Information Theory and Related Fileds

Lecture 1: Information-Theoretic Quantities

Lei Yu

Nankai University

Online Short Course at Beijing Normal University

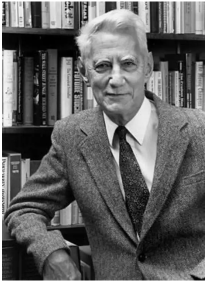
Outline

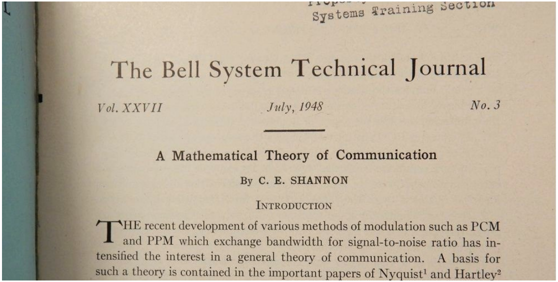
Outline

Birth of Information Theory

Information theory was essentially established

- by Claude Shannon in 1948
- in the paper "A Mathematical Theory of Communication"
- via introducing (information) entropy (borrowed from a similar notion in thermodynamics)





Systems Training Section

The Bell System Technical Journal

Vol. XXVII July, 1948 No. 3

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2]

Communication System

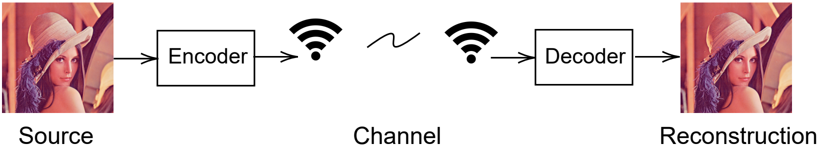


Figure: Practical Communication

Communication System

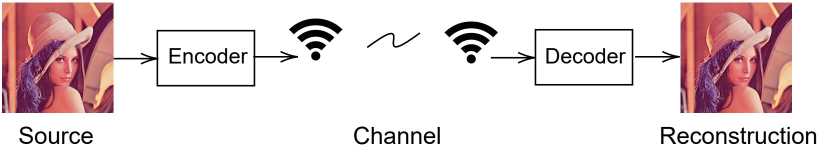


Figure: Practical Communication

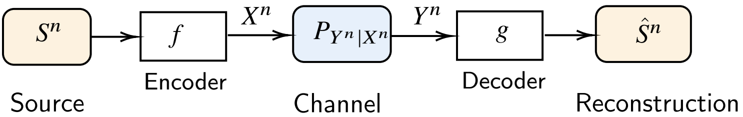


Figure: Mathematical Model, where notation $Z^n := (Z_1, Z_2, ..., Z_n)$ denotes a random vector

Source and Channel

- Source S^n (or P_{S^n}): A random vector S^n [or a stochastic process $S^\infty := (S_1, S_2, ...)$]

Source and Channel

- Source S^n (or P_{S^n}): A random vector S^n [or a stochastic process $S^\infty := (S_1, S_2, ...)$]
- Memoryless source S (or P_S): A source S^n consisting of i.i.d. copies of S
  - In other words, S^n follows the product distribution $P_S^{\otimes n}$

Source and Channel

- Source S^n (or P_{S^n}): A random vector S^n [or a stochastic process $S^\infty := (S_1, S_2, \ldots)$]
- Memoryless source S (or P_S): A source S^n consisting of i.i.d. copies of S
  - In other words, S^n follows the product distribution $P_S^{\otimes n}$
- Channel $P_{Y^n|X^n}$: A random transformation which outputs Y^n s.t. $Y^n|X^n \sim P_{Y^n|X^n}$ when the input is X^n

Source and Channel

- Source S^n (or P_{S^n}): A random vector S^n [or a stochastic process $S^\infty := (S_1, S_2, ...)$]
- Memoryless source S (or P_S): A source S^n consisting of i.i.d. copies of S
  - In other words, S^n follows the product distribution $P_S^{\otimes n}$
- Channel $P_{Y^n|X^n}$: A random transformation which outputs Y^n s.t. $Y^n|X^n \sim P_{Y^n|X^n}$ when the input is X^n
- Memoryless channel $P_{Y|X}$: A channel with product conditional distribution $P_{Y|X}^{\otimes n}$

Source and Channel

- Source S^n (or P_{S^n}): A random vector S^n [or a stochastic process $S^\infty := (S_1, S_2, ...)$]
- Memoryless source S (or P_S): A source S^n consisting of i.i.d. copies of S
  - In other words, S^n follows the product distribution $P_S^{\otimes n}$
- Channel $P_{Y^n|X^n}$: A random transformation which outputs Y^n s.t. $Y^n|X^n \sim P_{Y^n|X^n}$ when the input is X^n
- Memoryless channel $P_{Y|X}$: A channel with product conditional distribution $P_{Y|X}^{\otimes n}$
  - In other words, component Y_i is generated only by X_i via $P_{Y|X}$, i.e.,

$$\begin{array}{ccccc} X_1 & \longrightarrow & P_{Y|X} & \longrightarrow & Y_1 \\ ... & & ... & & ... \\ X_n & \longrightarrow & P_{Y|X} & \longrightarrow & Y_n \end{array}$$

  - If the inputs $X_1, X_2, ...$ are i.i.d., then the outputs $Y_1, Y_2, ...$ are also i.i.d.

Encoder and Decoder

Convention: For a r.v. X, we always use the calligraphic font $\mathcal{X}$ to denote the range of X.

Encoder and Decoder

Convention: For a r.v. X, we always use the calligraphic font $\mathcal{X}$ to denote the range of X.

- Encoder: $f : \mathcal{S}^n \to \mathcal{X}^n$, a map from source S^n to channel input X^n

Encoder and Decoder

Convention: For a r.v. X, we always use the calligraphic font $\mathcal{X}$ to denote the range of X.

