Chapter 2
Information Measures

Department of Information Engineering
The Chinese University of Hong Kong

2.1 Independence and Markov Chain

Notations

X	discrete random variable taking values in $\mathcal{X}$
$\{p_X(x)\}$	probability distribution for X
$\mathcal{S}_X$	support of X

- If $\mathcal{S}_X = \mathcal{X}$, we say that p is strictly positive.
- Non-strictly positive distributions are dangerous.

Definition 2.1 Two random variables X and Y are independent, denoted by $X \perp Y$, if

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

for all x and y (i.e., for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$).

Definition 2.2 (Mutual Independence) For $n \geq 3$, random variables $X_1, X_2, \cdots, X_n$ are mutually independent if

$$p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2) \cdots p(x_n)$$

for all $x_1, x_2, \cdots, x_n$.

Definition 2.3 (Pairwise Independence) For $n \geq 3$, random variables $X_1, X_2, \cdots, X_n$ are pairwise independent if X_i and X_j are independent for all $1 \leq i < j \leq n$.

Definition 2.4 (Conditional Independence) For random variables X, Y, and Z, X is independent of Z conditioning on Y, denoted by $X \perp Z|Y$, if

$$p(x, y, z)p(y) = p(x, y)p(y, z)$$

for all x, y, and z, or equivalently,

$$p(x, y, z) = \begin{cases} \frac{p(x,y)p(y,z)}{p(y)} = p(x, y)p(z|y) & \text{if } p(y) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 2.5 For random variables X, Y, and Z, $X \perp Z | Y$ if and only if

$$p(x, y, z) = a(x, y) b(y, z)$$

for all x, y, and z such that $p(y) > 0$.

Proposition 2.6 (Markov Chain) For random variables $X_1, X_2, \cdots, X_n$, where $n \geq 3$, $X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if

$$\begin{aligned} & p(x_1, x_2, \cdots, x_n) p(x_2) p(x_3) \cdots p(x_{n-1}) \\ = \quad & p(x_1, x_2) p(x_2, x_3) \cdots p(x_{n-1}, x_n) \end{aligned}$$

for all $x_1, x_2, \cdots, x_n$, or equivalently,

$$p(x_1, x_2, \cdots, x_n) =$$

$$\begin{cases} p(x_1, x_2) p(x_3|x_2) \cdots p(x_n|x_{n-1}) & \text{if } p(x_2), p(x_3), \cdots, p(x_{n-1}) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 2.7 $X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if and only if $X_n \to X_{n-1} \to \cdots \to X_1$ forms a Markov chain.

Proposition 2.8 $X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if and only if

$$\begin{aligned} & X_1 \to X_2 \to X_3 \\ & (X_1, X_2) \to X_3 \to X_4 \\ & \qquad\qquad \vdots \\ & (X_1, X_2, \cdots, X_{n-2}) \to X_{n-1} \to X_n \end{aligned}$$

form Markov chains.

Proposition 2.9 $X_1 \to X_2 \to \cdots \to X_n$ forms a Markov chain if and only if

$$p(x_1, x_2, \cdots, x_n) = f_1(x_1, x_2) f_2(x_2, x_3) \cdots f_{n-1}(x_{n-1}, x_n)$$

for all $x_1, x_2, \cdots, x_n$ such that $p(x_2), p(x_3), \cdots, p(x_{n-1}) > 0$.

Proposition 2.10 (Markov subchains) Let $\mathcal{N}_n = \{1, 2, \cdots, n\}$ and let $X_1 \to X_2 \to \cdots \to X_n$ form a Markov chain. For any subset α of $\mathcal{N}_n$, denote $(X_i, i \in \alpha)$ by X_α. Then for any disjoint subsets $\alpha_1, \alpha_2, \cdots, \alpha_m$ of $\mathcal{N}_n$ such that

$$k_1 < k_2 < \cdots < k_m$$

for all $k_j \in \alpha_j$, $j = 1, 2, \cdots, m$,

$$X_{\alpha_1} \to X_{\alpha_2} \to \cdots \to X_{\alpha_m}$$

forms a Markov chain. That is, a subchain of $X_1 \to X_2 \to \cdots \to X_n$ is also a Markov chain.

