Information Theory “Phase Zero”

Changyeol Lee (Yonsei University)

Cover and Thomas, Elements of Information Theory (2nd edition), Chapter 2~4

Cover and Thomas, Elements of Information Theory (2nd edition), Chapter 2

Entropy and Information

Entropy / Conditional Entropy

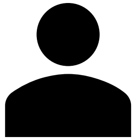
Relative Entropy / Conditional Relative Entropy

Mutual Information / Conditional Mutual Information

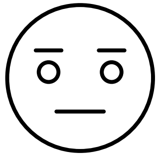
Chain Rules

Surprise

$$X \sim \begin{cases} a & 6/9 \\ b & 2/9 \\ c & 1/9 \end{cases}$$

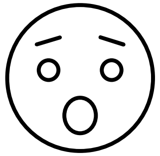


$X = a \text{ or } b \text{ or } c$



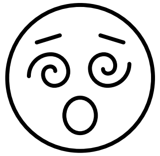
No *surprise*

$X = a$



Little *surprise*

$X = c$



More *surprise*

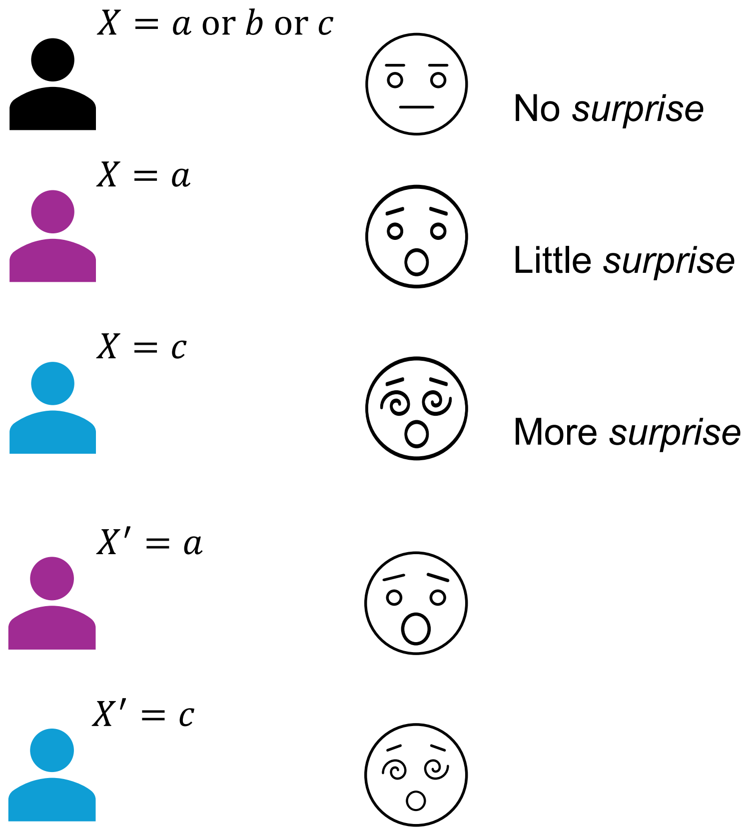
Surprise

Natural properties of *surprise*

- Event w/ prob. 1 = No surprise
- Rarer event = More surprise
- No jump in surprise

$$X \sim \begin{cases} a & 6/9 \\ b & 2/9 \\ c & 1/9 \end{cases}$$

$$X' \sim \begin{cases} a & 6/9 + \epsilon \\ b & 2/9 \\ c & 1/9 - \epsilon \end{cases}$$



Surprise

We say $S: (0,1] \to \mathbb{R}_{\geq 0}$ is a *surprise function* if it satisfies

- $S(1) = 0$
- S is (strictly) decreasing, i.e., $p < q \Rightarrow S(p) > S(q)$
- S is continuous
- $\boldsymbol{S(pq) = S(p) + S(q)}$, i.e., for two independent instantiations, S is additive

Which function can be a surprise function? $\boldsymbol{S(p) = -\log_2 p}$

w/ normalization $S(1/2) = 1$
i.e., we assume a fair coin flip gives a unit surprise

Any other possible function?

Surprise

Claim. $-\log_2 p$ is the only possible normalized surprise function.

proof)

- $S(p^n) = n \cdot S(p)$ for any $n \in \mathbb{N}$
- $S(p) = n \cdot S\left(p^{1/n}\right)$ by substituting p^n to p
- $S\left(p^{1/n}\right) = \frac{1}{n} \cdot S(p)$ by rearranging the terms
- $S\left(p^{m/n}\right) = m \cdot S\left(p^{1/n}\right) = \frac{m}{n} \cdot S(p)$ for any $n, m \in \mathbb{N}$
- $S(p^\alpha) = \alpha \cdot S(p)$ for any $\alpha \in \mathbb{Q}_{\geq 0}$.
- $S(p^\alpha) = \alpha \cdot S(p)$ for any $\alpha \in \mathbb{R}_{\geq 0}$ since S is continuous
- With normalization $S(1/2) = 1$, we have $S(2^{-\alpha}) = \alpha$.
- Every $p \in (0,1]$ can be represented as $2^{-\alpha}$ for some $\alpha \in \mathbb{R}_{\geq 0}$

Entropy

X: a discrete random variable over $\mathcal{X}$ with the probability mass function $p(\cdot)$.

The *entropy* of X is the expected surprise for X.

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) = \mathbb{E}_{X\sim p}[-\log_2 p(X)]$$

$$H(X) \text{ of } X \sim \begin{cases} a & 1/2 \\ b & 1/4 \\ c & 1/4 \end{cases} ?$$

- a measure of the uncertainty of X
- a measure of the (expected) amount of information required to describe X

* Sometimes we use $H(p)$ instead.

* $0\log 0 = 0$

* If the base is e, we say “the entropy is measured in *nats*”.

* If not specified, the base is always 2.

Fact. $H(X) \geq 0$ (since surprise ≥ 0)

Joint Entropy, Conditional Entropy

Joint Entropy

$$H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x,y) = \mathbb{E}_{(X,Y)\sim p}[-\log p(X,Y)]$$

Conditional Entropy

$$H(Y|X) = \sum_{x\in\mathcal{X}} p(x)H(Y|X=x)$$

$$= -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x)$$

$$= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(y|x) = -\mathbb{E}_{(X,Y)\sim p}[-\log p(Y|X)]$$

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

Chain Rule

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

proof 1)

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x)p(y|x) \\
&= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x) - \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}
$$

Chain Rule

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

proof 2)

Recall the entropy is the expected surprise.

$$\log p(x,y) = \log p(x) + \log p(y|x)$$

Chain Rule

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

Corollary. $H(X) - H(X|Y) = H(Y) - H(Y|X)$

Relative Entropy or Kullback-Leibler Divergence

Relative entropy or Kullback-Leibler divergence(distance) between p and q

$$D\big(p(x) \parallel q(x)\big) = D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_{X \sim p}\left[\log \frac{p(X)}{q(X)}\right]$$

- a measure of the inefficiency of assuming that the distribution of $X \sim p$ is q

* $0 \log \frac{0}{0} = 0,\ \ 0 \log \frac{0}{q} = 0,\ \ p \log \frac{p}{0} = \infty$ ($D(p \parallel q) = \infty$ if $\exists\, x \in \mathcal{X}$ s.t. $p(x) > 0$ and $q(x) = 0$.)

