

# Lecture 3: Information Inequalities

Tianpei Xie

Jan. 6th., 2023

## Contents

<b>1</b>	<b>Information Theory Basics</b>	<b>2</b>
1.1	Entropy, Relative Entropy, and Mutual Information . . . . .	2
1.2	Chain Rules for Entropy, Relative Entropy, and Mutual Information . . . . .	4
1.3	Log-Sum Inequalities and Convexity . . . . .	5
1.4	Data Processing Inequality . . . . .	5
1.5	Fano's Inequality . . . . .	6
<b>2</b>	<b>Information Inequalities</b>	<b>7</b>
2.1	Han's Inequality . . . . .	7
2.2	Applications of Han's Inequality . . . . .	9
	2.2.1 Combinatorial Entropies . . . . .	9
	2.2.2 Edge Isoperimetric Inequality on the Binary Hypercube . . . . .	9
2.3	$\Phi$ -Entropy . . . . .	9
2.4	Sub-Additivity of $\Phi$ -Entropy . . . . .	10
2.5	Duality and Variational Formulas . . . . .	12
2.6	Wasserstein Distance and Transportation Cost Inequality . . . . .	16
2.7	Pinsker's Inequality . . . . .	17
2.8	Birgé's Inequality and Multiple Testing Problem . . . . .	18

# 1 Information Theory Basics

## 1.1 Entropy, Relative Entropy, and Mutual Information

- **Definition (*Shannon Entropy*)** [Cover and Thomas, 2006]

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

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

- **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  $\mu$  on  $\mathcal{X} \times \mathcal{Y}$ . Then **the joint entropy** of  $(X, Y)$ , denoted as  $H(X, Y)$ , is defined as

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

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

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

- **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**)  $H(p) := \mathbb{E}_p [-\log p(X)]$  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

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

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

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

which is *the entropy of  $P$  relative to  $Q$* . Furthermore, if  $\mu$  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

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

- **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  $P_X \otimes P_Y$* :

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

If  $P_{(X,Y)}$  has a probability density function  $p(x, y)$  with respect to a base measure  $\mu$  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  $P, Q$ .

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

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

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

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  $X_1, X_2, \dots, X_n$  be drawn according to  $p(x_1, x_2, \dots, x_n)$ . Then

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

- **Proposition 1.5 (Sub-Additivity of Entropy)** [Cover and Thomas, 2006]  
Let  $X_1, X_2, \dots, X_n$  be drawn according to  $p(x_1, x_2, \dots, x_n)$ . Then

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

with equality if and only if the  $X_i$  are independent.

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

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

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

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

- **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

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

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

$$\text{KL}(P_{(X,Y)} \parallel Q_{(X,Y)}) = \text{KL}(P_X \parallel Q_X) + \mathbb{E}_X [\text{KL}(P_{Y|X} \parallel Q_{Y|X})] \quad (6)$$

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

$$\text{KL}(P \parallel Q) = \sum_{i=1}^n \mathbb{E}_{P_{1:i-1}} [\text{KL}(P_{X_i|1\dots i-1} \parallel Q_{X_i|1\dots i-1})] \quad (7)$$

### 1.3 Log-Sum Inequalities and Convexity

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

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

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]  
 $\text{KL}(p \parallel q)$  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  $\lambda \in [0, 1]$ ,

$$\text{KL}(\lambda p_1 + (1 - \lambda)p_2 \parallel \lambda q_1 + (1 - \lambda)q_2) \leq \lambda \text{KL}(p_1 \parallel q_1) + (1 - \lambda) \text{KL}(p_2 \parallel q_2) \quad (9)$$

- **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  $X \rightarrow Y \rightarrow Z$ ) 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  $X \rightarrow Y \rightarrow Z$  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  $X \rightarrow Y \rightarrow Z$ , 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  $X \rightarrow Y \rightarrow Z$ , 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 Fano’s Inequality

- **Remark** Suppose that we know a random variable  $Y$  and we wish to guess the value of a correlated random variable  $X$ . **Fano’s inequality** relates **the probability of error** in guessing the random variable  $X$  to its *conditional entropy*  $H(X|Y)$ . It will be crucial in proving the **converse** to Shannon’s **channel capacity theorem**.
- **Proposition 1.14** (*Fano’s Inequality*) [Cover and Thomas, 2006]  
Let  $X, Y$  be random variables on domain  $\mathcal{X}, \mathcal{Y}$  and  $\hat{X} = g(Y)$  is an estimate of  $X$  where  $g : \mathcal{Y} \rightarrow \mathcal{X}$  is measurable function. The probability of error is defined as

$$P_e = \mathbb{P} \left\{ \hat{X} \neq X \right\}.$$

Then we have

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y) \quad (10)$$

This inequality can be weakened to

$$1 + P_e \log |\mathcal{X}| \geq H(X|Y) \quad (11)$$

$$P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}. \quad (12)$$

- **Corollary 1.15** [Cover and Thomas, 2006]  
For any two random variables  $X, Y$ , let  $p = \mathbb{P} \{X \neq Y\}$ .