- Encoder: $f : \mathcal{S}^n \to \mathcal{X}^n$, a map from source S^n to channel input X^n
- Decoder: $g : \mathcal{Y}^n \to \hat{\mathcal{S}}^n$, a map from channel output Y^n to reconstruction $\hat{S}^n$

Encoder and Decoder

Convention: For a r.v. X, we always use the calligraphic font $\mathcal{X}$ to denote the range of X.

- Encoder: $f : \mathcal{S}^n \to \mathcal{X}^n$, a map from source S^n to channel input X^n
- Decoder: $g : \mathcal{Y}^n \to \hat{\mathcal{S}}^n$, a map from channel output Y^n to reconstruction $\hat{S}^n$
- So, for source P_{S^n}, channel $P_{Y^n|X^n}$, encoder f, and decoder g, the joint distribution induced by them is

$$P_{S^n X^n Y^n \hat{S}^n} = P_{S^n} P_{X^n|S^n} P_{Y^n|X^n} P_{\hat{S}^n|Y^n}$$

where $P_{X^n|S^n=s^n} = \delta_{f(s^n)}$ and $P_{\hat{S}^n|Y^n=y^n} = \delta_{g(y^n)}$ with δ_z denoting the Dirac measure at z

Goal of Information Theory

Question: Given a pair of source and channel, can the source be transmitted asymptotically losslessly over the channel in the sense that $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$ (as $n \to \infty$)?

Goal of Information Theory

Question: Given a pair of source and channel, can the source be transmitted asymptotically losslessly over the channel in the sense that $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$ (as $n \to \infty$)?

Solved by Shannon who introduced:

- Entropy of a discrete random variable X is

$$H(X) := \sum_x P_X(x) \log \frac{1}{P_X(x)}$$

Goal of Information Theory

Question: Given a pair of source and channel, can the source be transmitted asymptotically losslessly over the channel in the sense that $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$ (as $n \to \infty$)?

Solved by Shannon who introduced:

- Entropy of a discrete random variable X is

$$H(X) := \sum_x P_X(x) \log \frac{1}{P_X(x)}$$

- Mutual information between X and Y is

$$I(X;Y) := \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x) P_Y(y)}$$

Goal of Information Theory

Question: Given a pair of source and channel, can the source be transmitted asymptotically losslessly over the channel in the sense that $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$ (as $n \to \infty$)?

Solved by Shannon who introduced:

- Entropy of a discrete random variable X is

$$H(X) := \sum_x P_X(x) \log \frac{1}{P_X(x)}$$

- Mutual information between X and Y is

$$I(X;Y) := \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}$$

- Capacity of a memoryless channel $P_{Y|X}$ is

$$C(P_{Y|X}) := \max_{P_X} I(X;Y)$$

Solution

Theorem ([Shannon'48])

Consider memoryless source S and memoryless channel $P_{Y|X}$. There is a sequence of encoder-decoder pairs (f_n, g_n) such that $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$ (as $n \to \infty$) if $H(S) < C(P_{Y|X})$, and only if $H(S) \leq C(P_{Y|X})$.

Solution

Theorem ([Shannon'48])

Consider memoryless source S and memoryless channel $P_{Y|X}$. There is a sequence of encoder-decoder pairs (f_n, g_n) such that $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$ (as $n \to \infty$) if $H(S) < C(P_{Y|X})$, and only if $H(S) \leq C(P_{Y|X})$.

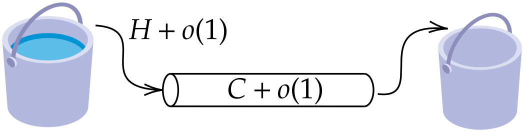
- To be proven in the first three lectures.

Solution

Theorem ([Shannon'48])

Consider memoryless source S and memoryless channel $P_{Y|X}$. There is a sequence of encoder-decoder pairs (f_n, g_n) such that $\mathbb{P}(S^n \neq \hat{S}^n) \to 0$ (as $n \to \infty$) if $H(S) < C(P_{Y|X})$, and only if $H(S) \leq C(P_{Y|X})$.

- To be proven in the first three lectures.
- Analogy: Water of volume $nH + o(n)$ can be transfered within time n via a pipe of capacity $C + o(1)$, if $H < C$ and only if $H \leq C$.



Outline

Entropy

Let X, Y be discrete random variables (r.v.'s). We adopt convention $0 \log 0 = 0$.

Entropy

Let X, Y be discrete random variables (r.v.'s). We adopt convention $0 \log 0 = 0$.

Definition

Recall that entropy of discrete X is

$$H(X) := \sum_{x} P_X(x) \log \frac{1}{P_X(x)} = \mathbb{E}_{P_X}\left[\log \frac{1}{P_X(X)}\right]$$

Joint entropy of X and Y is

$$H(X,Y) := \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_{XY}(x,y)}$$

Conditional entropy of Y given X is

$$H(Y|X) := \sum_{x,y} P_{XY}(x,y) \log \frac{1}{P_{Y|X}(y|x)} = \sum_{x} P_X(x) H(Y|X=x)$$

Facts and Examples

Fact

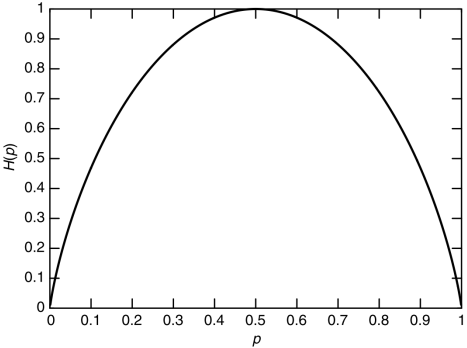
1) Nonnegativity: $H(X)$, $H(X,Y)$, $H(Y|X) \geq 0$ *(Moreover,* $H(Y|X) = 0$ *iff* $Y = g(X)$ *for some function)*
2) Chain rule: $H(X,Y) = H(X) + H(Y|X)$

