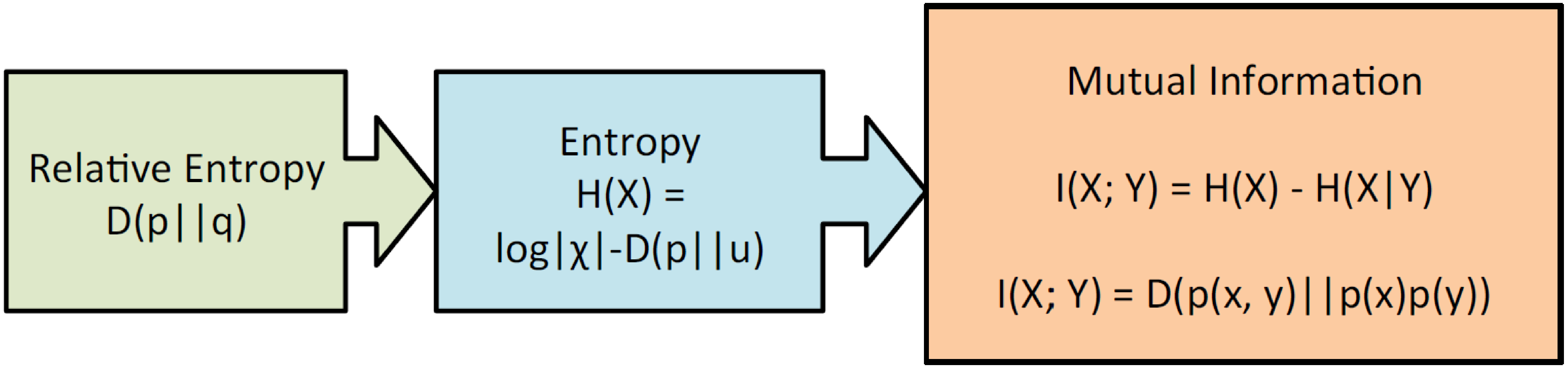
Lecture 10
Chain Rules and Inequalities

- Last lecture: entropy and mutual information
- This time
  - Chain rules
  - Jensen's inequality
  - Log-sum inequality
  - Concavity of entropy
  - Convex/concavity of mutual information

Logic order of things



Chain rule for entropy

- Last time, simple chain rule $H(X,Y) = H(X) + H(Y|X)$
- No matter how we play with chain rule, we get the same answer

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

- "entropy of two experiments"

Chain rule for entropy (general)

- Entropy for a collection of RV's is the sum of the conditional entropies
- More generally: $H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1)$
- Proof:

