Chapter 2

Information Measures for Discrete Systems

Po-Ning Chen, Professor

Department of Electrical and Computer Engineering

National Chiao Tung University

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

Self-information

- Self-information, denoted by $\mathcal{I}(E)$, is the information you gain by learning E has occurred.
- What properties should $\mathcal{I}(E)$ have?
  1. $\mathcal{I}(E)$ is a decreasing function of $\Pr(E)$, i.e., $\mathcal{I}(E) = I(\Pr(E))$.
     - The less likely an event is, the more information you have gained when you learn it has happened.
     - Here, $\mathcal{I}(\cdot)$ is a function defined over the event space, and $I(\cdot)$ is a function defined over $[0, 1]$.
  2. $I(\Pr(E))$ is continuous in $\Pr(E)$.
     - Intuitively, we should expect that a small change in $\Pr(E)$ will correspond to a small change in the uncertainty of E.
  3. If $E_1 \perp\!\!\!\perp E2$, then

     $$\mathcal{I}(E_1 \cap E_2) = \mathcal{I}(E_1) + \mathcal{I}(E_2) \text{ or } I(\Pr(E_1) \times \Pr(E_2)) = I(\Pr(E_1)) + I(\Pr(E_2)).$$

     - The amount of uncertainty we lose by learning that both E_1 and E_2 have occurred should be equal to the sum of the individual information losses for independent E_1 and E_2.
  4. $\mathcal{I}(E) \geq 0$. (Optional but automatically satisfied for the one-and-only function that satisfies the previous three properties.)

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 $0 \leq p \leq 1$;
3. $I(p_1 \times p_2) = I(p_1) + I(p_2)$;

is $I(p) = -C \cdot \log(p)$, where C is a positive constant.

Proof:

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

$$I\left(\frac{1}{n}\right) = -C \cdot \log\left(\frac{1}{n}\right),$$

where $C > 0$ is a constant.

Proof:

$(n = 1)$ Condition 3 $\Rightarrow$ $I(1) = I(1) + I(1)$ $\Rightarrow$ $I(1) = 0 = -C \cdot \log(1)$.

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

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

$\Rightarrow$ By condition 3

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

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

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

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

$$\log n^k \leq \log 2^r \leq \log n^{k+1} \;\Leftrightarrow\; \frac{k}{r} \leq \frac{\log(2)}{\log(n)} \leq \frac{k+1}{r}.$$

Therefore,

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

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

$$I\left(\frac{1}{n}\right) = C \cdot \log(n),$$

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

Step 2: Claim. $I(p) = -C \cdot \log(p)$ for positive rational number p, where $C > 0$ is a constant.

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

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

Step 3: For any $p \in [0, 1]$, it follows by continuity that

$$I(p) = \lim_{a \uparrow p,\ a \text{ rational}} I(a) = \lim_{b \downarrow p,\ b \text{ rational}} I(b) = -C \cdot \log(p).$$

□

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 $I(p)$.

- Information gained = uncertainty lost

Entropy

- Self-information

$$\mathcal{I}(x) \triangleq \log \frac{1}{P_X(x)},$$

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

- Entropy = expected self-information

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

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

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

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

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 $x \in \mathcal{X}$. □

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

Properties of Entropy

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

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

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 $x \in \mathcal{X}$).

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

Properties of Entropy

Proof:

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

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 $\mathcal{X}$. $\square$

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 $a > 0$. Again, this can be justified by "continuity.")

Comment: A tip for memorizing the log-sum inequality: log-first $\geq$ sum-first.

Joint entropy and conditional entropy

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

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

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

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

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

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

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

□

Corollary 2.11 (Chain rule for conditional entropy)

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

Properties of joint entropy 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 *fundamental inequality* or *log-sum inequality*.

