Lecture 3: Information Inequalities

Tianpei Xie

Jan. 6th., 2023

Contents

1 Information Theory Basics

1.1 Entropy, Relative Entropy, and Mutual Information

- **Definition** (***Shannon Entropy***) [Cover and Thomas, 2006]
  Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $X : \mathbb{R} \to \mathcal{X}$ be a random variable. Define $p(x)$ as *the probability density function* of X with respect to a base measure μ on $\mathcal{X}$. ***The Shannon Entropy*** is defined as

$$
\begin{aligned}
H(X) &:= \mathbb{E}_p\left[-\log p(X)\right] \\
&= \int_{\Omega} -\log p(X(\omega)) d\mathbb{P}(\omega) \\
&= -\int_{\mathcal{X}} p(x) \log p(x) d\mu(x)
\end{aligned}
$$

- **Definition** (***Conditional Entropy***) [Cover and Thomas, 2006]
  If a pair of random variables (X, Y) follows the joint probability density function $p(x, y)$ with respect to a base product measure μ on $\mathcal{X} \times \mathcal{Y}$. Then ***the joint entropy*** of (X, Y), denoted as $H(X, Y)$, is defined as

$$
H(X, Y) := \mathbb{E}_{X,Y}\left[-\log p(X, Y)\right] = -\int_{\mathcal{X}\times\mathcal{Y}} p(x, y) \log p(x, y) d\mu(x, y)
$$

  Then ***the conditional entropy*** $H(Y|X)$ is defined as

$$
\begin{aligned}
H(Y|X) &:= \mathbb{E}_{X,Y}\left[-\log p(Y|X)\right] = -\int_{\mathcal{X}\times\mathcal{Y}} p(x, y) \log p(y|x) d\mu(x, y) \\
&= \mathbb{E}_X\left[\mathbb{E}_Y\left[-\log p(Y|X)\right]\right] = \int_{\mathcal{X}} p(x) \left(-\int_{\mathcal{Y}} p(y|x) \log p(y|x) d\mu(y)\right) d\mu(x)
\end{aligned}
$$

- **Proposition 1.1** *(**Properties of Shannon Entropy**) [Cover and Thomas, 2006]*
  Let X, Y, Z be random variables.

  1. *(**Non-negativity**) $H(X) \geq 0$;*
  2. *(**Chain Rule**)*

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

     Furthermore,

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

  3. *(**Concavity**) $H(p) := \mathbb{E}_p\left[-\log p(X)\right]$ is a concave function in terms of p.d.f. p, i.e.*

$$
H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda) H(p_2)
$$

     for any two p.d.fs p_1, p_2 on $\mathcal{X}$ and any $\lambda \in [0, 1]$.

- **Definition** (***Relative Entropy / Kullback-Leibler Divergence***) [Cover and Thomas, 2006]
  Suppose that P and Q are *probability measures* on a measurable space $\mathcal{X}$, and P is *absolutely continuous* with respect to Q, then ***the relative entropy*** or ***the Kullback-Leibler divergence*** is defined as

$$\mathbb{KL}\left(P \,\|\, Q\right) := \mathbb{E}_P\left[\log\left(\frac{dP}{dQ}\right)\right] = \int_{\mathcal{X}} \log\left(\frac{dP(x)}{dQ(x)}\right) dP(x)$$

  where $\frac{dP}{dQ}$ is *the Radon-Nikodym derivative* of P with respect to Q. Equivalently, the KL-divergence can be written as

$$\mathbb{KL}\left(P \,\|\, Q\right) = \int_{\mathcal{X}} \left(\frac{dP(x)}{dQ(x)}\right) \log\left(\frac{dP(x)}{dQ(x)}\right) dQ(x)$$

  which is *the entropy of P relative to Q*. Furthermore, if μ is a base measure on $\mathcal{X}$ for which densities p and q with $dP = p(x)d\mu$ and $dQ = q(x)d\mu$ exist, then

$$\mathbb{KL}\left(P \,\|\, Q\right) = \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) d\mu(x)$$

- **Definition** (***Mutual Information***) [Cover and Thomas, 2006]
  Consider two random variables X, Y on $\mathcal{X} \times \mathcal{Y}$ with joint probability distribution $P_{(X,Y)}$ and marginal distribution P_X and P_Y. ***The mutual information*** $I(X;Y)$ is *the relative entropy* between *the joint distribution* $P_{(X,Y)}$ and *the product distribution* $P_X \otimes P_Y$:

$$I(X;Y) = \mathbb{KL}\left(P_{(X,Y)} \,\|\, P_X \otimes P_Y\right) = \mathbb{E}_{P_{(X,Y)}}\left[\log \frac{dP_{(X,Y)}}{dP_X \otimes dP_Y}\right]$$

  If $P_{(X,Y)}$ has a probability density function $p(x,y)$ with respect to a base measure μ on $\mathcal{X} \times \mathcal{Y}$, then

$$I(X;Y) = \int_{\mathcal{X}\times\mathcal{Y}} p(x,y) \log\left(\frac{p(x,y)}{p_X(x)p_Y(y)}\right) d\mu(x,y)$$

- **Proposition 1.2** *(**Properties of Relative Entropy and Mutual Information**) [Cover and Thomas, 2006]*
  Let X, Y be random variables.