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

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

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

$$\vdots$$

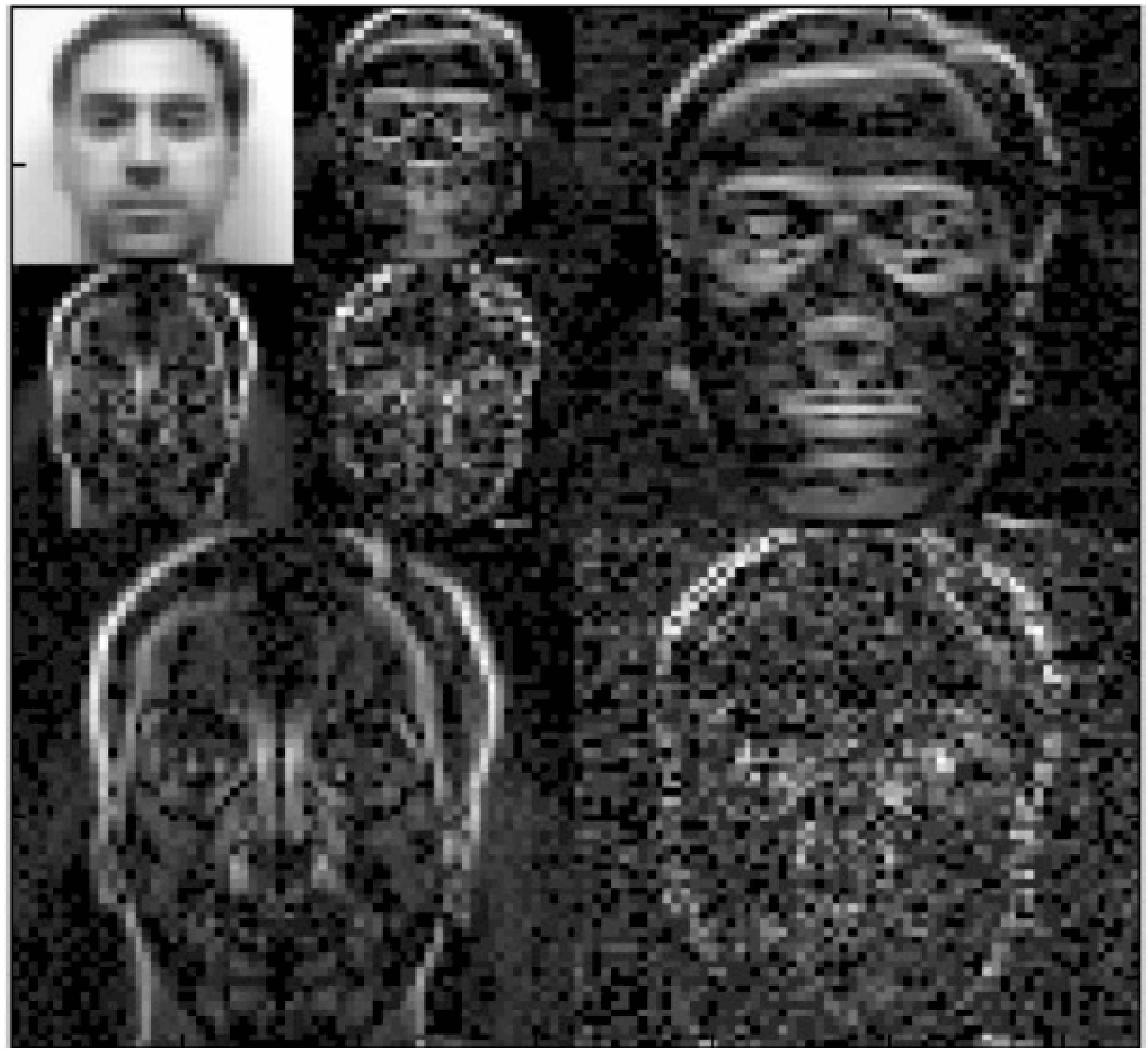
$H(X_1, X_2, \cdots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + H(X_4|X_1, X_2, X_3)$

$+ \cdots + H(X_n|X_1, X_2, \cdots, X_{n-1})$

$= \underline{\sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1})}$

Implication on image compression

$$H(X^n) = \sum_{i=1}^{n} H(X_i | \underbrace{X_{-j}}_{\text{everything seen before}})$$



Conditional mutual information

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

- Definition

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

- In our "asking native for weather" example
  - We want to infer X: rainy or sunny
  - Originally, we only know native's answer Y: yes or no. Value of native's answer $I(X;Y)$
  - If we also has a humidity meter with measurement Z. Value of native's answer $I(X;Y|Z)$

Chain rule for mutual information

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

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

- Chain rule for information

$$I(X_1, X_2, \cdots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \cdots, X_1)$$

- Proof:

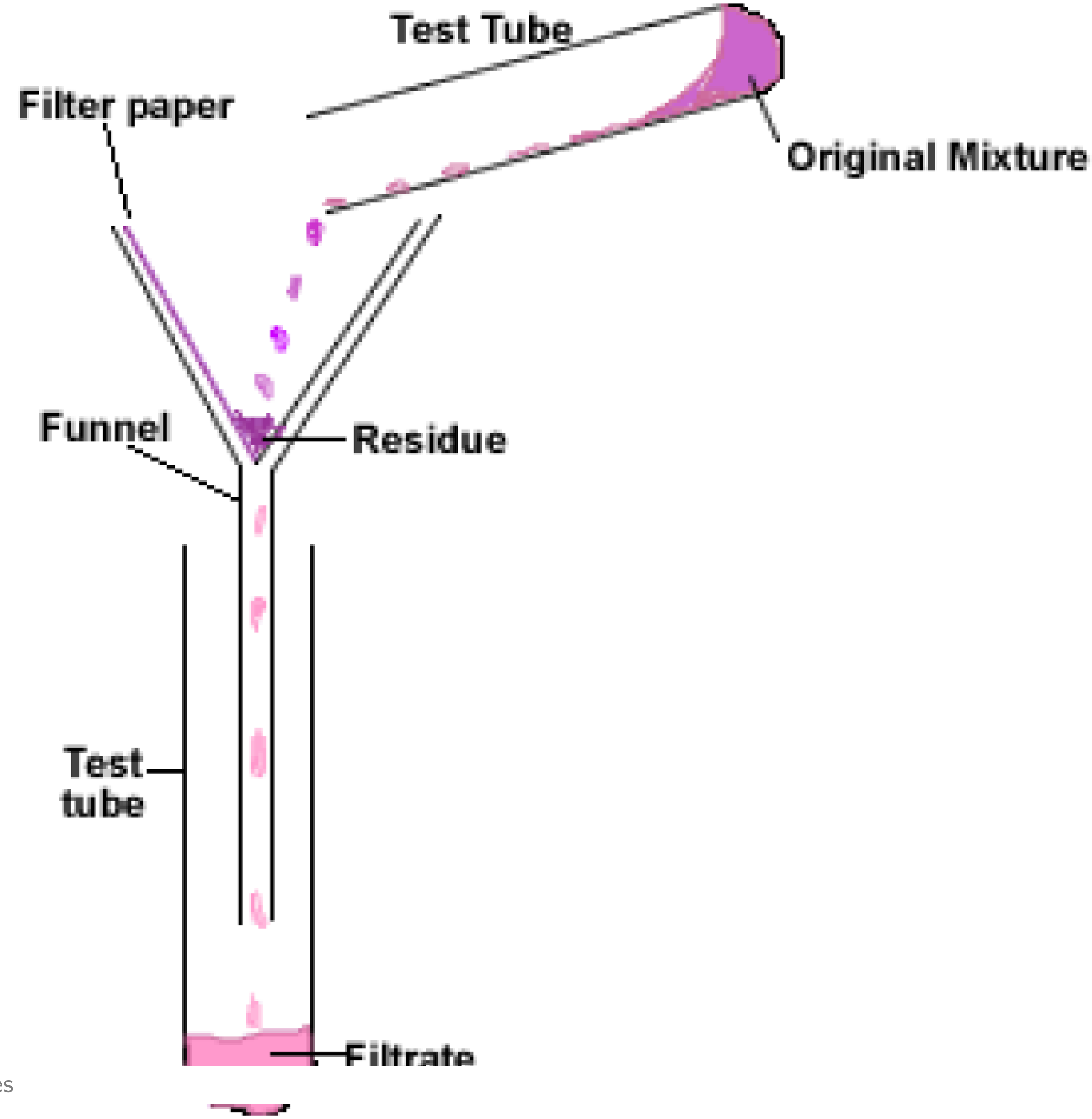
$\sum_{i=1}^{n} H(X_i | X_1, \cdots, X_{i-1}, Y)$

$$I(X_1, X_2, \cdots, X_n; Y) = H(X_1, \cdots, X_n) - H(X_1, \cdots, X_n | Y)$$

Apply chain rules for entropy on both sides.