$$H(p) + p \log |\mathcal{X}| \geq H(X|Y) \quad (13)$$

- **Corollary 1.16** [Cover and Thomas, 2006]  
Let  $P_e = \mathbb{P} \left\{ \hat{X} \neq X \right\}$ , and let  $\hat{X} : \mathcal{Y} \rightarrow \mathcal{X}$ ; then

$$H(P_e) + P_e (\log |\mathcal{X}| - 1) \geq H(X|Y) \quad (14)$$

- **Lemma 1.17** (*Bound of Error Probability via Shannon Entropy*) [Cover and Thomas, 2006]  
If  $X, X'$  are independent identically distributed random variables with entropy  $H(X)$ ,

$$\mathbb{P}\{X \neq X'\} \leq 1 - e^{-H(X)} \quad (15)$$

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

- **Corollary 1.18** (*Bound of Error Probability via Relative Entropy*) [Cover and Thomas, 2006]  
If  $X, X'$  are independent random variables in  $\mathcal{X}$  with distribution  $P$  and  $Q$ , respectively, and  $P \ll Q$

$$\mathbb{P}\{X \neq X'\} \leq 1 - e^{-H(P) - \text{KL}(P\|Q)}.$$

Similarly, if  $Q \ll P$ , then

$$\mathbb{P}\{X' \neq X\} \leq 1 - e^{-H(Q) - \text{KL}(Q\|P)}.$$

- **Remark** The error probability bound (15) states that *the **higher** the uncertainty* (i.e.  $H(X)$  increases), *the **lower** the probability that  $X = X'$* . Or, equivalently, *the **lower** (the Shannon and relative) entropy is, the **lower** the probability of error* for an estimate  $X'$  of  $X$ .

From *Fano's inequality* (10), we see that *the probability of error* for estimator  $\hat{X}$  based on observation  $Y$  is *bounded below* by *the conditional entropy  $H(X|Y)$*  of state  $X$  given observation  $Y$ . That is, we *cannot achieve lower error* of the estimation if uncertainty of state given observation ( $H(X|Y)$ ) is high.

## 2 Information Inequalities

### 2.1 Han's Inequality

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

$$\begin{aligned} H(X_1, X_2, \dots, X_n) &\leq \frac{1}{n-1} \sum_{i=1}^n H(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) \\ &\Leftrightarrow H(X) \leq \frac{1}{n-1} \sum_{i=1}^n H(X_{(-i)}) \end{aligned} \quad (17)$$

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

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

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

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

which is what we wanted to prove.  $\blacksquare$

- **Proposition 2.2 (*Han's Inequality for Relative Entropy*)** [Boucheron et al., 2013]  
Let  $(\mathcal{X}, \mathcal{B})$  be a measurable space, and  $P$  and  $Q$  be probability measures on  $\mathcal{X}^n$  such that  $P = P_1 \otimes \dots \otimes P_n$  is a **product measure**. We denote the element of  $\mathcal{X}^n$  by  $x = (x_1, \dots, x_n)$  and write  $x_{(-i)} := (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$  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  $\mu$  on  $\mathcal{X}$ .

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

Then

$$\text{KL}(Q \| P) \geq \frac{1}{n-1} \sum_{i=1}^n \text{KL}(Q_{(-i)} \| P_{(-i)}) \quad (18)$$

or equivalently,

$$\text{KL}(Q \| P) \leq \sum_{i=1}^n (\text{KL}(Q \| P) - \text{KL}(Q_{(-i)} \| P_{(-i)})) \quad (19)$$

**Proof:** From Han's inequality, we have

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

Since

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

and

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

it suffices to show that

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



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

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

Rearranging, we obtain

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

## 2.2 Applications of Han's Inequality

### 2.2.1 Combinatorial Entropies

### 2.2.2 Edge Isoperimetric Inequality on the Binary Hypercube

•

## 2.3 $\Phi$ -Entropy

- **Definition ( $\Phi$ -Entropy)** [Boucheron et al., 2013]

Let  $\Phi : [0, \infty) \rightarrow \mathbb{R}$  be a **convex** function, and assign, to every **non-negative integrable random variable**  $X$ , the  $\Phi$ -entropy of  $X$  is defined as

$$H_\Phi(X) = \mathbb{E} [\Phi(X)] - \Phi(\mathbb{E} [X]). \quad (20)$$

- **Remark** The  $\Phi$ -entropy is a **functional** of *distribution*  $P_X$  instead of a function of  $X$ .
- **Remark** By Jensen's inequality, the  $\Phi$ -entropy is *non-negative*

$$\begin{aligned}\Phi(\mathbb{E} [X]) &\leq \mathbb{E} [\Phi(X)] \\ \Rightarrow H_\Phi(X) &= \mathbb{E} [\Phi(X)] - \Phi(\mathbb{E} [X]) \geq 0.\end{aligned}$$

