Chapter 2

Information Measures for Discrete Systems

Po-Ning Chen, Professor

Institute of Communications Engineering

National Chiao Tung University

Hsin Chu, Taiwan 30010, R.O.C.

2.1.1 Self-information

- Self-information, denoted by $\mathcal{I}(E)$, is the information you gain by learning an event E has occurred.
- What properties should $\mathcal{I}(E)$ have?
  1. $\mathcal{I}(E)$ is a decreasing function of $p_E := \Pr(E)$, i.e., $\mathcal{I}(E) = I(p_E)$.
     - The less likely event E is, the more information is gained when one learns it has occurred.
     - Here, $\mathcal{I}(\cdot)$ is a function defined over the event space, and $I(\cdot)$ is a function defined over $[0, 1]$.
  2. $I(p_E)$ is continuous in p_E.
     - Intuitively, one should expect that a small change in p_E corresponds to a small change in the amount of information carried by E.
  3. If $E_1 \perp\!\!\!\perp E2$, where $\perp\!\!\!\perp \equiv$ independence, then $\mathcal{I}(E_1 \cap E_2) = \mathcal{I}(E_1) + \mathcal{I}(E_2)$, or equivalently, $I(p_{E_1} \times p_{E_2}) = I(p_{E_1}) + I(p_{E_2})$.
     - The amount of information one gains by learning that two independent events have jointly occurred should be equal to the sum of the amounts of information of each individual event.
  4. $\mathcal{I}(E) \geq 0$. (Optional but automatically satisfied for the one-and-only function that satisfies the previous three properties.)

2.1.1 Self-information

Theorem 2.1 The *only* function defined over $p \in (0, 1]$ and satisfying

1. $I(p)$ is monotonically decreasing in p;
2. $I(p)$ is a continuous function of p for $0 < p \leq 1$;
3. $I(p_1 \times p_2) = I(p_1) + I(p_2)$;

is $I(p) = -c \cdot \log_b(p)$, where c is a positive constant and the base b of the logarithm is any number larger than one.

Proof: The proof is completed in three steps.

Step 1: $I(p) = -c \cdot \log_b(p)$ is true for $p = 1/n$ for any positive integer n.

Step 2: $I(p) = -c \cdot \log_b(p)$ is true for positive rational number p.

Step 3: $I(p) = -c \cdot \log_b(p)$ is true for real-valued p.

2.1.1 Self-information

Step 1: Claim. For $n = 1, 2, 3, \ldots$,

$$I\left(\frac{1}{n}\right) = -c \cdot \log_b\left(\frac{1}{n}\right).$$

Proof:

$(n = 1)$ Condition 3 $\quad\Rightarrow\quad I(1) = I(1) + I(1) \quad\Rightarrow\quad I(1) = 0 = -c \cdot \log_b(1)$.

$(n > 1)$ For any positive integer r, $\exists$ non-negative integer k such that

$$n^k \le 2^r < n^{k+1} \Rightarrow I\left(\frac{1}{n^k}\right) \le I\left(\frac{1}{2^r}\right) < I\left(\frac{1}{n^{k+1}}\right) \text{ by Condition 1}$$

$\Rightarrow$ By Condition 3,

$$k \cdot I\left(\frac{1}{n}\right) \le r \cdot I\left(\frac{1}{2}\right) < (k+1) \cdot I\left(\frac{1}{n}\right).$$

Hence, since $I(1/n) > I(1) = 0$,

$$\frac{k}{r} \le \frac{I(1/2)}{I(1/n)} \le \frac{k+1}{r}.$$

On the other hand, by the monotonity of the logarithm, we obtain

$$\log_b n^k \le \log_b 2^r \le \log_b n^{k+1} \iff \frac{k}{r} \le \frac{\log_b(2)}{\log_b(n)} \le \frac{k+1}{r}.$$

Therefore,

$$\left| \frac{\log_b(2)}{\log_b(n)} - \frac{I(1/2)}{I(1/n)} \right| < \frac{1}{r}.$$

Since $n > 1$ is fixed, and r can be made arbitrarily large, we can let $r \to \infty$ to get:

$$I\left(\frac{1}{n}\right) = \frac{I(1/2)}{\log_b(2)} \cdot \log_b(n) = -c \cdot \log_b\left(\frac{1}{n}\right),$$

where $c = I(1/2)/\log_b(2) > 0$. This completes the proof of the claim.

Step 2: Claim. $I(p) = -c \cdot \log_b(p)$ for positive rational number p.

Proof: A rational number p can be represented by $p = r/s$, where r and s are both positive integers. Then Condition 3 gives that

$$I\left(\frac{1}{s}\right) = I\left(\frac{r}{s}\frac{1}{r}\right) = I\left(\frac{r}{s}\right) + I\left(\frac{1}{r}\right),$$

which, from Step 1, implies that

$$I(p) = I\left(\frac{r}{s}\right) = I\left(\frac{1}{s}\right) - I\left(\frac{1}{r}\right) = c \cdot \log_b s - c \cdot \log_b r = -c \cdot \log_b p.$$

Step 3: For any $p \in [0, 1]$, it follows by continuity (i.e., Condition 2) that

$$I(p) = \lim_{a \uparrow p,\ a \text{ rational}} I(a) = \lim_{b \downarrow p,\ b \text{ rational}} I(b) = -c \cdot \log_b(p). \qquad \square$$

Uncertainty and information

Summary:

- After observing event E with $\Pr(E) = p$, you gain information $I(p)$.
- Equivalently, after observing event E with $\Pr(E) = p$, you lose uncertainty $I(p)$.
- The amount of information gained = The amount of uncertainty lost

2.1.2 Entropy

- Self-information for outcome x (or elementary event $\{X = x\}$)

$$\mathcal{I}(x) := \log_b \frac{1}{P_X(x)},$$

where the constant c in the previous theorem is chosen to be 1.

- Entropy = expected self-information

$$H(X) := E[\mathcal{I}(X)] = \sum_{x \in \mathcal{X}} P_X(x) \log_b \frac{1}{P_X(x)}.$$

  – Units of entropy

    * $\log_2 =$ bits
    * $\log = \log_e = \ln =$ nats

  – Example. Binary entropy function.

$$\begin{aligned} H(X) &= -p \cdot \log p - (1-p)\log(1-p) \text{ nats} \\ &= -p \cdot \log_2 p - (1-p)\log_2(1-p) \text{ bits} \end{aligned}$$

for $P_X(1) = 1 - P_X(0) = p$.

2.1.3 Properties of entropy

Definition 2.2 (Entropy) The entropy of a discrete random variable X with pmf $P_X(\cdot)$ is denoted by $H(X)$ or $H(P_X)$ and defined by

$$H(X) := -\sum_{x\in\mathcal{X}} P_X(x)\cdot\log_2 P_X(x) \quad \text{(bits)}.$$

Assumption. The alphabet $\mathcal{X}$ of the random variable X is finite.

Lemma 2.4 (Fundamental inequality (FI)) For any $x > 0$ and $D > 1$, we have that

$$\log_D(x) \le \log_D(e)\cdot(x-1)$$

with equality if and only if (iff) $x = 1$.

Lemma 2.5 (Non-negativity) $H(X) \ge 0$. Equality holds iff X is deterministic (when X is deterministic, the uncertainty of X is obviously zero).

Proof: $0 \le P_X(x) \le 1$ implies that $\log_2[1/P_X(x)] \ge 0$ for every $x \in \mathcal{X}$. Hence,

$$H(X) = \sum_{x\in\mathcal{X}} P_X(x)\log_2\frac{1}{P_X(x)} \ge 0,$$

with equality holding iff $P_X(x) = 1$ for some $x \in \mathcal{X}$. □

2.1.3 Properties of entropy

Comment: When X is deterministic, the uncertainty of X is obviously zero.

Lemma 2.6 (Upper bound on entropy) If a random variable X takes values from a finite set $\mathcal{X}$, then

$$H(X) \le \log_2 |\mathcal{X}|,$$

where $|\mathcal{X}|$ denotes the size of the set $\mathcal{X}$. Equality holds iff X is equiprobable or uniformly distributed over $\mathcal{X}$ (i.e., $P_X(x) = \frac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$).

- *Interpretation:* Uniform distribution maximizes entropy.
- *Hint of proof*: Subtract one side of the inequality by the other side, and apply the *fundamental inequality* or *log-sum inequality*.