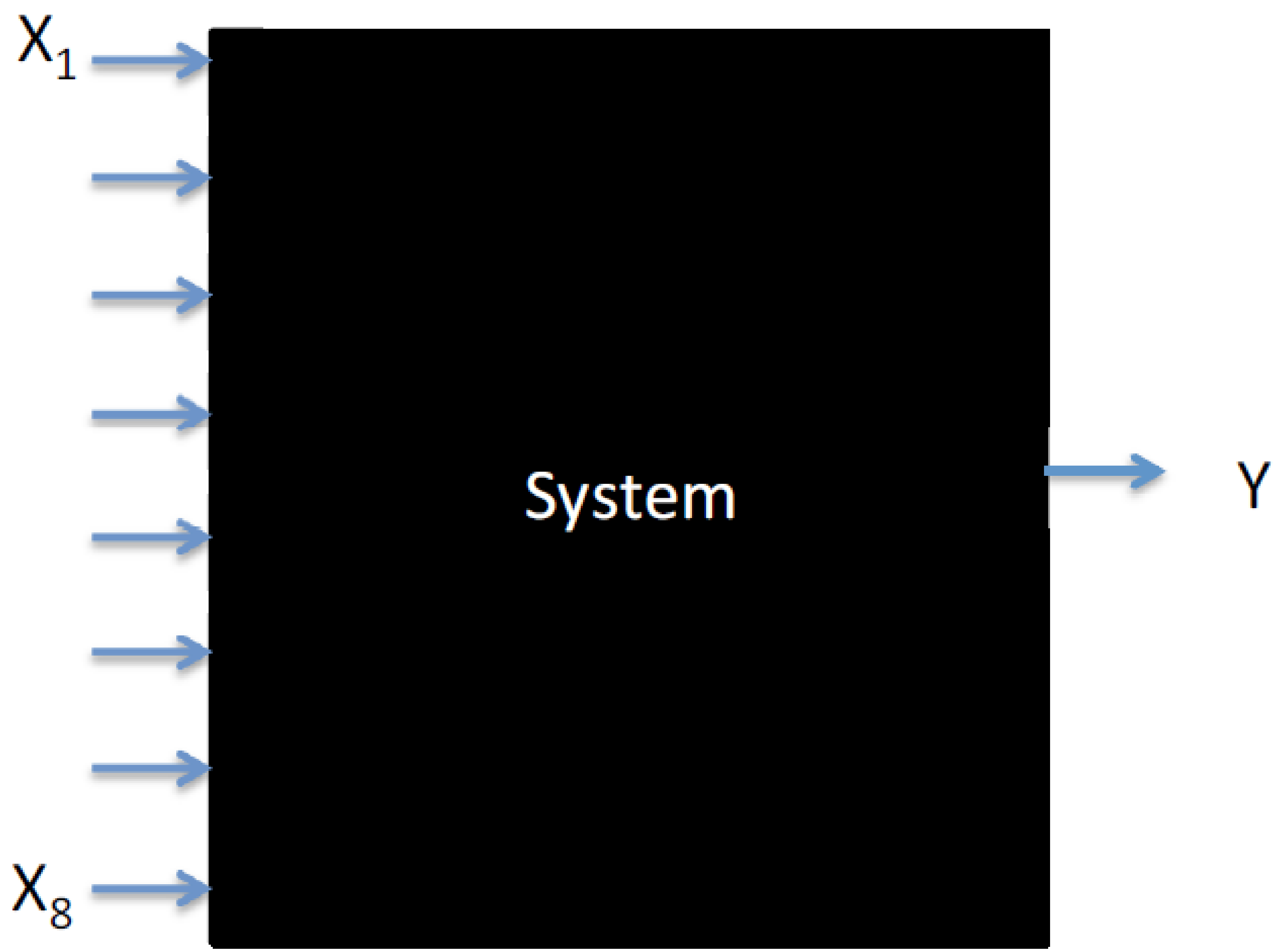
$\sum_{i=1}^{n} H(X_i | X_1, X_2, \cdots, X_{i-1})$

Interpretation 1: "Filtration of information"



Interpretation 2: resolvability

By observing Y, how many possible inputs $(X_1, \cdots, X_8)$ can be distinguished: resolvability of X_i as observed by Y



Conditional relative entropy

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

- Definition:

$$D(p(y|x)\|q(y|x)) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)}$$

- Chain rule for relative entropy

$$D(p(x,y)\|q(x,y)) = D(p(x)\|q(x)) + D(p(y|x)\|q(y|x))$$

- Distance between joint pdfs $=$ distances between margins $+$ distance between conditional pdfs

Why do we need inequalities in information theory?

Convexity

- A function $f(x)$ is convex over an interval (a, b) if for every $x, y \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

Strictly convex if equality holds only if $\lambda = 0$.