* $D(p \parallel q) \neq D(q \parallel p)$, i.e., no symmetricity (in general)

* $D(p \parallel q) + D(q \parallel r) \ngeq D(p \parallel r)$, i.e., no triangle inequality (in general)

* $D(p \parallel q) \geq 0$. Holds in equality if and only if $p = q$. (proof later)

Conditional Relative Entropy

Conditional relative entropy

$$D\big(p(y|x) \parallel q(y|x)\big) = \sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log \frac{p(y|x)}{q(y|x)} = \mathbb{E}_{(X,Y)\sim p}\left[\log \frac{p(Y|X)}{q(Y|X)}\right]$$

* (chain rule) $D\big(p(x,y) \parallel q(x,y)\big) = D\big(p(x) \parallel q(x)\big) + D\big(p(y|x) \parallel q(y|x)\big)$

Mutual Information

Mutual information

- a measure of the amount of information that one RV contains about another RV

$$
\begin{aligned}
I(X;Y) &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\
&= D\big(p(x,y) \parallel p(x)p(y)\big) \\
&= \mathbb{E}_{(X,Y)\sim p}\left[\log\frac{p(X,Y)}{p(X)p(Y)}\right]
\end{aligned}
$$

Entropy and Mutual Information

Mutual information

- a measure of the amount of information that one RV contains about another RV

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

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

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

the reduction in the uncertainty of X due to the knowledge of Y

Entropy and Mutual Information

Mutual information

- a measure of the amount of information that one RV contains about another RV

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

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

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

the reduction in the uncertainty of Y due to the knowledge of X

Entropy and Mutual Information

Mutual information

- a measure of the amount of information that one RV contains about another RV

$$
\begin{aligned}
I(X;Y) &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X) + H(Y) - H(X,Y) \qquad \text{(by chain rule)} \\
&= I(Y;X)
\end{aligned}
$$

* $I(X;X) = H(X)$ (Entropy is sometimes referred to as *self-information*)

Entropy and Mutual Information

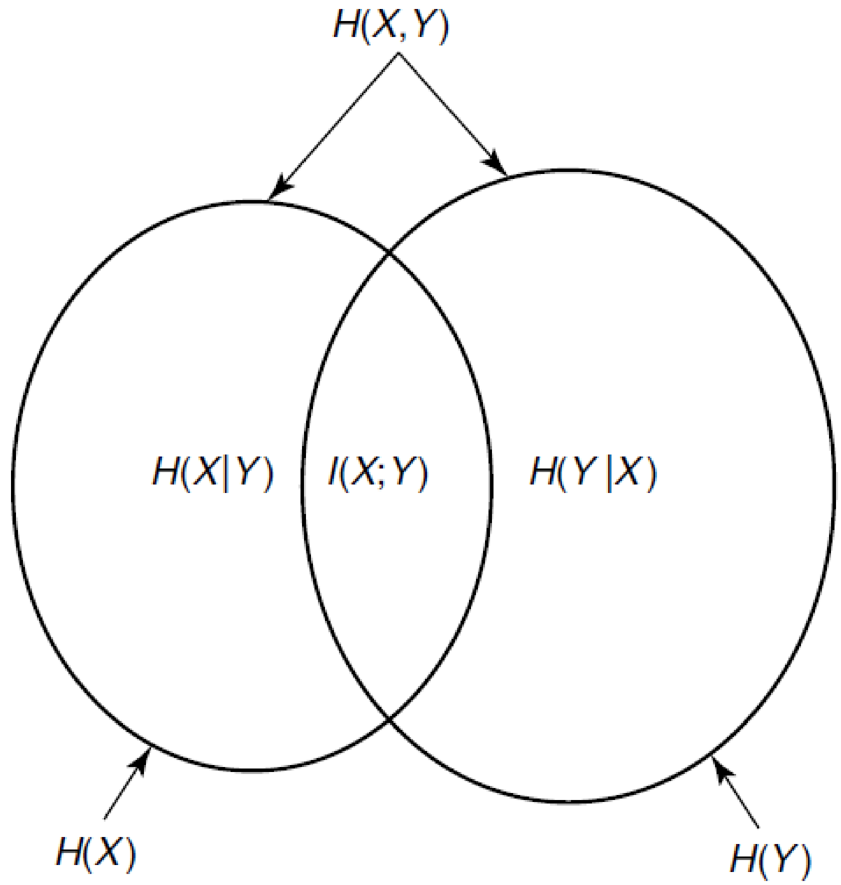


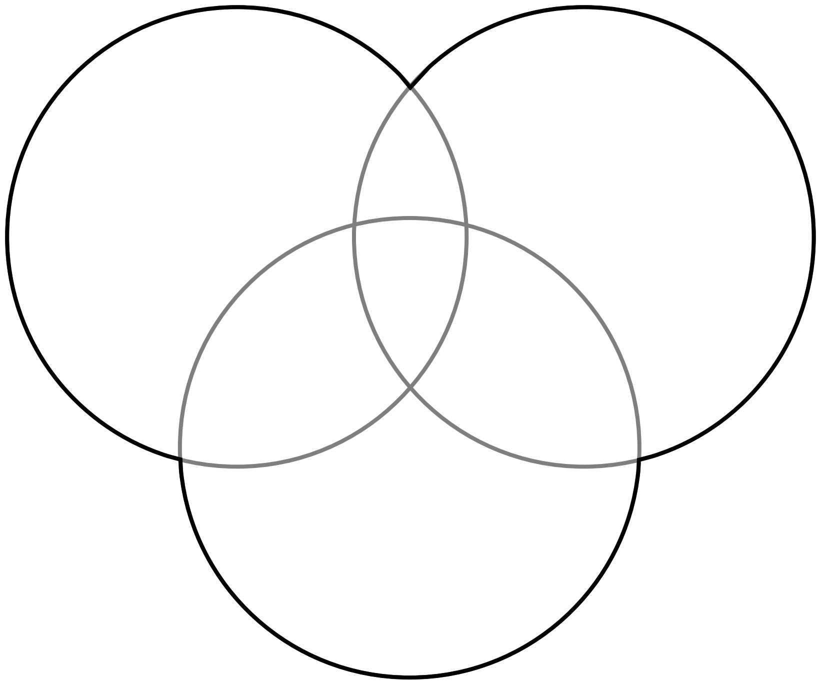
FIGURE 2.2. Relationship between entropy and mutual information.

Chain Rule (collection of random variables)

Theorem.

$$H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \cdots + H(X_n|X_{n-1}, \dots, X_2, X_1)$$

$H(X_1, X_2, X_3)$



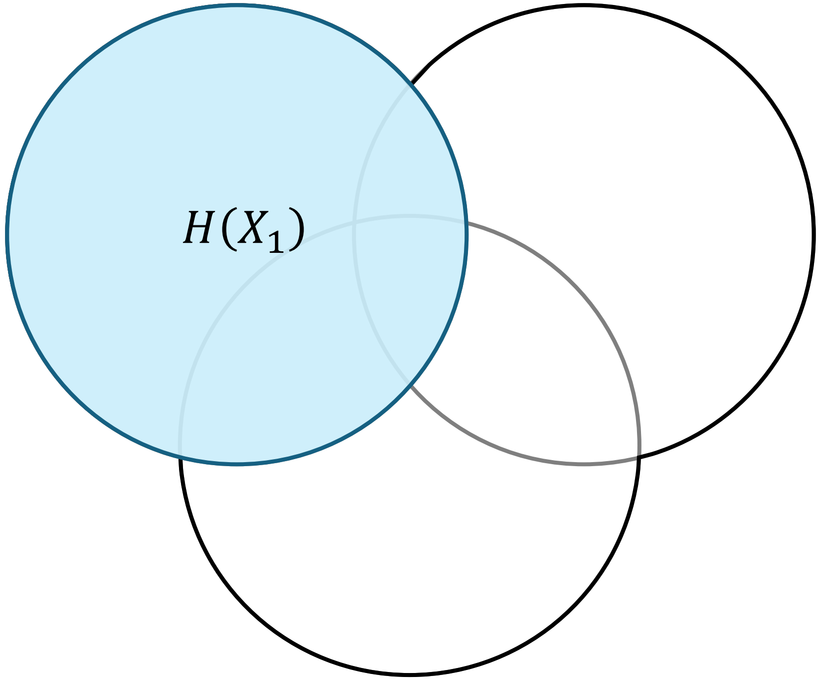
* Be careful! Venn diagram might mislead you!