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. □

Properties of joint entropy 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

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

□

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

Properties of joint entropy 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

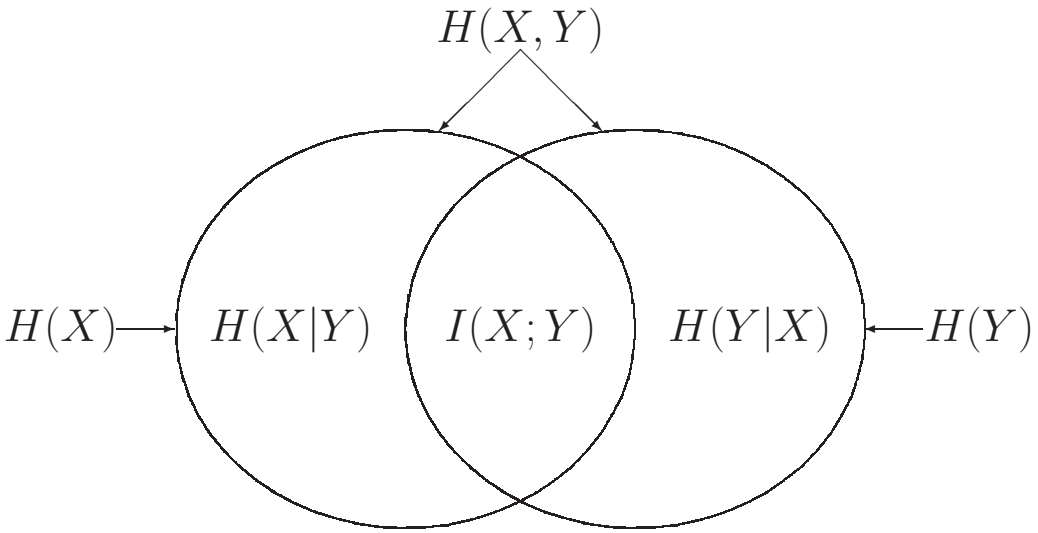
$$P_{X_1,X_2|Y_1,Y_2}(x_1, x_2|y_1, y_2) = P_{X_1|Y_1}(x_1|y_1)P_{X_2|Y_2}(x_2|y_2)$$

for all x_1, x_2, y_1 and y_2.

Mutual information

- Definition of mutual information

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



Relation between entropy and mutual information.

Properties of Mutual Information

Lemma 2.15

1. $I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{X,Y}(x,y) \log_2 \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}$.
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. $I(X;Y) \leq \min\{\log_2 |\mathcal{X}|, \log_2 |\mathcal{Y}|\}$.

Properties of Mutual Information

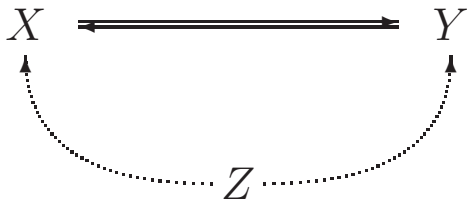
Lemma 2.16 (Chain rule for mutual information)

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

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

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

□



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

Higher dimensional extensions

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 $P_{X^n}(x^n) \triangleq P_{X_1,\cdots,X_n}(x_1,\ldots,x_n)$, where we use the common superscript notation to denote an n-tuple: $X^n \triangleq (X_1,\cdots,X_n)$ and $x^n \triangleq (x_1,\ldots,x_n)$. Then

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

where $H(X_i|X_{i-1},\ldots,X_1) \triangleq H(X_1)$ 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 $X^i \triangleq (X_1,\ldots,X_i)$.)

Theorem 2.18 (Chain rule for conditional entropy)

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

Higher dimensional extensions

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

Theorem 2.20 (Independence bound on entropy)

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

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

- This condition is equivalent to that X_i is independent of $(X_{i-1}, \ldots, X_1)$ for all i. Their equivalence can be easily proved by chain rule for probabilities, i.e., $P_{X^n}(x^n) = \prod_{i=1}^{n} P_{X_i|X_1^{i-1}}(x_i|x_1^{i-1})$, which is left to the readers as an exercise.

Higher dimensional extensions

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.

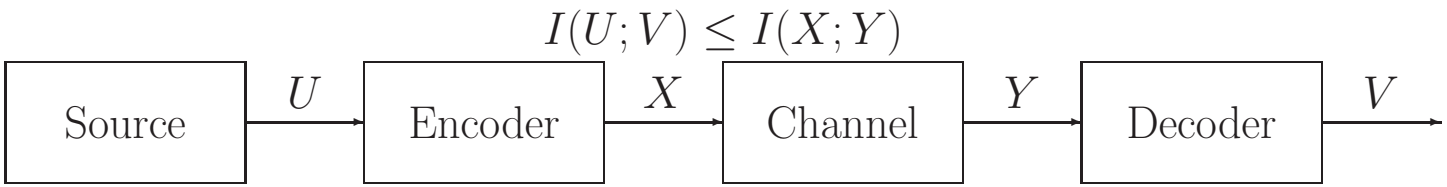
Data processing inequality

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: The Markov chain relationship $X \to Y \to Z$ means that X and Z are conditional independent given Y (cf. Appendix B); 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 $I(X;Y|Z) = 0$. □



"By processing, we can only lose the mutual information,
but the remained mutual information may be in a more *useful* form!"

