

# CS258: Information Theory

Fan Cheng

Shanghai Jiao Tong University

[http://www.cs.sjtu.edu.cn/~chengfan/  
chengfan@sjtu.edu.cn](http://www.cs.sjtu.edu.cn/~chengfan/chengfan@sjtu.edu.cn)

Spring, 2020

# Outline

- Differential Entropy
- AEP for Continuous Random Variable
- Relative Entropy and Mutual Information
- Property of Differential Information Measures
- Information inequalities and applications

# $I(X; Y)$ : Correlated Gaussian

(Mutual information between correlated Gaussian random variables with correlation  $\rho$ ) Let  $(X, Y) \sim \mathcal{N}(0, K)$ , where

$$K = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

$I(X; Y)$ ?

$$h(X) = h(Y) = \frac{1}{2} \log 2\pi e \sigma^2$$

$$h(X, Y) = \frac{1}{2} \log(2\pi e)^2 |K| = \frac{1}{2} \log(2\pi e)^2 \sigma^4 (1 - \rho^2)$$

$$I(X; Y) = h(X) + h(Y) - h(X, Y) = -\frac{1}{2} \log(1 - \rho^2)$$

- $\rho = 0$ ,  $X$  and  $Y$  are independent and  $I$  is 0
- $\rho = \pm 1$ ,  $X$  and  $Y$  are perfectly correlated and  $I$  is  $\infty$

# Maximum Entropy with Constraints

- Let the random variable  $X \in \mathcal{R}$  have mean  $\mu$  and variance  $\sigma^2$ . Then

$$h(X) \leq \frac{1}{2} \log 2\pi e \sigma^2$$

with equality iff  $X \sim \mathcal{N}(\mu, \sigma^2)$

- Let the random variable  $X \in \mathcal{R}$  satisfy  $EX^2 \leq \sigma^2$ . Then

$$h(X) \leq \frac{1}{2} \log 2\pi e \sigma^2$$

with equality iff  $X \sim \mathcal{N}(0, \sigma^2)$

1. Let  $X_G \sim \mathcal{N}(\mu, \sigma^2)$ . Consider

$$D(X||X_G) \geq 0$$

Then

$$\int f \log \frac{f}{g} \geq 0$$

$$h(X) = h(f) \leq -\int f \log g = -\int f \log \frac{1}{\sqrt{2\pi\sigma^2}} + f \left( -\frac{(x-\mu)^2}{2\sigma^2} \right)$$

$$h(X) \leq \frac{1}{2} \log 2\pi\sigma^2 + \frac{1}{2} = \frac{1}{2} \log 2\pi e \sigma^2$$

2.  $\text{Var}(X) = E(X^2) - E(X)^2 \leq \sigma^2 \Rightarrow \text{Case 1.}$

$E(X^2), \text{Var}(X)$  给定的情况下,  
高斯分布最大化微分熵

# Maximum Entropy

Consider the following problem: Maximize the entropy  $h(f)$  over all probability densities  $f$  satisfying

1.  $f(x) \geq 0$ , with equality outside the support

2.  $\int_S f(x)dx = 1$

(+ +)

3.  $\int_S f(x)r_i(x)dx = \alpha_i$  for  $1 \leq i \leq m$ . ( $r_i(x)$  is a function of  $x$ )

Thus,  $f$  is a density on support set  $S$  meeting certain moment constraints  $\alpha_1, \alpha_2, \dots, \alpha_m$ .

**Theorem 12.1.1 (Maximum entropy distribution)** Let

$$f^*(x) = f_\lambda(x) = e^{\lambda_0 + \sum_{i=1}^m \lambda_i r_i(x)}$$

$x \in S$ , where  $\lambda_0, \dots, \lambda_m$  are chosen so that  $f^*$  satisfies (+ +). Then  $f^*$  uniquely maximizes  $h(f)$  over all probability densities  $f$  satisfying constraints (+ +).

■ Let  $S = [a, b]$ , with no other constraints. Then the maximum entropy distribution is the uniform distribution over this range.