Chain Rule (collection of random variables)

Theorem.

$$H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \cdots + H(X_n|X_{n-1}, \dots, X_2, X_1)$$

$H(X_1, X_2, X_3)$



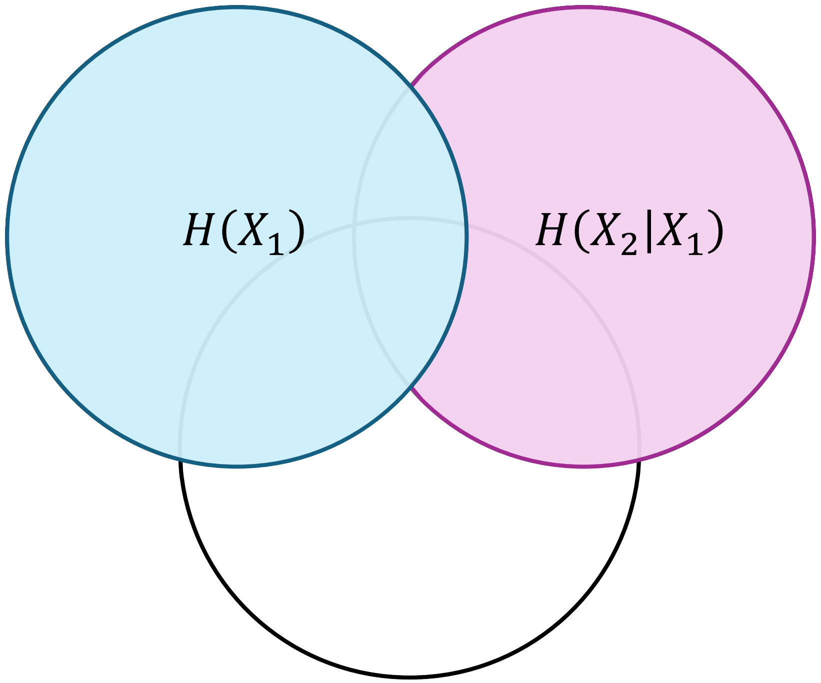
* Be careful! Venn diagram might mislead you!

Chain Rule (collection of random variables)

Theorem.

$$H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \cdots + H(X_n|X_{n-1}, \dots, X_2, X_1)$$

$H(X_1, X_2, X_3)$

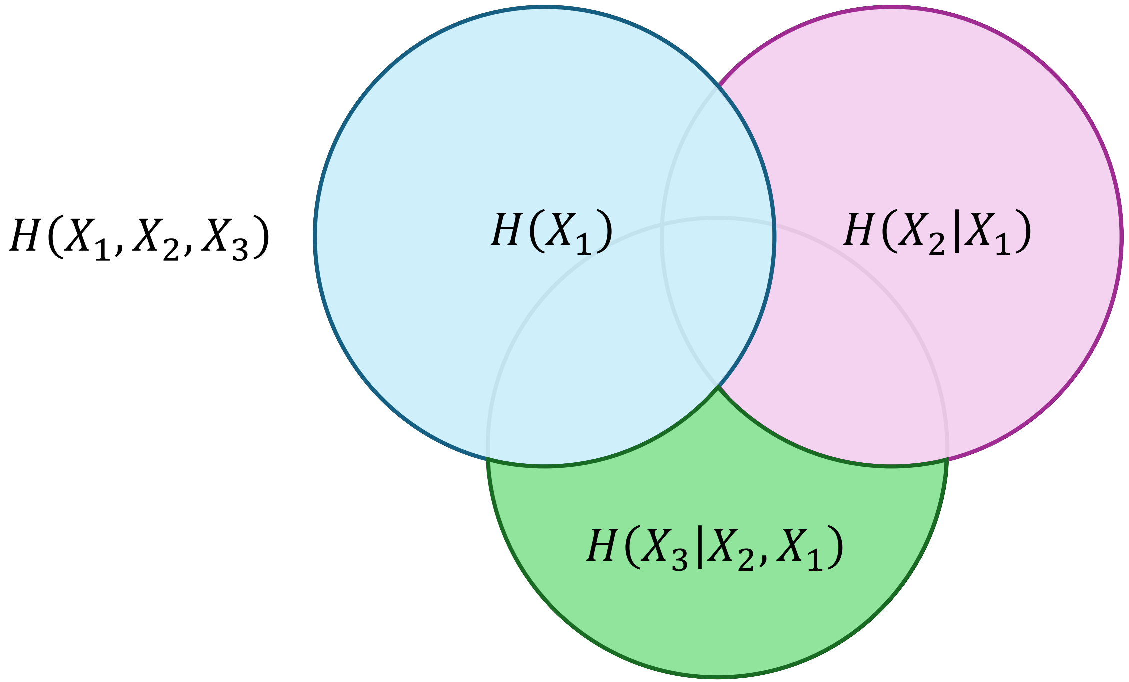


* Be careful! Venn diagram might mislead you!

Chain Rule (collection of random variables)

Theorem.

$$H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \cdots + H(X_n|X_{n-1}, \dots, X_2, X_1)$$



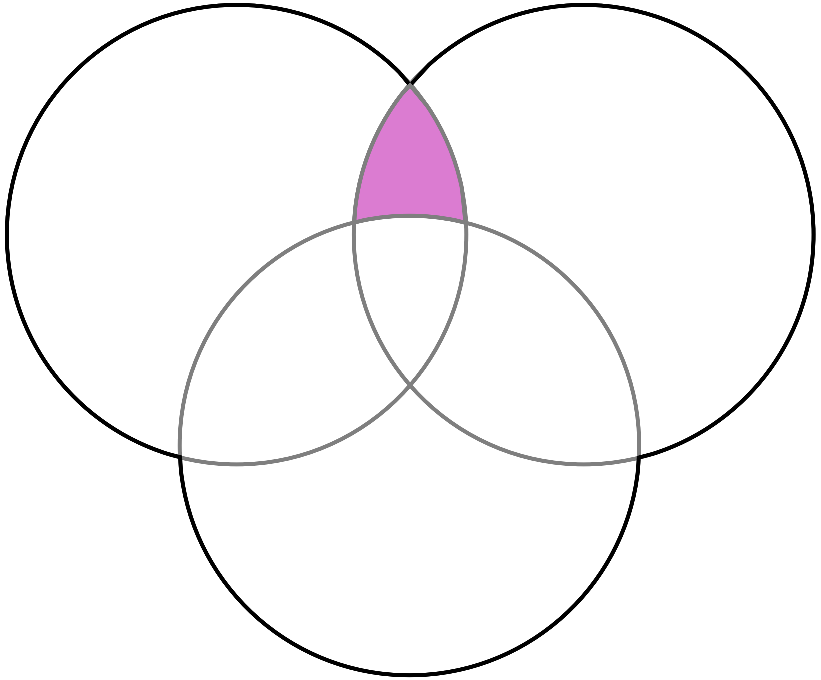
* Be careful! Venn diagram might mislead you!

