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. *(**Sub-Additivity**)*

$$
H(X,Y) \leq H(X) + H(Y)
$$

  4. *(**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

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

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

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

  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

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

- **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* $P_X \otimes P_Y$:

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

  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.*

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

     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. ***(Finite Cardinality Domain)*** *Let $|\mathcal{X}|$ be the number of elements in domain $\mathcal{X}$ and X is a discrete random variables in $\mathcal{X}$. Then the relative entropy of probability distribution p with respect to uniform distribution u on $\mathcal{X}$ is*

$$\begin{aligned} \mathbb{KL}\left(p \,\|\, u\right) &= \log|\mathcal{X}| - H(X) \geq 0 \\ \Rightarrow H(X) &\leq \log|\mathcal{X}| \end{aligned}$$

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

4. ***(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)*

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

5. ***(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

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

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** ***(Sub-Additivity of 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

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

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*

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

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

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

  In addition, let P and Q denote two joint distributions for $X_1, X_2, \ldots, X_n$, let $P_{1:i}$ and $Q_{1:i}$ denote the marginal distributions of $X_1, X_2, \ldots, X_i$ under P and Q, respectively. Let $P_{X_i|1\ldots i-1}$ and $Q_{X_i|1\ldots i-1}$ denote the conditional distribution of X_i with respect to $X_1, X_2, \ldots, X_{i-1}$ under P and under Q.

$$\mathbb{KL}\left(P \,\|\, Q\right) = \sum_{i=1}^{n} \mathbb{E}_{P_{1:i-1}}\left[\mathbb{KL}\left(P_{X_i|1\ldots i-1} \,\|\, Q_{X_i|1\ldots i-1}\right)\right] \tag{7}$$

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 $b_1, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log\frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right)\log\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{8}$$

  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 $\lambda \in [0, 1]$,*

$$\mathbb{KL}\left(\lambda p_1 + (1-\lambda)p_2 \,\|\, \lambda q_1 + (1-\lambda)q_2\right) \leq \lambda\mathbb{KL}\left(p_1 \,\|\, q_1\right) + (1-\lambda)\mathbb{KL}\left(p_2 \,\|\, q_2\right) \tag{9}$$

- **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.*

2 Information Inequalities

2.1 Han's Inequality

- **Proposition 2.1** *(**Han's Inequality**) [Cover and Thomas, 2006, Boucheron et al., 2013]*
  Let $X_1, X_2, \ldots, X_n$ be random variables. Then

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &\leq \frac{1}{n-1} \sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \qquad (10) \\ \Leftrightarrow H(X) &\leq \frac{1}{n-1} \sum_{i=1}^{n} H(X_{(-i)}) \end{aligned}$$

  Proof: For any $i = 1, \ldots, n$, by the definition of the conditional entropy and the fact that conditioning reduces entropy,

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_i | X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\ &\leq H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_i | X_1, \ldots, X_{i-1}). \end{aligned}$$

  Summing these n inequalities and using the chain rule for entropy, we get

$$\begin{aligned} nH(X_1, X_2, \ldots, X_n) &\leq \sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + \sum_{i=1}^{n} H(X_i | X_1, \ldots, X_{i-1}) \\ &= \sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) + H(X_1, X_2, \ldots, X_n) \end{aligned}$$

  which is what we wanted to prove. ∎

- **Proposition 2.2** *(**Han's Inequality for Relative Entropy**) [Boucheron et al., 2013]*
  *Let $(\mathcal{X}, \mathscr{B})$ be a measurable space, and P and Q be probability measures on $\mathcal{X}^n$ such that $P = P_1 \otimes \ldots \otimes P_n$ is a **product measure**. We denote the element of $\mathcal{X}^n$ by $x = (x_1, \ldots, x_n)$*