■  $S = [0, \infty)$  and  $EX = \mu$ . Then the entropy-maximizing distribution is

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \quad x \geq 0$$

■  $S = (-\infty, \infty)$ ,  $EX = \alpha_1$ , and  $EX^2 = \alpha_2$ . The maximum entropy distribution is  $\mathcal{N}(\alpha_1, \alpha_2 - \alpha_1^2)$

# Hadamard's Inequality

$K$  is a nonnegative definite symmetric  $n \times n$  matrix. Let  $|K|$  denote the determinant of  $K$ .

**Theorem (Hadamard)**  $|K| \leq \prod K_{ii}$ , with equality iff  $K_{ij} = 0, i \neq j$

Let  $X \sim \mathcal{N}(0, K)$ . Then

$$\frac{1}{2} \log(2\pi e)^n |K| = \mathbf{h}(X_1, X_2, \dots, X_n) \leq \sum \mathbf{h}(X_i) = \sum_{i=1}^n \frac{1}{2} \log 2\pi e |K_{ii}|$$

with equality iff  $X_1, X_2, \dots, X_n$  are independent (i.e.,  $K_{ij} = 0, i \neq j$ )

- $\log |K|$  is concave
- $\log(|K_n|/|K_{n-p}|)$  is concave in  $K_n$
- $|K_n|/|K_{n-1}|$  is concave in  $K_n$

- A general technique to deal with nonnegative definite symmetric matrix  $K$
- Ref. Ch. 17.9, 17.10, Cover

# Balanced Information Inequality

Differences between inequalities of the discrete entropy and differential entropy

- Both  $H(X, Y) \leq H(X) + H(Y)$  and  $h(X, Y) \leq h(X) + h(Y)$  are valid
- $H(X, Y) \geq H(X)$  but neither  $h(X, Y) \geq h(X)$  nor  $h(X, Y) \leq h(X)$  is valid

Take  $H(X, Y, Z) \leq \frac{1}{4}H(X) + \frac{1}{2}H(Y, Z) + \frac{3}{4}H(Z, X)$  for example.

Count the weights of random variables  $X, Y, Z$  in both sides

$$X: 1, 1; Y: 1, \frac{1}{2}; Z: 1, \frac{5}{4}$$

The net weights of  $X, Y, Z$  are  $0, \frac{1}{2}, -\frac{1}{4}$

**Balanced:** If the net weights of  $X, Y, Z$  are all zero.

$$h(X, Y) \leq h(X) + h(Y) \text{ and } h(X, Y, Z) \leq \frac{1}{2}h(X, Y) + \frac{1}{2}h(Y, Z) + \frac{1}{2}h(Z, X)$$

Let  $[n] := \{1, 2, \dots, n\}$ . For any  $\alpha \subseteq [n]$ , denote  $(X_i: i \in \alpha)$  by  $X_\alpha$ . For example,  $\alpha = \{1, 3, 4\}$ , we denote  $X_1, X_3, X_4$  by  $X_{\{1, 3, 4\}}$  for simplicity.

- We could write any information inequality in the form  $\sum_{\alpha} w_{\alpha} H(X_{\alpha}) \geq 0$  or  $\sum_{\alpha} w_{\alpha} h(X_{\alpha}) \geq 0$ .
- An information inequality is called **balanced** if for any  $i \in [n]$ , the net weight of  $X_i$  is zero.
- The linear continuous inequality  $\sum_{\alpha} w_{\alpha} h(X_{\alpha}) \geq 0$  is valid if and only if its corresponding **discrete counterpart**  $\sum_{\alpha} w_{\alpha} H(X_{\alpha}) \geq 0$  is valid and balanced.

Ref: Balanced Information Inequalities, T. H. Chan, IEEE Transactions on Information Theory, Vol. 49, No. 12, December 2003

# Han's Inequality

- Let  $(X_1, X_2, \dots, X_n)$  have a density, and for **every**  $S \subseteq \{1, 2, \dots, n\}$ , denoted by  $X(S)$  the subset  $\{X_i : i \in S\}$ . Let

$$h_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S: |S|=k} \frac{h(X(S))}{k}$$

$$g_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S: |S|=k} \frac{h(X(S)|X(S^c))}{k}$$