Facts and Examples

Fact

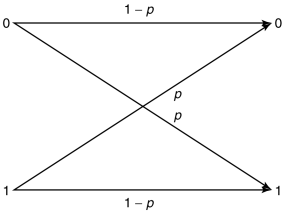
1) Nonnegativity: $H(X)$, $H(X,Y)$, $H(Y|X) \geq 0$ *(Moreover,* $H(Y|X) = 0$ *iff* $Y = g(X)$ *for some function)*
2) Chain rule: $H(X,Y) = H(X) + H(Y|X)$

- Binary Source: For $X \sim \mathrm{Bern}(p)$, $H(X) = H_2(p) := p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$, where H_2 is called binary entropy function



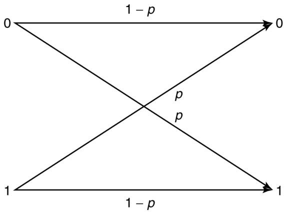
Facts and Examples (cont.)

- Binary Symmetric Channel $\mathrm{BSC}(p)$: $Y = X \oplus W$ with $W \sim \mathrm{Bern}(p)$ independent of X. For such channel, $H(Y|X) = H_2(p)$



Facts and Examples (cont.)

- Binary Symmetric Channel $\mathrm{BSC}(p)$: $Y = X \oplus W$ with $W \sim \mathrm{Bern}(p)$ independent of X. For such channel, $H(Y|X) = H_2(p)$



- Doubly Symmetric Binary Source (DSBS): $X \sim \mathrm{Bern}(\frac{1}{2})$ and Y is the output of $\mathrm{BSC}(p)$ with input X.

$$P_{XY} = \begin{array}{ccc} X\backslash Y & 0 & 1 \\ 0 & \frac{1-p}{2} & \frac{p}{2} \\ 1 & \frac{p}{2} & \frac{1-p}{2} \end{array}.$$

For such source, $H(X,Y) = 1 + H_2(p)$.

How to Interpret Entropy?

- How to measure the amount $\iota(A)$ of information (or uncertainty) contained in an event A?
  - $\iota(A)$ is the amount of information that we gain when knowing that A occurs

How to Interpret Entropy?

- How to measure the amount $\iota(A)$ of information (or uncertainty) contained in an event A?
  - $\iota(A)$ is the amount of information that we gain when knowing that A occurs
  - Without knowing A occurs, the probability of A is $P(A)$; but after knowing A occurs, the probability of A is 1

How to Interpret Entropy?

- How to measure the amount $\iota(A)$ of information (or uncertainty) contained in an event A?
  - $\iota(A)$ is the amount of information that we gain when knowing that A occurs
  - Without knowing A occurs, the probability of A is $P(A)$; but after knowing A occurs, the probability of A is 1
  - The ratio of these two probabilities is $\frac{1}{P(A)}$

How to Interpret Entropy?

- How to measure the amount $\iota(A)$ of information (or uncertainty) contained in an event A?
  - $\iota(A)$ is the amount of information that we gain when knowing that A occurs
  - Without knowing A occurs, the probability of A is $P(A)$; but after knowing A occurs, the probability of A is 1
  - The ratio of these two probabilities is $\frac{1}{P(A)}$
- Intuitively, $\iota(A)$ should satisfy following axioms:
  1. $\iota(A)$ is continuous in $P(A)$

How to Interpret Entropy?

- How to measure the amount $\iota(A)$ of information (or uncertainty) contained in an event A?
  - $\iota(A)$ is the amount of information that we gain when knowing that A occurs
  - Without knowing A occurs, the probability of A is $P(A)$; but after knowing A occurs, the probability of A is 1
  - The ratio of these two probabilities is $\frac{1}{P(A)}$
- Intuitively, $\iota(A)$ should satisfy following axioms:
  1. $\iota(A)$ is continuous in $P(A)$
  2. $\iota(A)$ is decreasing in $P(A)$ (or equivalently, increasing in $\frac{1}{P(A)}$)

How to Interpret Entropy?

- How to measure the amount $\iota(A)$ of information (or uncertainty) contained in an event A?
  - $\iota(A)$ is the amount of information that we gain when knowing that A occurs
  - Without knowing A occurs, the probability of A is $P(A)$; but after knowing A occurs, the probability of A is 1
  - The ratio of these two probabilities is $\frac{1}{P(A)}$
- Intuitively, $\iota(A)$ should satisfy following axioms:
  1. $\iota(A)$ is continuous in $P(A)$
  2. $\iota(A)$ is decreasing in $P(A)$ (or equivalently, increasing in $\frac{1}{P(A)}$)
  3. $\iota(A \cap B) = \iota(A) + \iota(B)$ for independent A, B (i.e., $P(A \cap B) = P(A)P(B)$)

How to Interpret Entropy?

- How to measure the amount $\iota(A)$ of information (or uncertainty) contained in an event A?
  - $\iota(A)$ is the amount of information that we gain when knowing that A occurs
  - Without knowing A occurs, the probability of A is $P(A)$; but after knowing A occurs, the probability of A is 1
  - The ratio of these two probabilities is $\frac{1}{P(A)}$
- Intuitively, $\iota(A)$ should satisfy following axioms:
  1. $\iota(A)$ is continuous in $P(A)$
  2. $\iota(A)$ is decreasing in $P(A)$ (or equivalently, increasing in $\frac{1}{P(A)}$)
  3. $\iota(A \cap B) = \iota(A) + \iota(B)$ for independent A, B (i.e., $P(A \cap B) = P(A)P(B)$)
- **Fact:** $\iota(A) = \log \frac{1}{P(A)}$ is the unique quantity (up to a positive constant factor) satisfying the three axioms above.

How to Interpret Entropy?