*and write $x_{(-i)} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ for the $(n-1)$-vector obtained by **leaving out the i-th component of** x (i.e. the i-th Jackknife sample of x). Denote $Q_{(-i)}$ and $P_{(-i)}$ the marginal distributions of Q and P. Let $p_{(-i)}$ and $q_{(-i)}$ denote the corresponding probability density function with respect to base measure μ on $\mathcal{X}$.*

$$
\begin{aligned}
q_{(-i)}(x_{(-i)}) &= \int_{y \in \mathcal{X}} q(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n) d\mu(y) \\
p_{(-i)}(x_{(-i)}) &= \int_{y \in \mathcal{X}} p(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n) d\mu(y) \\
&= \prod_{j \neq i} p_j(x_j).
\end{aligned}
$$

Then

$$
\mathbb{KL}\left(Q \parallel P\right) \geq \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{KL}\left(Q_{(-i)} \parallel P_{(-i)}\right) \tag{11}
$$

or equivalently,

$$
\mathbb{KL}\left(Q \parallel P\right) \leq \sum_{i=1}^{n} \left(\mathbb{KL}\left(Q \parallel P\right) - \mathbb{KL}\left(Q_{(-i)} \parallel P_{(-i)}\right)\right) \tag{12}
$$

Proof: From Han's inequality, we have

$$
-H(Q) \geq -\frac{1}{n-1} \sum_{i=1}^{n} H(Q_{(-i)}).
$$

Since

$$
\mathbb{KL}\left(Q \parallel P\right) = -H(Q) + \mathbb{E}_Q\left[-\log P(X)\right]
$$

and

$$
\mathbb{KL}\left(Q_{(-i)} \parallel P_{(-i)}\right) = -H(Q_{(-i)}) + \mathbb{E}_{Q_{(-i)}}\left[-\log P_{(-i)}(X_{(-i)})\right],
$$

it suffices to show that

$$
\mathbb{E}_Q\left[-\log P(X)\right] = \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}_{Q_{(-i)}}\left[-\log P_{(-i)}(X_{(-i)})\right].
$$

This may be seen easily by noting that by the product property of P, we have $p(x) = p_{(-i)}(x_{(-i)}) p_i(x_i)$ for all i, and also $p(x) = \prod_i p_i(x_i)$, and therefore

$$
\begin{aligned}
\mathbb{E}_Q\left[-\log P(X)\right] &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_Q\left[-\log P_{(-i)}(X_{(-i)}) - \log P_i(X_i)\right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_Q\left[-\log P_{(-i)}(X_{(-i)})\right] + \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_Q\left[-\log P_i(X_i)\right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_Q\left[-\log P_{(-i)}(X_{(-i)})\right] + \frac{1}{n} \mathbb{E}_Q\left[-\log P(X)\right].
\end{aligned}
$$

Rearranging, we obtain

$$\begin{aligned}\mathbb{E}_Q\left[-\log P(X)\right] &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}_Q\left[-\log P_{(-i)}(X_{(-i)})\right] \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}_{Q_{(-i)}}\left[-\log P_{(-i)}(X_{(-i)})\right]. \quad \blacksquare\end{aligned}$$

2.2 Applications of Han's Inequality

2.2.1 Combinatorial Entropies

2.2.2 Edge Isoperimetric Inequality on the Binary Hypercube

2.3 Φ-Entropy

- **Definition** (***Φ-Entropy***)[Boucheron et al., 2013]
  Let $\Phi : [0,\infty) \to \mathbb{R}$ be a ***convex*** function, and assign, to every ***non-negative*** *integrable random variable* X, ***the Φ-entropy*** of X is defined as

$$H_{\Phi}(X) = \mathbb{E}\left[\Phi(X)\right] - \Phi(\mathbb{E}\left[X\right]). \tag{13}$$

- **Remark** By Jenson's inequality, the Φ-entropy is *non-negative*

$$\begin{aligned}&\Phi(\mathbb{E}\left[X\right]) \le \mathbb{E}\left[\Phi(X)\right] \\ \Rightarrow\ &H_{\Phi}(X) = \mathbb{E}\left[\Phi(X)\right] - \Phi(\mathbb{E}\left[X\right]) \ge 0.\end{aligned}$$