- **Example (*Special Examples for  $\Phi$ -Entropy*)**

1. For  $\Phi(x) = x^2$ , the  $\Phi$ -entropy of  $X$  is the **variance** of  $X$ :

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

2. For  $\Phi(x) = -\log(x)$ , the  $\Phi$ -entropy of  $Y = e^{\lambda X}$  is the **logarithm of moment generating function** of  $X - \mathbb{E}[X]$ :

$$H_{\Phi}(e^{\lambda X}) = -\lambda \mathbb{E}[X] + \log \left( \mathbb{E} \left[ e^{\lambda X} \right] \right) = \log \mathbb{E} \left[ e^{\lambda(X - \mathbb{E}[X])} \right] := \psi_{X - \mathbb{E}[X]}(\lambda). \quad (21)$$

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

$$H_{\Phi}(X) = \text{Ent}(X) := \mathbb{E}[X \log X] - \mathbb{E}[X] \log(\mathbb{E}[X]). \quad (22)$$

Let  $(\Omega, \mathcal{B})$  be measurable space, and  $P$  and  $Q$  are probability measures on  $\Omega$  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  $\mathbb{E}_Q[X] = 1$ . Then the entropy of  $X$  is the relative entropy of  $P$  with respect to  $Q$ .

$$\text{Ent}(X) = \text{KL}(P \parallel Q) \quad (23)$$

## 2.4 Sub-Additivity of $\Phi$ -Entropy

- **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  $Z_1, Z_2, \dots, Z_n$  be independent random variables taking values in  $\mathcal{X}$ , and let  $f : \mathcal{X}^n \rightarrow [0, \infty)$  be a measurable function. Letting  $X = f(Z_1, Z_2, \dots, Z_n)$  such that  $\mathbb{E}[X \log X] < \infty$ , we have

$$\mathbb{E}[\Phi(X)] - \Phi(\mathbb{E}[X]) \leq \sum_{i=1}^n \mathbb{E}[\mathbb{E}_{(-i)}[\Phi(X)] - \Phi(\mathbb{E}_{(-i)}[X])], \quad (24)$$

where  $\mathbb{E}_{(-i)}[\cdot]$  is the conditional expectation operator conditioning on  $Z_{(-i)}$ . Introducing the notation  $\text{Ent}_{(-i)}(X) = \mathbb{E}_{(-i)}[\Phi(X)] - \Phi(\mathbb{E}_{(-i)}[X])$ , this can be re-written as

$$\mathbb{E}[\Phi(X)] - \Phi(\mathbb{E}[X]) \leq \mathbb{E} \left[ \sum_{i=1}^n \text{Ent}_{(-i)}(X) \right]. \quad (25)$$

**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  $\mathbb{E}[X] = 1$ . 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  $Z := (Z_1, Z_2, \dots, Z_n)$  and  $Q$  the corresponding probability measure. Then

$$\text{Ent}(X) := \mathbb{E}[X \log X] - \mathbb{E}[X] \log(\mathbb{E}[X]) = \text{KL}(P \parallel Q)$$

which, by Han's inequality for relative entropy

$$\text{Ent}(X) = \text{KL}(P \parallel Q) \leq \sum_{i=1}^n (\text{KL}(P \parallel Q) - \text{KL}(P_{(-i)} \parallel Q_{(-i)}))$$

However, straightforward calculation shows that

$$\sum_{i=1}^n (\text{KL}(P \parallel Q) - \text{KL}(P_{(-i)} \parallel Q_{(-i)})) = \sum_{i=1}^n \mathbb{E} [\mathbb{E}_{(-i)} [\Phi(X)] - \Phi(\mathbb{E}_{(-i)} [X])]$$

and the statement follows.  $\blacksquare$

**Proof: (Alternative Proof via Duality Formulation of Entropy)**

Denote the conditional expectation operator  $\mathbb{E}_{1:i}[\cdot] = \mathbb{E}[\cdot | Z_1, \dots, Z_i]$  for  $i = 1, \dots, n$  and the convention  $\mathbb{E}_0[\cdot] = \mathbb{E}[\cdot]$ . Noting that the operator  $\mathbb{E}_{1:n}[\cdot]$  is just identity when restricted to the set of  $(Z_1, \dots, Z_n)$ -measurable and integrable random variables, we have the decomposition

$$X (\log X - \log (\mathbb{E}[X])) = \sum_{i=1}^n X (\log (\mathbb{E}_{1:i}[X]) - \log (\mathbb{E}_{1:i-1}[X])).$$

Note that since  $Z_1, Z_2, \dots, Z_n$  are independent, we have  $\mathbb{E}_{(-i)} [\mathbb{E}_{1:i}[X]] = \mathbb{E}_{1:i-1}[X]$ . Now the duality formula given in Theorem 2.7 yields

$$\mathbb{E} [X (\log(T) - \log (\mathbb{E}[T]))] \leq \text{Ent}(X)$$

Setting  $T := \mathbb{E}_{1:i}[X]$ , and replacing expectation  $\mathbb{E}[\cdot]$  by conditional expectation  $\mathbb{E}_{(-i)}[\cdot]$

$$\mathbb{E}_{(-i)} [X (\log (\mathbb{E}_{1:i}[X]) - \log (\mathbb{E}_{(-i)} [\mathbb{E}_{1:i}[X]]))] \leq \text{Ent}_{(-i)}(X).$$