  1. *(**Non-negativity**) Let $p(x), q(x)$ be probability density function of P, Q.*

$$\mathbb{KL}\left(P \,\|\, Q\right) \geq 0$$

     with equality if and only if $p(x) = q(x)$ almost surely. Therefore, the mutual information is non-negative as well:

$$I(X;Y) \geq 0$$

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

  2. *(**Symmetry**) $I(X;Y) = I(Y;X)$*

3. ***(Information Gain via Conditioning)*** *The mutual information $I(X;Y)$ is the reduction in the uncertainty of X due to the knowledge of Y (and vice versa)*

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X) + H(Y) - H(X,Y)
\end{aligned} \tag{1}
$$

4. ***(Shannon Entropy as Self-Information)*** *$I(X;X) = H(X)$*

1.2 Chain Rules for Entropy, Relative Entropy, and Mutual Information

- **Proposition 1.3** ***(Conditioning Reduces Entropy)*** *[Cover and Thomas, 2006]*
  From non-negativity of mutual information, we see that the entropy of X is non-increasing when conditioning on Y

  $$H(X|Y) \le H(X) \tag{2}$$

  where equality holds if and only if X and Y are independent.

- **Proposition 1.4** ***(Chain Rule for Entropy)*** *[Cover and Thomas, 2006]*
  Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then

  $$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) \tag{3}$$

- **Proposition 1.5** ***(Independence Bound on Entropy)*** *[Cover and Thomas, 2006]*
  Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then

  $$H(X_1, X_2, \ldots, X_n) \le \sum_{i=1}^{n} H(X_i) \tag{4}$$

  with equality if and only if the X_i are independent.

- **Proposition 1.6** ***(Chain Rule for Mutual Information)*** *[Cover and Thomas, 2006]*
  Let $X_1, X_2, \ldots, X_n, Y$ be drawn according to $p(x_1, x_2, \ldots, x_n, y)$. Then

  $$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} H(X_i; Y | X_{i-1}, \ldots, X_1) \tag{5}$$

  *where **the conditional mutual information** is defined as*

  $$I(X;Y|Z) := H(X|Z) - H(X|Y,Z) = \mathbb{KL}\left(P_{(X,Y|Z)} \,\|\, P_{X|Z} \otimes P_{Y|Z}\right)$$

- **Proposition 1.7** ***(Chain Rule for Relative Entropy)*** *[Cover and Thomas, 2006]*
  *Let $P_{(X,Y)}$ and $Q_{(X,Y)}$ be two probability measures on product space $\mathcal{X} \times \mathcal{Y}$ and $P \ll Q$. Denote the marginal distributions P_X, Q_X and P_Y, Q_Y on $\mathcal{X}$ and $\mathcal{Y}$, respectively. $P_{Y|X}$ and $Q_{Y|X}$ are conditional distributions (Note that $P_{Y|X} \ll Q_{Y|X}$). Define **the conditional relative entropy** as*

  $$\mathbb{KL}\left(P_{Y|X} \,\|\, Q_{Y|X}\right) := \mathbb{E}_{P_{(X,Y)}}\left[\log\left(\frac{dP_{Y|X}}{dQ_{Y|X}}\right)\right].$$

  Then the relative entropy of joint distribution $P_{(X,Y)}$ with respect to $Q_{(X,Y)}$ is

  $$\mathbb{KL}\left(P_{(X,Y)} \,\|\, Q_{(X,Y)}\right) = \mathbb{KL}\left(P_X \,\|\, Q_X\right) + \mathbb{KL}\left(P_{Y|X} \,\|\, Q_{Y|X}\right) \tag{6}$$

1.3 Log-Sum Inequalities and Convexity

- **Proposition 1.8** *(**Log-Sum Inequalities**) [Cover and Thomas, 2006]*
  For non-negative numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{7}$$

  with equality if and only if $\frac{a_i}{b_i}$ is constant.

- **Proposition 1.9** *(**Joint Convexity of Relative Entropy**) [Cover and Thomas, 2006]*
  *$\mathbb{KL}\left(p \,\|\, q\right)$ is **convex** in the pair (p, q); that is, if (p_1, q_1) and (p_2, q_2) are two pairs of probability density functions, then for $\lambda \in [0, 1]$,*

$$\mathbb{KL}\left(\lambda p_1 + (1-\lambda)p_2 \,\|\, \lambda q_1 + (1-\lambda)q_2\right) \leq \lambda \mathbb{KL}\left(p_1 \,\|\, q_1\right) + (1-\lambda)\mathbb{KL}\left(p_2 \,\|\, q_2\right) \tag{8}$$

- **Proposition 1.10** *[Cover and Thomas, 2006]*
  *Let $(X, Y) \sim p(x, y) = p(x)p(y|x)$. The mutual information $I(X;Y)$ is a **concave** function of $p(x)$ for fixed $p(y|x)$ and a **convex** function of $p(y|x)$ for fixed $p(x)$.*

1.4 Data Processing Inequality

- **Definition** (*Data Processing Markov Chain*)
  Random variables X, Y, Z are said to ***form a Markov chain*** in that order (denoted by $X \to Y \to Z$) if the conditional distribution of Z depends only on Y and is ***conditionally independent*** of X. Specifically, X, Y, and Z form a Markov chain $X \to Y \to Z$ if the joint probability mass function can be written as

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

- **Proposition 1.11** *(**Data Processing Inequality**) [Cover and Thomas, 2006]*
  If $X \to Y \to Z$, then

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

- **Corollary 1.12** *[Cover and Thomas, 2006]*
  In particular, if $Z = g(Y)$, we have

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

- **Corollary 1.13** *[Cover and Thomas, 2006]*
  If $X \to Y \to Z$, then

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

  *Thus, the dependence of X and Y is **decreased** (or remains unchanged) by the observation of a "**downstream**" random variable Z.*

1.5 Combinatorial Entropies

2 Information Inequalities

2.1 Han's Inequality

2.2 Sub-Additivity of Entropy and Relative Entropy

2.3 Duality and Variational Formulas

2.4 Optimal Transport

2.5 Pinsker's Inequality

2.6 Birgé's Inequality

2.7 The Brunn-Minkowski Inequality

References

Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9.