Communication context of the data processing lemma.

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. Suppose X and Y are independent equiprobable binary random variables, taking values from $\{0,1\}$. Let $Z = X + Y$; hence, $Z \in \{0,1,2\}$. Then $I(X;Y) = 0$; but

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

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

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

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 $\hat{X} \triangleq g(Y)$ 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

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

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

Fano's inequality

Observation 2.27

- $P_e = 0$ implies $H(X|Y) = 0$.
- A weaker but simpler version of Fano's inequality can be directly obtained from (2.5.1) by noting that $h_b(P_e) \leq 1$:

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

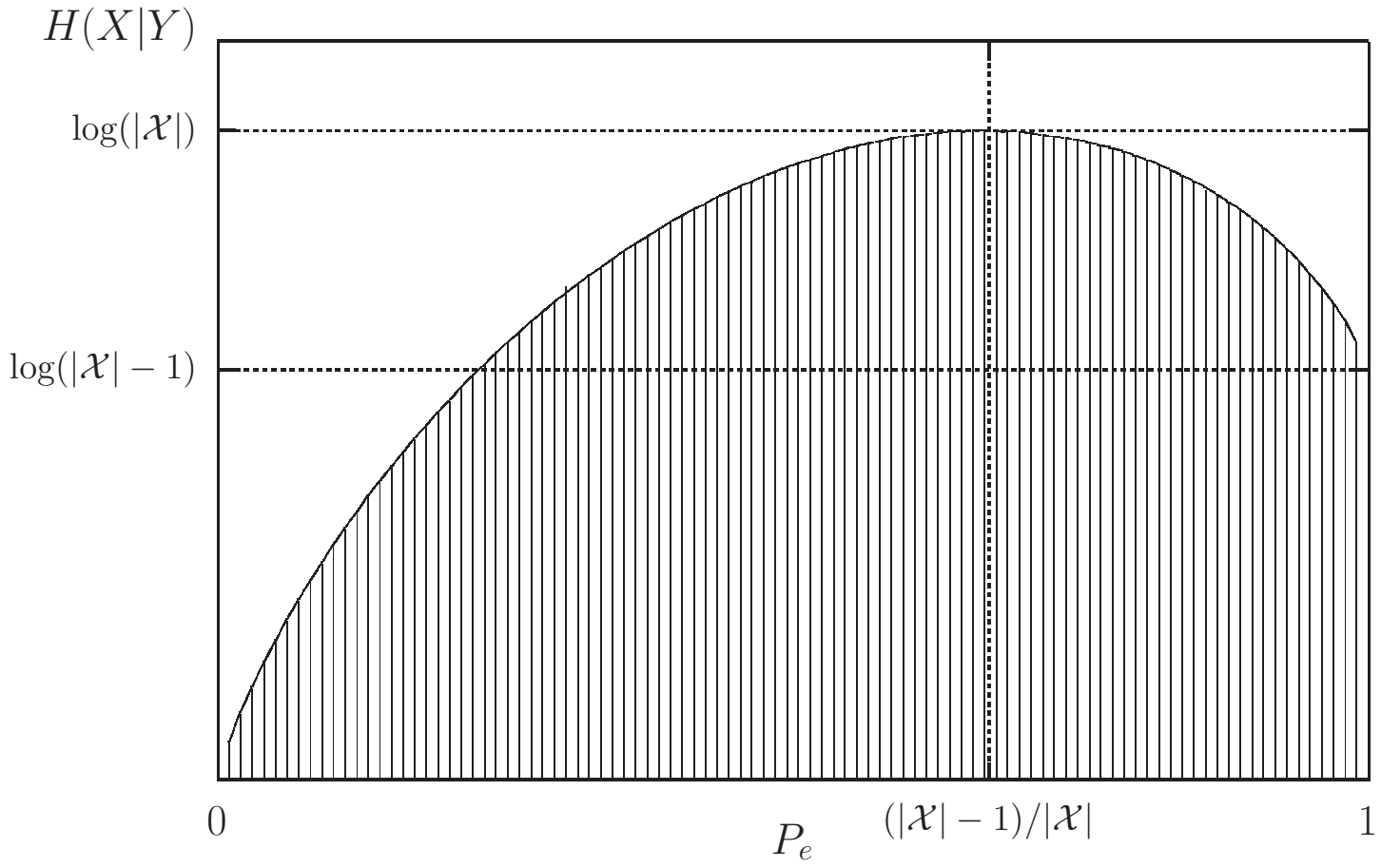
which in turn yields that

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

So, **Fano's inequality provides a lower bound to P_e.**

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.

Fano's inequality

(A quick) Proof of Lemma 2.26:

- Define a new random variable,

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

- 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. □

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; 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.

- For such case, see Example 2.28 in textbook.

Fano's inequality

Alternative and typical 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.,

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

Fano's inequality

- Setting 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

$$\begin{aligned} &H(X|\hat{X}) - h_b(P_e) - P_e \log_2(|\mathcal{X}|-1) \\ &= \underbrace{\sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\mathcal{X}:\hat{x}\neq x} P_{X,\hat{X}}(x,\hat{x}) \log_2 \frac{1}{P_{X|\hat{X}}(x|\hat{x})} + \sum_{x\in\mathcal{X}} P_{X,\hat{X}}(x,x) \log_2 \frac{1}{P_{X|\hat{X}}(x|x)}}_{H(X|\hat{X}} \\ &\quad - \underbrace{\left[\sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\mathcal{X}:\hat{x}\neq x} P_{X,\hat{X}}(x,\hat{x})\right]}_{P_e} \log_2 \frac{(|\mathcal{X}|-1)}{P_e} + \underbrace{\left[\sum_{x\in\mathcal{X}} P_{X,\hat{X}}(x,x)\right]}_{1-P_e} \log_2(1-P_e) \end{aligned} \tag{2.5.3}$$