Finally, taking expectations on both sides of the decomposition above yields

$$\begin{aligned} \mathbb{E} [X (\log X - \log (\mathbb{E}[X]))] &= \sum_{i=1}^n \mathbb{E} [\mathbb{E}_{(-i)} [X (\log (\mathbb{E}_{1:i}[X]) - \log (\mathbb{E}_{(-i)} [\mathbb{E}_{1:i}[X]]))] ] \\ &\leq \sum_{i=1}^n \mathbb{E} [\text{Ent}_{(-i)}(X)] \quad \blacksquare \end{aligned}$$

- **Remark** The Efron-Stein inequality is the special case of the inequality when  $\Phi(x) = x^2$ ,

$$\begin{aligned} \mathbb{E} [\Phi(X)] - \Phi(\mathbb{E}[X]) &\leq \sum_{i=1}^n \mathbb{E} [\mathbb{E}_{(-i)} [\Phi(X)] - \Phi(\mathbb{E}_{(-i)} [X])] \\ \Rightarrow \text{Var}(X) &\leq \sum_{i=1}^n \mathbb{E} [\text{Var}_{(-i)}(X)] \end{aligned}$$

- **Remark (Han's inequality from Sub-additivity of Entropy)** [Boucheron et al., 2013]  
It is interesting to notice that *Han's inequality* itself can be derived from *the sub-additivity of entropy*. In other words, for *discrete probability distributions*, the sub-additivity of entropy and Han's inequality are *equivalent*.

- **Remark (*Tensorization Property of Entropy*)** [Wainwright, 2019]

The inequality in (24) or (25) is also called **the tensorization property of entropy**.

Let  $\mu = \mu_1 \otimes \dots \otimes \mu_n$  where  $\mu_i$  be the probability distribution of  $Z_i$ . Thus  $\mu$  is the probability distribution of  $Z = (Z_1, \dots, Z_n)$  when  $Z_i$  are independent. The sub-additivity of entropy states that

$$\text{Ent}_{\mu_1 \otimes \dots \otimes \mu_n}(f) \leq \mathbb{E}_{\mu_1 \otimes \dots \otimes \mu_n} \left[ \sum_{i=1}^n \text{Ent}_{\mu_i}(f) \right]$$

where the subscript  $\mu_i$  indicates that the integration concerns the  $i$ -th variable only.

- **Proposition 2.4 (*Sub-Additivity of  $\Phi$ -Entropy*)** [Boucheron et al., 2013]

Let  $\mathcal{C}$  denote the class of functions  $\Phi : [0, \infty) \rightarrow \mathbb{R}$  that are **continuous** and **convex** on  $[0, \infty)$ , **twice differentiable** on  $(0, \infty)$ , and such that either  $\Phi$  is **affine** or  $\Phi''$  is **strictly positive** and  $1/\Phi''$  is **concave**. For all  $\Phi \in \mathcal{C}$ , the **entropy functional**  $H_\Phi$  is **sub-additive**. That is,

$$\begin{aligned} \mathbb{E} [\Phi(X)] - \Phi(\mathbb{E} [X]) &\leq \sum_{i=1}^n \mathbb{E} [\mathbb{E}_{(-i)} [\Phi(X)] - \Phi(\mathbb{E}_{(-i)} [X])], \\ \Leftrightarrow H_\Phi(X) &\leq \mathbb{E} \left[ \sum_{i=1}^n H_\Phi^{(-i)}(X) \right] \end{aligned} \quad (26)$$

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

- **Remark** The **sub-additivity property** of  $H_\Phi$  is equivalent to what we could call **the Jensen property**

$$\begin{aligned} H_\Phi \left( \int f(z, Z_2) d\mu_1(z) \right) &\leq \int H_\Phi(f(z, Z_2)) d\mu_1(z) \\ \Leftrightarrow H_\Phi(\mathbb{E}_{Z_1} [f(Z_1, Z_2)]) &\leq \mathbb{E}_{Z_1} [H_\Phi(f(Z_1, Z_2))] \end{aligned} \quad (27)$$

The proof of this property can be done by using the duality formulation of  $\Phi$ -entropy in Theorem 2.12.

