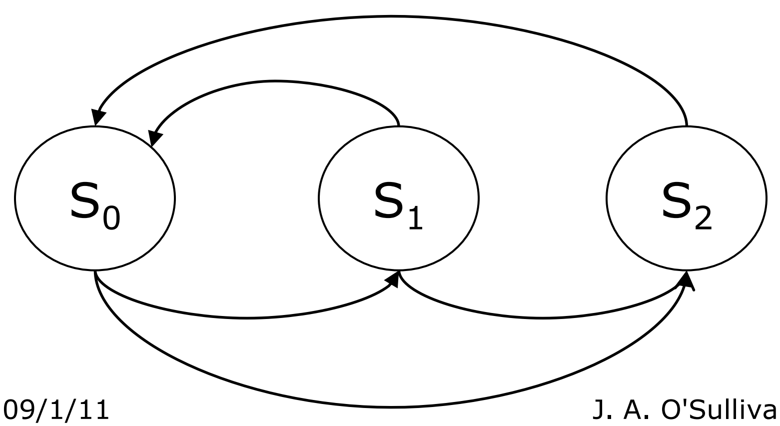


- http://www.thedailybeast.com/articles/2011/03/01/the-information-by-james-gleick-review-by-nicholas-carr.html
- http://boingboing.net/2011/03/24/james-gleicks-tour-d.html
- http://www.nytimes.com/2011/03/20/books/review/book-review-the-information-by-james-gleick.html?pagewanted=all
- http://around.com/the-information
- The Information *is so ambitious, illuminating and sexily theoretical that it will amount to aspirational reading for many of those who have the mettle to tackle it. Don't make the mistake of reading it quickly. Imagine luxuriating on a Wi-Fi-equipped desert island with Mr. Gleick's book, a search engine and no distractions.* The Information *is to the nature, history and significance of data what the beach is to sand.*
- —Janet Maslin, *The New York Times*

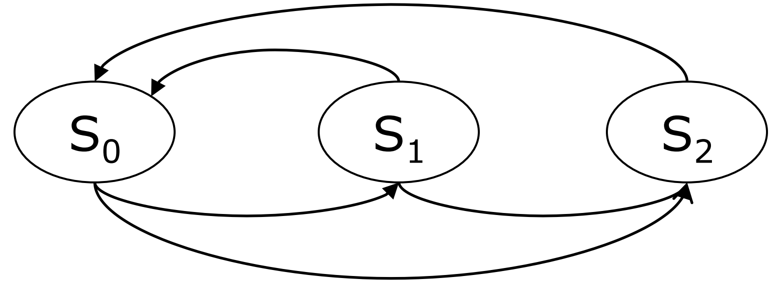


Markov Chains

- **Definition**: A stochastic process $\{X_i\}$ is a *Markov chain* if
$$P(X_{n+1} = x_{n+1} \mid X_n = x_n, \ldots, X_1 = x_1) = P(X_{n+1} = x_{n+1} \mid X_n = x_n)$$
- For a Markov chain,
$$p(x_1, x_2, \ldots, x_n) = p(x_1)\,p(x_2 \mid x_1) \ldots p(x_n \mid x_{n-1})$$
- **Definition**: A Markov chain is *time-invariant* if the transition probabilities do no depend on n.



Markov chain notation: time-invariant case



- X_n is called the state at time n.
  If $|\mathcal{X}|=m$ is finite, the probability transition matrix is

$$\mathbf{P} = \left\lfloor P(X_{n+1} = x_j \mid X_n = x_i) \right\rfloor$$

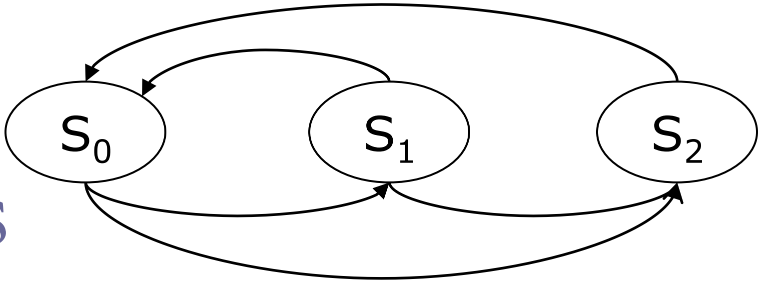
$$\mathbf{p}_n = \left[P(X_n = x_1) \quad P(X_n = x_2) \quad \ldots \quad P(X_n = x_m) \right]$$

$$\mathbf{p}_{n+1} = \mathbf{p}_n \mathbf{P}$$

If $\mathbf{p}_{n+1} = \mathbf{p}_n = \mu$, then μ is a stationary distribution.