- **Example** (***Special Examples for Φ-Entropy***)
  1. For $\Phi(x) = x^2$, *the Φ-entropy* of X is *the* ***variance*** of X:

$$H_{\Phi}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}\left[X\right])^2 = \mathrm{Var}(X).$$

  2. For $\Phi(x) = x\log x$, *the Φ-entropy* of X is defined as the ***entropy*** of X

$$\mathrm{Ent}(X) := \mathbb{E}\left[X\log X\right] - \mathbb{E}\left[X\right]\log\left(\mathbb{E}\left[X\right]\right). \tag{14}$$

  Let $(\Omega, \mathscr{B})$ be measurable space, and P and Q are probability measures on Ω with $P \ll Q$. Define a random variable X by the *Radon-Nikodym derivative* of P with respect to Q; that is,

$$X(\omega) := \begin{cases} \frac{dP}{dQ}(\omega) & Q(\omega) > 0 \\ 0 & \text{o.w.} \end{cases}.$$

  We see that X is Q-measurable and $dP = X\,dQ$ with $\mathbb{E}_Q\left[X\right] = 1$. Then the entropy of X is the relative entropy of P with respect to Q.

$$\mathrm{Ent}(X) = \mathbb{KL}\left(P \,\|\, Q\right) \tag{15}$$

2.4 Sub-Additivity of Φ-Entropy

- **Remark** (***Sub-Additivity of Shannon Entropy***)
  Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then
  $$H(X_1, X_2, \ldots, X_n) \le \sum_{i=1}^{n} H(X_i)$$
  with equality if and only if the X_i are independent.

- **Proposition 2.3** *(**Sub-Additivity of The Entropy**) [Boucheron et al., 2013]*
  Let $\Phi(x) = x \log x$, for $x > 0$ and $\Phi(0) = 0$. Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables taking values in $\mathcal{X}$, and let $f : \mathcal{X}^n \to [0, \infty)$. Letting $X = f(Z_1, Z_2, \ldots, Z_n)$, we have
  $$\mathbb{E}\left[\Phi(X)\right] - \Phi(\mathbb{E}\left[X\right]) \le \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}_{(-i)}\left[\Phi(X)\right] - \Phi(\mathbb{E}_{(-i)}\left[X\right])\right], \tag{16}$$
  where $\mathbb{E}_{(-i)}[\cdot]$ is the conditional expectation operator conditioning on $Z_{(-i)}$. Introducing the notation $Ent_{(-i)}(X) = \mathbb{E}_{(-i)}\left[\Phi(X)\right] - \Phi(\mathbb{E}_{(-i)}\left[X\right])$, this can be re-written as
  $$\mathbb{E}\left[\Phi(X)\right] - \Phi(\mathbb{E}\left[X\right]) \le \mathbb{E}\left[\sum_{i=1}^{n} Ent_{(-i)}(X)\right]. \tag{17}$$
  Proof: The proposition is a direct consequence of Han's inequality for relative entropies. First note that if the inequality is true for a random variable X, then it is also true for cX where c is a positive constant. Hence, we may assume that $\mathbb{E}\left[X\right] = 1$. Now define the probability measure P on $\mathcal{X}^n$ by its probability density function p given by
  $$p(z) = f(z)q(z), \quad \forall z \in \mathcal{X}^n$$
  where q denote the probability density of $Z := (Z_1, Z_2, \ldots, Z_n)$ and Q the corresponding probability measure. Then
  $$\mathrm{Ent}(X) := \mathbb{E}\left[X \log X\right] - \mathbb{E}\left[X\right] \log\left(\mathbb{E}\left[X\right]\right) = \mathbb{KL}\left(P \,\|\, Q\right)$$
  which, by Han's inequality for relative entropy
  $$\mathrm{Ent}(X) = \mathbb{KL}\left(P \,\|\, Q\right) \le \sum_{i=1}^{n} \left(\mathbb{KL}\left(P \,\|\, Q\right) - \mathbb{KL}\left(P_{(-i)} \,\|\, Q_{(-i)}\right)\right)$$
  However, straightforward calculation shows that
  $$\sum_{i=1}^{n} \left(\mathbb{KL}\left(P \,\|\, Q\right) - \mathbb{KL}\left(P_{(-i)} \,\|\, Q_{(-i)}\right)\right) = \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}_{(-i)}\left[\Phi(X)\right] - \Phi(\mathbb{E}_{(-i)}\left[X\right])\right]$$
  and the statement follows. $\blacksquare$