Example 2.11 Let $X_1 \to X_2 \to \cdots \to X_{10}$ form a Markov chain and $\alpha_1 = \{1, 2\}$, $\alpha_2 = \{4\}$, $\alpha_3 = \{6, 8\}$, and $\alpha_4 = \{10\}$ be subsets of $\mathcal{N}_{10}$. Then Proposition 2.10 says that

$$(X_1, X_2) \to X_4 \to (X_6, X_8) \to X_{10}$$

also forms a Markov chain.

Proposition 2.12 Let X_1, X_2, X_3, and X_4 be random variables such that $p(x_1, x_2, x_3, x_4)$ is strictly positive. Then

$$\left.\begin{array}{l} X_1 \perp X_4 | (X_2, X_3) \\ X_1 \perp X_3 | (X_2, X_4) \end{array}\right\} \Rightarrow X_1 \perp (X_3, X_4) | X_2.$$

- Not true if p is not strictly positive
- Let $X_1 = Y$, $X_2 = Z$, and $X_3 = X_4 = (Y, Z)$, where $Y \perp Z$
- Then $X_1 \perp X_4 | (X_2, X_3)$, $X_1 \perp X_3 | (X_2, X_4)$, but $X_1 \not\perp (X_3, X_4) | X_2$.
- p is not strictly positive because $p(x_1, x_2, x_3, x_4) = 0$ if $x_3 \neq (x_1, x_2)$ or $x_4 \neq (x_1, x_2)$.

2.2 Shannon's Information Measures

- Entropy
- Conditional entropy
- Mutual information
- Conditional mutual information

Definition 2.13 The entropy $H(X)$ of a random variable X is defined as

$$H(X) = -\sum_x p(x) \log p(x).$$

- Convention: summation is taken over $\mathcal{S}_X$.
- When the base of the logarithm is α, write $H(X)$ as $H_\alpha(X)$.
- Entropy measures the uncertainty of a discrete random variable.
- The unit for entropy is

$$\begin{array}{ll} \text{bit} & \text{if } \alpha = 2 \\ \text{nat} & \text{if } \alpha = e \\ D\text{-it} & \text{if } \alpha = D \end{array}$$

- $H(X)$ depends only on the distribution of X but not on the actual value taken by X, hence also write $H(p)$.
- A bit in information theory is different from a bit in computer science.

Entropy as Expectation

- Convention
$$Eg(X) = \sum_x p(x)g(x)$$
where summation is over $\mathcal{S}_X$.

- Linearity
$$E[f(X) + g(X)] = Ef(X) + Eg(X)$$

- Can write
$$H(X) = -E \log p(X) = -\sum_x p(x) \log p(x)$$

- In probability theory, when $Eg(X)$ is considered, usually $g(x)$ depends only on the value of x but not on $p(x)$.