- How to measure the amount $\iota(A)$ of information (or uncertainty) contained in an event A?
  - $\iota(A)$ is the amount of information that we gain when knowing that A occurs
  - Without knowing A occurs, the probability of A is $P(A)$; but after knowing A occurs, the probability of A is 1
  - The ratio of these two probabilities is $\frac{1}{P(A)}$
- Intuitively, $\iota(A)$ should satisfy following axioms:
  1. $\iota(A)$ is continuous in $P(A)$
  2. $\iota(A)$ is decreasing in $P(A)$ (or equivalently, increasing in $\frac{1}{P(A)}$)
  3. $\iota(A \cap B) = \iota(A) + \iota(B)$ for independent A, B (i.e., $P(A \cap B) = P(A)P(B)$)
- **Fact:** $\iota(A) = \log \frac{1}{P(A)}$ is the unique quantity (up to a positive constant factor) satisfying the three axioms above.

Definition

$\iota(A) = \log \frac{1}{P(A)}$ is called the self-information of A.

Fact

It holds that

$$H(X) = \sum_x P_X(x) \cdot \iota_X(x)$$

where $\iota_X(x) = \log \frac{1}{P_X(x)}$ *is the self-information of outcome* $X = x$.

How to Interpret Entropy? (cont.)

Fact

It holds that

$$H(X) = \sum_x P_X(x) \cdot \iota_X(x)$$

where $\iota_X(x) = \log \frac{1}{P_X(x)}$ *is the self-information of outcome* $X = x$*.*

- That is, the entropy of X is equal to the expected self-information of X. In other words, it is the average amount of information that we gain when measuring X.

How to Interpret Entropy? (cont.)

Fact

It holds that

$$H(X) = \sum_{x} P_X(x) \cdot \iota_X(x)$$

where $\iota_X(x) = \log \frac{1}{P_X(x)}$ *is the self-information of outcome* $X = x$*.*

- That is, the entropy of X is equal to the expected self-information of X. In other words, it is the average amount of information that we gain when measuring X.
- Another way to define entropy via axioms was given by Shannon in [Shannon'48] (omitted)

Mutual Information

Definition

Mutual information (MI) between X and Y is defined as

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}$$

Conditional mutual information between X and Y given W is defined as

$$I(X;Y|W) = \sum_{x,y,w} P_{XYW}(x,y,w) \log \frac{P_{XY|W}(x,y|w)}{P_{X|W}(x|w)P_{Y|W}(y|w)}$$

Facts

Fact

It holds that

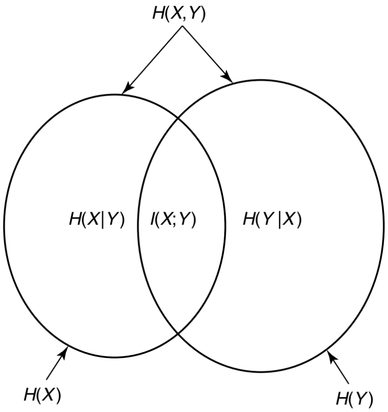
$$\begin{aligned} I(X;Y) = I(Y;X) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

and

$$I(X;g(X)) = H(g(X)).$$

The equalities above still hold if we put the condition "$|W$" in each quantity.

Venn Diagram



Relative Entropy

Definition (Relative Entropy for Discrete Distributions)

For a distribution P and a nonnegative measure μ, the relative entropy [a.k.a. Kullback–Leibler (KL) divergence] of P with respect to (w.r.t.) μ is defined as

$$D(P\|\mu) := \sum_x P(x) \log\left(\frac{P(x)}{\mu(x)}\right).$$

In particular, $D(P\|Q) := \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right)$.

Relative Entropy

Definition (Relative Entropy for Discrete Distributions)

For a distribution P and a nonnegative measure μ, the relative entropy [a.k.a. Kullback–Leibler (KL) divergence] of P with respect to (w.r.t.) μ is defined as

$$D(P\|\mu) := \sum_x P(x) \log \left(\frac{P(x)}{\mu(x)} \right).$$

In particular, $D(P\|Q) := \sum_x P(x) \log \left(\frac{P(x)}{Q(x)} \right)$.

Fact (Entropy and MI are special cases of relative entropy)

1) If μ is the counting measure, denoted by $\mathrm{Count}(\mathcal{X})$, then

$$D(P\| \mathrm{Count}(\mathcal{X})) = \sum_x P(x) \log P(x) = -H(X).$$

2) $D(P\| \mathrm{Unif}(\mathcal{X})) = \log |\mathcal{X}| - H(X)$.
3) It holds that $D(P_{XY}\|P_X \otimes P_Y) = I(X;Y)$.

Conditional Relative Entropy

Definition (Conditional Relative Entropy)

For probability measure P_X, transition probability measure $P_{Y|X}$, and transition measure $\mu_{Y|X}$, the relative entropy of $P_{Y|X}$ w.r.t. $\mu_{Y|X}$ conditioned on P_X is defined as

$$D(P_{Y|X}\|\mu_{Y|X}|P_X) := D(P_X P_{Y|X}\|P_X \mu_{Y|X}) = \mathbb{E}_{P_X}\left[D(P_{Y|X}\|\mu_{Y|X})\right].$$

Conditional Relative Entropy

Definition (Conditional Relative Entropy)

For probability measure P_X, transition probability measure $P_{Y|X}$, and transition measure $\mu_{Y|X}$, the relative entropy of $P_{Y|X}$ w.r.t. $\mu_{Y|X}$ conditioned on P_X is defined as

$$D(P_{Y|X} \| \mu_{Y|X} | P_X) := D(P_X P_{Y|X} \| P_X \mu_{Y|X}) = \mathbb{E}_{P_X}\left[D(P_{Y|X} \| \mu_{Y|X})\right].$$

Fact (Conditional entropy and MI are special cases of conditional relative entropy)

1) If $\mu_{Y|X=x}$ is the counting measure $\mathrm{Count}(\mathcal{Y})$ for every x, then

$$D(P_{Y|X} \| \mathrm{Count}(\mathcal{Y}) | P_X) = \sum_{x,y} P_{XY}(x,y) \log P_{Y|X}(y|x) = -H(Y|X).$$

2) $D(P_{Y|X} \| \mathrm{Unif}(\mathcal{Y}) | P_X) = \log|\mathcal{Y}| - H(Y|X)$.
3) It holds that $D(P_{Y|XW} \| P_{Y|W} | P_{XW}) = I(X;Y|W)$.