- **Remark** The Efron-Stein inequality is the special case of the inequality when $\Phi(x) = x^2$,
  $$\mathbb{E}\left[\Phi(X)\right] - \Phi(\mathbb{E}\left[X\right]) \le \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}_{(-i)}\left[\Phi(X)\right] - \Phi(\mathbb{E}_{(-i)}\left[X\right])\right].$$
  $$\Rightarrow \mathrm{Var}(X) \le \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Var}_{(-i)}(X)\right]$$

- **Proposition 2.4** ***(Sub-Additivity of Φ-Entropy)*** *[Boucheron et al., 2013]*
*Let $\mathcal{C}$ denote the class of functions $\Phi : [0, \infty) \to \mathbb{R}$ that are **continuous** and **convex** on $[0, \infty)$, **twice differentiable** on $(0, \infty)$, and such that either Φ is **affine** or Φ'' is **strictly positive** and $1/\Phi''$ is **concave**. For all $\Phi \in \mathcal{C}$, the **entropy functional** H_Φ is **sub-additive**. That is,*

$$\mathbb{E}\left[\Phi(X)\right] - \Phi(\mathbb{E}\left[X\right]) \le \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}_{(-i)}\left[\Phi(X)\right] - \Phi(\mathbb{E}_{(-i)}\left[X\right])\right], \tag{18}$$
$$\Leftrightarrow H_\Phi(X) \le \mathbb{E}\left[\sum_{i=1}^{n} H_\Phi^{(-i)}(X)\right]$$

where $H_\Phi^{(-i)}(X) := \mathbb{E}_{(-i)}\left[\Phi(X)\right] - \Phi(\mathbb{E}_{(-i)}\left[X\right])$ is the conditional entropy and, $\mathbb{E}_{(-i)}\left[\cdot\right]$ denotes conditional expectation conditioned on the $(n-1)$-vector $Z_{(-i)} := (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$.

- **Remark** ***The sub-additivity property*** of H_Φ is equivalent to what we could call ***the Jensen property***

$$H_\Phi\left(\int f(z, Z_2) d\mu_1(z)\right) \le \int H_\Phi(f(z, Z_2)) d\mu_1(z)$$
$$\Leftrightarrow H_\Phi\left(\mathbb{E}_{Z_1}\left[f(Z_1, Z_2)\right]\right) \le \mathbb{E}_{Z_1}\left[H_\Phi\left(f(Z_1, Z_2)\right)\right] \tag{19}$$

This implies that in order to prove sub-additivity of a Φ-entropy, it suffices to show that it has the Jensen property.

2.5 Duality and Variational Formulas

- **Lemma 2.5** *The **Legendre transform** (or **convex conjugate**) of $\Phi(x) = x\log(x)$ is e^{u-1}. That is,*

$$\sup_{x>0}\left\{u\,x - x\log(x)\right\} = e^{u-1}$$

Proof: Solve the supremum on the left-hand side by taking derivative of the objective function and setting it as zero:

$$\nabla g(x) = u - \log(x) - 1 = 0$$
$$\Rightarrow x^* = e^{u-1}$$
$$\Rightarrow \sup_{x}\left\{u\,x - x\log(x)\right\} = g(x^*) = u\,e^{u-1} - e^{u-1}(u-1) = e^{u-1} \qquad \blacksquare$$