Binary Entropy Function

- For $0 \leq \gamma \leq 1$, define the binary entropy function

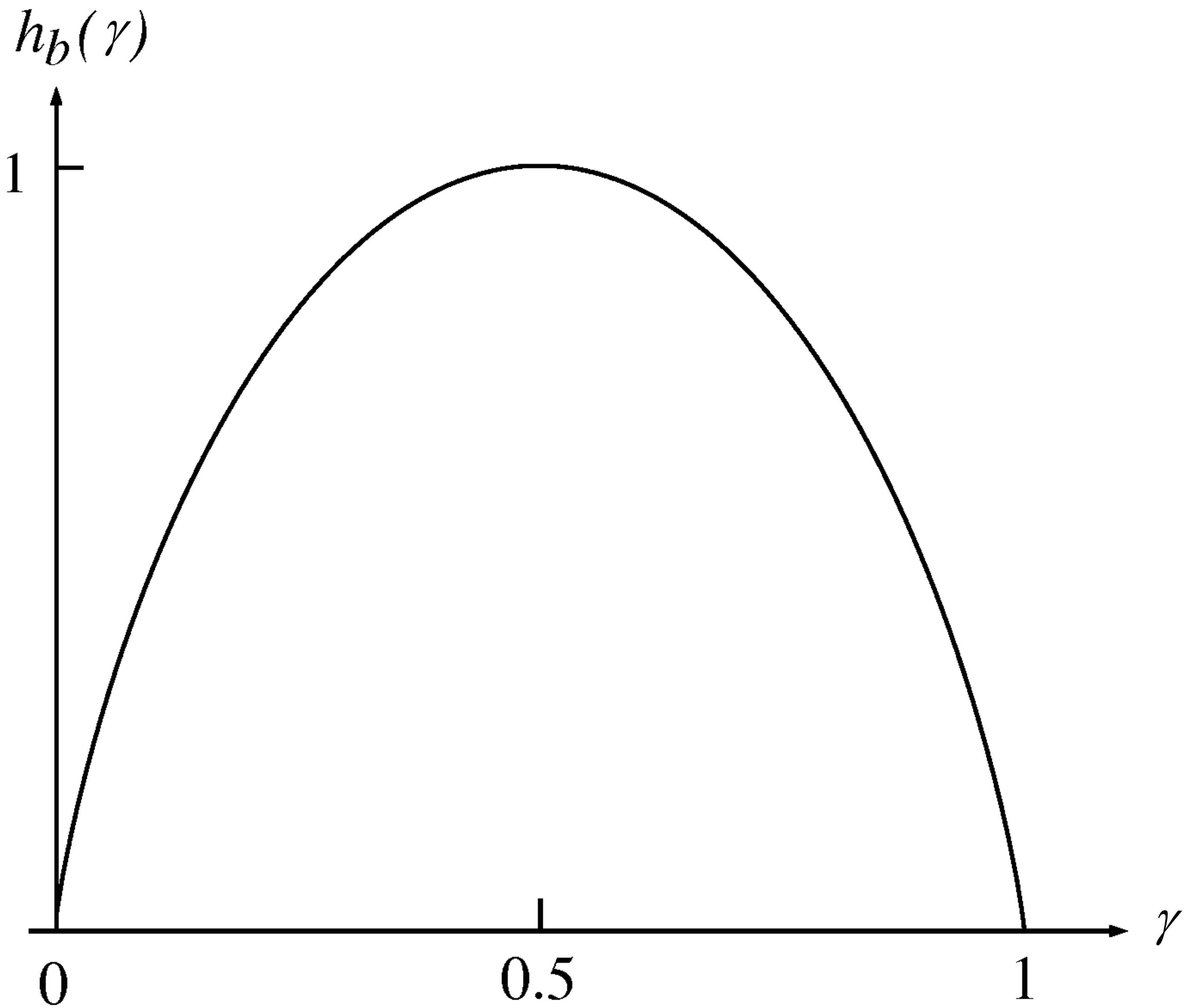
$$h_b(\gamma) = -\gamma \log \gamma - (1-\gamma) \log(1-\gamma)$$

  with the convention $0 \log 0 = 0$.

- For $X \sim \{\gamma, 1-\gamma\}$,

$$H(X) = h_b(\gamma).$$

- $h_b(\gamma)$ achieves the maximum value 1 when $\gamma = \frac{1}{2}$.



$h_b(\gamma)$
1
γ
0
0.5
1

Definition 2.14 The joint entropy $H(X, Y)$ of a pair of random variables X and Y is defined as

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

Definition 2.15 For random variables X and Y, the conditional entropy of Y given X is defined as

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) = -E \log p(Y|X).$$

- Write

$$H(Y|X) = \sum_x p(x) \left[-\sum_y p(y|x) \log p(y|x) \right].$$

- The inner sum is the entropy of Y conditioning on a fixed $x \in \mathcal{S}_X$, denoted by $H(Y|X = x)$.

- Thus

$$H(Y|X) = \sum_x p(x) H(Y|X = x),$$

- Similarly,

$$H(Y|X, Z) = \sum_z p(z) H(Y|X, Z = z),$$

  where

$$H(Y|X, Z = z) = -\sum_{x,y} p(x, y|z) \log p(y|x, z).$$

Proposition 2.16

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

and

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

Proof Consider

$$\begin{aligned} H(X,Y) &= -E \log p(X,Y) \\ &\overset{a)}{=} -E \log[p(X)p(Y|X)] \\ &\overset{b)}{=} -E \log p(X) - E \log p(Y|X) \\ &= H(X) + H(Y|X). \end{aligned}$$

a) summation is over $\mathcal{S}_{XY}$

b) linearity of expectation

Definition 2.17 For random variables X and Y, the mutual information between X and Y is defined as

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

Remark $I(X;Y)$ is symmetrical in X and Y.