Conditional Mutual Information

Conditional Mutual Information

$$I(X_1; X_2 | X_3) = H(X_1 | X_3) - H(X_1 | X_2, X_3)$$

$H(X_1, X_2, X_3)$



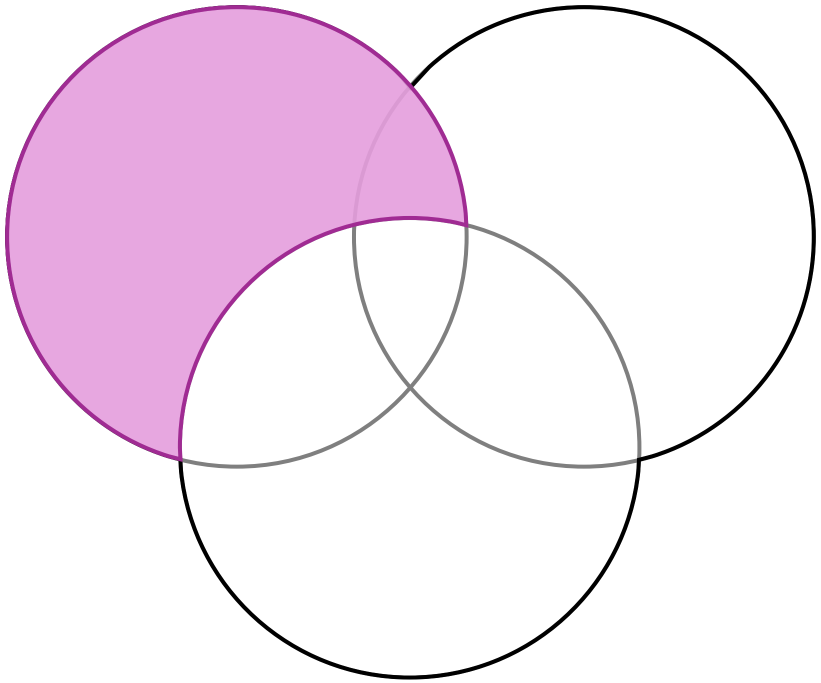
* Be careful! Venn diagram might mislead you!

Conditional Mutual Information

Conditional Mutual Information

$$I(X_1;X_2|X_3) = H(X_1|X_3) - H(X_1|X_2,X_3)$$

$H(X_1,X_2,X_3)$



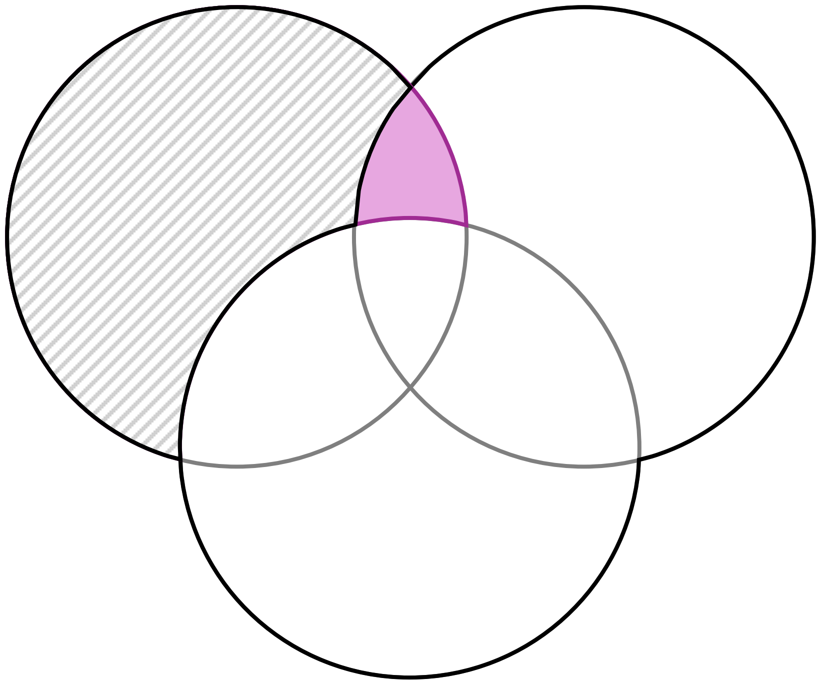
* Be careful! Venn diagram might mislead you!

Conditional Mutual Information

Conditional Mutual Information

$$I(X_1;X_2|X_3) = H(X_1|X_3) - H(X_1|X_2,X_3)$$

$H(X_1,X_2,X_3)$



* Be careful! Venn diagram might mislead you!

Conditional Mutual Information

Conditional Mutual Information

$$I(X_1; X_2 | X_3) = H(X_1 | X_3) - H(X_1 | X_2, X_3)$$

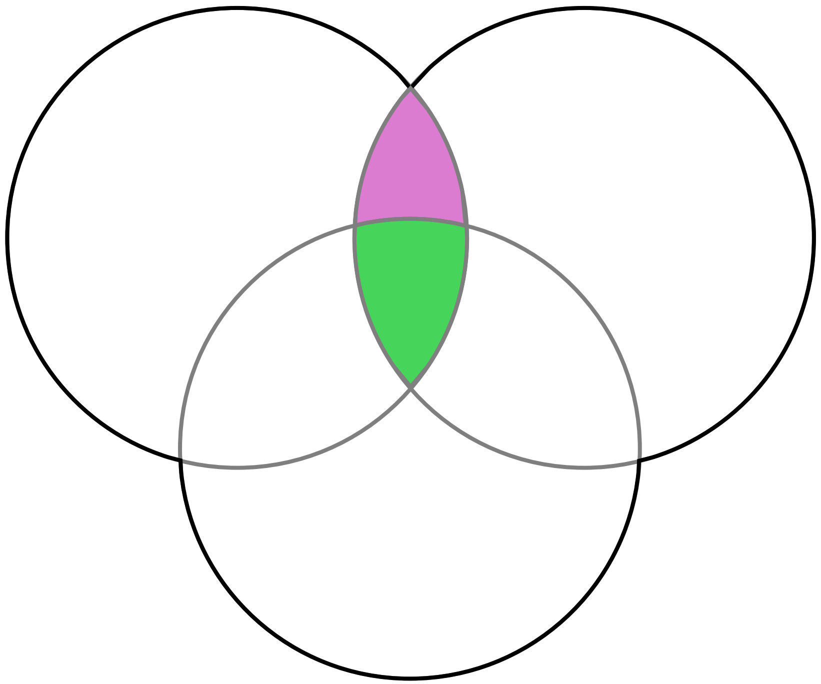
*(chain rule) $I(X_1, X_2, \dots, X_n; Y) = I(X_1; Y) + I(X_2; Y | X_1) + I(X_3; Y | X_2, X_1) + \cdots + I(X_n; Y | X_{n-1}, \dots, X_2, X_1)$

* Be careful! Venn diagram might mislead you!

MISLEADING Representation of Entropies

Claim. $I(X;Y|Z) \leq I(X;Y)$ holds by Venn diagram.

This claim is not always true! Then… is the claim always false?



Consider two independent fair coins X, Y. Let $Z = X + Y$.

We have

$$I(X;Y) = 0$$

and,

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(X|Z) - 0.$$

When $Z \neq 1$, X is determined to one value, i.e., no surprise. Therefore

$$H(X|Z) = \Pr[Z = 1]\, H(X|Z = 1) = 1/2$$

Some Inequalities

Information Inequalities

Data-processing Inequalities

Fano’s Inequalities

Information Inequality

Theorem. $D(p \parallel q) \geq 0$ with equality if and only if $p = q$.

$$-D(p \parallel q) = \mathbb{E}_{X\sim p}\left[\log \frac{q(X)}{p(X)}\right]$$

(by Jensen's inequality)

$$\leq \log \mathbb{E}_{X\sim p}\left[\frac{q(X)}{p(X)}\right]$$

Since $\log$ is strictly concave,
$=$ implies $q(x)/p(x) = c$ for all $x \in \mathrm{supp}(p)$
for some constant c.

$$= \log \sum_{x\in \mathrm{supp}(p)} q(x)$$

$$\leq \log \sum_{x\in \mathrm{supp}(q)} q(x)$$

$=$ implies $\mathrm{supp}(q) = \mathrm{supp}(p)$, which implies $c = 1$.

$$= \log 1 = 0$$

Trivial that if $p = q$, then $D(p \parallel q) = 0$.

We show if $D(p \parallel q) = 0$, then $p = q$.

Information Inequality

Theorem. $D(p \parallel q) \geq 0$ with equality if and only if $p = q$.

Corollary. $D\big(p(y|x) \parallel q(y|x)\big) \geq 0$ with equality if and only if $p(y|x) = q(y|x)$ for all x, y s.t. $p(x) > 0$.