- **Remark** If $\Phi(X) = X\log(X)$ is integrable, and $\mathbb{E}\left[e^U\right] = 1$, we have

$$UX \le X\log(X) + \frac{1}{e}e^U.$$

Therefore, U_+X is integrable, and one can always define $\mathbb{E}\left[UX\right] = \mathbb{E}\left[U_+X\right] - \mathbb{E}\left[U_-X\right]$ for positive and negative part of U. Thus the $\mathbb{E}\left[UX\right]$ is well-defined.

- **Theorem 2.6** ***(Duality Formula of Entropy)*** *[Boucheron et al., 2013]*
*Let X be a non-negative random variable defined on a probability space $(\Omega, \mathscr{A}, P)$ such that $\mathbb{E}\left[\Phi(X)\right] < \infty$. Then we have **the duality formula***

$$\mathit{Ent}(X) = \sup_{U \in \mathcal{U}} \mathbb{E}\left[U\,X\right] \tag{20}$$

where the supremum is taken over the set $\mathcal{U}$ *of all random variables* $U : \Omega \to \mathbb{R} \cup \{\infty\}$ *with* $\mathbb{E}\left[e^U\right] = 1$*. Moreover, if* U *is such that* $\mathbb{E}\left[UX\right] \leq$ *Ent*(X) *for all non-negative random variable* X *such that* $\Phi(X)$ *is integrable and* $\mathbb{E}\left[X\right] = 1$*, then* $\mathbb{E}\left[e^U\right] \leq 1$*.*

Proof: Note that for any random variable U such that $\mathbb{E}\left[e^U\right] = 1$, we have

$$
\begin{aligned}
\mathrm{Ent}(X) - \mathbb{E}_P\left[UX\right] &= \mathbb{E}_P\left[X\log(X)\right] - \mathbb{E}_P\left[X\right]\log(\mathbb{E}_P\left[X\right]) - \mathbb{E}_P\left[UX\right] \\
&= \mathbb{E}_P\left[X(\log(X) - U)\right] - \mathbb{E}_P\left[X\right]\log(\mathbb{E}_P\left[X\right]) \\
&= \mathbb{E}_P\left[X\log(Xe^{-U})\right] - \mathbb{E}_P\left[X\right]\log(\mathbb{E}_P\left[X\right]) \\
&= \mathbb{E}_{e^U P}\left[Xe^{-U}\log(Xe^{-U})\right] - \mathbb{E}_{e^U P}\left[Xe^{-U}\right]\log(\mathbb{E}_{e^U P}\left[Xe^{-U}\right]) \\
&= \mathrm{Ent}_{e^U P}(Xe^{-U})
\end{aligned}
$$

Note that due to $\mathbb{E}\left[e^U\right] = 1$, $\int e^U dP = 1$, thus $e^U P$ is a proper probability measure. This shows that

$$
\begin{aligned}
&\mathrm{Ent}_{e^U P}(Xe^{-U}) \geq 0 \\
&\quad \Rightarrow \mathrm{Ent}(X) \geq \mathbb{E}_P\left[UX\right]
\end{aligned}
$$

with equality whenever $e^U = X/\mathbb{E}_P\left[X\right]$. This proves the duality formula.

Conversely, let U be such that $\mathbb{E}_P\left[UX\right] \leq \mathrm{Ent}(X)$ for all non-negative random variables such that $\Phi(X)$ is integrable. If $\mathbb{E}\left[e^U\right] = 0$, then there is nothing to prove Otherwise, given a positive integer n large enough to ensure that $x_n = \mathbb{E}\left[e^{\min\{U,n\}}\right] > 0$, one may define $X_n = e^{\min\{U,n\}}/x_n$, so that $\mathbb{E}\left[X_n\right] = 1$, which leads to

$$
\mathbb{E}\left[UX_n\right] \leq \mathrm{Ent}(X_n),
$$

and therefore

$$
\begin{aligned}
\frac{1}{x_n}\mathbb{E}\left[Ue^{\min\{U,n\}}\right] &\leq \mathrm{Ent}(e^{\min\{U,n\}}/x_n) \\
&= \frac{1}{x_n}\left[\mathbb{E}\left[\min\{U,n\}\, e^{\min\{U,n\}}\right] - \log(x_n)\right]
\end{aligned}
$$