- When  $n = 3$ ,

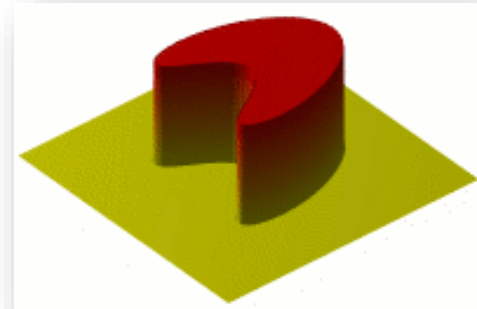
$$\begin{aligned} h_1^{(3)} &= \frac{H(X_1) + H(X_2) + H(X_3)}{3} \geq h_2^{(3)} = \frac{H(X_1, X_2) + H(X_2, X_3) + H(X_3, X_1)}{6} \\ &\geq h_3^{(3)} = H(X_1, X_2, X_3)/3 \\ g_1^{(3)} &= \frac{H(X_1|X_2, X_3) + H(X_2|X_1, X_3) + H(X_3|X_1, X_2)}{3} \\ &\leq g_2^{(3)} = \frac{H(X_1, X_2|X_3) + H(X_2, X_3|X_1) + H(X_3, X_1|X_2)}{6} \\ &\leq g_3^{(3)} = H(X_1, X_2, X_3)/3 \end{aligned}$$

**Han's inequality:**

$$h_1^{(n)} \geq h_2^{(n)} \dots \geq h_n^{(n)} = H(X_1, X_2, \dots, X_n)/n = g_n^{(n)} \geq \dots \geq g_2^{(n)} \geq g_1^{(n)}$$



# Information of Heat



- Heat equation (**Fourier**): Let  $x$  be the position and  $t$  be the time,

$$\frac{\partial}{\partial t} f(x, t) = \frac{1}{2} \frac{\partial^2}{\partial x^2} f(x, t)$$

- Let  $X$  be any random variable with a density  $f(x)$ . Let  $Z$  be an independent normally distributed random variable with zero mean and unit variance,  $Z \sim \mathcal{N}(0,1)$ . Let

$$Y_t = X + \sqrt{t}Z$$

The **probability density function**  $f(y; t)$  ( $f(y; t)$  is a function in  $y$ , not  $t$ ) of  $Y_t$  **satisfies heat equation**

$$f(y; t) = \int f(x) \frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}} dx$$

Gaussian channel  $\leftrightarrow$  Heat equation

# Entropy and Fisher Information

**Fisher information:** Let  $X$  be any random variable with density  $f(x)$ . Its Fisher information is given by

$$I(X) = \int_{-\infty}^{+\infty} f(x) \left[ \frac{\frac{\partial}{\partial x} f(x)}{f(x)} \right]^2 dx$$

- Let  $X$  be any random variable with a density  $f(x)$ . Let  $Z$  be an independent normally distributed random variable with zero mean and unit variance. Let  $Y_t = X + \sqrt{t}Z$

$$\frac{\partial}{\partial t} h(Y_t) = \frac{1}{2} I(Y_t)$$

- Let  $f(y, t)$  (or  $f$ ) be the p.d.f of  $Y_t$

$$\frac{\partial}{\partial t} h(Y_t) = \frac{1}{2} I(Y_t) = \frac{1}{2} \int \frac{f_y^2}{f} dy \geq 0$$

$$\frac{\partial^2}{\partial t^2} h(Y_t) = -\frac{1}{2} \int f \left( \frac{f_{yy}}{f} - \frac{f_y^2}{f^2} \right)^2 dy \leq 0$$

- When  $X$  is Gaussian  $\mathcal{N}(0,1)$ ,

$$h(Y_t) = \frac{1}{2} \log 2\pi e(1 + t)$$

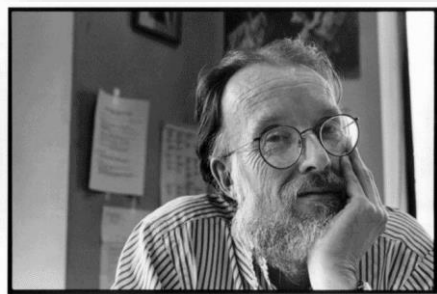
All the derivatives alternate in signs: +, -, +, -, ...