$$
\begin{aligned}
&= \sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\mathcal{X}:\hat{x}\neq x} P_{X,\hat{X}}(x,\hat{x}) \log_2 \frac{P_e}{P_{X|\hat{X}}(x|\hat{x})(|\mathcal{X}|-1)} \\
&\qquad + \sum_{x\in\mathcal{X}} P_{X,\hat{X}}(x,x) \log_2 \frac{1-P_e}{P_{X|\hat{X}}(x|x)} \qquad (2.5.4) \\
&\leq \log_2(e) \sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\mathcal{X}:\hat{x}\neq x} P_{X,\hat{X}}(x,\hat{x}) \left[\frac{P_e}{P_{X|\hat{X}}(x|\hat{x})(|\mathcal{X}|-1)} - 1 \right] \quad \text{(FI inequality)} \\
&\qquad + \log_2(e) \sum_{x\in\mathcal{X}} P_{X,\hat{X}}(x,x) \left[\frac{1-P_e}{P_{X|\hat{X}}(x|x)} - 1 \right] \\
&= \log_2(e) \left[\frac{P_e}{(|\mathcal{X}|-1)} \sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\mathcal{X}:\hat{x}\neq x} P_{\hat{X}}(\hat{x}) - \sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\mathcal{X}:\hat{x}\neq x} P_{X,\hat{X}}(x,\hat{x}) \right] \\
&\qquad + \log_2(e) \left[(1-P_e) \sum_{x\in\mathcal{X}} P_{\hat{X}}(x) - \sum_{x\in\mathcal{X}} P_{X,\hat{X}}(x,x) \right] \\
&= \log_2(e) \left[\frac{P_e}{(|\mathcal{X}|-1)} (|\mathcal{X}|-1) - P_e \right] + \log_2(e) \left[(1-P_e) - (1-P_e) \right] \\
&= 0
\end{aligned}
$$

□

Fundamentals in 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 possibly drawn according to either null hypothesis distribution P_{X^n} or alternative hypothesis distribution $P_{\hat{X}^n}$. They are usually denoted by:
  - $H_0 \ : \ P_{X^n}$
  - $H_1 \ : \ P_{\hat{X}^n}$.
- Decision

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

- Acceptance regions

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

- Error types

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

Fundamentals in Hypothesis Testing

1. Bayesian hypothesis testing.

$\phi(\cdot)$ 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 null hypothesis and alternative hypothesis, 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 fixed test level.