Proposition 2.18 The mutual information between a random variable X and itself is equal to the entropy of X, i.e., $I(X;X) = H(X)$.

Proposition 2.19

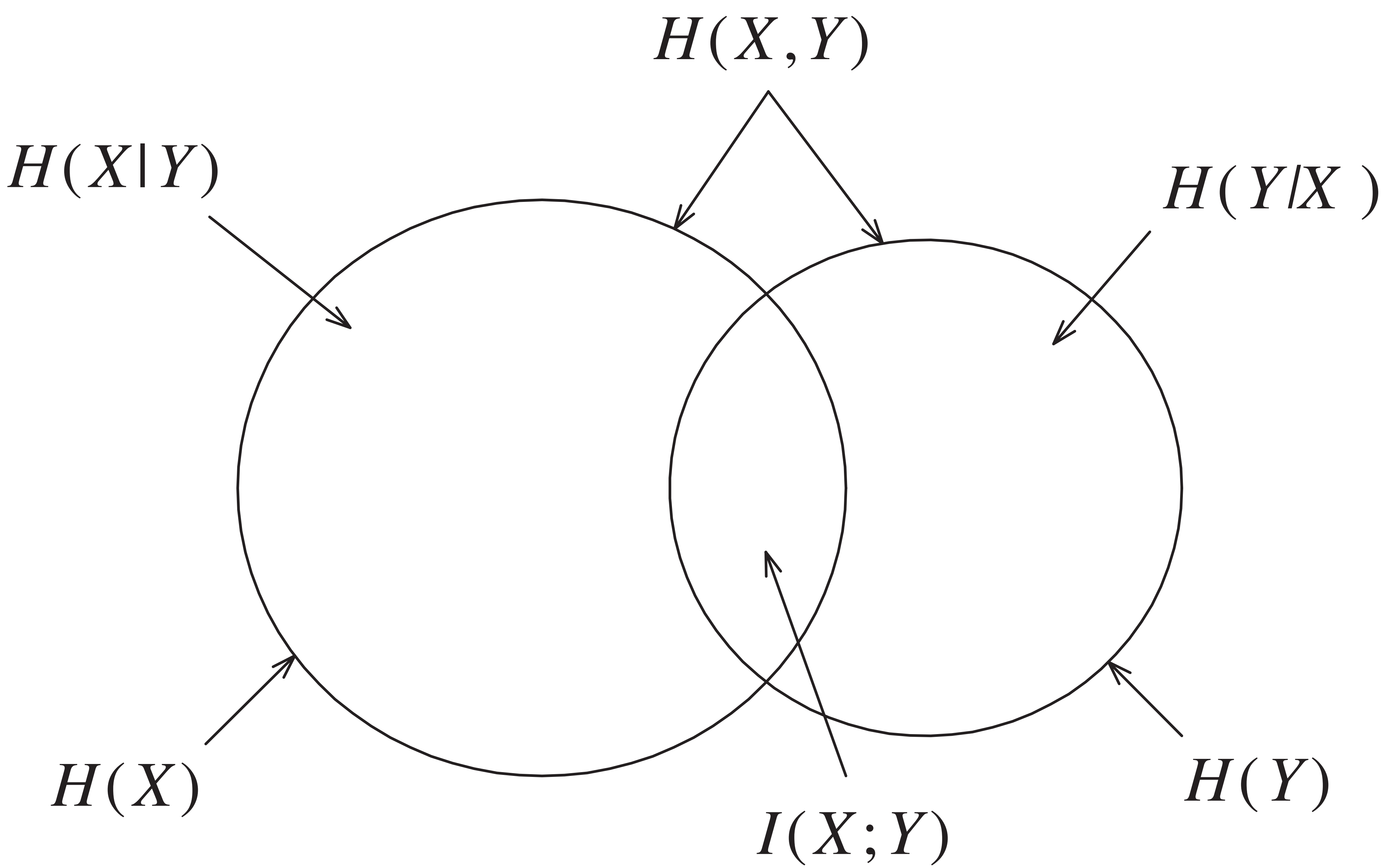
$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y), \\ I(X;Y) &= H(Y) - H(Y|X), \end{aligned}$$

and

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

provided that all the entropies and conditional entropies are finite.