2.1.3 Properties of entropy

Proof:

$$
\begin{aligned}
\log_2|\mathcal{X}| - H(X) &= \log_2|\mathcal{X}| \times \left(\sum_{x\in\mathcal{X}} P_X(x)\right) - \left(-\sum_{x\in\mathcal{X}} P_X(x)\log_2 P_X(x)\right)\\
&= \sum_{x\in\mathcal{X}} P_X(x) \times \log_2|\mathcal{X}| + \sum_{x\in\mathcal{X}} P_X(x)\log_2 P_X(x)\\
&= \sum_{x\in\mathcal{X}} P_X(x)\log_2\big(|\mathcal{X}| \times P_X(x)\big)\\
&\geq \sum_{x\in\mathcal{X}} P_X(x)\cdot\log_2(e)\left(1 - \frac{1}{|\mathcal{X}| \times P_X(x)}\right)\\
&= \log_2(e)\sum_{x\in\mathcal{X}}\left(P_X(x) - \frac{1}{|\mathcal{X}|}\right)\\
&= \log_2(e)\cdot(1-1) = 0
\end{aligned}
$$

where the inequality follows from the FI Lemma, with equality iff $(\forall\ x \in \mathcal{X})$, $|\mathcal{X}| \times P_X(x) = 1$, which means $P_X(\cdot)$ is a uniform distribution on $\mathcal{X}$. □

2.1.3 Properties of entropy

Lemma 2.7 (Log-sum inequality) For non-negative numbers, a_1, a_2, ..., a_n and b_1, b_2, ..., b_n,

$$\sum_{i=1}^{n}\left(a_i \log_D \frac{a_i}{b_i}\right) \geq \left(\sum_{i=1}^{n} a_i\right) \log_D \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{2.1.1}$$

with equality holding iff $(\forall\ 1 \leq i \leq n)$ $(a_i/b_i) = (a_1/b_1)$, a constant independent of i.

(By convention, $0 \cdot \log_D(0) = 0$, $0 \cdot \log_D(0/0) = 0$ and $a \cdot \log_D(a/0) = \infty$ if $a > 0$. This can be justified by "continuity.")

- *Comment:* A tip for memorizing the log-sum inequality: log-first $\geq$ sum-first.
- *Hint of proof*: Subtract one side of the inequality by the other side, and apply the *fundamental inequality*.

2.1.4 Joint entropy and conditional entropy

Definition 2.8 (Joint entropy) The joint entropy $H(X,Y)$ of random variables (X,Y) is defined by

$$\begin{aligned} H(X,Y) &:= -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\cdot\log_2 P_{X,Y}(x,y) \\ &= E[-\log_2 P_{X,Y}(X,Y)]. \end{aligned}$$

Definition 2.9 (Conditional entropy) Given two jointly distributed random variables X and Y, the conditional entropy $H(Y|X)$ of Y given X is defined by

$$H(Y|X) := \sum_{x\in\mathcal{X}} P_X(x)\left(-\sum_{y\in\mathcal{Y}} P_{Y|X}(y|x)\cdot\log_2 P_{Y|X}(y|x)\right) \tag{2.1.5}$$

where $P_{Y|X}(\cdot|\cdot)$ is the conditional pmf of Y given X.

2.1.4 Joint entropy and conditional entropy

Theorem 2.10 (Chain rule for entropy)

$$H(X,Y) = H(X) + H(Y|X). \tag{2.1.6}$$

Proof: Since

$$P_{X,Y}(x,y) = P_X(x)P_{Y|X}(y|x),$$

we directly obtain that

$$\begin{aligned} H(X,Y) &= E[-\log_2 P_{X,Y}(X,Y)] \\ &= E[-\log_2 P_X(X)] + E[-\log_2 P_{Y|X}(Y|X)] \\ &= H(X) + H(Y|X). \end{aligned}$$

$\square$

Corollary 2.11 (Chain rule for conditional entropy)

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

2.1.5 Properties of joint and conditional entropy

Lemma 2.12 (Conditioning never increases entropy) Side information Y decreases the uncertainty about X:

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

with equality holding iff X and Y are independent. In other words, "conditioning" reduces entropy.

- *Interpretation*: Only when X is independent of Y, the pre-given Y will be of no help in determining X.
- *Hint of proof*: Subtract one side of the inequality by the other side, and apply the *fundamental inequality* or *log-sum inequality*.

2.1.5 Properties of joint and conditional entropy

Proof:

$$
\begin{aligned}
H(X) - H(X|Y) &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\cdot\log_2\frac{P_{X|Y}(x|y)}{P_X(x)} \\
&= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\cdot\log_2\frac{P_{X|Y}(x|y)P_Y(y)}{P_X(x)P_Y(y)} \\
&= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\cdot\log_2\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} \\
&\geq \left(\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)\right)\log_2\frac{\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X,Y}(x,y)}{\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_X(x)P_Y(y)} \\
&= 0
\end{aligned}
$$

where the inequality follows from the log-sum inequality, with equality holding iff

$$
\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} = \text{constant} \quad \forall\ (x,y)\in\mathcal{X}\times\mathcal{Y}.
$$

Since probability must sum to 1, the above constant equals 1, which is exactly the case of X being independent of Y. □

2.1.5 Properties of joint and conditional entropy

Lemma 2.13 Entropy is additive for independent random variables; i.e.,

$$H(X,Y) = H(X) + H(Y) \quad \text{for independent } X \text{ and } Y.$$

Proof: By the previous lemma, independence of X and Y implies $H(Y|X) = H(Y)$. Hence

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

$\square$

- In general, $H(X,Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$.

2.1.5 Properties of joint and conditional entropy

Lemma 2.14 Conditional entropy is lower additive; i.e.,

$$H(X_1, X_2|Y_1, Y_2) \leq H(X_1|Y_1) + H(X_2|Y_2).$$

Equality holds iff

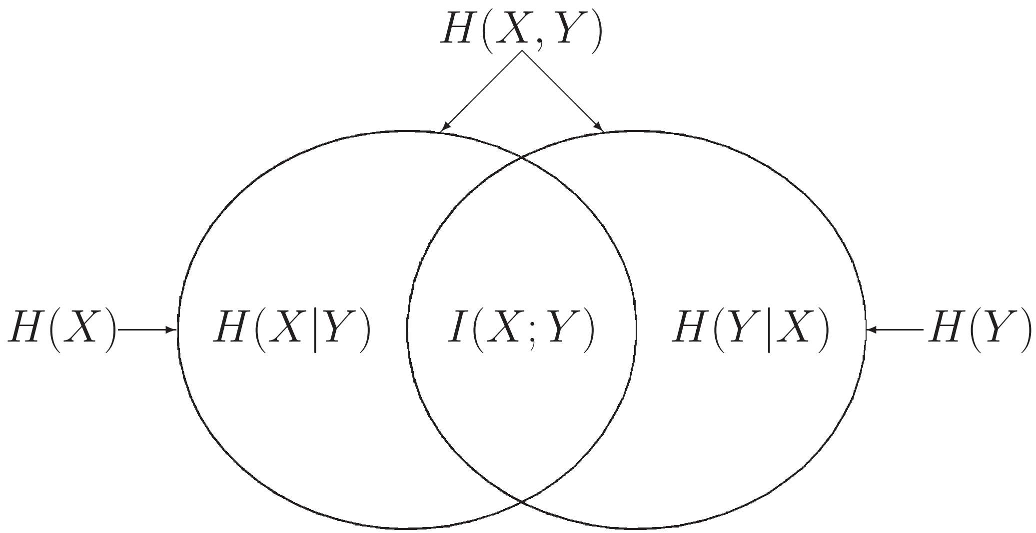
$$P_{X_1,X_2|Y_1,Y_2}(x_1, x_2|y_1, y_2) = P_{X_1|Y_1}(x_1|y_1)P_{X_2|Y_2}(x_2|y_2)$$

for all x_1, x_2, y_1 and y_2.

2.2 Mutual information

- Definition of mutual information

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



Relation between entropy and mutual information.

2.2.1 Properties of mutual information

Lemma 2.15

1. $I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x,y) \log_2 \dfrac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}$.
2. $I(X;Y) = I(Y;X)$.
3. $I(X;Y) = H(X) + H(Y) - H(X,Y)$.
4. $I(X;Y) \leq H(X)$ with equality holding iff X is a function of Y (i.e., $X = f(Y)$ for some function $f(\cdot)$).
5. $I(X;Y) \geq 0$ with equality holding iff X and Y are independent.
6. $I(X;Y) \leq \min\{\log_2 |\mathcal{X}|, \log_2 |\mathcal{Y}|\}$.