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

$$\alpha_n \leq \varepsilon.$$

The mathematical expression for Neyman-Pearson testing is:

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

Lemma 2.47 (Neyman-Pearson Lemma) For a simple hypothesis testing problem, define an acceptance region for null hypothesis through *likelihood ratio* as

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

and let

$$\alpha_n^* \triangleq P_{X^n}\left\{\mathcal{A}_n^c(\tau)\right\} \quad \text{and} \quad \beta_n^* \triangleq P_{\hat{X}^n}\left\{\mathcal{A}_n(\tau)\right\}.$$

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

$$\alpha_n \leq \alpha_n^* \ \Rightarrow \ \beta_n \geq \beta_n^*.$$

Proof: Let $\mathcal{B}$ be a choice of acceptance region for 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 imply the desired result. □

Divergence and variational distance

- Based on its optimality, the log-likelihood ratio naturally becomes a good measure for hypothesis testing.

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

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

- In order to be consistent with the units (in bits) adopted for entropy and mutual information, we will also use the base-2 logarithm for divergence unless otherwise specified.

- Relative entropy rate or Kullback-Leibler divergence rate

$$\frac{1}{n}D(X^n\|\hat{X}^n) \;=\; \frac{1}{n}D(P_{X^n}\|P_{\hat{X}^n}) \triangleq E_{X^n}\left[\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\hat{X}^n}(X^n)}\right].$$

  Here, "rate" means "normalized by number of samples," i.e., n. We can similarly have *entropy rate* and *(mutual-)information rate*.

Divergence and variational distance

- E.g., i.i.d. observations.

$$\begin{aligned}\frac{1}{n}D(X^n\|\hat{X}^n) &= E_{X^n}\left[\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\hat{X}^n}(X^n)}\right] \\ &= \sum_{i=1}^{n}\frac{1}{n}E_X\left[\log\frac{P_X(X_i)}{P_{\hat{X}}(X_i)}\right] \\ &= D(X\|\hat{X}).\end{aligned}$$

- Why name it divergence?
  - A measure on the deviation between two distributions

Divergence and variational distance

- Why name it relative entropy?
  - A measure of the inefficiency of mistakenly assuming the distribution is $P_{\hat{X}}$ when the true distribution is P_X.
  - E.g.
    - If we know the true distribution P_X of a source, then (by Shannon) we can construct a lossless data compression code with average codeword length achieving entropy $H(P_X)$.
    - If, however, we mistakenly thought the "true" distribution is $P_{\hat{X}}$ and employ the "best" code corresponding to $P_{\hat{X}}$, then the resultant average codeword length becomes

$$\sum_{x \in \mathcal{X}} [-P_X(x) \cdot \log P_{\hat{X}}(x)].$$

Divergence and variational distance

– *Relative entropy* is a measure on the system cost paid (e.g., more storage consumed) due to the deed of mis-classifying system statistics.

$$\begin{aligned} & \sum_{x\in\mathcal{X}}[-P_X(x)\cdot\log P_{\hat{X}}(x)] - H(X) \\ = & \sum_{x\in\mathcal{X}}[-P_X(x)\cdot\log P_{\hat{X}}(x)] - \sum_{x\in\mathcal{X}}[-P_X(x)\cdot\log P_X(x)] \\ = & \sum_{x\in\mathcal{X}} P_X(x)\log\frac{P_X(x)}{P_{\hat{X}}(x)} = D(X\|\hat{X}). \end{aligned}$$

- 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 $x \in \mathcal{X}$,

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

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

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.

Proof: The observation follows directly from the definitions of divergence and mutual information. □

Definition 2.32 (Refinement of distribution) Given 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 $\mathcal{U} = \{1, 2, \cdots, k\}$ 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.