Properties of convex functions

- If a function f has second order derivative ≥ 0 (> 0), the function is convex (strictly convex).
- Vector valued function: Hessian matrix is nonnegative definite.
- Examples: $x^2, e^x, |x|, x \log x (x \geq 0), \|x\|^2$.
- A function f is concave if $-f$ is convex.
- Linear function $ax + b$ is both convex and concave.

How to show a function is convex

- By definition: $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ (function must be continuous)

$g(x) = f(Ax+b)$

- Verify $f''(x) \geq 0$ (or nonnegative definite)

$\Rightarrow g(\lambda x + (1-\lambda)y) = f(A[\lambda x + (1-\lambda)y] + b)$

$= f(\lambda(Ax+b) + (1-\lambda)(Ay+b))$

$\leq \lambda f(Ax+b) + (1-\lambda) f(Ay+b) = g(x)\lambda + g(y)(1-\lambda)$

- By composition rules:
  - Composition of affine function $f(Ax+b)$ is convex if f is convex
  - Composition with a scalar function: $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$,

$$f(x) = h(g(x)),$$

then f is convex if

1. g convex, h convex, $\tilde{h}$ nondecreasing
2. g concave, h convex, $\tilde{h}$ nonincreasing

Extended-value extension $\tilde{f}(x) = f(x), x \in \mathcal{X}$, otherwise is ∞

Jensen's inequality

- Due to Danish mathematician Johan Jensen, 1906
- Widely used in mathematics and information theory
- Convex transformation of a mean $\leq$ mean after convex transformation

Jensen's inequality theorem

Theorem. (Jensen's inequality) If f is a convex function,

$$\mathbb{E}f(X) \geq f(\mathbb{E}X).$$

$f(x) = X^2$: LHS : $E[X^2]$

RHS: $(E[X])^2$

If f strictly convex, equality holds when

$$X = \text{constant}.$$

Step 1: If f is convex

$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle$

Step 2: $x^* = E[X]$

Proof: Let $x^* = \mathbb{E}X$. Expand $f(x)$ by Taylor's Theorem at x^*:

$\Rightarrow$ LHS: $E[f(X)]$

$\Rightarrow$ RHS: $f(E[X]) + \langle \nabla f(x^*), E[X - x^*] \rangle$

$= f(E[X])$

$$f(x) = f(x^*) + f'(x^*)(x - x^*) + \frac{f''(z)}{2}(x - x^*)^2, \quad z \in (x, x^*)$$

f convex: $f''(z) \geq 0$. So $f(x) \geq f(x^*) + f'(x^*)(x - x^*)$. Take expectation on both sides:

$$\mathbb{E}f(X) \geq f(x^*).$$

Consequences of Jensen's inequality

- $f(x) = x^2, \mathbb{E}X^2 \geq [\mathbb{E}X]^2$: variance is nonnegative
- $f(x) = e^x, \mathbb{E}e^x \geq e^{\mathbb{E}(x)}$
- Arithmetic mean $\geq$ Geometric mean $\geq$ Harmonic mean

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \geq \sqrt[n]{x_1 x_2 \cdots x_n} \geq \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Proof using Jensen's inequality: $f(x) = x \log x$ is convex.

Information inequality

$$\underbrace{D}_{\text{relative entropy}}(p\|q) \geq 0,$$

equality iff $p(x) = q(x)$ for all x.

Proof:

$$\begin{aligned}
D(p\|q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} && \text{definition.} \\
&= -\sum_x p(x) \log \frac{q(x)}{p(x)} = E_{x\sim p}\left[-\log \frac{q(x)}{p(x)}\right] \\
&\geq \log \sum_x p(x) \frac{q(x)}{p(x)} && x \mapsto -\log x \text{ is convex.} \\
&= \log \sum_x q(x) = 0.
\end{aligned}$$

Nonnegativity of mutual information

- $I(X;Y) \geq 0$, equality iff X and Y are independent. Since $I(X;Y) = D(p(x,y) \| p(x)p(y))$.
- Conditional relative entropy and mutual information are also nonnegative

Conditioning reduces entropy

Information cannot hurt:

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

$$= \sum_y p(y) H(X|Y=y)$$

- Since $I(X;Y) = H(X) - H(X|Y) \geq 0$
- Knowing another RV Y only reduces uncertainty in X on average
- $H(X|Y=y)$ may be larger than $H(X)$: in court, knowing a new evidence sometimes can increase uncertainty

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

Independence bound on entropy

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

equality iff X_i independent.