2.2.1 Properties of mutual information

Lemma 2.16 (Chain rule for mutual information)

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

Proof: Without loss of generality, we only prove the second equality:

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

□

$$\begin{tikzcd}
X \arrow[rr, shift left] & & Y \arrow[ll, shift left] \\
& Z \arrow[ul, dotted, bend left] \arrow[ur, dotted, bend right] &
\end{tikzcd}$$

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

2.3 Properties of entropy and mutual information for multiple random variables

Theorem 2.17 (Chain rule for entropy) Let X_1, X_2, ..., X_n be drawn according to $P_{X^n}(x^n) := P_{X_1,\cdots,X_n}(x_1,\ldots,x_n)$, where we use the common superscript notation to denote an n-tuple: $X^n := (X_1,\ldots,X_n)$ and $x^n := (x_1,\ldots,x_n)$. Then

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

where $H(X_i|X_{i-1},\ldots,X_1) := H(X_1)$ for $i = 1$. (The above chain rule can also be written as:

$$H(X^n) = \sum_{i=1}^{n} H(X_i|X^{i-1}),$$

where $X^i := (X_1,\ldots,X_i)$.)

Theorem 2.18 (Chain rule for conditional entropy)

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

2.3 Properties of entropy and mutual information for multiple random variables

Theorem 2.19 (Chain rule for mutual information)

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

where $I(X_i; Y | X_{i-1}, \ldots, X_1) := I(X_1; Y)$ for $i = 1$.

Theorem 2.20 (Independence bound on entropy)

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

Equality holds iff all the X_i's are independent from each other.

- This condition is equivalent to requiring that X_i be independent of $(X_{i-1}, \ldots, X_1)$ for all i. The equivalence can be directly proved using the chain rule for joint probabilities, i.e., $P_{X^n}(x^n) = \prod_{i=1}^{n} P_{X_i|X_1^{i-1}}(x_i|x_1^{i-1})$.

2.3 Properties of entropy and mutual information for multiple random variables

Theorem 2.21 (Bound on mutual information) If $\{(X_i, Y_i)\}_{i=1}^n$ is a process satisfying the conditional independence assumption $P_{Y^n|X^n} = \prod_{i=1}^n P_{Y_i|X_i}$, then

$$I(X_1, \ldots, X_n; Y_1, \ldots, Y_n) \leq \sum_{i=1}^n I(X_i; Y_i)$$

with equality holding iff $\{X_i\}_{i=1}^n$ are independent.

2.4 Data processing inequality

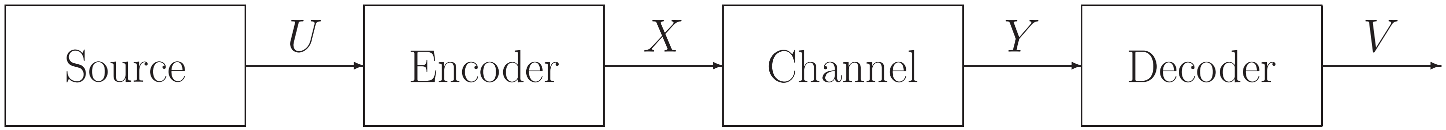
Lemma 2.22 (Data processing inequality) (This is also called the *data processing lemma.*) If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$.

Proof: Since $X \to Y \to Z$, we directly have that $I(X;Z|Y) = 0$. By the chain rule for mutual information,

$$\begin{aligned} I(X;Z) + I(X;Y|Z) &= I(X;Y,Z) && (2.4.1) \\ &= I(X;Y) + I(X;Z|Y) \\ &= I(X;Y). && (2.4.2) \end{aligned}$$

Since $I(X;Y|Z) \geq 0$, we obtain that $I(X;Y) \geq I(X;Z)$ with equality holding iff $I(X;Y|Z) = 0$. $\square$

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



"By processing, we can only reduce the (mutual) information, but the processed information may be in a more *useful* form!"

Communication context of the data processing lemma.

2.4 Data processing inequality

Corollary 2.23 For jointly distributed random variables X and Y and any function $g(\cdot)$, we have $X \to Y \to g(Y)$ and

$$I(X;Y) \geq I(X;g(Y)).$$

Corollary 2.24 If $X \to Y \to Z$, then

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

- *Interpretation*: For Z, all the information about X is obtained from Y; hence, giving Z will not help increasing the "mutual information" between X and Y.

- Without the condition of $X \to Y \to Z$, both $I(X;Y|Z) \leq I(X;Y)$ and $I(X;Y|Z) > I(X;Y)$ could happen.

  E.g. let X and Y be independent equiprobable binary zero-one random variables, and let $Z = X + Y$; hence, $Z \in \{0, 1, 2\}$. Then $I(X;Y) = 0$; but

$$\begin{aligned}
&I(X;Y|Z)\\
&\quad = H(X|Z) - H(X|Y,Z) = H(X|Z)\\
&\quad = P_Z(0)H(X|Z=0) + P_Z(1)H(X|Z=1) + P_Z(2)H(X|Z=2)\\
&\quad = 0 + 0.5 + 0 = 0.5 \text{ bit.}
\end{aligned}$$

2.4 Data processing inequality

Corollary 2.25 If $X_1 \to X_2 \to \cdots \to X_n$, then for any i, j, k, l such that $1 \leq i \leq j \leq k \leq l \leq n$, we have that

$$I(X_i; X_l) \leq I(X_j; X_k).$$

2.5 Fano's inequality

Lemma 2.26 (Fano's inequality) Let X and Y be two random variables, correlated in general, with alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, where $\mathcal{X}$ is finite but $\mathcal{Y}$ can be countably infinite. Let $\hat{X} := g(Y)$ be an estimate of X from observing Y, where $g : \mathcal{Y} \to \mathcal{X}$ is a given estimation function. Define the probability of error as

$$P_e := \Pr[\hat{X} \neq X].$$

Then the following inequality holds

$$H(X|Y) \leq h_b(P_e) + P_e \cdot \log_2(|\mathcal{X}| - 1), \tag{2.5.1}$$

where $h_{\mathrm{b}}(x) := -x \log_2 x - (1-x) \log_2(1-x)$ for $0 \leq x \leq 1$ is the binary entropy function.

2.5 Fano's inequality

Observation 2.27

- If $P_e = 0$ for some estimator $g(\cdot)$, then $H(X|Y) = 0$.
- A weaker but simpler version of Fano's inequality can be directly obtained from (2.5.1) by noting that $h_b(P_e) \le 1$:

$$H(X|Y) \le 1 + P_e \cdot \log_2(|\mathcal{X}| - 1), \tag{2.5.2}$$

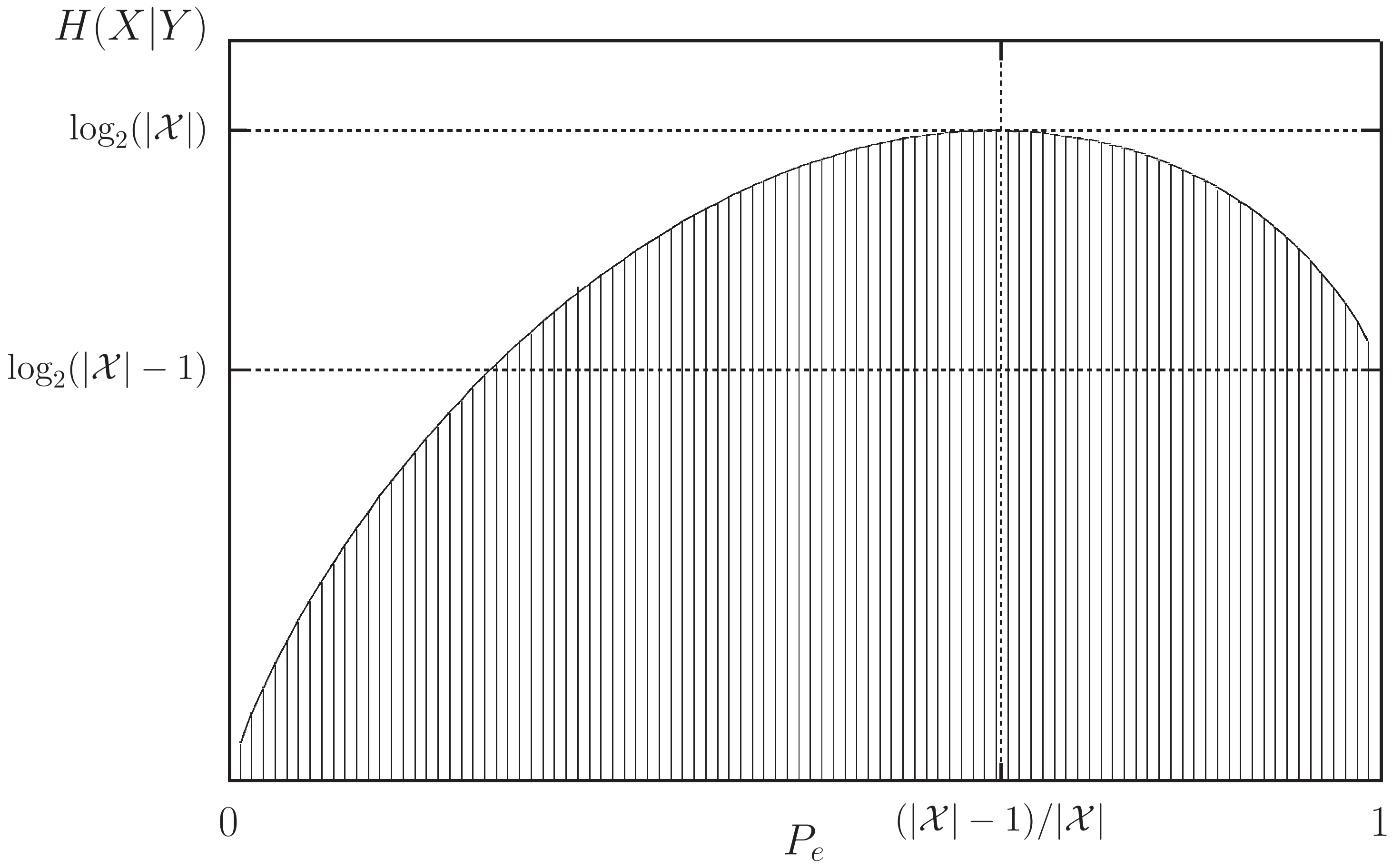
which in turn yields that

$$P_e \ge \frac{H(X|Y) - 1}{\log_2(|\mathcal{X}| - 1)} \quad (\text{ for } |\mathcal{X}| > 2).$$

So, **Fano's inequality provides a lower bound to P_e (for arbitrary estimators).**

2.5 Fano's inequality

- In fact, Fano's inequality yields both upper and lower bounds on P_e in terms of $H(X|Y)$.



Permissible $(P_e, H(X|Y))$ region due to Fano's inequality.

2.5 Fano's inequality

(A quick) Proof of Lemma 2.26:

- Define a new random variable,

$$E := \begin{cases} 1, & \text{if } g(Y) \neq X \\ 0, & \text{if } g(Y) = X \end{cases}.$$

- Then using the chain rule for conditional entropy, we obtain

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

- Observe that E is a function of X and Y; hence, $H(E|X, Y) = 0$.
- Since conditioning never increases entropy, $H(E|Y) \leq H(E) = h_b(P_e)$.
- The remaining term, $H(X|E, Y)$, can be bounded as follows:

$$\begin{aligned} H(X|E, Y) &= \Pr[E = 0]H(X|Y, E = 0) + \Pr[E = 1]H(X|Y, E = 1) \\ &\leq (1 - P_e) \cdot 0 + P_e \cdot \log_2(|\mathcal{X}| - 1), \end{aligned}$$

  since $X = g(Y)$ for $E = 0$, and given $E = 1$, we can upper bound the conditional entropy by the logarithm of the number of remaining outcomes, i.e., $(|\mathcal{X}| - 1)$.

- Combining these results completes the proof. $\square$

2.5 Fano's inequality

- Fano's inequality cannot be improved in the sense that the lower bound, $H(X|Y)$, can be achieved for some specific cases (See Example 2.28 in the text); so it is a sharp bound.

Definition. A bound is said to be *sharp* if the bound is achievable for *some specific* cases. A bound is said to be *tight* if the bound is achievable for *all* cases.

2.5 Fano's inequality

Alternative proof of Fano's inequality:

- Noting that $X \to Y \to \hat{X}$ form a Markov chain, we directly obtain via the data processing inequality that

$$I(X;Y) \geq I(X;\hat{X}),$$

which implies that

$$H(X|Y) \leq H(X|\hat{X}).$$

- Thus, if we show that $H(X|\hat{X})$ is no larger than the right-hand side of (2.5.1), the proof of (2.5.1) is complete. I.e.,

$$H(X|\hat{X}) \leq h_b(P_e) + P_e \cdot \log_2(|\mathcal{X}| - 1),$$

2.5 Fano's inequality

- Noting that

$$P_e = \sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\mathcal{X}:\hat{x}\neq x} P_{X,\hat{X}}(x,\hat{x})$$

and

$$1 - P_e = \sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\mathcal{X}:\hat{x}=x} P_{X,\hat{X}}(x,\hat{x}) = \sum_{x\in\mathcal{X}} P_{X,\hat{X}}(x,x),$$

we obtain that

$$\begin{aligned}
&H(X|\hat{X}) - h_b(P_e) - P_e \log_2(|\mathcal{X}|-1)\\
&= \underbrace{\sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\mathcal{X}:\hat{x}\neq x} P_{X,\hat{X}}(x,\hat{x}) \log_2 \frac{1}{P_{X|\hat{X}}(x|\hat{x})} + \sum_{x\in\mathcal{X}} P_{X,\hat{X}}(x,x) \log_2 \frac{1}{P_{X|\hat{X}}(x|x)}}_{H(X|\hat{X})}\\
&\quad - \underbrace{\left[\sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\mathcal{X}:\hat{x}\neq x} P_{X,\hat{X}}(x,\hat{x})\right]}_{P_e} \log_2 \frac{(|\mathcal{X}|-1)}{P_e} + \underbrace{\left[\sum_{x\in\mathcal{X}} P_{X,\hat{X}}(x,x)\right]}_{1-P_e} \log_2(1-P_e)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\mathcal{X}:\hat{x}\neq x} P_{X,\hat{X}}(x,\hat{x}) \log_2 \frac{P_e}{P_{X|\hat{X}}(x|\hat{x})(|\mathcal{X}|-1)} \\
&\qquad + \sum_{x\in\mathcal{X}} P_{X,\hat{X}}(x,x) \log_2 \frac{1-P_e}{P_{X|\hat{X}}(x|x)} \qquad (2.5.3) \\
&\le \log_2(e) \sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\mathcal{X}:\hat{x}\neq x} P_{X,\hat{X}}(x,\hat{x}) \left[\frac{P_e}{P_{X|\hat{X}}(x|\hat{x})(|\mathcal{X}|-1)} - 1\right] \quad \text{(FI lemma)} \\
&\qquad + \log_2(e) \sum_{x\in\mathcal{X}} P_{X,\hat{X}}(x,x) \left[\frac{1-P_e}{P_{X|\hat{X}}(x|x)} - 1\right] \\
&= \log_2(e) \left[\frac{P_e}{(|\mathcal{X}|-1)} \sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\mathcal{X}:\hat{x}\neq x} P_{\hat{X}}(\hat{x}) - \sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\mathcal{X}:\hat{x}\neq x} P_{X,\hat{X}}(x,\hat{x})\right] \\
&\qquad + \log_2(e) \left[(1-P_e) \sum_{x\in\mathcal{X}} P_{\hat{X}}(x) - \sum_{x\in\mathcal{X}} P_{X,\hat{X}}(x,x)\right] \\
&= \log_2(e) \left[\frac{P_e}{(|\mathcal{X}|-1)}(|\mathcal{X}|-1) - P_e\right] + \log_2(e)\left[(1-P_e)-(1-P_e)\right] \\
&= 0
\end{aligned}
$$

$\square$

2.6 Divergence and variational distance

Definition 2.29 (Divergence) Given two discrete random variables X and $\hat{X}$ defined over a common alphabet $\mathcal{X}$, the divergence or the *Kullback-Leibler divergence or distance* (other names are *relative entropy* and *discrimination*) is denoted by $D(X\|\hat{X})$ or $D(P_X\|P_{\hat{X}})$ and defined by

$$D(X\|\hat{X}) = D(P_X\|P_{\hat{X}}) := E_X\left[\log_2 \frac{P_X(X)}{P_{\hat{X}}(X)}\right] = \sum_{x\in\mathcal{X}} P_X(x)\log_2\frac{P_X(x)}{P_{\hat{X}}(x)}.$$

Why name it relative entropy?

- $D(X\|\hat{X})$ is also called *relative entropy* since it can be regarded as a measure of the inefficiency of mistakenly assuming that the distribution of a source is $P_{\hat{X}}$ when the true distribution is P_X.
- Specifically, if we mistakenly thought that the "true" distribution is $P_{\hat{X}}$ and employ the "best" code corresponding to $P_{\hat{X}}$, then the resultant average codeword length becomes
$$\sum_{x\in\mathcal{X}}[-P_X(x)\cdot\log_2 P_{\hat{X}}(x)].$$
As a result, the *relative* difference between the resultant average codeword length and $H(X)$ is the *relative entropy* $D(X\|\hat{X})$.