Information Diagram



Definition 2.20 For random variables X, Y and Z, the mutual information between X and Y conditioning on Z is defined as

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

Remark $I(X;Y|Z)$ is symmetrical in X and Y.

Similar to entropy, we have

$$I(X;Y|Z) = \sum_{z} p(z) I(X;Y|Z=z),$$

where

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

Proposition 2.21 The mutual information between a random variable X and itself conditioning on a random variable Z is equal to the conditional entropy of X given Z, i.e., $I(X;X|Z) = H(X|Z)$.

Proposition 2.22

$$
\begin{aligned}
I(X;Y|Z) &= H(X|Z) - H(X|Y,Z), \\
I(X;Y|Z) &= H(Y|Z) - H(Y|X,Z),
\end{aligned}
$$

and

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

provided that all the conditional entropies are finite.

Remark All Shannon's information measures are special cases of conditional mutual information.

2.3 Continuity of Shannon's Information Measures for Fixed Finite Alphabets

- All Shannon's information measures are continuous when the alphabets are fixed and finite.
- For countable alphabets, Shannon's information measures are everywhere discontinuous.

Definition 2.23 Let p and q be two probability distributions on a common alphabet $\mathcal{X}$. The variational distance between p and q is defined as

$$V(p,q) = \sum_{x \in \mathcal{X}} |p(x) - q(x)|.$$

The entropy function is continuous at p if

$$\lim_{p' \to p} H(p') = H\left(\lim_{p' \to p} p'\right) = H(p),$$

or equivalently, for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$|H(p) - H(q)| < \epsilon$$

for all $q \in \mathcal{P}_{\mathcal{X}}$ satisfying

$$V(p,q) < \delta,$$

2.4 Chain Rules

Proposition 2.24 (Chain Rule for Entropy)

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

Proposition 2.25 (Chain Rule for Conditional Entropy)

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

Proposition 2.26 (Chain Rule for Mutual Information)

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

Proposition 2.27 (Chain Rule for Conditional Mutual Information)

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

Proof of Proposition 2.25

$$
\begin{aligned}
& H(X_1, X_2, \cdots, X_n|Y) \\
& = H(X_1, X_2, \cdots, X_n, Y) - H(Y) \\
& = H((X_1, Y), X_2, \cdots, X_n) - H(Y) \\
& \stackrel{a)}{=} H(X_1, Y) + \sum_{i=2}^{n} H(X_i|X_1, \cdots, X_{i-1}, Y) - H(Y) \\
& = H(X_1|Y) + \sum_{i=2}^{n} H(X_i|X_1, \cdots, X_{i-1}, Y) \\
& = \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1}, Y),
\end{aligned}
$$

where a) follows from Proposition 2.24 (chain rule for entropy).

Alternative Proof of Proposition 2.25

$$
\begin{aligned}
&H(X_1, X_2, \cdots, X_n|Y) \\
&= \sum_y p(y) H(X_1, X_2, \cdots, X_n|Y = y) \\
&= \sum_y p(y) \sum_{i=1}^n H(X_i|X_1, \cdots, X_{i-1}, Y = y) \\
&= \sum_{i=1}^n \sum_y p(y) H(X_i|X_1, \cdots, X_{i-1}, Y = y) \\
&= \sum_{i=1}^n H(X_i|X_1, \cdots, X_{i-1}, Y),
\end{aligned}
$$

2.5 Informational Divergence

Definition 2.28 The informational divergence between two probability distributions p and q on a common alphabet $\mathcal{X}$ is defined as

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

where E_p denotes expectation with respect to p.

- Convention:
    1. summation over $\mathcal{S}_p$
    2. $c \log \frac{c}{0} = \infty$ for $c > 0$ — if $D(p\|q) < \infty$, then $\mathcal{S}_p \subset \mathcal{S}_q$.
- $D(p\|q)$ measures the "distance" between p and q.
- $D(p\|q)$ is not symmetrical in p and q, so $D(\cdot\|\cdot)$ is not a true metric.
- $D(\cdot\|\cdot)$ does not satisfy the triangular inequality.
- Also called *relative entropy* or the *Kullback-Leibler distance*.

