EE 231A: Information Theory
Lecture 2

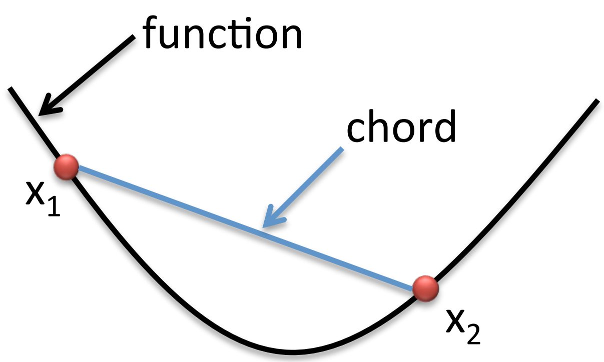
A. Convexity
B. Jensen's Inequality and its applications
C. Markov Chains and the Data Processing Inequality
D. Log-Sum inequality and its Applications

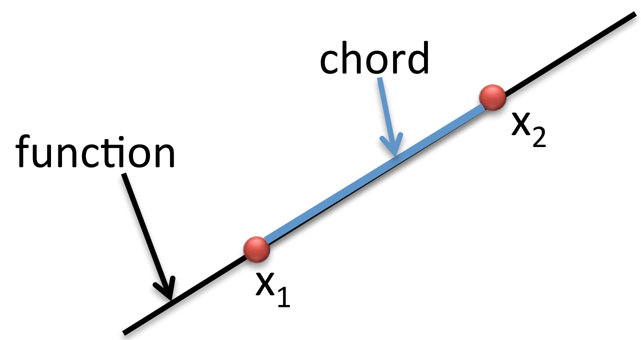
EE 231A: Information Theory
Lecture 2

A. Convexity
B. Jensen’s Inequality and its applications
C. Markov Chains and the Data Processing Inequality
D. Log-Sum inequality and its Applications

Convexity Definition

- A convex function lies on or *below* any chord

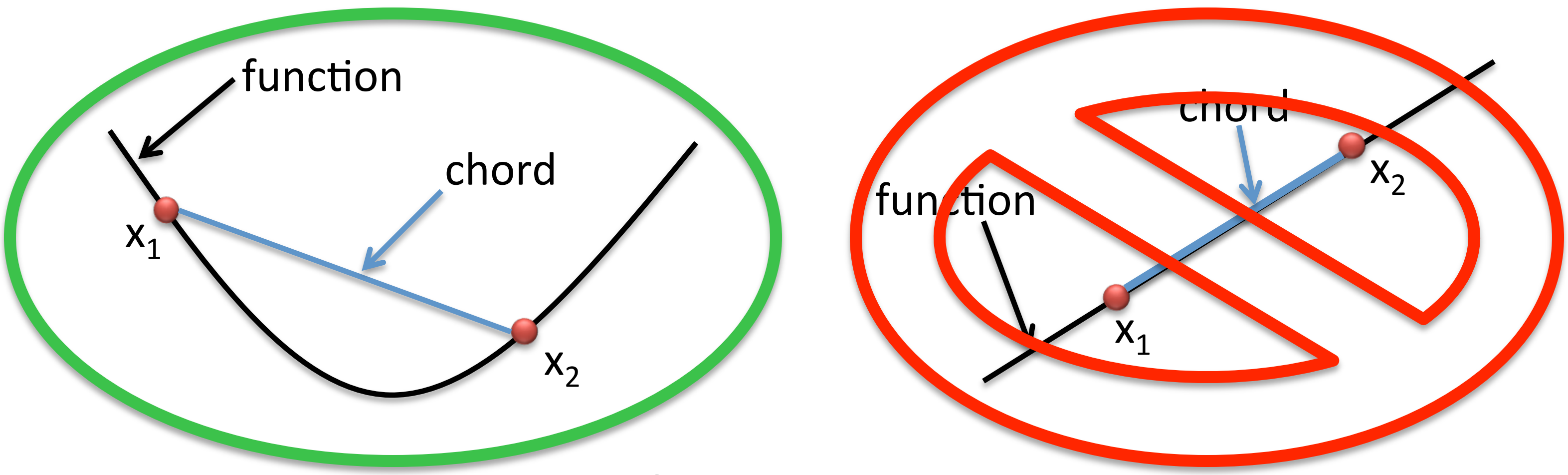




- A strictly convex function lies strictly *below* any chord except at the intersection.