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 $i \in \{1, 2, \cdots, k\}$

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

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,

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

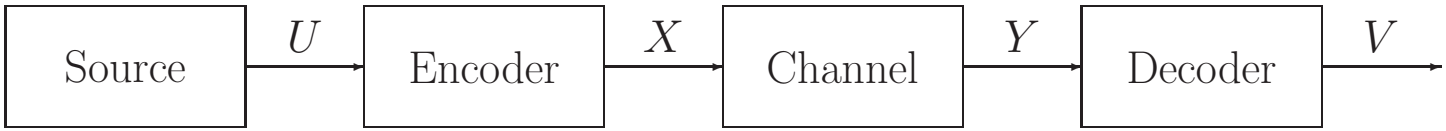
Divergence and variational distance

with equality iff

$$(\forall\, i)(\forall\, x \in \mathcal{U}_i)\ \frac{P_X(x)}{P_{\hat{X}}(x)} = \frac{P_U(i)}{P_{\hat{U}}(i)}.$$

□

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



"By processing, we can only lose the mutual information,
but the remained mutual information may be in a more *useful* form!"

Communication context of the data processing lemma.

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

Divergence and variational distance

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

Divergence and variational distance

- One drawback of adopting the divergence as a measure between two distributions is that it does not meet the symmetry requirement of a metric, i.e., it is possible

$$D(P_X \| P_{\hat{X}}) \neq D(P_{\hat{X}} \| P_X)$$

  - A *distance* or *metric*
    1. non-negativity;
    2. being zero iff two points coincide;
    3. symmetry;
    4. triangle inequality.

- Due to this reason, some other measures, such as *variational distance*, are sometimes used instead.

Divergence and variational distance

Definition 2.35 (Variational distance) The *variational distance* (or $\mathcal{L}_1$-distance) between two distributions P_X and $P_{\hat{X}}$ with common alphabet $\mathcal{X}$ is defined by

$$\|P_X - P_{\hat{X}}\| \stackrel{\triangle}{=} \sum_{x \in \mathcal{X}} |P_X(x) - P_{\hat{X}}(x)|.$$

Lemma 2.36 The variational distance satisfies

$$\begin{aligned} \|P_X - P_{\hat{X}}\| &= 2 \cdot \sup_{E \subset \mathcal{X}} \left| P_X(E) - P_{\hat{X}}(E) \right| \\ &= 2 \cdot \sum_{x \in \mathcal{X}: P_X(x) > P_{\hat{X}}(x)} \left[P_X(x) - P_{\hat{X}}(x) \right]. \end{aligned}$$

Proof: We first show that $\|P_X - P_{\hat{X}}\| = 2 \cdot \sum_{x \in \mathcal{X}: P_X(x) > P_{\hat{X}}(x)} \left[P_X(x) - P_{\hat{X}}(x) \right]$.

Setting $\mathcal{A} \stackrel{\triangle}{=} \{x \in \mathcal{X} \ : \ P_X(x) > P_{\hat{X}}(x)\}$, we have

$$\begin{aligned}
\|P_X - P_{\hat{X}}\| &= \sum_{x \in \mathcal{X}} |P_X(x) - P_{\hat{X}}(x)| \\
&= \sum_{x \in \mathcal{A}} |P_X(x) - P_{\hat{X}}(x)| + \sum_{x \in \mathcal{A}^c} |P_X(x) - P_{\hat{X}}(x)| \\
&= \sum_{x \in \mathcal{A}} \left[P_X(x) - P_{\hat{X}}(x) \right] + \sum_{x \in \mathcal{A}^c} \left[P_{\hat{X}}(x) - P_X(x) \right] \\
&= \sum_{x \in \mathcal{A}} \left[P_X(x) - P_{\hat{X}}(x) \right] + P_{\hat{X}}\left(\mathcal{A}^c\right) - P_X\left(\mathcal{A}^c\right) \\
&= \sum_{x \in \mathcal{A}} \left[P_X(x) - P_{\hat{X}}(x) \right] + P_X(\mathcal{A}) - P_{\hat{X}}(\mathcal{A}) \\
&= \sum_{x \in \mathcal{A}} \left[P_X(x) - P_{\hat{X}}(x) \right] + \sum_{x \in \mathcal{A}} \left[P_X(x) - P_{\hat{X}}(x) \right] \\
&= 2 \cdot \sum_{x \in \mathcal{A}} \left[P_X(x) - P_{\hat{X}}(x) \right]
\end{aligned}$$

where $\mathcal{A}^c$ denotes the complement set of $\mathcal{A}$.