Outline

Chain Rule

Theorem (Chain rule)

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

which still holds with substitution $Q \leftarrow \mu$ (arbitrary nonnegative measure). In particular,

$$H(X^n) = \sum_{i=1}^{n} H(X_i|X^{i-1}) \ \text{and}\ I(X^n; Y) = \sum_{i=1}^{n} I(X_i; Y|X^{i-1}).$$

Chain Rule

Theorem (Chain rule)

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

which still holds with substitution $Q \leftarrow \mu$ (arbitrary nonnegative measure). In particular,

$$H(X^n) = \sum_{i=1}^{n} H(X_i|X^{i-1}) \text{ and } I(X^n;Y) = \sum_{i=1}^{n} I(X_i;Y|X^{i-1}).$$

Proof: By the disintegration theorem, $P_{X^n} = \prod_{i=1}^n P_{X_i|X^{i-1}}$ and $Q_{X^n} = \prod_{i=1}^n Q_{X_i|X^{i-1}}$. Then,

$$\begin{aligned} D(P_{X^n} \| Q_{X^n}) = \mathbb{E}_P \left[\log \left(\frac{P(X^n)}{Q(X^n)} \right) \right] &= \sum_{i=1}^{n} \mathbb{E}_P \left[\log \left(\frac{P(X_i|X^{i-1})}{Q(X_i|X^{i-1})} \right) \right] \\ &= \sum_{i=1}^{n} D(P_{X_i|X^{i-1}} \| Q_{X_i|X^{i-1}} | P_{X^{i-1}}). \end{aligned}$$

Nonnegativity

Theorem (Nonnegativity)

For probability measures P, Q*, it holds that* $D(P\|Q) \geq 0$ *with equality iff* $P = Q$*. In particular,* $I(X;Y) \geq 0$ *with quality iff* X *and* Y *are independent.*

Nonnegativity

Theorem (Nonnegativity)

For probability measures P, Q, it holds that $D(P\|Q) \geq 0$ with equality iff $P = Q$. In particular, $I(X;Y) \geq 0$ with quality iff X and Y are independent.

Proof:

$$\begin{aligned}
D(P\|Q) &= \sum_x Q(x)\frac{P(x)}{Q(x)} \log\left(\frac{P(x)}{Q(x)}\right) \\
&= \mathbb{E}_Q\left[\varphi\left(\frac{P(X)}{Q(X)}\right)\right] && \varphi(t) = t\log t \\
&\geq \varphi\left(\mathbb{E}_Q\left[\frac{P(X)}{Q(X)}\right]\right) && \text{by convexity of } \varphi \text{ and Jensen's inequality} \\
&= \varphi(1) = 0
\end{aligned}$$

Consequence 1: Conditioning Reducing Entropy

Question: What can be derived from Chain Rule and Nonnegativity?

Consequence 1: Conditioning Reducing Entropy

Question: What can be derived from Chain Rule and Nonnegativity?

Corollary (Conditioning reduces entropy)

$H(X) \geq H(X|Y)$ *with equality iff* X *and* Y *are independent.*

Consequence 1: Conditioning Reducing Entropy

Question: What can be derived from Chain Rule and Nonnegativity?

Corollary (Conditioning reduces entropy)

$H(X) \geq H(X|Y)$ *with equality iff* X *and* Y *are independent.*

Proof: $I(X;Y) = H(X) - H(X|Y) \geq 0$

Consequence 2: Joint Relative Entropy is Larger

Theorem (Joint relative entropy is larger than marginal ones)

$$D(P_{XY}\|Q_{XY}) \geq D(P_Y\|Q_Y) \quad (and\ D(P_{XY}\|Q_{XY}) \geq D(P_X\|Q_X)),$$

which still holds with $Q \leftarrow \mu$.

Consequence 2: Joint Relative Entropy is Larger

Theorem (Joint relative entropy is larger than marginal ones)

$$D(P_{XY}\|Q_{XY}) \geq D(P_Y\|Q_Y) \quad (\textit{and } D(P_{XY}\|Q_{XY}) \geq D(P_X\|Q_X)),$$

which still holds with $Q \leftarrow \mu$.

Proof:

$$\begin{aligned} D(P_{XY}\|Q_{XY}) &= D(P_Y\|Q_Y) + D(P_{X|Y}\|Q_{X|Y}|P_Y) && \text{by chain rule} \\ &\geq D(P_Y\|Q_Y) && \text{by nonnegativity} \end{aligned}$$

Two Special Cases

Corollary (Data processing inequality (DPI))

Given a channel $P_{Y|X}$, it holds that

$$D(P_X \| Q_X) \geq D(P_Y \| Q_Y),$$

where $P_X \to P_{Y|X} \to P_Y$ and $Q_X \to P_{Y|X} \to Q_Y$.

Two Special Cases

Corollary (Data processing inequality (DPI))

Given a channel $P_{Y|X}$, it holds that

$$D(P_X \| Q_X) \geq D(P_Y \| Q_Y),$$

where $P_X \to P_{Y|X} \to P_Y$ and $Q_X \to P_{Y|X} \to Q_Y$.

Corollary (Conditioning increases relative entropy)

Given a distribution P_X, it holds that

$$D(P_{Y|X} \| Q_{Y|X} | P_X) \geq D(P_Y \| Q_Y), \tag{1}$$

where $P_X \to P_{Y|X} \to P_Y$ and $P_X \to Q_{Y|X} \to Q_Y$.