Lemma 2.29 (Fundamental Inequality) For any $a > 0$,

$$\ln a \leq a - 1$$

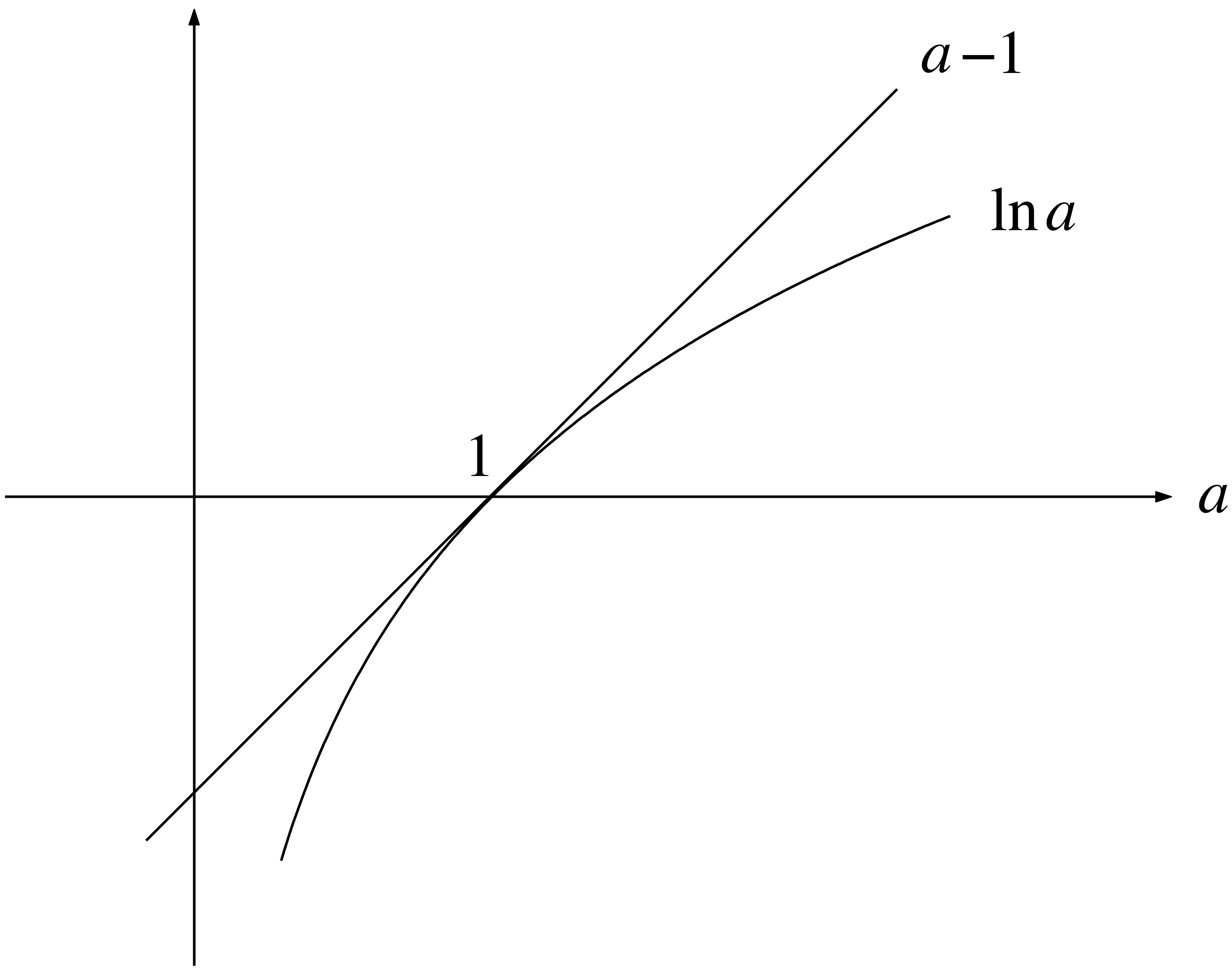
with equality if and only if $a = 1$.

Corollary 2.30 For any $a > 0$,

$$\ln a \geq 1 - \frac{1}{a}$$

with equality if and only if $a = 1$.

$$\ln a \leq a - 1$$



Theorem 2.31 (Divergence Inequality) For any two probability distributions p and q on a common alphabet $\mathcal{X}$,

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

with equality if and only if $p = q$.