## 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,

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

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

$$\begin{aligned} \nabla g(x) &= u - \log(x) - 1 = 0 \\ \Rightarrow x^* &= e^{u-1} \\ \Rightarrow \sup_x \{u x - x \log(x)\} &= g(x^*) = u e^{u-1} - e^{u-1}(u-1) = e^{u-1} \quad \blacksquare \end{aligned}$$

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

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

Therefore,  $U_+X$  is integrable, and one can always define  $\mathbb{E}[UX] = \mathbb{E}[U_+X] - \mathbb{E}[U_-X]$  for positive and negative part of  $U$ . Thus the  $\mathbb{E}[UX]$  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  $(\Omega, \mathcal{A}, P)$  such that  $\mathbb{E}[\Phi(X)] < \infty$ . Then we have **the duality formula**

$$\text{Ent}(X) = \sup_{U \in \mathcal{U}} \mathbb{E}[UX] \quad (28)$$

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

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

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

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

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

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

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

$$\mathbb{E}[UX_n] \leq \text{Ent}(X_n),$$

and therefore

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

Hence

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

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

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

$$\text{Ent}(X) = \sup_T \mathbb{E} [X (\log(T) - \log(\mathbb{E}[T]))] \quad (29)$$

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

**Proof:** From (28), taking  $U = \log \frac{T}{\mathbb{E}[T]}$ , so that  $\mathbb{E}[e^U] = \mathbb{E}\left[\frac{T}{\mathbb{E}[T]}\right] = 1$ . This gives us (29). ■

- **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  $\lambda \in \mathbb{R}$ ,

$$\log \mathbb{E}_Q [e^{\lambda(X - \mathbb{E}[X])}] = \sup_{P \ll Q} \{\lambda(\mathbb{E}_P[X] - \mathbb{E}_Q[X]) - \text{KL}(P \parallel Q)\}, \quad (30)$$

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  $\psi_X(\lambda) := \log \mathbb{E}_Q[e^{\lambda X}]$ . Note that  $\mathbb{E}_Q[Y] = 1$  and  $\mathbb{E}[e^U] = 1$ . It follows from the duality formula that

$$\begin{aligned} \text{KL}(P \parallel Q) = \text{Ent}(Y) &\geq \mathbb{E}[UY] = \mathbb{E}[\lambda(X - \mathbb{E}_Q[X])Y] - \psi_{X - \mathbb{E}_Q[X]}(\lambda) \\ &= \lambda(\mathbb{E}_P[X] - \mathbb{E}_Q[X]) - \psi_{X - \mathbb{E}_Q[X]}(\lambda) \end{aligned}$$

or equivalently

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

therefore

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

Conversely, setting

$$U = \lambda(X - \mathbb{E}_Q[X]) - \sup_{P \ll Q} \{\lambda(\mathbb{E}_P[X] - \mathbb{E}_Q[X]) - \text{KL}(P \parallel Q)\}$$

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

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

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

$$\log \mathbb{E}_Q [e^{\lambda(X - \mathbb{E}_Q[X])}] \leq \sup_{P \ll Q} \{\lambda(\mathbb{E}_P[X] - \mathbb{E}_Q[X]) - \text{KL}(P \parallel Q)\}. \quad \blacksquare$$

- **Corollary 2.9 (Duality Formula of Kullback-Leibler Divergence)** [Cover and Thomas, 2006, Boucheron et al., 2013]

Let  $P$  and  $Q$  be two probability distributions on the same space. Then

$$\text{KL}(P \parallel Q) = \sup_X \{\mathbb{E}_P[X] - \log \mathbb{E}_Q[e^X]\}, \quad (31)$$

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

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

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

If  $P \not\ll Q$ , then there exists an event  $A$  such that  $P(A) > 0 = Q(A)$ ,  $\mathbb{KL}(P \parallel Q) = \infty$ , and choosing  $X_n = n\mathbb{1}\{A\}$  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,  $\mathbb{KL}(P \parallel Q)$  is the **convex dual** of the functional  $X \rightarrow \log \mathbb{E}_Q[e^X]$ .
- **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  $f : I \rightarrow \mathbb{R}$  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

$$\mathbb{E} [f(X) - f(\mathbb{E}[X])] = \inf_{a \in I} \mathbb{E} [f(X) - f(a) - f'(a)(X - a)] \quad (32)$$

**Proof:** Let  $a \in I$ . The difference between the expected Bregman divergence from  $a$  and the expected Bregman divergence from  $\mathbb{E}[X]$

$$\mathbb{E} [f(X) - f(\mathbb{E}[X]) - f'(\mathbb{E}[X])(X - \mathbb{E}[X])] = \mathbb{E} [f(X) - f(\mathbb{E}[X])]$$

satisfies

$$\begin{aligned} &\mathbb{E} [f(X) - f(a) - f'(a)(X - a)] - \mathbb{E} [f(X) - f(\mathbb{E}[X]) - f'(\mathbb{E}[X])(X - \mathbb{E}[X])] \\ &= \mathbb{E} [f(X) - f(a) - f'(a)(X - a)] - \mathbb{E} [f(X) - f(\mathbb{E}[X])] \\ &= \mathbb{E} [-(f(a) - f(\mathbb{E}[X])) - f'(a)(X - a)] \\ &= f(\mathbb{E}[X]) - f(a) - f'(a)(\mathbb{E}[X] - a) \end{aligned}$$

The last expression is the Bregman divergence of  $f$  from  $a$  to  $\mathbb{E}[X]$ . As  $f$  is **convex**, it is **nonnegative**. ■

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

$$\text{Ent}(X) = \inf_{u > 0} \mathbb{E} [X (\log(X) - \log(u)) - (X - u)] \quad (33)$$