Corollary. $I(X;Y) \geq 0$ with equality if and only if X and Y are independent.

Corollary. $I(X;Y|Z) \geq 0$ with equality if and only if X and Y are conditionally independent given Z.

Corollary. $H(X|Y) \leq H(X)$, i.e., *conditioning only reduces entropy*.

proof) $I(X;Y) = H(X) - H(X|Y) \geq 0$.

Theorem. $H(X) \leq \log|\mathcal{X}|$ with equality if and only if p is the uniform distribution.

proof) Let $u(x) = 1/|\mathcal{X}|$ be the uniform distribution.

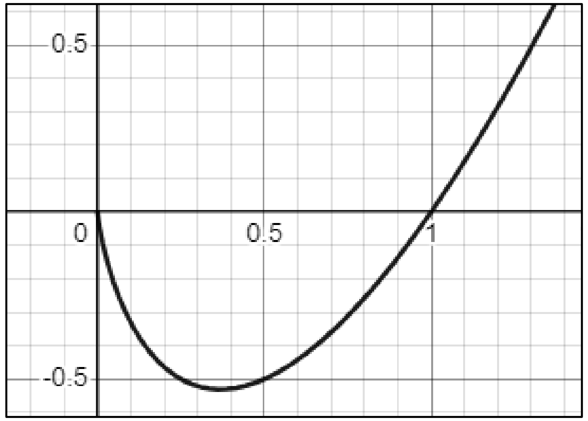
$$D(p \parallel u) = \mathbb{E}_{X \sim p}\left[\log \frac{p(X)}{u(X)}\right] = \log|\mathcal{X}| - H(X) \geq 0$$

Convexity of Relative Entropy

distance btw averaged distribution ≤ average of distance btw distributions

Theorem. $D(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2) \le \lambda D(p_1 \parallel q_1) + (1-\lambda)D(p_2 \parallel q_2)$ for all $\lambda \in [0,1]$.

proof) Fix any $x \in \mathcal{X}$.



Let $P_1 := \lambda p_1(x),\ P_2 := (1-\lambda)p_2(x),\ Q_1 := \lambda q_1(x),\ Q_2 := (1-\lambda)q_2(x)$.

Let $f(x) = x\log x$. Observe that f is (strictly) convex. ($f''(x) = \frac{1}{x \ln 2} > 0$.)

$$(P_1+P_2)\log\frac{P_1+P_2}{Q_1+Q_2} = (Q_1+Q_2)\cdot\frac{P_1+P_2}{Q_1+Q_2}\log\frac{P_1+P_2}{Q_1+Q_2}$$
$$= (Q_1+Q_2) f\left(\frac{P_1+P_2}{Q_1+Q_2}\right)$$

$$\frac{P_1+P_2}{Q_1+Q_2} = \frac{Q_1}{Q_1+Q_2}\cdot\frac{P_1}{Q_1} + \frac{Q_2}{Q_1+Q_2}\cdot\frac{P_2}{Q_2}$$

By Jensen's inequality,

$$(Q_1+Q_2) f\left(\frac{P_1+P_2}{Q_1+Q_2}\right) \le Q_1\cdot f\left(\frac{P_1}{Q_1}\right) + Q_2 \cdot f\left(\frac{P_2}{Q_2}\right) = P_1\log\frac{P_1}{Q_1} + P_2\log\frac{P_2}{Q_2}.$$

Concavity of Entropy

entropy of averaged distribution ≥ average of entropy of distributions

Theorem. $H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2)$ for all $\lambda \in [0,1]$.

proof)

Recall that

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

where u is the uniform distribution.

The theorem follows from the convexity of D.

Convexity/Concavity of Mutual Information

Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. Write $\alpha(x) = p(x)$ and $\beta(x, y) = p(y|x)$. Then (α, β) specifies p.

Theorem. (Mutual information concave in α) $\lambda \cdot I(X_1; Y_1) + (1 - \lambda) \cdot I(X_2; Y_2) \leq I(X_3; Y_3)$ where $(X_1, Y_1) \sim (\alpha_1, \beta), (X_2, Y_2) \sim (\alpha_2, \beta)$ and $(X_3, Y_3) \sim (\lambda\alpha_1 + (1 - \lambda)\alpha_2, \beta)$.

proof)

Let B_λ be the biased coin which takes 1 w/ prob. λ and 0 w/ prob. $1 - \lambda$.

Let X be the RV whose distribution is α_1 if $B_\lambda = 1$, o/w, α_2.

Let Y be the RV conditioned on X with distribution β.

$$
\begin{aligned}
I(X_3; Y_3) &= I(B_\lambda, X; Y) \\
&= I(B_\lambda; Y) + I(X; Y|B_\lambda) && \text{(by chain rule)} \\
&\geq I(X; Y|B_\lambda) && \text{(by information inequality)} \\
&= \lambda \cdot I(X; Y|B_\lambda = 1) + (1 - \lambda) \cdot I(X; Y|B_\lambda = 0) \\
&= \lambda \cdot I(X_1; Y_1) + (1 - \lambda) \cdot I(X_2; Y_2)
\end{aligned}
$$

Convexity/Concavity of Mutual Information

Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. Write $\alpha(x) = p(x)$ and $\beta(x, y) = p(y|x)$. Then (α, β) specifies p.

Theorem. (Mutual information convex in β) $\lambda \cdot I(X_1; Y_1) + (1-\lambda) \cdot I(X_2; Y_2) \geq I(X_3; Y_3)$ where $(X_1, Y_1) \sim (\alpha, \beta_1), (X_2, Y_2) \sim (\alpha, \beta_2)$ and $(X_3, Y_3) \sim (\alpha, \lambda\beta_1 + (1-\lambda)\beta_2)$.

proof)

Let B_λ be the biased coin which takes 1 w/ prob. λ and 0 w/ prob. $1-\lambda$.

Let X be the RV whose distribution is α. (Independent from B_λ.)

Let Y be the RV conditioned on X with distribution β_1 if $B_\lambda = 1$, o/w, β_2.

$$I(B_\lambda, Y; X) = I(Y; X) + I(B_\lambda; X|Y) \qquad \text{(by chain rule)}$$

$$\geq I(Y; X) = I(X_3; Y_3) \qquad \text{(by information inequality)}$$

$$I(B_\lambda, Y; X) = I(B_\lambda; X) + I(Y; X|B_\lambda) = 0 + I(Y; X|B_\lambda)$$

$$= \lambda \cdot I(Y; X|B_\lambda = 1) + (1-\lambda) \cdot I(Y; X|B_\lambda = 0)$$

$$= \lambda \cdot I(X_1; Y_1) + (1-\lambda) \cdot I(X_2; Y_2)$$

Data-processing Inequality

We say random variables X, Y, Z *form a Markov chain* $X \to Y \to Z$ if $p(x,y,z) = p(x)p(y|x)p(z|y)$.

* $X \to Y \to Z$ if and only if X and Z are conditionally independent given Y.

* $X \to Y \to Z$ implies $Z \to Y \to X$.

Theorem. If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$.

proof)

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

Since X and Z are conditionally independent given Y, $I(X;Z|Y) = 0$.

Corollary. If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Y|Z)$.

* Holds with equality if and only if $I(X;Z) = 0$, i.e., X and Z are independent.

Corollary. If $X \to Y \to Z$, then $H(X|Y) \leq H(X|Z)$.

Fano's Inequality

Given Y, we wish to guess the value of X.

- If we can estimate X with 0 probability of error, then $H(X|Y) = 0$, i.e., no uncertainty.

- If we can estimate X with "low" probability of error, then $H(X|Y)$ is "small".

Let $\hat{X} = g(Y)$ be the estimate of X and takes on values in $\hat{\mathcal{X}}$.