- From chain rule:

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

Maximum entropy

Uniform distribution has maximum entropy among all distributions with finite discrete support.

Theorem. $H(X) \leq \log |\mathcal{X}|$, where $\mathcal{X}$ is the number of elements in the set. Equality iff X has uniform distribution.

Proof: Let U be a uniform distributed RV, $u(x) = 1/|\mathcal{X}|$

$$0 \leq D(p\|u) = \sum p(x) \log \frac{p(x)}{u(x)} \tag{1}$$

$$= \sum p(x) \log |\mathcal{X}| - (-\sum p(x) \log p(x)) = \log |\mathcal{X}| - H(X) \tag{2}$$

$$\sum_x p(x) \log \frac{p(x)}{1/|X|} = \sum_x p(x) [\log p(x) + \log |X|]$$

$$= \underbrace{\sum_x p(x) \log p(x)}_{-H(x)} + \log |X| \underbrace{\sum_x p(x)}_{=1} = \log |X| - H(X)$$

Log sum inequality

- Consequence of concavity of log
- **Theorem.** For nonnegative $a_1, \cdots, a_n$ and $b_1, \cdots, b_n$

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

$f\left(\frac{a_i}{b_i}\right) = \frac{a_i}{b_i} \log \frac{a_i}{b_i}$

  Equality iff $a_i / b_i =$ constant.

- Proof by Jensen's inequality using convexity of $f(x) = x \log x$.

Application of log-sum inequality

- Very handy in proof: e.g., prove $D(p\|q) \geq 0$:

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

$$\geq \left(\sum_x p(x)\right) \log \frac{\sum_x p(x)}{\sum_x q(x)} = 1 \log 1 = 0.$$

Convexity of relative entropy

Theorem. $D(p\|q)$ is convex in the pair (p,q): given two pairs of pdf,

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

for all $0 \leq \lambda \leq 1$.

Proof: By definition and log-sum inequality

$$\begin{aligned}
&D(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \\
&\sum_{x\in\mathcal{X}} = (\lambda p_1 + (1-\lambda)p_2) \log \frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2} \\
&\sum_{x\in\mathcal{X}} \leq \lambda p_1 \log \frac{\lambda p_1}{\lambda q_1} + (1-\lambda) \log \frac{(1-\lambda)p_2}{(1-\lambda)q_2} \\
&= \lambda D(p_1\|q_1) + (1-\lambda)D(p_2\|q_2)
\end{aligned}$$

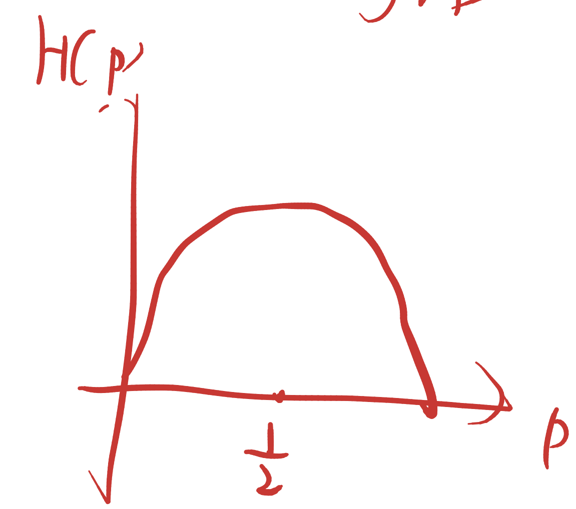
log-sum inequality

Concavity of entropy

$\mathcal{X} = \{0, 1\}$

p $\quad H(p) = -p \log p$

$- (1-p) \log(1-p)$



Entropy

$$H(p) = -\sum_i p_i \log p_i$$

is concave in p

Proof 1:

$$\begin{aligned}
H(p) = -\sum_{i \in \mathcal{X}} p_i \log p_i &= -\sum_{i \in \mathcal{X}} p_i \log \frac{p_i}{u_i} u_i \\
&= -\sum_{i \in \mathcal{X}} p_i \log \frac{p_i}{u_i} - \sum_{i \in \mathcal{X}} p_i \log u_i \\
&= -D(p\|u) - \log \frac{1}{|\mathcal{X}|} \sum_{i \in \mathcal{X}} p_i \\
&= \log |\mathcal{X}| - D(p\|u)
\end{aligned}$$

Concavity of entropy (Proof 2)

Proof 2: We want to prove

$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2)$.