- **Theorem 2.12 (Duality Formula of General  $\Phi$ -Entropy)** [Boucheron et al., 2013]  
Let  $\mathcal{C}$  denote the class of functions  $\Phi : [0, \infty) \rightarrow \mathbb{R}$  that are **continuous** and **convex** on  $[0, \infty)$ , **twice differentiable** on  $(0, \infty)$ , and such that either  $\Phi$  is **affine** or  $\Phi''$  is **strictly positive** and  $1/\Phi''$  is **concave**. Denote  $\text{conv}(L_1^+)$  as the **convex set of non-negative and integrable** random variables  $X$ . Let  $\Phi \in \mathcal{C}$  and  $X \in \text{conv}(L_1^+)$ . If  $\Phi(X)$  is integrable, then

$$H_\Phi(X) = \sup_{T \in \text{conv}(L_1^+), T \neq 0} \{ \mathbb{E} [(\Phi'(T) - \Phi'(\mathbb{E}[T])) (X - T) + \Phi(T)] - \Phi(\mathbb{E}[T]) \}. \quad (34)$$

The supremum is achieved when  $T = X$  (or  $T = 1$  if  $X = 0$ ).

Another variational formulation of  $\Phi$ -entropy via Bregman divergence is

$$H_\Phi(X) = \inf_{u>0} \mathbb{E} [\Phi(X) - \Phi(u) - \Phi'(u)(X - u)]. \quad (35)$$

## 2.6 Wasserstein Distance and Transportation Cost Inequality

- **Proposition 2.13** (*Wasserstein Distance and Transportation Cost Inequality*) [Boucheron et al., 2013]

Let  $X$  be a real-valued integrable random variable. Let  $\phi$  be a **convex** and **continuously differentiable** function on a (possibly unbounded) interval  $[0, b)$  and assume that  $\phi(0) = \phi'(0) = 0$ . Define, for every  $x \geq 0$ , the **Legendre transform**  $\phi^*(x) = \sup_{\lambda \in (0, b)} (\lambda x - \phi(\lambda))$ , and let, for every  $t \geq 0$ ,  $\phi^{*-1}(t) = \inf\{x \geq 0 : \phi^*(x) > t\}$ , i.e. the **generalized inverse** of  $\phi^*$ . Then the following two statements are equivalent:

1. for every  $\lambda \in (0, b)$ ,

$$\psi_{X - \mathbb{E}[X]}(\lambda) \leq \phi(\lambda)$$

where  $\psi_X(\lambda) := \log \mathbb{E}_Q [e^{\lambda X}]$  is the logarithm of moment generating function;

2. for any probability measure  $P$  absolutely continuous with respect to  $Q$  such that  $\text{KL}(P \parallel Q) < \infty$ ,

$$\mathbb{E}_P[X] - \mathbb{E}_Q[X] \leq \phi^{*-1}(\text{KL}(P \parallel Q)). \quad (36)$$

In particular, given  $\nu > 0$ ,  $X$  follows a sub-Gaussian distribution, i.e.

$$\psi_{X - \mathbb{E}[X]}(\lambda) \leq \frac{\nu \lambda^2}{2}$$

for every  $\lambda > 0$  **if and only if** for any probability measure  $P$  absolutely continuous with respect to  $Q$  and such that  $\text{KL}(P \parallel Q) < \infty$ ,

$$\mathbb{E}_P[X] - \mathbb{E}_Q[X] \leq \sqrt{2\nu \text{KL}(P \parallel Q)}. \quad (37)$$

**Proof:** As a direct consequence of Corollary 2.8, we see that (1) holds if and only if for every distribution  $P \ll Q$ ,

$$\begin{aligned} & \psi_{X - \mathbb{E}[X]}(\lambda) \leq \phi(\lambda) \\ \Leftrightarrow & \lambda(\mathbb{E}_P[X] - \mathbb{E}_Q[X]) - \text{KL}(P \parallel Q) \leq \phi(\lambda), & \forall P \ll Q \\ \Leftrightarrow & \mathbb{E}_P[X] - \mathbb{E}_Q[X] \leq \frac{\phi(\lambda) + \text{KL}(P \parallel Q)}{\lambda}, & \forall P \ll Q, \lambda \in (0, b) \\ \Leftrightarrow & \mathbb{E}_P[X] - \mathbb{E}_Q[X] \leq \inf_{\lambda \in (0, b)} \left\{ \frac{\text{KL}(P \parallel Q) + \phi(\lambda)}{\lambda} \right\} & \forall P \ll Q \end{aligned}$$

Note that

$$\phi^{*-1}(t) = \inf_{\lambda \in (0, b)} \left[ \frac{t + \phi(\lambda)}{\lambda} \right]$$



Setting  $t = \mathbb{KL}(P \parallel Q)$ , we have

$$\begin{aligned} \psi_{X-\mathbb{E}[X]}(\lambda) &\leq \phi(\lambda) \\ \Leftrightarrow \mathbb{E}_P[X] - \mathbb{E}_Q[X] &\leq \phi^{*-1}(\mathbb{KL}(P \parallel Q)). \end{aligned}$$

which shows that (i) is equivalent to (ii). Applying the previous result with  $\phi(\lambda) = \lambda^2 \nu / 2$  for every  $\lambda > 0$  leads to the stated special case of equivalence since then  $\phi^{*-1}(t) = \sqrt{2\nu t}$ .  $\blacksquare$

- **Remark (*The Quadratic Transportation Cost Inequality / The Information Inequality*)** [Boucheron et al., 2013, Wainwright, 2019]

The inequality (36) and (37) are called *information inequality* in [Wainwright, 2019] due to the role of Kullback-Leibler Divergence in information theory.