Strict Convexity

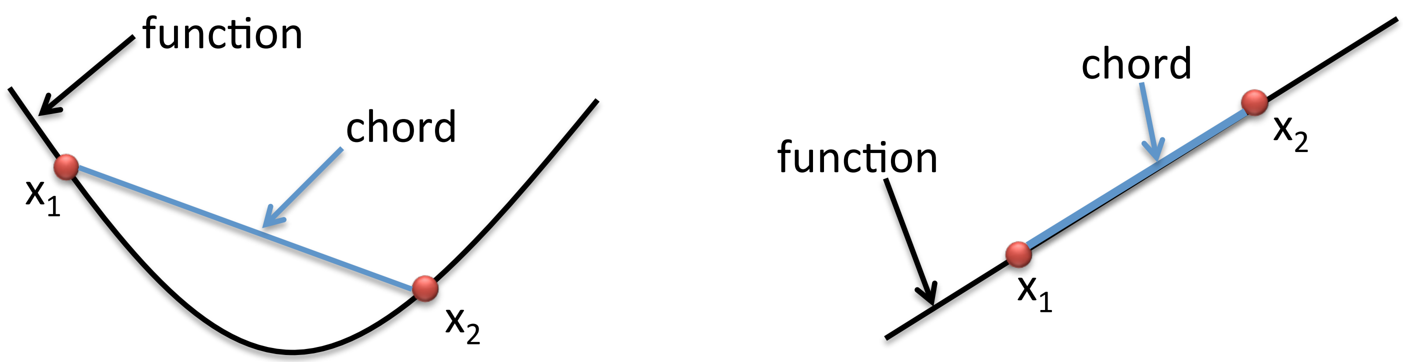
- A convex function lies on or *below* any chord



- A strictly convex function lies strictly *below* any chord except at the intersection.

Concavity

- A convex function lies on or *below* any chord



- A strictly convex function lies strictly *below* any chord except at the intersection.
- Concave, strictly concave: replace *below* with *above*.

Formal convexity definition

- $f(x)$ is convex over (a,b) if for every $x_1, x_2 \in (a,b)$
  $0 \le \lambda \le 1$

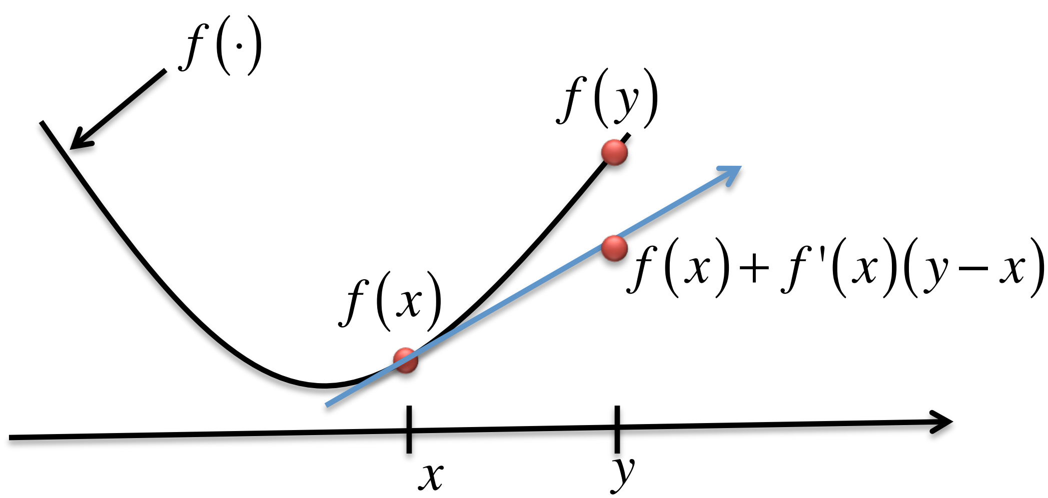
$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$$

- $f(x)$ is concave if $-f(x)$ is convex.

Convexity and the first derivative

- If $f(x)$ has a first derivative $f'(x)$, then the function is convex if and only if:

$$f(y) \geq f(x) + f'(x)(y - x)$$



Convexity and the second derivative

- If $f(x)$ has a second derivative that is non-negative everywhere, then the function is convex.