If for all $n \geq 1$, $\mathbf{p}_n = \mu$, then the Markov chain is a stationary stochastic process.

Markov Chain Properties



- Definition: If for all i and j, there is a k such that

  $$\left(\mathbf{P}^k\right)_{i,j} > 0$$

  the Markov chain is *irreducible* (*connected*).
  If there is a k such that

  $$\left(\mathbf{P}^k\right)_{i,j} > 0$$

  for all i and j, the Markov chain is *strongly connected*
  (*irreducible and aperiodic*).
- Comment: strongly connected → irreducible (connected)

Three-State Example

- q=r=1➔irreducible, periodic with period 3
- q=1;r=0.5➔ irreducible and aperiodic (strongly connected)

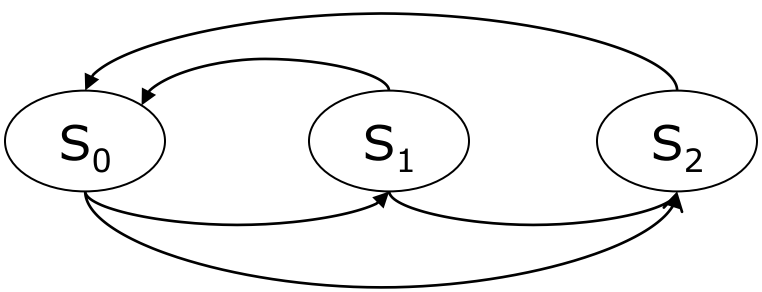
$$\mathbf{P} = \begin{bmatrix} 0 & q & 1-q \\ 1-r & 0 & r \\ 1 & 0 & 0 \end{bmatrix}$$

$$q = r = 1 \Rightarrow \quad \mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}; \quad \mathbf{P}^2 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix};$$

$$\mathbf{P}^3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \quad \mathbf{P}^4 = \mathbf{P}$$

$$q = 1; r = \frac{1}{2} \Rightarrow \quad \mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 \end{bmatrix}; \quad \mathbf{P}^2 = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 \end{bmatrix};$$

$$\mathbf{P}^3 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}; \quad \mathbf{P}^4 = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}; \quad \mathbf{P}^5 = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{3}{8} & \frac{1}{2} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}$$



Two-State Example

- Let $$\mathbf{P} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$
- If either $\alpha=0$ or $\beta=0$, the Markov chain is not connected. For $\alpha\neq0$ and $\beta\neq0$, the stationary distribution is $$\mu = \begin{bmatrix} \dfrac{\beta}{\alpha+\beta} & \dfrac{\alpha}{\alpha+\beta} \end{bmatrix}$$
- If $\alpha=\beta=1$, the Markov chain is connected, but not strongly connected.

Entropy rates of Markov chains

- **Theorem:** Let $\{X_i\}$ be a stationary Markov chain. Then the entropy rate is

$$H(\mathcal{X}) = -\sum_i \sum_j \mu_i P_{ij} \log P_{ij}$$

- **Proof:**

$$\begin{aligned}
H(\mathcal{X}) &= \lim_{n\to\infty} H(X_n \mid X_{n-1},\ldots,X_1) \\
&= H(X_n \mid X_{n-1}) \\
&= \sum_i P(X_{n-1}=x_i)\left[-\sum_j P(X_n = x_j \mid X_{n-1}=x_i)\log P(X_n = x_j \mid X_{n-1}=x_i)\right] \\
&= -\sum_i \sum_j \mu_i P_{ij} \log P_{ij}
\end{aligned}$$

Washington University in St. Louis
School of Engineering & Applied Science

Entropy rates of Markov chains

- **Theorem**: Let $\{X_i\}$ be a time-invariant Markov chain that is irreducible and aperiodic. Then the entropy rate is

$$H(\mathcal{X}) = -\sum_i \sum_j \mu_i P_{ij} \log P_{ij}$$

where μ is the stationary distribution.