- No assumption $\hat{\mathcal{X}} = \mathcal{X}$
- g can be random

Theorem. For any estimator $\hat{X}$ s.t. $X \to Y \to \hat{X}$, we have

$$H(P_e) + P_e \log|\mathcal{X}| \geq H\left(X \middle| \hat{X}\right) \geq H(X|Y)$$

where $P_e = \Pr\left[\hat{X} \neq X\right]$ is the probability of error.

Weaker statement.

Why $H(P_e) \leq 1$?

$$1 + P_e \log|\mathcal{X}| \geq H(X|Y) \iff P_e \geq \frac{H(X|Y) - 1}{\log|\mathcal{X}|}.$$

Fano's Inequality

Theorem. For any estimator $\hat{X}$ s.t. $X \to Y \to \hat{X}$, we have

$$H(P_e) + P_e \log|\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y)$$

where $P_e = \Pr[\hat{X} \neq X]$ is the probability of error.

data-processing inequality

If $X \to Y \to Z$, then $H(X|Y) \leq H(X|Z)$

proof of first inequality)

Let $E = \mathbb{1}[\hat{X} \neq X]$ be the binary RV.

$$H(E, X|\hat{X}) = H(X|\hat{X}) + H(E|X, \hat{X}) = H(X|\hat{X})$$

$$= H(E|\hat{X}) + H(X|E, \hat{X}) \leq H(P_e) + P_e \log|\mathcal{X}|$$

- $H(E|X, \hat{X}) = 0$
- $H(E|\hat{X}) \leq H(E) = H(P_e)$ unconditioning increases entropy
- $H(X|E, \hat{X}) = \Pr[E = 1]\, H(X|E = 1, \hat{X}) \leq P_e \cdot H(X) \leq P_e \log|\mathcal{X}|$.

  uniform distribution maximizes entropy

* The first inequality holds without the condition $X \to Y \to \hat{X}$.

Fano's Inequality

Theorem. For any estimator $\hat{X}$ s.t. $X \to Y \to \hat{X}$ and $\mathcal{X} = \hat{\mathcal{X}}$, we have

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y)$$

where $P_e = \Pr[\hat{X} \neq X]$ is the probability of error.

Weaker statement.

$$P_e \geq \frac{H(X|Y) - 1}{\log(|\mathcal{X}| - 1)}$$

Fano’s Inequality

Remark. Fano’s inequality s sharp.

Suppose no knowledge of Y, i.e., guess X without any information.

Let our (deterministic) estimator be x^* where $p(x^*) = \max_{x \in \mathcal{X}} p(x)$.

Fano’s inequality says

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X).$$

If $p(\cdot)$ restricted to $\mathcal{X} \backslash \{x^*\}$ were a uniform distribution, i.e., $p(x) = \frac{1-p(x^*)}{|\mathcal{X}|-1}$ for all $x \neq x^*$,

this holds with equality.

More Inequalities Related to Probability of Error and Entropy

Lemma. If X and X' are independent identically distributed,

$$\Pr[X = X'] \geq 2^{-H(X)}$$

with equality if and only if X has a uniform distribution.

proof) Note that 2^x is (strictly) convex.

By Jensen's inequality,

$$2^{-H(X)} = 2^{\mathbb{E}[\log p(X)]} \leq \mathbb{E}\left[2^{\log p(X)}\right] = \mathbb{E}[p(X)] = \sum_{x \in \mathcal{X}} p^2(x) = \Pr[X = X'].$$

Corollary. If $X \sim p$ and $X' \sim q$ are independent and $\mathcal{X} = \mathcal{X}'$,

$$\Pr[X = X'] \geq 2^{-H(p) - D(p\|q)}$$

$$\Pr[X = X'] \geq 2^{-H(q) - D(q\|p)}$$

AEP

Asymptotic Equipartition Property

Typical Set

Simple Data Compression

Weak Law of Large Numbers

Let $Z_1, Z_2, \dots, Z_n$ be a sequence of i.i.d RVs with mean μ and variance σ^2.

Let $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$ be the sample mean.

Weak law of large numbers.

$$\Pr[|\bar{Z}_n - \mu| > \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}$$

or

$$\Pr[|\bar{Z}_n - \mu| > \epsilon] \to 0 \quad \text{as} \quad n \to \infty$$

proof)

Note $\mathbb{E}[\bar{Z}_n] = \mu$ and $\mathrm{Var}(\bar{Z}_n) = \sigma^2/n$. (Each Z_i has variance σ^2/n^2.)

Apply Chebyshev's inequality.

AEP (Asymptotic Equipartition Property)

Consider a sequence of i.i.d RVs $X_1, X_2, \dots, X_n$.

AEP

$$-\frac{1}{n}\log p(X_1, X_2, \dots, X_n) \to H(X) \text{ in probability}$$

AEP (Asymptotic Equipartition Property)

Consider a sequence of i.i.d RVs $X_1, X_2, \dots, X_n$.

Consider a sequence of RVs $Z_1, Z_2, \dots, Z_n$ (also i.i.d.) such that $Z_i \coloneqq -\log p(X_i)$ for all $i = 1, \dots, n$.

Let $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i = -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i)$. Note that $\mathbb{E}[\bar{Z}_n] = H(X)$.

AEP

$$\bar{Z}_n \to H(X) \text{ in probability}$$

AEP (Asymptotic Equipartition Property)

AEP (more formally). For any $\epsilon > 0$, there exists n_0 such that for all $n \geq n_0$,

$$\Pr[|\bar{Z}_n - H(X)| > \epsilon] \leq \epsilon$$

or equivalently,

$$\Pr[|\bar{Z}_n - H(X)| > \epsilon] \to 0 \quad \text{as} \quad n \to \infty$$

proof)

Direct application of weak law of large numbers gives the following:

$$\Pr[|\bar{Z}_n - H(X)| > \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}$$

where σ^2 is the variance of Z_i.

Let $n_0 = \frac{\sigma^2}{\epsilon^3}$. Then for all $n \geq n_0$,

$$\frac{\sigma^2}{n\epsilon^2} \leq \frac{\sigma^2}{n_0\epsilon^2} \leq \epsilon.$$

AEP (Asymptotic Equipartition Property)

AEP. For any $\epsilon > 0$, for all sufficiently large n,

$$\Pr\left[\left|-\frac{1}{n}\log p(X_1, X_2, \dots, X_n) - H(X)\right| > \epsilon\right] \le \epsilon$$

$$\Pr\left[\left|\frac{1}{n}\log p(X_1, X_2, \dots, X_n) + H(X)\right| > \epsilon\right] \le \epsilon$$

$$\Pr\left[\left|\frac{1}{n}\log p(X_1, X_2, \dots, X_n) + H(X)\right| < \epsilon\right] \ge 1 - \epsilon$$

$$\Pr\left[-\epsilon < \frac{1}{n}\log p(X_1, X_2, \dots, X_n) + H(X) < \epsilon\right] \ge 1 - \epsilon$$

AEP (Asymptotic Equipartition Property)

AEP. For any $\epsilon > 0$, for all sufficiently large n,

$$\Pr\left[-\epsilon < \frac{1}{n}\log p(X_1, X_2, \ldots, X_n) + H(X) < \epsilon\right] \geq 1 - \epsilon$$

$$\Pr\left[-H(X) - \epsilon < \frac{1}{n}\log p(X_1, X_2, \ldots, X_n) < -H(X) + \epsilon\right] \geq 1 - \epsilon$$

$$\Pr[-n(H(X) + \epsilon) < \log p(X_1, X_2, \ldots, X_n) < -n(H(X) - \epsilon)] \geq 1 - \epsilon$$

$$\Pr\left[2^{-n(H(X)+\epsilon)} < p(X_1, X_2, \ldots, X_n) < 2^{-n(H(X)-\epsilon)}\right] \geq 1 - \epsilon$$

"Almost all events are almost equally surprising".