Hence

$$
\log(x_n) \leq 0
$$

and taking the limit when $n \to \infty$, we show by monotonicity that $\mathbb{E}\left[e^U\right] \leq 1$. ∎

- **Theorem 2.7** ***(Alternative Duality Formula of Entropy)*** *[Boucheron et al., 2013]*

$$
Ent(X) = \sup_T \mathbb{E}\left[X\left(\log(T) - \log\left(\mathbb{E}\left[T\right]\right)\right)\right] \tag{21}
$$

where the supremum is taken over all non-negative and integrable random variables.

Proof: From (20), taking $U = \log\frac{T}{\mathbb{E}[T]}$, so that $\mathbb{E}\left[e^U\right] = \mathbb{E}\left[\frac{T}{\mathbb{E}[T]}\right] = 1$. This gives us (21). ∎

- **Corollary 2.8** ***(Duality Formula of Log Moment Generating Function)*** *[Cover and Thomas, 2006, Boucheron et al., 2013]*
  Let X be a real-valued integrable random variable. Then for every $\lambda \in \mathbb{R}$,

$$\log \mathbb{E}_Q \left[e^{\lambda(X - \mathbb{E}[X])} \right] = \sup_{P \ll Q} \left\{ \lambda \left(\mathbb{E}_P [X] - \mathbb{E}_Q [X] \right) - \mathbb{KL} \left(P \,\|\, Q \right) \right\}, \tag{22}$$

  where the supremum is taken over all probability measures P absolutely continuous with respect to Q, and $\mathbb{E}_P [\cdot]$ denotes integration with respect to the measure P (recall that $\mathbb{E}_Q [\cdot]$ is integration with respect to Q).

  Proof: Let $P \ll Q$. Taking $Y := \frac{dP}{dQ}$ and $U := \lambda(X - \mathbb{E}_Q [X]) - \psi_{X - \mathbb{E}_Q[X]}(\lambda)$ where $\psi_X(\lambda) := \log \mathbb{E}_Q \left[e^{\lambda X} \right]$. Note that $\mathbb{E}_Q [Y] = 1$ and $\mathbb{E} \left[e^U \right] = 1$. It follows from the duality formula that

$$\begin{aligned} \mathbb{KL} \left(P \,\|\, Q \right) = \operatorname{Ent}(Y) \geq \mathbb{E} [U\, Y] &= \mathbb{E} [\lambda(X - \mathbb{E}_Q [X])Y] - \psi_{X - \mathbb{E}_Q[X]}(\lambda) \\ &= \lambda(\mathbb{E}_P [X] - \mathbb{E}_Q [X]) - \psi_{X - \mathbb{E}_Q[X]}(\lambda) \end{aligned}$$

  or equivalently

$$\psi_{X - \mathbb{E}_Q[X]}(\lambda) \geq \lambda(\mathbb{E}_P [X] - \mathbb{E}_Q [X]) - \mathbb{KL} \left(P \,\|\, Q \right),$$

  therefore

$$\log \mathbb{E}_Q \left[e^{\lambda(X - \mathbb{E}_Q[X])} \right] \geq \sup_{P \ll Q} \left\{ \lambda(\mathbb{E}_P [X] - \mathbb{E}_Q [X]) - \mathbb{KL} \left(P \,\|\, Q \right) \right\}.$$