- If $f(x)$ has a second derivative that is positive everywhere, then the function is strictly convex.

Proof that $f''(x) \geq 0$ means $f(x)$ is convex.

$$f''(x) \geq 0 \Rightarrow \quad f(y) \geq f(x) + f'(x)(y - x)$$

Proof that $f''(x) \geq 0$ means $f(x)$ is convex.

$$f(y) = f(x) + \int_x^y f'(t)\,dt$$

Proof that $f''(x) \geq 0$ means $f(x)$ is convex.

$$f(y) = f(x) + \int_x^y f'(t)\,dt$$

$$= f(x) + \int_x^y \left[f'(x) + \int_x^t f''(u)\,du \right] dt$$

Proof that $f''(x) \geq 0$ means $f(x)$ is convex.

$$f(y) = f(x) + \int_x^y f'(t)\,dt$$

$$= f(x) + \int_x^y \left[f'(x) + \int_x^t f''(u)\,du \right] dt$$

$$= f(x) + \int^y f'(x)\,dt + \int^y \int^t f''(u)\,du\,dt$$

Proof that $f''(x) \geq 0$ means $f(x)$ is convex.

$$
\begin{aligned}
f(y) &= f(x) + \int_x^y f'(t)\,dt \\
&= f(x) + \int_x^y \left[f'(x) + \int_x^t f''(u)\,du \right] dt \\
&= f(x) + \int_x^y f'(x)\,dt + \int_x^y \int_x^t f''(u)\,du\,dt \\
&= f(x) + f'(x)(y-x) + \underbrace{\int_x^y \int_x^t f''(u)\,du\,dt}_{\geq 0}
\end{aligned}
$$

Proof that $f''(x) \geq 0$ means $f(x)$ is convex.

$$
\begin{aligned}
f(y) &= f(x) + \int_x^y f'(t)\,dt \\
&= f(x) + \int_x^y \left[f'(x) + \int_x^t f''(u)\,du \right] dt \\
&= f(x) + \int_x^y f'(x)\,dt + \int_x^y \int_x^t f''(u)\,du\,dt \\
&= f(x) + f'(x)(y-x) + \underbrace{\int_x^y \int_x^t f''(u)\,du\,dt}_{\geq 0} \\
&\geq f(x) + f'(x)(y-x)
\end{aligned}
$$

Examples of convexity

- $x \log x$ is convex for $x \geq 0$.

$$\frac{d}{dx} x \ln x = \ln x + 1$$

$$\frac{d^2}{dx^2} x \ln x = \frac{d}{dx}(\ln x + 1) = \frac{1}{x}$$

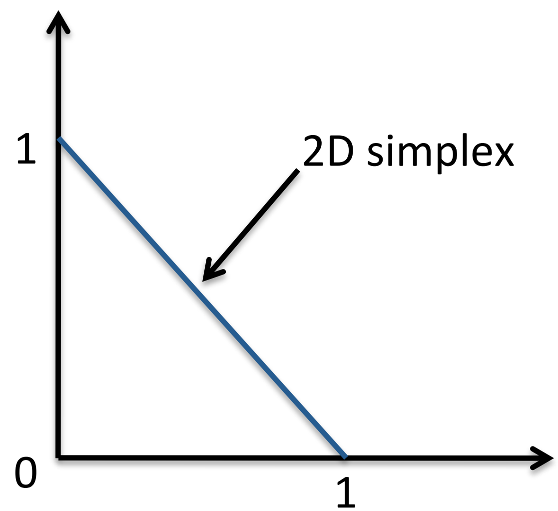
- $\log x$ is concave for $x \geq 0$.