Typical Set

The typical set $A_\epsilon^{(n)}$ w.r.t. p is the set of sequence $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ such that

$$2^{-n(H(X)+\epsilon)} < p(\mathbf{x}) < 2^{-n(H(X)-\epsilon)}.$$

Trivially, $\Pr\left[\mathbf{X} \in A_\epsilon^{(n)}\right] \geq 1 - \epsilon.$

Theorem. $(1-\epsilon)2^{n(H(X)-\epsilon)} \leq \left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\epsilon)}$ for sufficiently large n.

proof)

$$\text{(upper bound) } 1 = \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}) \geq \sum_{\mathbf{x} \in A_\epsilon^{(n)}} p(\mathbf{x}) \geq 2^{-n(H(X)+\epsilon)} \left|A_\epsilon^{(n)}\right|$$

$$\text{(lower bound) } 1 - \epsilon \leq \sum_{\mathbf{x} \in A_\epsilon^{(n)}} p(\mathbf{x}) \leq 2^{-n(H(X)-\epsilon)} \left|A_\epsilon^{(n)}\right|$$

Consequence of AEP: Data Compression

Find a short description (i.e., binary string representation) for sequences of i.i.d RVs $X_1, X_2, \dots, X_n$.

Algorithm.

1. Divide sequences in $\mathcal{X}^n$ into $A_\epsilon^{(n)}$ and $A_\epsilon^{(n)} \backslash \mathcal{X}^n$.
2. Index all $\mathbf{x} \in A_\epsilon^{(n)}$ using $\lceil n(H(X)+\epsilon) \rceil + 1$ bits with most significant bit set to 0.
3. Index all $\mathbf{x} \notin A_\epsilon^{(n)}$ using $\lceil n \log|\mathcal{X}| \rceil + 1$ bits with most significant bit set to 1.

Expected length ℓ of the codeword

$$\sum_{\mathbf{x}\in\mathcal{X}^n} p(\mathbf{x})\ell(\mathbf{x}) = \sum_{\mathbf{x}\in A_\epsilon^{(n)}} p(\mathbf{x})(\lceil n(H(X)+\epsilon)\rceil + 1) + \sum_{\mathbf{x}\notin A_\epsilon^{(n)}} p(\mathbf{x})(\lceil n\log|\mathcal{X}|\rceil + 1)$$

$$\leq (n(H(X)+\epsilon)+2)\Pr\left[\mathbf{X}\in A_\epsilon^{(n)}\right] + (n\log|\mathcal{X}|+2)\Pr\left[\mathbf{X}\notin A_\epsilon^{(n)}\right]$$

$$\leq n(H(X)+\epsilon) + \epsilon n\log|\mathcal{X}| + 2$$

$$= n\left(H(X)+\epsilon+\epsilon\log|\mathcal{X}|+\frac{2}{n}\right) = n(H(X)+\epsilon')$$

High Probability and Small Set

$A_\epsilon^{(n)}$ has size $\approx 2^{nH(X)}$ but contains most of the probability.

Is there much smaller set with most of the probability?

For each n, let $B_\delta^{(n)} \subseteq \mathcal{X}^n$ be a smallest set with $\Pr\left[\mathbf{X} \in B_\delta^{(n)}\right] \geq 1 - \delta$.

Observe $\Pr\left[\mathbf{X} \in A_\epsilon^{(n)} \cap B_\delta^{(n)}\right] \geq 1 - \Pr\left[\mathbf{X} \notin A_\epsilon^{(n)}\right] - \Pr\left[\mathbf{X} \notin B_\delta^{(n)}\right] \geq 1 - \epsilon - \delta$.

Moreover,

$$\Pr\left[\mathbf{X} \in A_\epsilon^{(n)} \cap B_\delta^{(n)}\right] \leq \left|A_\epsilon^{(n)} \cap B_\delta^{(n)}\right| 2^{-n(H(X)-\epsilon)} \leq \left|B_\delta^{(n)}\right| 2^{-n(H(X)-\epsilon)}$$

$\mathbf{x} \in A_\epsilon^{(n)} \Rightarrow p(\mathbf{x}) < 2^{-n(H(X)-\epsilon)}$

By rearranging, we obtain

$$\left|B_\delta^{(n)}\right| \geq (1 - \epsilon - \delta) 2^{n(H(X)-\epsilon)} \approx 2^{nH(X)}$$

$A_\epsilon^{(n)}$ vs $B_\delta^{(n)}$

Suppose we have a biased coin X with probability 0.6.

$$H(X) = -0.6\log 0.6 - 0.4\log 0.4 \approx 0.97$$

Consider when $n = 25$ and $\epsilon = 0.1$.

Recall $A_\epsilon^{(n)} = \left\{ \mathbf{x} \in \mathcal{X}^n \;\middle|\; H(X) - \epsilon < -\frac{1}{n}\log p(\mathbf{x}) < H(X) + \epsilon \right\}$

$$A_{0.1}^{(25)} = \left\{ \mathbf{x} \in \mathcal{X}^{25} \;\middle|\; 0.87 < -\frac{1}{n}\log p(\mathbf{x}) < 1.07 \right\}$$

$A_\epsilon^{(n)}$ vs $B_\delta^{(n)}$

$$A_{0.1}^{(25)} = \left\{ \mathbf{x} \in \mathcal{X}^{25} \,\middle|\, 0.87 < -\frac{1}{n}\log p(\mathbf{x}) < 1.07 \right\}$$

For $\mathbf{x}$ with #1=0, $-\frac{1}{n}\log p(\mathbf{x}) = -\frac{1}{25}\log 0.4^{25} = -\log 0.4 \approx 1.32$

For $\mathbf{x}$ with #1=1, $-\frac{1}{n}\log p(\mathbf{x}) = -\frac{1}{25}\log 0.4^{24} 0.6 \approx 1.29$

…

For $\mathbf{x}$ with #1=10, $-\frac{1}{n}\log p(\mathbf{x}) = -\frac{1}{25}\log 0.4^{15} 0.6^{10} \approx 1.08$

For $\mathbf{x}$ with #1=11, $-\frac{1}{n}\log p(\mathbf{x}) = -\frac{1}{25}\log 0.4^{16} 0.6^{11} \approx 1.06$

…

For $\mathbf{x}$ with #1=19, $-\frac{1}{n}\log p(\mathbf{x}) = -\frac{1}{25}\log 0.4^{6} 0.6^{19} \approx 0.88$

For $\mathbf{x}$ with #1=20, $-\frac{1}{n}\log p(\mathbf{x}) = -\frac{1}{25}\log 0.4^{5} 0.6^{20} \approx 0.85$

…

$A_\epsilon^{(n)}$ vs $B_\delta^{(n)}$

Suppose we have a biased coin X with probability 0.6.

$$H(X) = -0.6\log 0.6 - 0.4\log 0.4 \approx 0.97$$

Consider when $n = 25$ and $\epsilon = 0.1$.

Recall $A_\epsilon^{(n)} = \left\{\mathbf{x} \in \mathcal{X}^n \;\middle|\; H(X) - \epsilon < -\frac{1}{n}\log p(\mathbf{x}) < H(X) + \epsilon\right\}$

$$\boldsymbol{A_{0.1}^{(25)} = \{\mathbf{x} \in \mathcal{X}^{25} \mid 11 \le \#1 \text{ in } \mathbf{x} \le 19\}}$$

Recall $B_\delta^{(n)}$ is a smallest set with $\Pr\left[\mathbf{X} \in B_\delta^{(n)}\right] \ge 1 - \delta$.

To find $B_{0.1}^{(25)}$, keep selecting $\mathbf{x} \in \mathcal{X}^n$ with highest prob. until we reach a total probability of 0.9.

$A_\epsilon^{(n)}$ vs $B_\delta^{(n)}$

$B_{0.1}^{(25)}$ is a smallest set with $\Pr\left[\mathbf{X} \in B_{0.1}^{(25)}\right] \geq 0.9$.