2.6 Divergence and variational distance

- Computation conventions from continuity

$$0 \cdot \log \frac{0}{p} = 0 \quad \text{and} \quad p \cdot \log \frac{p}{0} = \infty \quad \text{for } p > 0.$$

Lemma 2.30 (Non-negativity of divergence)

$$D(X\|\hat{X}) \geq 0,$$

with equality iff $P_X(x) = P_{\hat{X}}(x)$ for all $x \in \mathcal{X}$ (i.e., the two distributions are equal).

Proof:

$$\begin{aligned} D(X\|\hat{X}) &= \sum_{x\in\mathcal{X}} P_X(x) \log_2 \frac{P_X(x)}{P_{\hat{X}}(x)} \\ &\geq \left(\sum_{x\in\mathcal{X}} P_X(x)\right) \log_2 \frac{\sum_{x\in\mathcal{X}} P_X(x)}{\sum_{x\in\mathcal{X}} P_{\hat{X}}(x)} \\ &= 0, \end{aligned}$$

where the second step follows from the log-sum inequality with equality holding iff for every $x \in \mathcal{X}$,

$$\frac{P_X(x)}{P_{\hat{X}}(x)} = \frac{\sum_{a\in\mathcal{X}} P_X(a)}{\sum_{b\in\mathcal{X}} P_{\hat{X}}(b)} = 1,$$

or equivalently $P_X(x) = P_{\hat{X}}(x)$ for all $x \in \mathcal{X}$. $\square$

2.6 Divergence and variational distance

Lemma 2.31 (Mutual information and divergence)

$$I(X;Y) = D(P_{X,Y} \| P_X \times P_Y),$$

where $P_{X,Y}(\cdot,\cdot)$ is the joint distribution of the random variables X and Y and $P_X(\cdot)$ and $P_Y(\cdot)$ are the respective marginals.

Definition 2.32 (Refinement of distribution) Given the distribution P_X on $\mathcal{X}$, divide $\mathcal{X}$ into k mutually disjoint sets, $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_k$, satisfying

$$\mathcal{X} = \bigcup_{i=1}^{k} \mathcal{U}_i.$$

Define a new distribution P_U on $\mathcal{U} = \{1, 2, \ldots, k\}$ as

$$P_U(i) = \sum_{x \in \mathcal{U}_i} P_X(x).$$

Then P_X is called a *refinement* (or more specifically, a *k-refinement*) of P_U.

2.6 Divergence and variational distance

Lemma 2.33 (Refinement cannot decrease divergence) Let P_X and $P_{\hat{X}}$ be the refinements (k-refinements) of P_U and $P_{\hat{U}}$ respectively. Then

$$D(P_X \| P_{\hat{X}}) \geq D(P_U \| P_{\hat{U}}).$$

Proof: By the **log-sum inequality**, we obtain that for any $i \in \{1, 2, \ldots, k\}$

$$\begin{aligned} \sum_{x \in \mathcal{U}_i} P_X(x) \log_2 \frac{P_X(x)}{P_{\hat{X}}(x)} &\geq \left(\sum_{x \in \mathcal{U}_i} P_X(x) \right) \log_2 \frac{\sum_{x \in \mathcal{U}_i} P_X(x)}{\sum_{x \in \mathcal{U}_i} P_{\hat{X}}(x)} \\ &= P_U(i) \log_2 \frac{P_U(i)}{P_{\hat{U}}(i)}, \end{aligned} \quad (2.6.1)$$

with equality iff

$$\frac{P_X(x)}{P_{\hat{X}}(x)} = \frac{P_U(i)}{P_{\hat{U}}(i)}$$

for all $x \in \mathcal{U}$.

Hence,

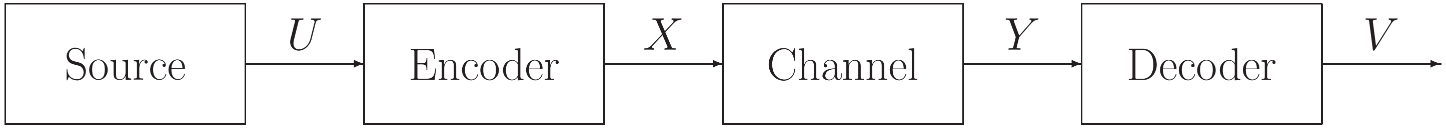
$$
\begin{aligned}
D(P_X \| P_{\hat{X}}) &= \sum_{i=1}^{k} \sum_{x \in \mathcal{U}_i} P_X(x) \log_2 \frac{P_X(x)}{P_{\hat{X}}(x)} \\
&\geq \sum_{i=1}^{k} P_U(i) \log_2 \frac{P_U(i)}{P_{\hat{U}}(i)} \\
&= D(P_U \| P_{\hat{U}}),
\end{aligned}
$$

with equality iff

$$
\frac{P_X(x)}{P_{\hat{X}}(x)} = \frac{P_U(i)}{P_{\hat{U}}(i)}
$$

for all i and $x \in \mathcal{U}_i$. $\square$

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



"By processing, we can only reduce the (mutual) information,
but the processed information may be in a more *useful* form!"

Communication context of the data processing lemma.

2.6 Divergence and variational distance

- Processing of information can be modeled as a (many-to-one) mapping, and refinement is actually the reverse operation.
- Recall that the *data processing lemma* shows that mutual information can never increase due to *processing*. Hence, if one wishes to increase mutual information, he should "anti-process" (or refine) the involved statistics.
- From Lemma 2.31, the mutual information can be viewed as the divergence of a joint distribution against the product distribution of the marginals. It is therefore reasonable to expect that a similar effect due to *processing* (or a reverse effect due to *refinement*) should also apply to *divergence*. This is shown in the next lemma.

- Processing only decreases mutual information and divergence.
- Only by refinement can mutual information and divergence be increased.

2.6 Divergence and variational distance

- Divergence is not a *distance*, a drawback in certain applications.

Given a non-empty set A, the function $d\colon A \times A \to [0, \infty)$ is called a *distance* or *metric* if it satisfies the following properties.

1. Non-negativity: $d(a, b) \geq 0$ for every $a, b \in A$ with equality holding iff $a = b$.
2. ~~Symmetry~~: $d(a, b) = d(b, a)$ for every $a, b \in A$.
3. ~~Triangular inequality~~: $d(a, b) + d(b, c) \geq d(a, c)$ for every $a, b, c \in A$.

Definition 2.35 (Variational distance) The *variational distance* (also known as the $\mathcal{L}_1$-distance, the total variation distance, the statistical distance) between two distributions P_X and $P_{\hat{X}}$ with common alphabet $\mathcal{X}$ is defined by

$$\|P_X - P_{\hat{X}}\| := \sum_{x \in \mathcal{X}} \left| P_X(x) - P_{\hat{X}}(x) \right|.$$

Lemma 2.36 The variational distance satisfies

$$\|P_X - P_{\hat{X}}\| = 2 \cdot \sum_{x \in \mathcal{X}\colon P_X(x) > P_{\hat{X}}(x)} \left(P_X(x) - P_{\hat{X}}(x) \right) = 2 \cdot \sup_{E \subset \mathcal{X}} \left| P_X(E) - P_{\hat{X}}(E) \right|.$$

2.6 Divergence and variational distance

Lemma 2.37 (Variational distance vs divergence: Pinsker's inequality)

$$D(X\|\hat{X}) \geq \frac{\log_2(e)}{2} \cdot \|P_X - P_{\hat{X}}\|^2.$$

This result is referred to as Pinsker's inequality.