The inequality (37) is related to what is usually termed a *quadratic transportation cost inequality*. If  $\Omega$  is a *metric space*, the probability measure  $Q$  is said to satisfy a *quadratic transportation cost inequality* if the last inequality holds for every  $X$  which is *Lipschitz* on  $\Omega$  with *Lipschitz norm* at most 1.

$$\mathcal{W}(P, Q) = \sup_{X \in \text{Lip}_1} \{ \mathbb{E}_P[X] - \mathbb{E}_Q[X] \} \leq \sqrt{2\nu \mathbb{KL}(P \parallel Q)}. \quad (38)$$

where  $\text{Lip}_1 = \{f \in \mathbb{R}^\Omega : |f(x) - f(y)| \leq L d(x, y), L \leq 1\}$  and  $d$  is the metric in  $\Omega$ . Here  $\mathcal{W}(P, Q)$  is *the Wasserstein distance* between  $P$  and  $Q$  induced by metric  $d$ .

## 2.7 Pinsker's Inequality

- **Definition (*Total Variation / Variational Distance*)**

Let  $P, Q$  be two probability measures on measurable space  $(\Omega, \mathcal{F})$ . The *total variation* or *variational distance* between  $P$  and  $Q$  is defined by

$$V(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)| \quad (39)$$

- **Remark (*Equivalent Formulation of Total Variation*)**

It is a well-known and simple fact that the total variation is half the  $L_1$ -distance, that is, if  $\mu$  is a *common dominating measure* of  $P$  and  $Q$  and  $p(x) = dP/d\mu$  and  $q(x) = dQ/d\mu$  denote their respective densities, then

$$V(P, Q) := P(A^*) - Q(A^*) = \frac{1}{2} \int_{\Omega} |p(x) - q(x)| d\mu(x), \quad (40)$$

where  $A^* = \{x : p(x) \geq q(x)\}$ .

- **Remark (*Total Variation via Optimal Coupling of Two Measures*)**

We note that another important interpretation of *the variational distance* is related to *the best coupling of the two measures*

$$V(P, Q) = \min P \{X \neq Y\} \quad (41)$$

where the minimum is taken over *all pairs of joint distributions* for the random variables  $(X, Y)$  whose marginal distributions are  $X \sim P$  and  $Y \sim Q$ .

- **Remark (*Applications of Pinsker's Inequality*)**

The importance of *Pinsker's inequality* in statistics stems from the fact that it provides a **lower bound** for the **error** of certain hypothesis testing problems.

We use Pinsker's inequality for a completely different purpose, namely for establishing a transportation cost inequality that may be used to prove concentration inequalities.

- **Proposition 2.14 (*Pinsker's Inequality*)** [Cover and Thomas, 2006, Boucheron et al., 2013]

Let  $P, Q$  be two probability distributions on measurable space  $(\Omega, \mathcal{F})$  such that  $P \ll Q$ . Then

$$V(P, Q)^2 \leq \frac{1}{2} \text{KL}(P \parallel Q). \quad (42)$$

**Proof:** Define the random variable  $X$  such that  $dP = XdQ$  and let  $A^* = \{X \geq 1\}$  be the set achieving the maximum in the definition of the total variation between  $P$  and  $Q$ . Then, setting  $Z = \mathbf{1}\{A^*\}$ ,

$$V(P, Q) := P(A^*) - Q(A^*) = \mathbb{E}_P[Z] - \mathbb{E}_Q[Z].$$

It follows from Hoeffding's lemma that

$$\psi_{Z - \mathbb{E}[Z]}(\lambda) \leq \frac{\lambda^2}{8}$$

which by transportation cost inequality for sub-Gaussian variables we have

$$\mathbb{E}_P[Z] - \mathbb{E}_Q[Z] \leq \sqrt{\frac{1}{2} \text{KL}(P \parallel Q)}. \quad \blacksquare$$

- **Remark (*Total Variation as 1-Wasserstein Distance*)**

The total variation between  $P$  and  $Q$  is **the Wasserstein distance** induced by **the Hamming distance**  $d(x, y) = \#\{i : x_i \neq y_i\}$ .

$$V(P, Q) = \mathcal{W}_1(P, Q).$$

Thus the *Pinsker's inequality* (42) is the special case of *transportation cost inequality* (36).