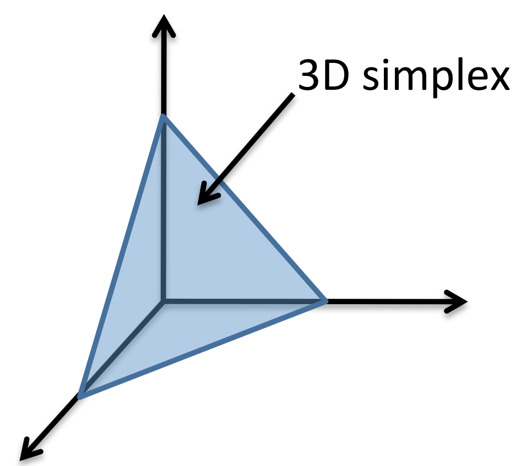
$$\frac{d}{dx} \ln x = \frac{1}{x}$$

$$\frac{d^2}{dx^2} \ln x = \frac{d}{dx}\left(\frac{1}{x}\right) = \frac{-1}{x^2}$$

Recall the probability simplex

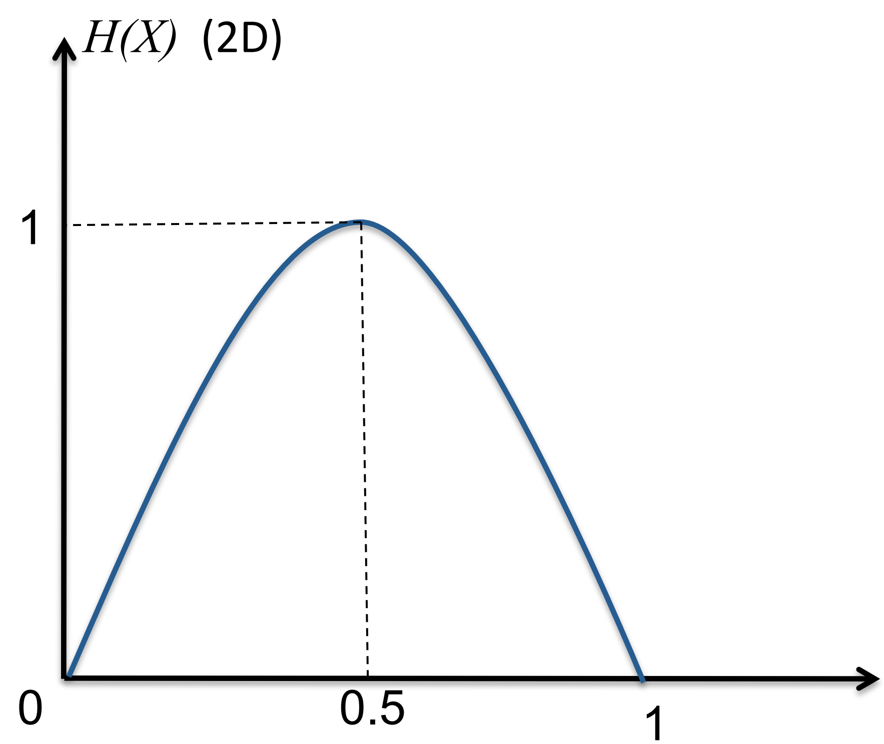
- Recall that the probability simplex is the set of valid PMF's . It's often represented as a triangle because it is a triangle in three dimensions.





$H(X)$ is concave over the probability simplex.

(Proven in part D of this lecture…)



EE 231A: Information Theory
Lecture 2

A. Convexity
B. Jensen's Inequality and its applications
C. Markov Chains and the Data Processing Inequality
D. Log-Sum inequality and its Applications

Jensen's inequality

- For x an r.v., and f a convex function,

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

- If f is a strictly convex function, then equality will occur only when $x = E[X]$.

Proof of Jensen's inequality

- Proof:
  - For two mass points, it's simply convexity:

$$E[f(X)] = P_x(x_1) f(x_1) + P_x(x_2) f(x_2)$$

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

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

$$= f(E[X])$$

- This is the base case for an induction proof.

Proof of Jensen's inequality (cont.)

- Suppose Jensen holds for k-1 mass points in a PMF. Consider the k mass points P_i , $\sum_{i=1}^{k} P_i = 1$.
- Now create a k-1 mass point PMF by neglecting the last point and normalizing.
- Let $P_i^{'} = P_i / (1 - P_k)$

$$
\begin{aligned}
E[f(X)] = \sum_{i=1}^{k} P_i f(x_i) &= P_k f(x_k) + (1 - P_k) \sum_{i=1}^{k-1} P_i^{'} f(x_i) \\
&\geq P_k f(x_k) + (1 - P_k) f(\sum_{i=1}^{k-1} P_i^{'} x_i) \\
&\geq f(P_k x_k + (1 - P_k) \sum_{i=1}^{k-1} P_i^{'} x_i) \\
&= f(\sum_{i=1}^{k} P_i x_i) = f(E[X])
\end{aligned}
$$

Applications of Jensen

- 1. $E[-\log(T)] \geq -\log ET$

  (Jensen on the convex function -$\log$).