Proof:

1. With $\mathcal{A} := \{x \in \mathcal{X} \colon P_X(x) > P_{\hat{X}}(x)\}$, we have from the previous lemma that

$$\|P_X - P_{\hat{X}}\| = 2[P_X(\mathcal{A}) - P_{\hat{X}}(\mathcal{A})].$$

2. Define two random variables U and $\hat{U}$ as:

$$U = \begin{cases} 1, & \text{if } X \in \mathcal{A}, \\ 0, & \text{if } X \in \mathcal{A}^c, \end{cases} \quad \text{and} \quad \hat{U} = \begin{cases} 1, & \text{if } \hat{X} \in \mathcal{A}, \\ 0, & \text{if } \hat{X} \in \mathcal{A}^c. \end{cases}$$

   Then P_X and $P_{\hat{X}}$ are refinements (2-refinements) of P_U and $P_{\hat{U}}$, respectively. From Lemma 2.33, we obtain that

$$D(P_X \| P_{\hat{X}}) \geq D(P_U \| P_{\hat{U}}).$$

3. The proof is complete if we show that

$$\begin{aligned} D(P_U \| P_{\hat{U}}) &\geq 2\log_2(e)[P_X(\mathcal{A}) - P_{\hat{X}}(\mathcal{A})]^2 \\ &= 2\log_2(e)[P_U(1) - P_{\hat{U}}(1)]^2. \end{aligned}$$

   For ease of notations, let $p = P_U(1)$ and $q = P_{\hat{U}}(1)$. Then to prove the above inequality is equivalent to show that

$$p \cdot \ln\frac{p}{q} + (1-p) \cdot \ln\frac{1-p}{1-q} \geq 2(p-q)^2.$$

□

Define

$$f(p,q) := p \cdot \ln \frac{p}{q} + (1-p) \cdot \ln \frac{1-p}{1-q} - 2(p-q)^2,$$

and observe that

$$\frac{df(p,q)}{dq} = (p-q)\left(4 - \frac{1}{q(1-q)}\right) \le 0 \quad \text{for } q \le p.$$

Thus, $f(p,q)$ is non-increasing in q for $q \le p$. Also note that $f(p,q) = 0$ for $q = p$. Therefore,

$$f(p,q) \ge 0 \quad \text{for } q \le p.$$

The proof is completed by noting that

$$f(p,q) \ge 0 \quad \text{for } q \ge p,$$

since $f(1-p, 1-q) = f(p,q)$.

Lemma 2.39 If $D(P_X \| P_{\hat{X}}) < \infty$, then

$$D(P_X \| P_{\hat{X}}) \le \frac{\log_2(e)}{\displaystyle\min_{\{x:\, P_X(x)>0\}} \min\{P_X(x), P_{\hat{X}}(x)\}} \cdot \|P_X - P_{\hat{X}}\|.$$

Definition 2.40 (Conditional divergence) Given three discrete random variables, X, $\hat{X}$ and Z, where X and $\hat{X}$ have a common alphabet $\mathcal{X}$, we define the conditional divergence between X and $\hat{X}$ given Z by

$$\begin{aligned} D(X \| \hat{X} | Z) = D(P_{X|Z} \| P_{\hat{X}|Z} | P_Z) &:= \sum_{z \in \mathcal{Z}} P_Z(z) \sum_{x \in \mathcal{X}} P_{X|Z}(x|z) \log \frac{P_{X|Z}(x|z)}{P_{\hat{X}|Z}(x|z)} \\ &= \sum_{z \in \mathcal{Z}} \sum_{x \in \mathcal{X}} P_{X,Z}(x,z) \log \frac{P_{X|Z}(x|z)}{P_{\hat{X}|Z}(x|z)}. \end{aligned}$$

Similarly, the conditional divergence between $P_{X|Z}$ and $P_{\hat{X}}$ given P_Z is defined as

$$D(P_{X|Z} \| P_{\hat{X}} | P_Z) := \sum_{z \in \mathcal{Z}} P_Z(z) \sum_{x \in \mathcal{X}} P_{X|Z}(x|z) \log \frac{P_{X|Z}(x|z)}{P_{\hat{X}}(z)}.$$

Lemma 2.41 (Conditional mutual information and conditional divergence) Given three discrete random variables X, Y and Z with alphabets $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, respectively, and joint distribution $P_{X,Y,Z}$, we have

$$\begin{aligned} I(X;Y|Z) &= D(P_{X,Y|Z} \| P_{X|Z}P_{Y|Z} | P_Z) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} P_{X,Y,Z}(x,y,z)\log_2 \frac{P_{X,Y|Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)}, \end{aligned}$$

where $P_{X,Y|Z}$ is the conditional joint distribution of X and Y given Z, and $P_{X|Z}$ and $P_{Y|Z}$ are the conditional distributions of X and Y, respectively, given Z.

2.6 Divergence and variational distance

Lemma 2.42 (Chain rule for divergence) Let P_{X^n} and Q_{X^n} be two joint distributions on $\mathcal{X}^n$. We have that

$$D(P_{X_1,X_2}\|Q_{X_1,X_2}) = D(P_{X_1}\|Q_{X_1}) + D(P_{X_2|X_1}\|Q_{X_2|X_1}|P_{X_1}),$$

and more generally,

$$D(P_{X^n}\|Q_{X^n}) = \sum_{i=1}^{n} D(P_{X_i|X^{i-1}}\|Q_{X_i|X^{i-1}}|P_{X^{i-1}}),$$

where $D(P_{X_i|X^{i-1}}\|Q_{X_i|X^{i-1}}|P_{X^{i-1}}) := D(P_{X_1}\|Q_{X_1})$ for $i = 1$.

2.6 Divergence and variational distance

Lemma 2.43 (Conditioning never decreases divergence) For three discrete random variables, X, $\hat{X}$ and Z, where X and $\hat{X}$ have a common alphabet $\mathcal{X}$, we have that

$$D(P_{X|Z}\|P_{\hat{X}|Z}|P_Z) \geq D(P_X\|P_{\hat{X}}).$$

Proof:

$$\begin{aligned}
& D(P_{X|Z}\|P_{\hat{X}|Z}|P_Z) - D(P_X\|P_{\hat{X}}) \\
= & \sum_{z\in\mathcal{Z}}\sum_{x\in\mathcal{X}} P_{X,Z}(x,z)\cdot\log_2\frac{P_{X|Z}(x|z)}{P_{\hat{X}|Z}(x|z)} - \sum_{x\in\mathcal{X}} P_X(x)\cdot\log_2\frac{P_X(x)}{P_{\hat{X}}(x)} \\
= & \sum_{z\in\mathcal{Z}}\sum_{x\in\mathcal{X}} P_{X,Z}(x,z)\cdot\log_2\frac{P_{X|Z}(x|z)}{P_{\hat{X}|Z}(x|z)} - \sum_{x\in\mathcal{X}}\left(\sum_{z\in\mathcal{Z}} P_{X,Z}(x,z)\right)\cdot\log_2\frac{P_X(x)}{P_{\hat{X}}(x)} \\
= & \sum_{z\in\mathcal{Z}}\sum_{x\in\mathcal{X}} P_{X,Z}(x,z)\cdot\log_2\frac{P_{X|Z}(x|z)P_{\hat{X}}(x)}{P_{\hat{X}|Z}(x|z)P_X(x)}
\end{aligned}$$

$$\begin{aligned}
&\geq \sum_{z\in\mathcal{Z}}\sum_{x\in\mathcal{X}} P_{X,Z}(x,z)\cdot\log_2(e)\left(1-\frac{P_{\hat{X}|Z}(x|z)P_X(x)}{P_{X|Z}(x|z)P_{\hat{X}}(x)}\right) \quad \text{(by the FI Lemma)}\\
&= \log_2(e)\left(1-\sum_{x\in\mathcal{X}}\frac{P_X(x)}{P_{\hat{X}}(x)}\sum_{z\in\mathcal{Z}}P_Z(z)P_{\hat{X}|Z}(x|z)\right)\\
&= \log_2(e)\left(1-\sum_{x\in\mathcal{X}}\frac{P_X(x)}{P_{\hat{X}}(x)}P_{\hat{X}}(x)\right)\\
&= \log_2(e)\left(1-\sum_{x\in\mathcal{X}}P_X(x)\right)=0,
\end{aligned}$$

with equality holding iff for all x and z,

$$\frac{P_X(x)}{P_{\hat{X}}(x)}=\frac{P_{X|Z}(x|z)}{P_{\hat{X}|Z}(x|z)}.$$

$\square$

2.6 Divergence and variational distance

Lemma 2.44 (Independent side information does not change divergence) If X is independent of Z and $\hat{X}$ is independent of $\hat{Z}$, where X and Z share a common alphabet with $\hat{X}$ and $\hat{Z}$, respectively, then

$$D(P_{X|Z}\|P_{\hat{X}|\hat{Z}}|P_Z) = D(P_X\|P_{\hat{X}}).$$

Corollary 2.45 (Additivity of divergence under independence) If X is independent of Z and $\hat{X}$ is independent of $\hat{Z}$, where X and Z share a common alphabet with $\hat{X}$ and $\hat{Z}$, respectively, then

$$D(P_{X,Z}\|P_{\hat{X},\hat{Z}}) = D(P_X\|P_{\hat{X}}) + D(P_Z\|P_{\hat{Z}}).$$

Lemma 2.46

1. $H(P_X)$ is a concave function of P_X, namely

$$H(\lambda P_X + (1-\lambda)P_{\widetilde{X}}) \geq \lambda H(P_X) + (1-\lambda)H(P_{\widetilde{X}})$$

for all $\lambda \in [0, 1]$. Equality holds iff $P_X(x) = P_{\widetilde{X}}(x)$ for all x.