## 2.8 Birgé's Inequality and Multiple Testing Problem

- **Remark** We will use the *Pinsker's inequality* to derive a **lower bound** on **the probability of error** in *multiple testing problem*.

- **Proposition 2.15 (*Sharper Information Inequality for Total Variation*)** [Boucheron et al., 2013]

Let  $P, Q$  be two probability distributions on measurable space  $(\Omega, \mathcal{F})$  such that  $P \ll Q$ .

$$\sup_{A \in \mathcal{F}} h(P(A), Q(A)) \leq \text{KL}(P \parallel Q) \quad (43)$$

where  $h(p, q) = \text{KL}(p \parallel q) = q \log(q/p) + (1 - q) \log((1 - q)/(1 - p))$  when  $p, q \in [0, 1]$  are parameters of Bernoulli random variables.

**Proof:** For any  $p \in [0, 1]$ , let

$$\phi_p(\lambda) = \log \left( p \left( e^\lambda - 1 \right) + 1 \right)$$

denote the logarithm of the moment generating function of the Bernoulli( $p$ ) distribution where  $\lambda \in \mathbb{R}$ . By the duality formulation of relative entropy, for  $X = \mathbb{1}\{A\}$ ,

$$\begin{aligned} \text{KL}(P \parallel Q) &\geq \mathbb{E}_P[\lambda \mathbb{1}\{A\}] - \log \mathbb{E}_Q[e^{\lambda \mathbb{1}\{A\}}] \\ \Rightarrow \text{KL}(P \parallel Q) &\geq \sup_{\lambda \geq 0} \{ \lambda P(A) - \phi_{Q(A)}(\lambda) \}. \end{aligned}$$

The proposition follows by noting that for any  $a \in [0, 1]$ ,

$$h(a, p) = \sup_{\lambda \geq 0} \{ \lambda a - \phi_p(\lambda) \}. \quad \blacksquare$$

- **Remark** Note that

$$h(P(A), Q(A)) \geq 2(P(A) - Q(A))^2.$$

Thus the proposition above implies the Pinsker's inequality.

- **Remark** *The variational representation of relative entropy* may be used to establish *lower bounds* for *the probability of error* in *multiple testing problems*. The next result is a sharper version of *Fano's inequality*, a classical tool from information theory.

**Proposition 2.16** (*Birgé's Inequality*) [Boucheron et al., 2013]

Let  $P_0, P_1, \dots, P_N$  be probability distributions on measurable space  $(\Omega, \mathcal{F})$  and let  $A_0, A_1, \dots, A_N \in \mathcal{F}$  be pairwise disjoint events. If  $a = \min_{i=0, \dots, N} P_i(A_i) \geq 1/(N+1)$ ,

$$a \leq h\left(a, \frac{1-a}{N}\right) \leq \frac{1}{N} \sum_{i=1}^N \text{KL}(P_i \parallel P_0) \quad (44)$$

**Proof:** By the variational representation of relative entropy, for any  $i = 0, \dots, N$ ,

$$\sup_{\lambda > 0} \left\{ \mathbb{E}_{P_i}[\lambda \mathbb{1}\{A_i\}] - \log \mathbb{E}_{P_0}[e^{\lambda \mathbb{1}\{A_i\}}] \right\} \leq \text{KL}(P_i \parallel P_0).$$

See that

$$\begin{aligned} 1 - a &= 1 - \min_{i=0, \dots, N} P_i(A_i) \\ &\geq 1 - P_0(A_0) \geq \sum_{i=1}^N P_0(A_i). \end{aligned}$$

For any  $\lambda > 0$ ,

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^N \mathbb{KL}(P_i \parallel P_0) &\geq \frac{1}{N} \sum_{i=1}^N \left\{ \lambda P_i(A_i) - \log \mathbb{E}_{P_0} \left[ e^{\lambda \mathbb{1}_{\{A_i\}}} \right] \right\} \\
&\geq \frac{1}{N} \sum_{i=1}^N \left\{ \lambda a - \log \left( P_0(A_i) (e^\lambda - 1) + 1 \right) \right\} \\
&= \lambda a - \frac{1}{N} \sum_{i=1}^N \log \left( P_0(A_i) (e^\lambda - 1) + 1 \right) \\
&\geq \lambda a - \log \left( \frac{1}{N} \sum_{i=1}^N \left( P_0(A_i) (e^\lambda - 1) + 1 \right) \right) \quad (\text{by convexity of } -\log(x)) \\
&= \lambda a - \log \left( \left( \frac{1}{N} \sum_{i=1}^N P_0(A_i) \right) (e^\lambda - 1) + 1 \right) \\
&\geq \lambda a - \log \left( \frac{1 - P_0(A_0)}{N} (e^\lambda - 1) + 1 \right) \\
&\geq \lambda a - \log \left( \frac{1 - a}{N} (e^\lambda - 1) + 1 \right)
\end{aligned}$$

Note that the supremum of the right-hand side with respect to  $\lambda$  is  $h\left(a, \frac{1-a}{N}\right)$ . ■

## 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.
- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.