2. Relative entropy is always positive

$$D(p \| q) = E_{p(x)}\left[\log \frac{p(x)}{q(x)}\right]$$

Note: $p(x)$ and $q(x)$ are both PMFs, but in this application of Jensen $p(x)$ is the "true PMF". Moreover, when $p(x)/g(x)$ appears inside the expectation, they might as well be any function of x. $T=q(x)/p(x)$ is a new random variable.

2. Relative entropy is nonnegative.

–
$$
\begin{aligned}
D(p \| q) &= E[-\log(T)] \\
&\geq -\log E[T] \\
&= -\log E_{p(x)}\left[\frac{q(x)}{p(x)}\right] \\
&= -\log \sum_x p(x) \frac{q(x)}{p(x)} \\
&= -\log \sum_x q(x) \\
&= 0
\end{aligned}
$$

– Equality only when $T=q(x)/p(x)$ is a deterministic constant, i.e. when p and q are the same distribution.

3. Mutual Information is nonnegative.

$I(X;Y) \geq 0$ why?

$$\begin{aligned} I(X;Y) &= D(p(x,y) \,\|\, p(x)p(y)) \\ &\geq 0 \end{aligned}$$

4. Entropy upper bound

- 4. $H(X) \le \log\left(|\mathcal{X}|\right)$
  - $|\mathcal{X}|$ is the cardinality of the alphabet $\mathcal{X}$.

4. Entropy upper bound

- Proof of 4:
  - Let u be the uniform distribution on $\mathcal{X}$.

$$E_p[-\log u(x)] = E_p[-\log \frac{1}{|\mathcal{X}|}] = \log|\mathcal{X}|$$

$$\begin{aligned} D(p \| u) &= E_p\left[\log \frac{p(x)}{u(x)}\right] \\ &= E_p[\log p(x) - \log u(x)] \\ &= E_p[-\log u(x)] - E_p[-\log p(x)] \\ &= \log|\mathcal{X}| - H(x) \geq 0 \end{aligned}$$

5. Conditioning reduces entropy.

$$H(X \mid Y) \le H(X)$$

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

$$\ge 0$$

6. The joint entropy is less than the sum of the marginal entropies

$$H(X_1, X_2, ..., X_n) \le \sum_{i=1}^{n} H(X_i)$$

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

$$\le \sum_{i=1}^{n} H(X_i)$$

EE 231A: Information Theory
Lecture 2

A. Convexity
B. Jensen's Inequality and its applications
C. Markov Chains and the Data Processing Inequality
D. Log-Sum inequality and its Applications

Markov Chains

- Random variables X,Y,Z form a Markov chain in that order

$$X \to Y \to Z$$

if $P(Z \mid X,Y) = P(Z \mid Y)$.

- In other words, if the conditional probability of Z given X,Y is the same as the conditional probability of Z given only Y.

Conditional Independence in Markov Chains

$$X \to Y \to Z \quad \Rightarrow \quad I(X;Z \mid Y) = 0$$

- Proof:

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

Proof of conditional independence

$$H(Z \mid X,Y) = \sum_y \sum_x p(x,y) H(Z \mid X=x, Y=y)$$

$$= \sum_y \sum_x p(x,y) \sum_z -p(Z=z \mid X=x, Y=y) \log p(Z=z \mid X=x, Y=y)$$

$$= \sum_y \sum_x p(x,y) \sum_z -p(Z=z \mid Y=y) \log p(Z=z \mid Y=y)$$

$$= \sum_y \underbrace{\left(\sum_z -p(Z \mid Y=y) \log p(Z \mid Y=y) \right)}_{H(Z|Y=y)} \underbrace{\sum_x p(x,y)}_{p(y)}$$

$$= \sum_y p(y) H(Z \mid Y=y)$$

$$= H(Z \mid Y)$$

Conditioning and mutual information

- That was an example where conditioning reduced mutual information.
- Conditioning may also increase mutual information.
- Consider X and Y, two independent binary random variables and let

$$Z = X \oplus Y$$

- $I(X;Y) = 0, I(X;Y \mid Z) = 1$

Functions of r.v.'s and Markov Chains

- If $Z = f(Y)$, then $X \to Y \to Z$
- If $X \to Y \to Z$, $I(X;Y) \geq I(X;Z)$ Why?