2. Noting that $I(X;Y)$ can be re-written as $I(P_X, P_{Y|X})$, where

$$I(P_X, P_{Y|X}) := \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x)P_X(x)\log_2 \frac{P_{Y|X}(y|x)}{\sum_{a\in\mathcal{X}} P_{Y|X}(y|a)P_X(a)},$$

then

- $I(X;Y)$ is a concave function of P_X (for fixed $P_{Y|X}$), i.e.,

$$I(\lambda P_X + (1-\lambda)P_{\widetilde{X}}, P_{Y|X}) \geq \lambda I(P_X, P_{Y|X}) + (1-\lambda)I(P_{\widetilde{X}}, P_{Y|X})$$

with equality holding iff

$$P_Y(y) = \sum_{x\in\mathcal{X}} P_X(x)P_{Y|X}(y|x) = \sum_{x\in\mathcal{X}} P_{\widetilde{X}}(x)P_{Y|X}(y|x) = P_{\widetilde{Y}}(y)$$

for all $y \in \mathcal{Y}$, and

2.7 Convexity/concavity of information measures

- $I(X;Y)$ is a convex function of $P_{Y|X}$ (for fixed P_X), i.e.,

$$\lambda I(P_X, P_{Y|X}) + (1-\lambda) I(P_X, P_{\tilde{Y}|X}) \geq I(P_X, \lambda P_{Y|X} + (1-\lambda) P_{\tilde{Y}|X})$$

with equality holding iff

$$(\forall\, x \in \mathcal{X})\ \frac{P_{Y|X}(y|x)}{P_{\tilde{Y}|X}(y|x)} = L(y).$$

$$\begin{aligned} & P_{Y|X}(y|x) = L(y) P_{\tilde{Y}|X}(y|x) \\ \Rightarrow\ & \sum_{x\in\mathcal{X}} P_X(x) P_{Y|X}(y|x) = L(y) \sum_{x\in\mathcal{X}} P_X(x) P_{\tilde{Y}|X}(y|x) \\ \Rightarrow\ & P_Y(y) = L(y) P_{\tilde{Y}}(y) \\ \Rightarrow\ & L(y) = \frac{P_Y(y)}{P_{\tilde{Y}}(y)} \end{aligned}$$

2.7 Convexity/concavity of information measures

3. $D(P_X \| P_{\hat{X}})$ is convex in the pair $(P_X, P_{\hat{X}})$; i.e., if $(P_X, P_{\hat{X}})$ and $(Q_X, Q_{\hat{X}})$ are two pairs of pmfs, then

$$\begin{aligned} & D(\lambda P_X + (1-\lambda)Q_X \| \lambda P_{\hat{X}} + (1-\lambda)Q_{\hat{X}}) \\ & \le \ \lambda \cdot D(P_X \| P_{\hat{X}}) + (1-\lambda) \cdot D(Q_X \| Q_{\hat{X}}), \end{aligned} \tag{2.7.1}$$

with equality holding iff

$$(\forall\ x \in \mathcal{X})\ \frac{P_X(x)}{P_{\hat{X}}(x)} = \frac{Q_X(x)}{Q_{\hat{X}}(x)}.$$

Thus, $D(P_X \| P_{\hat{X}})$ is convex with respect to both the first argument P_X and the second argument $P_{\hat{X}}$.

2.8 Fundamentals of hypothesis testing

- Simple hypothesis testing problem
  - whether a coin is fair or not
  - whether a product is successful or not
- **Problem description:** Let $X_1, \ldots, X_n$ be a sequence of observations which is drawn according to either a "null hypothesis" distribution P_{X^n} or an "alternative hypothesis" distribution $P_{\hat{X}^n}$. The hypotheses are usually denoted by:
  - $H_0 \ : \ P_{X^n}$
  - $H_1 \ : \ P_{\hat{X}^n}$.
- Decision mapping

$$\phi(x^n) = \begin{cases} 0, & \text{if distribution of } X^n \text{ is classified to be } P_{X^n}; \\ 1, & \text{if distribution of } X^n \text{ is classified to be } P_{\hat{X}^n}. \end{cases}.$$

- Acceptance regions

$$\begin{aligned} &\text{Acceptance region for } H_0 \ : \ \{x^n \in \mathcal{X}^n \ : \ \phi(x^n) = 0\} \\ &\text{Acceptance region for } H_1 \ : \ \{x^n \in \mathcal{X}^n \ : \ \phi(x^n) = 1\}. \end{aligned}$$

2.8 Fundamentals of hypothesis testing

- Error types

$$\begin{aligned}\text{Type I error} &: \alpha_n = \alpha_n(\phi) = P_{X^n}\left(\{x^n \in \mathcal{X}^n \ : \ \phi(x^n) = 1\}\right)\\ \text{Type II error} &: \beta_n = \beta_n(\phi) = P_{\hat{X}^n}\left(\{x^n \in \mathcal{X}^n \ : \ \phi(x^n) = 0\}\right).\end{aligned}$$

2.8 Fundamentals of hypothesis testing

1. Bayesian hypothesis testing.

Here, $\phi(\cdot)$ is chosen so that the Bayesian cost

$$\pi_0\alpha_n + \pi_1\beta_n$$

is minimized, where π_0 and π_1 are the prior probabilities for the null and alternative hypotheses, respectively. The mathematical expression for Bayesian testing is:

$$\min_{\{\phi\}} \left[\pi_0\alpha_n(\phi) + \pi_1\beta_n(\phi)\right].$$

2. Neyman-Pearson hypothesis testing subject to a fixed test level.

Here, $\phi(\cdot)$ is chosen so that the type II error β_n is minimized subject to a constant bound on the type I error; i.e.,

$$\alpha_n \le \varepsilon$$

where $\varepsilon > 0$ is fixed. The mathematical expression for Neyman-Pearson testing is:

$$\min_{\{\phi:\ \alpha_n(\phi)\le\varepsilon\}} \beta_n(\phi).$$

2.8 Fundamentals of hypothesis testing

Lemma 2.48 (Neyman-Pearson Lemma) For a simple hypothesis testing problem, define an acceptance region for the null hypothesis through the *likelihood ratio* as

$$\mathcal{A}_n(\tau) := \left\{ x^n \in \mathcal{X}^n : \frac{P_{X^n}(x^n)}{P_{\hat{X}^n}(x^n)} > \tau \right\},$$

and let

$$\alpha_n^* := P_{X^n} \left\{ \mathcal{A}_n^c(\tau) \right\}$$

and

$$\beta_n^* := P_{\hat{X}^n} \left\{ \mathcal{A}_n(\tau) \right\}.$$

Then for type I error α_n and type II error β_n associated with another choice of acceptance region for the null hypothesis, we have

$$\alpha_n \le \alpha_n^* \implies \beta_n \ge \beta_n^*.$$

Proof: Let $\mathcal{B}$ be a choice of acceptance region for the null hypothesis. Then

$$\begin{aligned}\alpha_n + \tau\beta_n &= \sum_{x^n \in \mathcal{B}^c} P_{X^n}(x^n) + \tau \sum_{x^n \in \mathcal{B}} P_{\hat{X}^n}(x^n) \\ &= \sum_{x^n \in \mathcal{B}^c} P_{X^n}(x^n) + \tau \left[1 - \sum_{x^n \in \mathcal{B}^c} P_{\hat{X}^n}(x^n)\right] \\ &= \tau + \sum_{x^n \in \mathcal{B}^c} \left[P_{X^n}(x^n) - \tau P_{\hat{X}^n}(x^n)\right]. \qquad (2.8.1)\end{aligned}$$

Observe that (2.8.1) is minimized by choosing $\mathcal{B} = \mathcal{A}_n(\tau)$. Hence,

$$\alpha_n + \tau\beta_n \geq \alpha_n^* + \tau\beta_n^*,$$

which immediately implies the desired result. $\square$

Lemma 2.49 (Chernoff-Stein lemma) For a sequence of i.i.d. observations X^n which is possibly drawn from either the null hypothesis distribution P_{X^n} or the alternative hypothesis distribution $P_{\hat{X}^n}$, the best type II error satisfies

$$\lim_{n\to\infty} -\frac{1}{n}\log_2 \beta_n^*(\varepsilon) = D(P_X\|P_{\hat{X}}),$$

for any $\varepsilon \in (0,1)$, where $\beta_n^*(\varepsilon) = \min_{\alpha_n\le\varepsilon}\beta_n$, and α_n and β_n are the type I and type II errors, respectively.