Consequence 3: Convexity

- If further, $P_X = \mathrm{Bern}(\lambda)$, then (1) implies

$$(1-\lambda)D(P_0\|Q_0) + \lambda D(P_1\|Q_1) \geq D((1-\lambda)P_0 + \lambda P_1\|(1-\lambda)Q_0 + \lambda Q_1)$$

where $P_i := P_{Y|X=i}$, $Q_i := Q_{Y|X=i}$, $i = 0, 1$.

Consequence 3: Convexity

- If further, $P_X = \mathrm{Bern}(\lambda)$, then (1) implies

$$(1-\lambda)D(P_0\|Q_0) + \lambda D(P_1\|Q_1) \geq D((1-\lambda)P_0 + \lambda P_1\|(1-\lambda)Q_0 + \lambda Q_1)$$

where $P_i := P_{Y|X=i}$, $Q_i := Q_{Y|X=i}$, $i = 0, 1$.

Theorem

$(P, Q) \mapsto D(P\|Q)$ is convex, which still holds with $Q \leftarrow \mu$.

Consequence 3: Convexity

- If further, $P_X = \mathrm{Bern}(\lambda)$, then (1) implies

$$(1-\lambda)D(P_0\|Q_0) + \lambda D(P_1\|Q_1) \geq D((1-\lambda)P_0 + \lambda P_1\|(1-\lambda)Q_0 + \lambda Q_1)$$

where $P_i := P_{Y|X=i}$, $Q_i := Q_{Y|X=i}$, $i = 0, 1$.

Theorem

$(P, Q) \mapsto D(P\|Q)$ *is convex, which still holds with* $Q \leftarrow \mu$.

Corollary

$P_X \mapsto H(X)$, $P_{XY} \mapsto H(X|Y)$, *and* $P_X \mapsto I(X;Y)$ *are all concave, and* $P_{Y|X} \mapsto I(X;Y)$ *is convex.*

Consequence 3: Convexity

- If further, $P_X = \mathrm{Bern}(\lambda)$, then (1) implies

$$(1-\lambda)D(P_0\|Q_0) + \lambda D(P_1\|Q_1) \geq D((1-\lambda)P_0 + \lambda P_1\|(1-\lambda)Q_0 + \lambda Q_1)$$

where $P_i := P_{Y|X=i}$, $Q_i := Q_{Y|X=i}$, $i = 0, 1$.

Theorem

$(P, Q) \mapsto D(P\|Q)$ *is convex, which still holds with* $Q \leftarrow \mu$.

Corollary

$P_X \mapsto H(X)$, $P_{XY} \mapsto H(X|Y)$, *and* $P_X \mapsto I(X;Y)$ *are all concave, and* $P_{Y|X} \mapsto I(X;Y)$ *is convex.*

Proof: 1) $H(X) = -D(P\|\,\mathrm{Count}(\mathcal{X}))$. 2) $H(Y|X) = -D(P_{Y|X}\|\,\mathrm{Count}(\mathcal{Y})|P_X)$.

Consequence 3: Convexity

- If further, $P_X = \mathrm{Bern}(\lambda)$, then (1) implies

$$(1-\lambda)D(P_0\|Q_0) + \lambda D(P_1\|Q_1) \geq D((1-\lambda)P_0 + \lambda P_1\|(1-\lambda)Q_0 + \lambda Q_1)$$

where $P_i := P_{Y|X=i}$, $Q_i := Q_{Y|X=i}$, $i = 0, 1$.

Theorem

$(P, Q) \mapsto D(P\|Q)$ is convex, which still holds with $Q \leftarrow \mu$.

Corollary

$P_X \mapsto H(X)$, $P_{XY} \mapsto H(X|Y)$, and $P_X \mapsto I(X;Y)$ are all concave, and $P_{Y|X} \mapsto I(X;Y)$ is convex.

Proof: 1) $H(X) = -D(P\|\,\mathrm{Count}(\mathcal{X}))$. 2) $H(Y|X) = -D(P_{Y|X}\|\,\mathrm{Count}(\mathcal{Y})|P_X)$. 3) (a) $I(X;Y) = H(Y) - H(Y|X)$; (b) $P_X \mapsto H(Y)$ is concave (since $P_Y \mapsto H(Y)$ is concave and $P_X \mapsto P_Y$ is linear); (c) $P_X \mapsto H(Y|X)$ is linear.

Consequence 3: Convexity

- If further, $P_X = \mathrm{Bern}(\lambda)$, then (1) implies

$$(1-\lambda)D(P_0\|Q_0) + \lambda D(P_1\|Q_1) \geq D((1-\lambda)P_0 + \lambda P_1\|(1-\lambda)Q_0 + \lambda Q_1)$$

where $P_i := P_{Y|X=i}$, $Q_i := Q_{Y|X=i}$, $i = 0, 1$.

Theorem

$(P, Q) \mapsto D(P\|Q)$ is convex, which still holds with $Q \leftarrow \mu$.

Corollary

$P_X \mapsto H(X)$, $P_{XY} \mapsto H(X|Y)$, and $P_X \mapsto I(X;Y)$ are all concave, and $P_{Y|X} \mapsto I(X;Y)$ is convex.

Proof: 1) $H(X) = -D(P\|\mathrm{Count}(\mathcal{X}))$. 2) $H(Y|X) = -D(P_{Y|X}\|\mathrm{Count}(\mathcal{Y})|P_X)$.
3) (a) $I(X;Y) = H(Y) - H(Y|X)$; (b) $P_X \mapsto H(Y)$ is concave (since $P_Y \mapsto H(Y)$ is concave and $P_X \mapsto P_Y$ is linear); (c) $P_X \mapsto H(Y|X)$ is linear.
4) $I(X;Y) = D(P_{XY}\|P_X \otimes P_Y) = D(P_{Y|X}\|P_Y|P_X)$. Given P_X, $P_{Y|X} \mapsto P_Y$ is linear.