A neat idea: introduce auxiliary RV:

$$\theta = \begin{cases} 1, & \text{w.p. } \lambda \\ 2, & \text{w.p. } 1-\lambda. \end{cases}$$

Let $Z = X_\theta$, distribution of Z is $\lambda p_1 + (1-\lambda)p_2$. Conditioning reduces entropy:

$$H(Z) \geq H(Z|\theta)$$

By their definitions

$$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2).$$

Concavity and convexity of mutual information

Mutual information $I(X;Y)$ is:

- a concave function of $p(x)$ for fixed $p(y|x)$
- convex function of $p(y|x)$ for fixed $p(x)$

Mixing two gases of equal entropy results in a gas with higher entropy.

Proof of mutual information properties

Proof: write $I(X;Y)$ as a function of $p(x)$ and $p(y|x)$:

$$I(X;Y) = \sum_{x,y} p(x)p(y|x)\log\frac{p(y|x)}{p(y)}$$

$$= \sum_{x,y} p(x)p(y|x)\log p(y|x)$$

$$-\sum_{y}\left\{\sum_{x} p(x)p(y|x)\right\}\log\left\{\sum_{x} p(y|x)p(x)\right\}$$

- (a): Fixing $p(y|x)$, first linear in $p(x)$, second term concave in $p(x)$
- (b): Fixing $p(x)$, complicated in $p(y|x)$. Instead of verify it directly, we will relate it to something we know.

Strategy for convexity proof

Our strategy is to introduce auxiliary RV $\tilde{Y}$ with a mixing distribution

$$p(\tilde{y}|x) = \lambda p_1(y|x) + (1-\lambda)p_2(y|x).$$

To prove convexity, we need to prove:

$$I(X;\tilde{Y}) \leq \lambda I_{p_1}(X;Y) + (1-\lambda)I_{p_2}(X;Y)$$

Since

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

We want to use the fact that $D(p\|q)$ is convex in the pair (p,q).

Completing the convexity proof

What we need is to find out the pdfs:

$$p(\tilde{y}) = \sum_x [\lambda p_1(y|x)p(x) + (1-\lambda)p_2(y|x)p(x)] = \lambda p_1(y) + (1-\lambda)p_2(y)$$

$p(\hat{y}) = \sum_x p(x)\, p(\hat{y}|x)$.

We also need

$$p(x,\tilde{y}) = p(\tilde{y}|x)p(x) = \lambda p_1(x,y) + (1-\lambda)p_2(x,y)$$

Finally, we get, from convexity of $D(p\|q)$:

$$D(p(x,\tilde{y})\|p(x)p(\tilde{y}))$$

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

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

$$= \lambda I_{p_1}(X;Y) + (1-\lambda)I_{p_2}(X;Y)$$

Summary of proof techniques

- Conditioning $p(x, y) = p(x|y)p(y)$, sometimes do this iteratively
- Use Jensen's inequality – identify what is the "average"

$$f(\mathbb{E}X) \leq \mathbb{E}f(X)$$

- Prove convexity: several approaches
- Introduce auxiliary random variable – e.g. uniform RV U, indexing RV θ

Summary of important results

- Mutual information is nonnegative
- Conditioning reduces entropy
- Uniform distribution maximizes entropy
- Properties:
  - $D(p\|q)$ is convex in (p, q)
  - Entropy $H(p)$ is concave in p
  - Mutual information $I(X;Y)$ concave in $p(x)$ (fixing $p(y \mid x)$), and convex in $p(y \mid x)$ (fixing $p(x)$)