- Select $\mathbf{x}$ with #1=25 / cumulative total probability $0.6^{25} \approx 0.000003$
- Select $\mathbf{x}$ with #1=24 / cumulative total probability $\approx 0.000003 + 0.000047 = 0.00005$

…

- Select $\mathbf{x}$ with #1=13 / cumulative total probability ≈ 0.846
- Select $\mathbf{x}$ with #1=12 / cumulative total probability ≈ 0.922

$A_\epsilon^{(n)}$ vs $B_\delta^{(n)}$

Suppose we have a biased coin X with probability 0.6.

$$H(X) = -0.6\log 0.6 - 0.4\log 0.4 \approx 0.97$$

Consider when $n = 25$ and $\epsilon = 0.1$.

Recall $A_\epsilon^{(n)} = \left\{ \mathbf{x} \in \mathcal{X}^n \;\middle|\; H(X) - \epsilon < -\frac{1}{n}\log p(\mathbf{x}) < H(X) + \epsilon \right\}$

$$\boldsymbol{A_{0.1}^{(25)} = \left\{ \mathbf{x} \in \mathcal{X}^{25} \mid \mathbf{11 \le \#1\ in\ x \le 19} \right\}}$$

Recall $B_\delta^{(n)}$ is a smallest set with $\Pr\left[\mathbf{X} \in B_\delta^{(n)}\right] \ge 1 - \delta$.

$$\left\{ \mathbf{x} \in \mathcal{X}^{25} \mid \mathbf{\#1\ in\ x \ge 13} \right\} \subset \boldsymbol{B_{0.1}^{(25)}} \subsetneq \left\{ \mathbf{x} \in \mathcal{X}^{25} \mid \mathbf{\#1\ in\ x \ge 12} \right\}$$

$$\Pr\left[\mathbf{X} \in A_{0.1}^{(25)} \cap B_{0.1}^{(25)}\right] \approx 0.87$$

Remark. The bound $(1-\epsilon)2^{n(H(X)-\epsilon)} \le \left|A_\epsilon^{(n)}\right| \le 2^{n(H(X)+\epsilon)}$ is (very) loose.

$\left|A_{0.1}^{(25)}\right| = 26{,}366{,}510$

lower bound = $3{,}742{,}308$ and upper bound = $114{,}438{,}718$.

Cover and Thomas, Elements of Information Theory (2nd edition), Chapter 4

Entropy Rate

Entropy of RVs from a stationary process

Markov chain

Stochastic Process

Stochastic process $\{X_i\}$: an indexed sequence of RVs with arbitrary dependence

Stationary stochastic process: joint distribution of any subset is invariant w.r.t. shifts in index

$$\Pr[(X_1, X_2, \dots, X_n) = (x_1, x_2, \dots, x_n)] = \Pr[(X_{1+\ell}, X_{2+\ell}, \dots, X_{n+\ell}) = (x_1, x_2, \dots, x_n)]$$

Entropy Rate

Definition 1 (entropy per symbol).

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \dots, X_n)$$ when the limit exists

Definition 2 (conditional entropy of the last).

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \dots, X_1)$$ when the limit exists

Theorem. For a stationary stochastic process, $H(\mathcal{X}) = H'(\mathcal{X})$.

proof)

Observe $H(X_n \mid X_{n-1}, X_{n-2}, \dots, X_1)$ only decreases when n increases. (Since $H \geq 0$, limit exists)

$$H(X_n \mid X_{n-1}, X_{n-2}, \dots, X_1) = H(X_{n+1} \mid X_n, X_{n-1}, \dots, X_2) \geq H(X_{n+1} \mid X_n, X_{n-1}, \dots, X_2, X_1)$$

stationarity conditioning property

By Cesáro mean, $\lim_{n\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \dots, X_1) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_{i-1}, \dots, X_1)$.

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

General AEP

AEP

For any i.i.d. process, in probability,

$$-\frac{1}{n}\log p(X_1, \dots, X_n) \to H(X)$$

General AEP

For any stationary *ergodic* process, with probability 1,

$$-\frac{1}{n}\log p(X_1, \dots, X_n) \to H(\mathcal{X})$$

Markov Chain

Markov chain (or *process*): dependence only on the one just before it

$$\Pr[X_{n+1} = x_{n+1} \mid (X_1, X_2, \dots, X_n) = (x_1, x_2, \dots, x_n)] = \Pr[X_{n+1} = x_{n+1} \mid X_n = x_n]$$

* Here we assume Markov chain is time invariant, i.e.,

$$\Pr[X_{n+1} = b \mid X_n = a] = \Pr[X_2 = b \mid X_1 = a]$$

Fundamental Theorem of Markov Chain.

A finite, irreducible and aperiodic Markov chain

- has the unique stationary distribution and

- any distribution converges to the stationary distribution.

Stationary distribution: $\mu = \mu^T P$

Irreducible: Transition graph P is strongly connected component.

Aperiodic: GCD(all closed directed walk from v to v w/ prob.>0)=1.

Stationary Markov Chain

With initial dist. as stationary dist. μ, Markov chain is a stationary process.

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \dots, X_1) = \lim_{n\to\infty} H(X_n \mid X_{n-1}) = H(X_2 \mid X_1)$$

Markovity stationarity

We have

$$H(\mathcal{X}) = H(X_2 \mid X_1) = \sum_{i,j} \mu(i) H(X_2 | X_1 = i)$$

$$= -\sum_i \mu(i) \sum_j P_{ij} \log P_{ij}$$

$$= -\sum_{i,j} \mu(i) P_{ij} \log P_{ij}$$

Functions of Markov Chain

Let $\{X_i\}$ be a stationary Markov chain.

Consider $\{Y_i\}$ where $Y_i = \phi(X_i)$.

Note $\{Y_i\}$ does not necessarily form a Markov chain.

Consider a Markov chain with $P_{ac} = P_{ca} = P_{bb} = 1$.

Observe the uniform distribution is a stationary distribution.

Now consider a function ϕ such that $\phi(a) = \phi(b) = s$ and $\phi(c) = t$.

$$\Pr[\, Y_3 = s \mid Y_2 = s \,] = \frac{1}{2}$$

$$\Pr[\, Y_3 = s \mid Y_2 = s, Y_1 = s \,] = 1$$

Functions of Markov Chain

Let $\{X_i\}$ be a stationary Markov chain.

Consider $\{Y_i\}$ where $Y_i = \phi(X_i)$.

Note $\{Y_i\}$ does not necessarily form a Markov chain.

Therefore, to compute $H(\mathcal{Y})$, need to compute $H(Y_n \mid Y_{n-1}, Y_{n-2}, \dots, Y_1)$.

How to know $H(Y_n \mid Y_{n-1}, Y_{n-2}, \dots, Y_1) \approx H(\mathcal{Y})$ for any n?

Recall that it converges from above.

$$\cdots \geq H(Y_n \mid Y_{n-1}, Y_{n-2}, \dots, Y_1) \geq H(Y_{n+1} \mid Y_n, Y_{n-1}, \dots, Y_1) \geq \cdots \geq H(\mathcal{Y})$$

Lemma. $H(\mathcal{Y}) \geq H(Y_n \mid Y_{n-1}, Y_{n-2}, \dots, Y_1, X_1)$.

$$H(Y_n \mid Y_{n-1}, Y_{n-2}, \dots, Y_1, X_1) \leq H(\mathcal{Y}) \leq H(Y_n \mid Y_{n-1}, Y_{n-2}, \dots, Y_1)$$

Theorem.

$$\lim_{n\to\infty} H(Y_n \mid Y_{n-1}, Y_{n-2}, \dots, Y_1, X_1) = H(\mathcal{Y}) = \lim_{n\to\infty} H(Y_n \mid Y_{n-1}, Y_{n-2}, \dots, Y_1)$$

If ϕ is random, this is related to a *hidden Markov chain* (HMM)

Thank You