Divergence and variational distance

We next prove that $\|P_X - P_{\hat{X}}\| = 2 \cdot \sup_{E \subset \mathcal{X}} \left| P_X(E) - P_{\hat{X}}(E) \right|$ by showing that each quantity is greater than or equal to the other. For any set $E \subset \mathcal{X}$, we can write

$$\begin{aligned}
\|P_X - P_{\hat{X}}\| &= \sum_{x \in \mathcal{X}} |P_X(x) - P_{\hat{X}}(x)| \\
&= \sum_{x \in E} |P_X(x) - P_{\hat{X}}(x)| + \sum_{x \in E^c} |P_X(x) - P_{\hat{X}}(x)| \\
&\geq \left| \sum_{x \in E} \left[P_X(x) - P_{\hat{X}}(x) \right] \right| + \left| \sum_{x \in E^c} \left[P_X(x) - P_{\hat{X}}(x) \right] \right| \\
&= |P_X(E) - P_{\hat{X}}(E)| + |P_X(E^c) - P_{\hat{X}}(E^c)| \\
&= |P_X(E) - P_{\hat{X}}(E)| + |P_{\hat{X}}(E) - P_X(E)| \\
&= 2 \cdot |P_X(E) - P_{\hat{X}}(E)|.
\end{aligned}$$

Thus $\|P_X - P_{\hat{X}}\| \geq 2 \cdot \sup_{E \subset \mathcal{X}} \left| P_X(E) - P_{\hat{X}}(E) \right|$.

Divergence and variational distance

Conversely, we have that

$$\begin{aligned}
2 \cdot \sup_{E \subset \mathcal{X}} \left| P_X(E) - P_{\hat{X}}(E) \right| &\geq 2 \cdot |P_X(\mathcal{A}) - P_{\hat{X}}(\mathcal{A})| \\
&= |P_X(\mathcal{A}) - P_{\hat{X}}(\mathcal{A})| + |P_{\hat{X}}(\mathcal{A}^c) - P_X(\mathcal{A}^c)| \\
&= \left| \sum_{x \in \mathcal{A}} \left[P_X(x) - P_{\hat{X}}(x) \right] \right| + \left| \sum_{x \in \mathcal{A}^c} \left[P_{\hat{X}}(x) - P_X(x) \right] \right| \\
&= \sum_{x \in \mathcal{A}} |P_X(x) - P_{\hat{X}}(x)| + \sum_{x \in \mathcal{A}^c} |P_X(x) - P_{\hat{X}}(x)| \\
&= \| P_X - P_{\hat{X}} \|.
\end{aligned}$$

Therefore, $\| P_X - P_{\hat{X}} \| = 2 \cdot \sup_{E \subset \mathcal{X}} \left| P_X(E) - P_{\hat{X}}(E) \right|$. $\square$

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.

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

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

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

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

In other words, it is the expected value with respect to $P_{X,Z}$ of the log-likelihood ratio $\log\frac{P_{X|Z}}{P_{\hat{X}|Z}}$.

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 $P_{X,Y,Z}$, then

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

where $P_{X,Y|Z}$ is 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.

Lemma 2.42 (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

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

Lemma 2.43 (Chain rule for 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

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

Proof:

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

with equality holding iff for all x and z,

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

□

Lemma 2.44 (Independent side information does not change divergence) If $(X, \hat{X})$ is independent of $(Z, \hat{Z})$, then

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

where

$$D(P_{X|Z}\|P_{\hat{X}|\hat{Z}}) \stackrel{\triangle}{=} \sum_{x\in\mathcal{X}}\sum_{z\in\mathcal{Z}} P_{X,Z}(x,z)\log_2\frac{P_{X|Z}(x|z)}{P_{\hat{X}|\hat{Z}}(x|z)}.$$

Lemma 2.45 (Additivity of divergence under independence)

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

provided that $(X, \hat{X})$ is independent of $(Z, \hat{Z})$.

Convexity/concavity of information measure

Lemma 2.46

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

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

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

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

   then $I(X;Y)$ is a concave function of P_X (for fixed $P_{Y|X}$), and a convex function of $P_{Y|X}$ (for fixed P_X).

3. $D(P_X\|P_{\hat{X}})$ is convex with respect to both the first argument P_X and the second argument $P_{\hat{X}}$. It is also 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 probability mass functions, then

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

   for all $\lambda \in [0,1]$.

Key Notes

- Conditions 1, 2 and 3 for self-information, and how these conditions correspond to mathematical expressions (proof is not that important)
- 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