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y \mid Z) \\ &= I(X;Y) + I(X;Z \mid Y) \end{aligned}$$

- X and Z are conditionally independent given Y, so $I(X;Z \mid Y) = 0$.

$$\begin{aligned} I(X;Y) &= I(X;Z) + I(X;Y \mid Z) \\ &\geq I(X;Z) \end{aligned}$$

Data Processing Inequality

$$\left.\begin{array}{l}\text{• In particular, if } Z = f(Y) \\[1em] \qquad I(X;Y) \geq I(X;f(Y))\end{array}\right\}\ \begin{array}{l}\text{Data}\\\text{Processing}\\\text{Inequality}\end{array}$$

- Also $X \to Y \to Z$ implies $I(X;Y \mid Z) \leq I(X;Y)$
  - Another example where conditioning reduces mutual information.

$$I(X;Y) = I(X;Z) + I(X;Y \mid Z)$$

$$\begin{aligned} I(X;Y \mid Z) &= I(X;Y) - \underbrace{I(X;Z)}_{\geq 0} \\ &\leq I(X;Y) \end{aligned}$$

EE 231A: Information Theory
Lecture 2

A. Convexity
B. Jensen's Inequality and its applications
C. Markov Chains and the Data Processing Inequality
D. Log-Sum inequality and its Applications

Log-Sum Inequality

- For $a_1, \ldots, a_n$ and $b_1, \ldots, b_n \geq 0$

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

Proof of log-sum inequality

– Apply Jensen to $f(t) = t \log t$ with $t_i = \frac{a_i}{b_i}, p(t_i) = \frac{b_i}{\sum_j b_j}$

– Jensen: $E[f(t)] \geq f(E[t])$

$$E[f(t)] = \sum_i p(t_i) f(t_i)$$

$$= \sum_i \underbrace{\frac{b_i}{\sum_j b_j}}_{p(t_i)} \underbrace{\frac{a_i}{b_i} \log \frac{a_i}{b_i}}_{f(t_i)}$$

$$= \frac{1}{\sum_j b_j} \sum_i \left(a_i \log \frac{a_i}{b_i} \right)$$

$$E[t] = \sum_i t_i p(t_i) = \sum_i \frac{a_i}{b_i} \frac{b_i}{\sum_j b_j}$$

$$= \frac{\sum_i a_i}{\sum_j b_j}$$

$$f(E[t]) = \frac{\sum_i a_i}{\sum_j b_j} \log \frac{\sum_i a_i}{\sum_j b_j}$$

Applications of the log-sum inequality

1. $D(p||q) \geq 0$

- Proof:

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

$$\text{Set } \sum a_i = 1 \text{ and } \sum b_i = 1$$

– l.h.s. is $D(p||q)$, r.h.s. is zero.

2. $D(p||q)$ is convex in the pair (p,q)

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

Proof of (2)

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

$$= \sum_i (\lambda p_{1,i} + (1-\lambda) p_{2,i}) \log \frac{\lambda p_{1,i} + (1-\lambda) p_{2,i}}{\lambda q_{1,i} + (1-\lambda) q_{2,i}}$$

$$\leq \sum_i (\lambda p_{1,i}) \log \frac{\lambda p_{1,i}}{\lambda q_{1,i}} + \sum_i (1-\lambda) p_{2,i} \log \frac{(1-\lambda) p_{2,i}}{(1-\lambda) q_{2,i}}$$

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

$$\lambda p_{1,i} = a_1, (1-\lambda) p_{2,i} = a_2$$

$$\lambda q_{1,i} = b_1, (1-\lambda) q_{2,i} = b_2$$

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

$\sum_{i=1}^{2} a_i \log \frac{\sum_{i=1}^{2} a_i}{\sum_{i=1}^{2} b_i}$ is r.h.s of log-sum

3. $H(p)$ is a concave function of p.

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

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

4. Concavity and Convexity of Mutual Information

- $I(X;Y)$ is concave in $p(x)$ for fixed $p(y|x)$

- $I(X;Y)$ is convex in $p(y|x)$ for fixed $p(x)$