# Higher Order Derivatives of $h(Y_t)$

**(Cheng 2015)** Let  $X$  be any random variable with a density  $f(x)$ . Let  $Z$  be an independent normally distributed random variable with zero mean and unit variance. Let  $Y_t = X + \sqrt{t}Z$  and  $f(y, t)$  (or  $f$ ) be the p.d.f of  $Y_t$ . Then

$$\frac{\partial^3}{\partial t^3} h(Y_t) \geq 0 \text{ and } \frac{\partial^4}{\partial t^4} h(Y_t) \leq 0$$

Conjecture: When  $n$  is even,  $\frac{\partial^n}{\partial t^n} h(Y_t) \leq 0$ , otherwise  $\frac{\partial^n}{\partial t^n} h(Y_t) \geq 0$



“This suggests that....., etc., but I could not prove it” (1966)

H. P. McKean

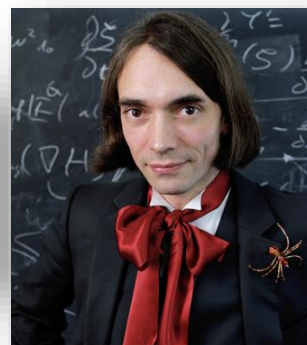
A review of mathematical topics in collisional kinetic theory

Cédric Villani

completed: October 4, 2001

revised for publication: May 9, 2002

most recent corrections: June 7, 2006



C. Villani

2010 Fields medalist

$$\frac{\partial^3}{\partial t^3} h(Y_t) = \frac{1}{2} \int f \left( \frac{f_3}{f} - \frac{f_1 f_2}{f^2} + \frac{1}{3} \frac{f_1^3}{f^3} \right)^2 + \frac{f_1^6}{45 f^5} dy$$

$$\begin{aligned} \frac{\partial^4}{\partial t^4} h(Y_t) &= -\frac{1}{2} \int f \left( \frac{f_4}{f} - \frac{6}{5} \frac{f_1 f_3}{f^2} - \frac{7}{10} \frac{f_2^2}{f^2} + \frac{8}{5} \frac{f_1^2 f_2}{f^3} - \frac{1}{2} \frac{f_1^4}{f^4} \right)^2 \\ &\quad + f \left( \frac{2}{5} \frac{f_1 f_3}{f^2} - \frac{1}{3} \frac{f_1^2 f_2}{f^3} + \frac{9}{100} \frac{f_1^4}{f^4} \right)^2 \\ &\quad + f \left( -\frac{4}{100} \frac{f_1^2 f_2}{f^3} + \frac{4}{100} \frac{f_1^4}{f^4} \right)^2 \\ &\quad + \frac{1}{300} \frac{f_2^4}{f^3} + \frac{56}{90000} \frac{f_1^4 f_2^2}{f^5} + \frac{13}{70000} \frac{f_1^8}{f^7} dy \end{aligned}$$

Ref: F. Cheng and Y. Geng, “Higher Order Derivatives in Costa's Entropy Power Inequality”

# EPI and FII

(**Shannon 1948, Entropy power inequality (EPI)**) If  $X$  and  $Y$  are independent random  $n$ -vectors with densities, then

$$e^{\frac{2}{n}h(X+Y)} \geq e^{\frac{2}{n}h(X)} + e^{\frac{2}{n}h(Y)}$$

$$e^{2h(X+Y)} \geq e^{2h(X)} + e^{2h(Y)} \quad (n = 1)$$

- Fisher information inequality (FII)

$$\frac{1}{I(X+Y)} \geq \frac{1}{I(X)} + \frac{1}{I(Y)}$$

- Most profound result in Shannon's 1948 paper
- EPI can imply some very fundamental results
  - Uncertainty principle
  - Young's inequality
  - Nash's inequality
  - Cramer-Rao bound

## Reference

- T. Cover, "Information theoretic inequalities," 1990
- O. Rioul, "Information Theoretic Proofs of Entropy Power Inequalities," 2011

# Summary

Cover: 8.9, 12.1, 17.6, 17.7, 17.9, 17.10