Proof:

Forward Part: In this part, we prove that there exists an acceptance region for the null hypothesis such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log_2 \beta_n(\varepsilon) \ge D(P_X\|P_{\hat{X}}).$$

2.8 Fundamentals of hypothesis testing

Step 1: Divergence typical set. For any $\delta > 0$, define the divergence typical set as

$$\mathcal{A}_n(\delta) := \left\{ x^n \in \mathcal{X}^n : \left| \frac{1}{n} \log_2 \frac{P_{X^n}(x^n)}{P_{\hat{X}^n}(x^n)} - D(P_X \| P_{\hat{X}}) \right| < \delta \right\}.$$

Note that any sequence x^n in this set satisfies

$$P_{\hat{X}^n}(x^n) \leq P_{X^n}(x^n) 2^{-n(D(P_X \| P_{\hat{X}}) - \delta)}.$$

Step 2: Computation of type I error. The observations being i.i.d., we have by the weak law of large numbers that

$$P_{X^n}(\mathcal{A}_n(\delta)) \to 1 \qquad \text{as } n \to \infty.$$

Hence,

$$\alpha_n = P_{X^n}(\mathcal{A}_n^c(\delta)) < \varepsilon$$

for sufficiently large n.

Step 3: Computation of type II error.

$$\begin{aligned}
\beta_n(\varepsilon) &= P_{\hat{X}^n}(\mathcal{A}_n(\delta)) \\
&= \sum_{x^n \in \mathcal{A}_n(\delta)} P_{\hat{X}^n}(x^n) \\
&\leq \sum_{x^n \in \mathcal{A}_n(\delta)} P_{X^n}(x^n) 2^{-n(D(P_X \| P_{\hat{X}}) - \delta)} \\
&= 2^{-n(D(P_X \| P_{\hat{X}}) - \delta)} \sum_{x^n \in \mathcal{A}_n(\delta)} P_{X^n}(x^n) \\
&= 2^{-n(D(P_X \| P_{\hat{X}}) - \delta)} (1 - \alpha_n).
\end{aligned}$$

Hence,

$$-\frac{1}{n}\log_2 \beta_n(\varepsilon) \geq D(P_X \| P_{\hat{X}}) - \delta + \frac{1}{n}\log_2(1 - \alpha_n),$$

which implies that

$$\liminf_{n\to\infty} -\frac{1}{n}\log_2 \beta_n(\varepsilon) \geq D(P_X \| P_{\hat{X}}) - \delta.$$

The above inequality is true for any $\delta > 0$; therefore,

$$\liminf_{n\to\infty} -\frac{1}{n}\log_2 \beta_n(\varepsilon) \geq D(P_X \| P_{\hat{X}}).$$

Converse Part: We next prove that for any acceptance region $\mathcal{B}_n$ for the null hypothesis satisfying the type I error constraint, i.e.,

$$\alpha_n(\mathcal{B}_n) = P_{X^n}(\mathcal{B}_n^c) \le \varepsilon,$$

its type II error $\beta_n(\mathcal{B}_n)$ satisfies

$$\limsup_{n\to\infty} -\frac{1}{n}\log_2 \beta_n(\mathcal{B}_n) \le D(P_X \| P_{\hat{X}}).$$

We have

$$\begin{aligned}
\beta_n(\mathcal{B}_n) = P_{\hat{X}^n}(\mathcal{B}_n) &\ge P_{\hat{X}^n}(\mathcal{B}_n \cap \mathcal{A}_n(\delta)) \\
&\ge \sum_{x^n \in \mathcal{B}_n \cap \mathcal{A}_n(\delta)} P_{\hat{X}^n}(x^n) \\
&\ge \sum_{x^n \in \mathcal{B}_n \cap \mathcal{A}_n(\delta)} P_{X^n}(x^n) 2^{-n(D(P_X\|P_{\hat{X}})+\delta)} \\
&= 2^{-n(D(P_X\|P_{\hat{X}})+\delta)} P_{X^n}(\mathcal{B}_n \cap \mathcal{A}_n(\delta)) \\
&\ge 2^{-n(D(P_X\|P_{\hat{X}})+\delta)} \left[1 - P_{X^n}(\mathcal{B}_n^c) - P_{X^n}\left(\mathcal{A}_n^c(\delta)\right)\right] \\
&= 2^{-n(D(P_X\|P_{\hat{X}})+\delta)} \left[1 - \alpha_n(\mathcal{B}_n) - P_{X^n}\left(\mathcal{A}_n^c(\delta)\right)\right] \\
&\ge 2^{-n(D(P_X\|P_{\hat{X}})+\delta)} \left[1 - \varepsilon - P_{X^n}\left(\mathcal{A}_n^c(\delta)\right)\right].
\end{aligned}$$

Hence,

$$-\frac{1}{n}\log_2 \beta_n(\mathcal{B}_n) \le D(P_X\|P_{\hat{X}}) + \delta + \frac{1}{n}\log_2\left[1 - \varepsilon - P_{X^n}\left(\mathcal{A}_n^c(\delta)\right)\right],$$

which, upon noting that $\lim_{n\to\infty} P_{X^n}\left(\mathcal{A}_n^c(\delta)\right) = 0$ (by the weak law of large numbers), implies that

$$\limsup_{n\to\infty} -\frac{1}{n}\log_2 \beta_n(\mathcal{B}_n) \le D(P_X\|P_{\hat{X}}) + \delta.$$

The above inequality is true for any $\delta > 0$; therefore,

$$\limsup_{n\to\infty} -\frac{1}{n}\log_2 \beta_n(\mathcal{B}_n) \le D(P_X\|P_{\hat{X}}).$$

$\square$

2.9 Rényi's information measures

Definition 2.50 (Rényi's entropy) Given a parameter $\alpha > 0$ with $\alpha \neq 1$, and given a discrete random variable X with alphabet $\mathcal{X}$ and distribution P_X, its Rényi entropy of order α is given by

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_{x \in \mathcal{X}} P_X(x)^\alpha \right). \tag{2.9.1}$$

- As in case of the Shannon entropy, the base of the logarithm determines the units.
- If the base is D, Rényi's entropy is in D-ary units.
- Other notations for $H_\alpha(X)$ are $H(X;\alpha)$, $H_\alpha(P_X)$ and $H(P_X;\alpha)$.

2.9 Rényi's information measures

Definition 2.51 (Rényi's divergence) Given a parameter $0 < \alpha < 1$, and two discrete random variables X and $\hat{X}$ with common alphabet $\mathcal{X}$ and distribution P_X and $P_{\hat{X}}$, respectively, then the Rényi divergence of order α between X and $\hat{X}$ is given by

$$D_\alpha(X\|\hat{X}) = \frac{1}{\alpha - 1} \log\left(\sum_{x\in\mathcal{X}} \left[P_X^\alpha(x) P_{\hat{X}}^{1-\alpha}(x)\right]\right). \tag{2.9.2}$$

- This definition can be extended to $\alpha > 1$ if $P_{\hat{X}}(x) > 0$ for all $x \in \mathcal{X}$.
- Other notations for $D_\alpha(X\|\hat{X})$ are $D(X\|\hat{X};\alpha)$, $D_\alpha(P_X\|P_{\hat{X}})$ and $D(P_X\|P_{\hat{X}};\alpha)$.

Lemma 2.52 When $\alpha \to 1$, we have the following:

$$\lim_{\alpha\to 1} H_\alpha(X) = H(X) \tag{2.9.3}$$

and

$$\lim_{\alpha\to 1} D_\alpha(X\|\hat{X}) = D(X\|\hat{X}). \tag{2.9.4}$$

2.9 Rényi's information measures

Observation 2.54 (α-mutual information)

- While Rényi did not propose a mutual information of order α that generalizes Shannon's mutual information, there are at least three different possible definitions of such measure due to Sibson (1969), Arimoto (1975) and Csiszár (1995), respectively.

Key Notes

- Conditions 1, 2 and 3 for self-information, and how these conditions correspond to mathematical expressions
- Definition of entropy, joint entropy and mutual information. Also definitions of their conditional counterparts.
- Physical interpretations of each property
  - Subtraction proofs using fundamental inequality and log-sum inequality
- Venn diagram for entropy and mutual information
- Chain rules and independent bounds (Operational meaning)
- Data processing lemma (Operational meaning)
- Why divergence is also named "relative entropy"
- Representing mutual information in terms of divergence
- Refinement and Processing
- Variational distance and divergence
- Side information and divergence
- Convexity and concavity of information measures
- Extension of information measures such as Rényi's information measures