- **Proof**:

$$\begin{aligned} H(\mathcal{X}) &= \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1) \\ &= \lim_{n\to\infty} H(X_n \mid X_{n-1}) \\ &= \lim_{n\to\infty} \sum_i P(X_{n-1} = x_i) \left[-\sum_j P_{ij} \log P_{ij} \right] \end{aligned}$$

Proof continued

- All that remains to be shown is that

$$\begin{cases} P(X = x_i) \to \mu_i \text{ as } n \to \infty, \text{ or} \\ \mathbf{p}_n \to \mu. \end{cases}$$

- Note that

$$\mathbf{p}_n = \mathbf{p}_{n-1}\mathbf{P} \text{ and } \mu = \mu\mathbf{P}. \text{ Thus,}$$

$$D(\mu \| \mathbf{p}_n) = \sum_{j=1}^{m} \mu_j \log \frac{\mu_j}{p_n(x_j)}$$

$$= \sum_{j=1}^{m} \left(\sum_{i=1}^{m} \mu_i P_{ij} \right) \log \frac{\left(\sum_{i=1}^{m} \mu_i P_{ij} \right)}{\left(\sum_{i=1}^{m} p_{n-1}(x_i) P_{ij} \right)}$$

$$\leq \sum_{j=1}^{m} \sum_{i=1}^{m} \mu_i P_{ij} \log \frac{\mu_i P_{ij}}{p_{n-1}(x_i) P_{ij}} = D(\mu \| \mathbf{p}_{n-1})$$

- The inequality is the log sum inequality; get equality iff $\mu_i P_{ij} = p_{n-1}(x_i) P_{ij}$ for all i and j, or $\mu_i = p_{n-1}(x_i)$ for P strongly connected.
- This is an information-theoretic proof of convergence.

Markov Chains and Time

- Let $X_1, X_2, \ldots, X_n \ldots$ be a Markov chain. Suppose that $p(x_n|x_{n-1})$ does not depend on n (time). Let μ_n be a distribution at time n. Then

1. The relative entropy between two distributions decreases with n
2. The relative entropy between a distribution and a stationary distribution decreases with n
3. Entropy increases with n if the stationary distribution is uniform (2nd Law of Thermodynamics)

The relative entropy between two distributions decreases with *n*.

Suppose two possible probability distributions at time n are given

$p_n(i) = P(X_n = x_i)$ and $\pi_n(i)$.

There are two corresponding probability distributions at time $n+1$

$$p_{n+1}(j) = \sum_{i=1}^{m} p_n(i) P_{ij} \quad \text{and} \quad \pi_{n+1}(j) = \sum_{i=1}^{m} \pi_n(i) P_{ij}.$$

The goal is to prove that $D(p_{n+1} \| \pi_{n+1}) \le D(p_n \| \pi_n)$. To show this,

$$D(p_n(i)P_{ij} \| \pi_n(i)P_{ij}) = \sum_{i=1}^{m}\sum_{j=1}^{m} p_n(i)P_{ij} \log \frac{p_n(i)P_{ij}}{\pi_n(i)P_{ij}}$$

$$= \sum_{i=1}^{m}\left[\left(\sum_{j=1}^{m} P_{ij}\right) p_n(i) \log \frac{p_n(i)}{\pi_n(i)}\right] = D(p_n \| \pi_n)$$

$$D(p_n(i)P_{ij} \| \pi_n(i)P_{ij}) = \sum_{i=1}^{m}\sum_{j=1}^{m} p_{n+1}(j) \frac{p_n(i)P_{ij}}{p_{n+1}(j)} \left[\log \frac{p_n(i)P_{ij} / p_{n+1}(j)}{\pi_n(i)P_{ij} / \pi_{n+1}(j)} + \log \frac{p_{n+1}(j)}{\pi_{n+1}(j)} \right]$$

$$\ge D(p_{n+1} \| \pi_{n+1})$$

The relative entropy between a distribution and a stationary distribution decreases with *n*

Let $\pi_n(i) = \mu_i$ be the stationary distribution. Then

$$\pi_{n+1}(j) = \sum_{i=1}^{m} \mu_i P_{ij} = \mu_j$$ and from the previous

result,

$$D(p_{n+1} \| \mu) \le D(p_n \| \mu).$$

2nd Law of Thermodynamics:

Entropy increases with *n* if the stationary distribution is uniform

Let $\pi_n(i) = \mu_i = \frac{1}{|\mathcal{X}|}$ be the stationary distribution. Then

$$D(p_n \| \mu) = \sum_{i=1}^{|\mathcal{X}|} p_n(i) \log \frac{p_n(i)}{1/|\mathcal{X}|} = \log|\mathcal{X}| - H(p_n)$$

$$D(p_{n+1} \| \mu) \le D(p_n \| \mu) \Rightarrow H(p_{n+1}) \ge H(p_n)$$