  Conversely, setting

$$U = \lambda \left(X - \mathbb{E}_Q [X] \right) - \sup_{P \ll Q} \left\{ \lambda(\mathbb{E}_P [X] - \mathbb{E}_Q [X]) - \mathbb{KL} \left(P \,\|\, Q \right) \right\}$$

  for every non-negative random variable Y such that $\mathbb{E} [Y] = 1$,

$$\mathbb{E} [UY] \leq \operatorname{Ent}(Y).$$

  Hence, $\mathbb{E} \left[e^U \right] \leq 1$ by duality theorem, which means that

$$\log \mathbb{E}_Q \left[e^{\lambda \left(X - \mathbb{E}_Q[X] \right)} \right] \leq \sup_{P \ll Q} \left\{ \lambda(\mathbb{E}_P [X] - \mathbb{E}_Q [X]) - \mathbb{KL} \left(P \,\|\, Q \right) \right\}. \qquad \blacksquare$$

- **Corollary 2.9** ***(Duality Formula of K-L divergence)*** *[Cover and Thomas, 2006, Boucheron et al., 2013]*
  Let P and Q be two probability distributions on the same space. Then

$$\mathbb{KL} \left(P \,\|\, Q \right) = \sup_X \left\{ \mathbb{E}_P [X] - \log \mathbb{E}_Q \left[e^X \right] \right\}, \tag{23}$$

  where the supremum is taken over all random variables such that $\mathbb{E}_Q [\exp (X)] < \infty$.

  Proof: If $P \ll Q$, $\mathbb{KL} \left(P \,\|\, Q \right) = \operatorname{Ent}(dP/dQ)$ and the corollary follows from the alternative formulation of the duality formula. Let $Y = dP/dQ$ and $X = \log(T)$ so that

$$\begin{aligned} \mathbb{KL} \left(P \,\|\, Q \right) = \operatorname{Ent}(Y) &= \sup_T \mathbb{E} \left[dP/dQ \left(\log(T) - \log \left(\mathbb{E} [T] \right) \right) \right] \\ &= \sup_X \left\{ \mathbb{E}_P [X] - \log \mathbb{E}_Q \left[e^X \right] \right\}. \end{aligned}$$

If $P \not\ll Q$, then there exists an event A such that $P(A) > 0 = Q(A)$, $\mathbb{KL}\left(P \,\|\, Q\right) = \infty$, and choosing $X_n = n\mathbb{1}\left\{A\right\}$ and letting n tend to infinity, we observe that the supremum on the right-hand side is infinite. ■

- **Remark** This corollary asserts that if Q remains fixed, $\mathbb{KL}\left(P \,\|\, Q\right)$ is *the* ***convex dual*** *of the functional* $X \to \log \mathbb{E}_Q\left[e^X\right]$.

- **Theorem 2.10** ***(The Expected Value Minimizes Expected Bregman Divergence)*** *[Boucheron et al., 2013]*
Let $I \subseteq \mathbb{R}$ be an open interval and let $f : I \to \mathbb{R}$ be ***convex*** *and* ***differentiable****. For any $x, y \in I$,* ***the Bregman divergence*** *of f from x to y is $f(y) - f(x) - f'(x)(y - x)$. Let X be an I-valued random variable. Then*

$$\mathbb{E}\left[f(X) - f(\mathbb{E}\left[X\right])\right] = \inf_{a \in I} \mathbb{E}\left[f(X) - f(a) - f'(a)(X - a)\right] \tag{24}$$

- **Corollary 2.11** ***(Duality Formula of Entropy via Bregman Divergence)*** *[Boucheron et al., 2013]*
Let X be a non-negative random variable such that $\mathbb{E}\left[\Phi(X)\right] < \infty$. Then

$$Ent(X) = \inf_{u > 0} \mathbb{E}\left[X\left(\log(X) - \log(u)\right) - (X - u)\right] \tag{25}$$

2.6 Optimal Transport

2.7 Pinsker's Inequality

2.8 Birgé's Inequality

2.9 The Brunn-Minkowski Inequality

References

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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