Theorem 2.32 (Log-Sum Inequality) For positive numbers $a_1, a_2, \cdots$ and nonnegative numbers $b_1, b_2, \cdots$ such that $\sum_i a_i < \infty$ and $0 < \sum_i b_i < \infty$,

$$\sum_i a_i \log \frac{a_i}{b_i} \geq \left(\sum_i a_i \right) \log \frac{\sum_i a_i}{\sum_i b_i}$$

with the convention that $\log \frac{a_i}{0} = \infty$. Moreover, equality holds if and only if $\frac{a_i}{b_i} = \mathit{constant}$ for all i.

Example:

$$a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2} \geq (a_1 + a_2) \log \frac{a_1 + a_2}{b_1 + b_2}.$$

Divergence Inequality vs Log-Sum Inequality

- The divergence inequality implies the log-sum inequality.
- The log-sum inequality also implies the divergence inequality.
- The two inequalities are equivalent.

Theorem 2.33 (Pinsker's Inequality)

$$D(p\|q) \geq \frac{1}{2\ln 2} V^2(p,q).$$

- If $D(p\|q)$ or $D(q\|p)$ is small, then so is $V(p,q)$.
- For a sequence of probability distributions q_k, as $k \to \infty$, if $D(p\|q_k) \to 0$ or $D(q_k\|p) \to 0$, then $V(p,q_k) \to 0$.
- That is, "convergence in divergence" is a stronger notion than "convergence in variational distance."

2.6 The Basic Inequalities

Theorem 2.34 For random variables X, Y, and Z,

$$I(X;Y|Z) \geq 0,$$

with equality if and only if X and Y are independent when conditioning on Z.

Corollary All Shannon's information measures are nonnegative, because they are all special cases of conditional mutual information.

Proposition 2.35 $H(X) = 0$ if and only if X is deterministic.

Proposition 2.36 $H(Y|X) = 0$ if and only if Y is a function of X.

Proposition 2.37 $I(X;Y) = 0$ if and only if X and Y are independent.

2.7 Some Useful Information Inequalities

Theorem 2.38 (Conditioning Does Not Increase Entropy)

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

with equality if and only if X and Y are independent.

- Similarly, $H(Y|X, Z) \leq H(Y|Z)$.
- Warning: $I(X; Y|Z) \leq I(X; Y)$ does not hold in general.

Theorem 2.39 (Independence Bound for Entropy)

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

with equality if and only if X_i, $i = 1, 2, \cdots, n$ are mutually independent.

Theorem 2.40

$$I(X;Y,Z) \geq I(X;Y),$$

with equality if and only if $X \to Y \to Z$ forms a Markov chain.

Lemma 2.41 If $X \to Y \to Z$ forms a Markov chain, then

$$I(X;Z) \leq I(X;Y)$$

and

$$I(X;Z) \leq I(Y;Z).$$

Corollary

- If $X \to Y \to Z$, then $H(X|Z) \geq H(X|Y)$.
- Suppose Y is an observation of X. Then further processing of Y can only increase the uncertainty about X on the average.

Theorem 2.42 (Data Processing Theorem) If $U \to X \to Y \to V$ forms a Markov chain, then

$$I(U;V) \leq I(X;Y).$$

Fano's Inequality

Theorem 2.43 For any random variable X,

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

where $|\mathcal{X}|$ denotes the size of the alphabet $\mathcal{X}$. This upper bound is tight if and only if X is distributed uniformly on $\mathcal{X}$.

Corollary 2.44 The entropy of a random variable may take any nonnegative real value.

Remark The entropy of a random variable

- is finite if its alphabet is finite.
- can be finite or infinite if its alphabet is finite (see Examples 2.45 and 2.46).

Theorem 2.47 (Fano's Inequality) Let X and $\hat{X}$ be random variables taking values in the same alphabet $\mathcal{X}$. Then

$$H(X|\hat{X}) \leq h_b(P_e) + P_e \log(|\mathcal{X}| - 1),$$

where h_b is the binary entropy function.

Corollary 2.48 $H(X|\hat{X}) < 1 + P_e \log |\mathcal{X}|$.

Interpretation

- For finite alphabet, if $P_e \to 0$, then $H(X|\hat{X}) \to 0$.
- This may NOT hold for countably infinite alphabet (see Example 2.49).