Consequence 4: Superadditivity

Theorem (Superadditivity of Relative Entropy)

$$D(P_{X^n} \| Q_X^{\otimes n}) \geq \sum_{i=1}^{n} D(P_{X_i} \| Q_X),$$

which still holds with $Q \leftarrow \mu$.

Consequence 4: Superadditivity

Theorem (Superadditivity of Relative Entropy)

$$D(P_{X^n} \| Q_X^{\otimes n}) \geq \sum_{i=1}^{n} D(P_{X_i} \| Q_X),$$

which still holds with $Q \leftarrow \mu$.

Proof:

$$D(P_{X^n} \| Q_X^{\otimes n}) = \sum_{i=1}^{n} D(P_{X_i|X^{i-1}} \| Q_X | P_{X^{i-1}}) \geq \sum_{i=1}^{n} D(P_{X_i} \| Q_X).$$

Consequence 4: Superadditivity

Theorem (Superadditivity of Relative Entropy)

$$D(P_{X^n} \| Q_X^{\otimes n}) \geq \sum_{i=1}^{n} D(P_{X_i} \| Q_X),$$

which still holds with $Q \leftarrow \mu$.

Proof:

$$D(P_{X^n} \| Q_X^{\otimes n}) = \sum_{i=1}^{n} D(P_{X_i|X^{i-1}} \| Q_X | P_{X^{i-1}}) \geq \sum_{i=1}^{n} D(P_{X_i} \| Q_X).$$

Corollary (Subadditivity of Entropy)

$H(X^n) \leq \sum_{i=1}^{n} H(X_i)$ *with equality iff the* X_i*'s are independent.*

Consequence 5: Maximum Entropy

Theorem

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

where $|\mathcal{X}|$ *denotes the number of elements in the range of* X*, with equality iff* X *is uniformly distributed over* $\mathcal{X}$*.*

Consequence 5: Maximum Entropy

Theorem

$$H(X) \le \log |\mathcal{X}|$$

where $|\mathcal{X}|$ denotes the number of elements in the range of X, with equality iff X is uniformly distributed over $\mathcal{X}$.

Proof: $D(P \| \mathrm{Unif}(\mathcal{X})) = \log |\mathcal{X}| - H(X) \ge 0.$

Consequence 6: DPI for Mutual Information

- We say X, Y, Z forms a Markov chain (denoted by $X \leftrightarrow Y \leftrightarrow Z$) if X, Z are conditionally independent given Y.

Consequence 6: DPI for Mutual Information

- We say X, Y, Z forms a Markov chain (denoted by $X \leftrightarrow Y \leftrightarrow Z$) if X, Z are conditionally independent given Y.

Theorem (Data-processing inequality for mutual information)

If $X \leftrightarrow Y \leftrightarrow Z$, then $I(X;Y) \geq I(X;Z)$ and $I(X;Y) \geq I(X;Y|Z)$. In particular, if $Z = g(Y)$, we have $I(X;Y) \geq I(X;g(Y))$ and $I(X;Y) \geq I(X;Y|g(Y))$.

Consequence 6: DPI for Mutual Information

- We say X, Y, Z forms a Markov chain (denoted by $X \leftrightarrow Y \leftrightarrow Z$) if X, Z are conditionally independent given Y.

Theorem (Data-processing inequality for mutual information)

If $X \leftrightarrow Y \leftrightarrow Z$, then $I(X;Y) \geq I(X;Z)$ and $I(X;Y) \geq I(X;Y|Z)$. In particular, if $Z = g(Y)$, we have $I(X;Y) \geq I(X;g(Y))$ and $I(X;Y) \geq I(X;Y|g(Y))$.

Proof:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - H(X|Y,Z) \qquad \text{by conditional independence } P_{X|YZ} = P_{X|Y} \\ &= I(X;Y,Z) \\ &= I(X;Z) + I(X;Y|Z) \end{aligned}$$

By nonnegativity of mutual information, we obtain the desired results.

Summary: Properties of D

$$D(P\|\,\mathrm{Count}(\mathcal{X})) = -H(X).$$
$$D(P\|\,\mathrm{Unif}(\mathcal{X})) = \log|\mathcal{X}| - H(X)$$
$$D(P_{XY}\|P_X\otimes P_Y) = I(X;Y)$$
$$D(P_{X^n}\|Q_{X^n})= \sum_{i=1}^{n} D(P_{X_i|X^{i-1}}\|Q_{X_i|X^{i-1}}|P_{X^{i-1}})$$
$$D(P\|Q)\geq 0 \text{ with equality iff } P = Q$$
$$D(P_{XY}\|Q_{XY}) \geq \max\{D(P_Y\|Q_Y)\, D(P_X\|Q_X)\}$$
$$D(P_X\|Q_X) \geq D(P_Y\|Q_Y), \text{ if } P_X \to P_{Y|X} \to P_Y \text{ and } Q_X \to P_{Y|X} \to Q_Y$$
$$D(P_{Y|X}\|Q_{Y|X}|P_X) \geq D(P_Y\|Q_Y), \text{ if } P_X \to P_{Y|X} \to P_Y \text{ and } P_X \to Q_{Y|X} \to Q_Y$$
$$D(P_{X^n}\|Q_X^{\otimes n}) \geq \sum_{i=1}^{n} D(P_{X_i}\|Q_X)$$
$$(P,Q) \mapsto D(P\|Q) \text{ is convex.}$$

Summary: Properties of H

$$
\begin{aligned}
H(X) &\geq 0 \\
H(X|Y) &\leq H(X) \text{ with equality iff } X \perp Y \\
H(g(X)|X) &= 0 \\
H(X,Y) &= H(X) + H(Y|X) \\
H(X^n) &= \sum_{i=1}^{n} H(X_i|X^{i-1}) \\
P_X \mapsto H(X) &\text{ and } P_{XY} \mapsto H(X|Y) \text{ are concave.}
\end{aligned}
$$

Summary: Properties of I

$$
\begin{aligned}
I(X;Y) &\geq 0 \text{ with equality iff } X \perp Y \\
I(X;Y) &= H(X) - H(X|Y) \\
I(X;Y) &= H(Y) - H(Y|X) \\
I(X;Y) &= H(X) + H(Y) - H(X,Y) \\
I(X;Y) &= I(Y;X) \\
I(X;g(X)) &= H(g(X)) \\
I(X^n;Y) &= \sum_{i=1}^{n} I(X_i;Y|X^{i-1}), \\
I(X;Y) &\geq \max\{I(X;Z),\, I(X;Y|Z)\} \text{ if } X \leftrightarrow Y \leftrightarrow Z \\
P_X &\mapsto I(X;Y) \text{ is concave} \\
P_{Y|X} &\mapsto I(X;Y) \text{ is convex.}
\end{aligned}
$$

Outline

Relative Entropy in Abstract Spaces

- We now consider arbitrary measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$.
- We say P is absolutely continuous w.r.t. μ, written as $P \ll \mu$, if for any measurable A, $P(A) = 0$ whenever $\mu(A) = 0$.

Relative Entropy in Abstract Spaces

- We now consider arbitrary measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$.
- We say P is absolutely continuous w.r.t. μ, written as $P \ll \mu$, if for any measurable A, $P(A) = 0$ whenever $\mu(A) = 0$.

Definition

For a probability measure P and a nonnegative (σ-finite) measure μ on $(\mathcal{X}, \Sigma_{\mathcal{X}})$ such that $P \ll \mu$, the relative entropy of P w.r.t. μ is

$$D(P\|\mu) := \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right) \mathrm{d}P$$

where $\frac{\mathrm{d}P}{\mathrm{d}\mu}$ is the Radon–Nikodym derivative.

Relative Entropy in Abstract Spaces

- We now consider arbitrary measurable space $(\mathcal{X}, \Sigma_{\mathcal{X}})$.
- We say P is absolutely continuous w.r.t. μ, written as $P \ll \mu$, if for any measurable A, $P(A) = 0$ whenever $\mu(A) = 0$.

Definition

For a probability measure P and a nonnegative (σ-finite) measure μ on $(\mathcal{X}, \Sigma_{\mathcal{X}})$ such that $P \ll \mu$, the relative entropy of P w.r.t. μ is

$$D(P\|\mu) := \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right) \mathrm{d}P$$

where $\frac{\mathrm{d}P}{\mathrm{d}\mu}$ is the Radon–Nikodym derivative.

- If μ is the Lebesgue measure, then $h(X) := -D(P\|\mu) = -\int q(x) \log q(x) \mathrm{d}x$ is the differential entropy of $X \sim P$.

Relative Entropy in Abstract Spaces

- We now consider arbitrary measurable space $(\mathcal{X}, \Sigma_\mathcal{X})$.
- We say P is absolutely continuous w.r.t. μ, written as $P \ll \mu$, if for any measurable A, $P(A) = 0$ whenever $\mu(A) = 0$.

Definition

For a probability measure P and a nonnegative (σ-finite) measure μ on $(\mathcal{X}, \Sigma_\mathcal{X})$ such that $P \ll \mu$, the relative entropy of P w.r.t. μ is

$$D(P\|\mu) := \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}\mu}\right)\mathrm{d}P$$

where $\frac{\mathrm{d}P}{\mathrm{d}\mu}$ is the Radon–Nikodym derivative.

- If μ is the Lebesgue measure, then $h(X) := -D(P\|\mu) = -\int q(x)\log q(x)\mathrm{d}x$ is the differential entropy of $X \sim P$.
- We define the mutual information via $I(X;Y) := D(P_{XY}\|P_X \otimes P_Y)$

Alternative Definition via Discretization

- All statements (except the ones involving entropies) given on previous slides still hold for the measurable spaces satisfying certain regularity conditions.

Alternative Definition via Discretization

- All statements (except the ones involving entropies) given on previous slides still hold for the measurable spaces satisfying certain regularity conditions.
- In particular, the DPI still holds, i.e., for any measurable function f,

$$D(P_X \| Q_X) \geq D(P_{f(X)} \| Q_{f(X)}).$$

Alternative Definition via Discretization

- All statements (except the ones involving entropies) given on previous slides still hold for the measurable spaces satisfying certain regularity conditions.
- In particular, the DPI still holds, i.e., for any measurable function f,

$$D(P_X \| Q_X) \geq D(P_{f(X)} \| Q_{f(X)}).$$

Theorem (Alternative Definition via Discretization)

It holds that

$$D(P_X \| Q_X) = \sup_f D(P_{f(X)} \| Q_{f(X)}),$$

where the supremum is over all functions f that take only finitely many values.

Alternative Definition via Discretization

- All statements (except the ones involving entropies) given on previous slides still hold for the measurable spaces satisfying certain regularity conditions.
- In particular, the DPI still holds, i.e., for any measurable function f,

$$D(P_X \| Q_X) \geq D(P_{f(X)} \| Q_{f(X)}).$$

Theorem (Alternative Definition via Discretization)

It holds that

$$D(P_X \| Q_X) = \sup_f D(P_{f(X)} \| Q_{f(X)}),$$

where the supremum is over all functions f that take only finitely many values.

Proof of "$\leq$" part is based on discretization of $\frac{\mathrm{d}P_X}{\mathrm{d}Q_X}$ [Van Erven–Harremos, 2014]

References

1. Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3), 379-423.
2. Cover, T. M. (1999). Elements of information theory. John Wiley & Sons.
3. Van Erven, T., and Harremos, P. (2014). Rényi divergence and Kullback-Leibler divergence. IEEE Transactions on Information Theory, 60(7), 3797-3820.
4. Polyanskiy, Y., and Wu, Y. (2022+). Information theory: from coding to learning. Cambridge University Press (book draft)
5. Yu, L. The entropy method. In Preparation.

Thank you for your attention!