

# ECE253/CSE208 Introduction to Information Theory

## Lecture 4: Convexity and Inequalities

Dr. Yu Zhang

ECE Department

University of California, Santa Cruz

- Chap 2 of *Elements of Information Theory (2nd Edition)* by Thomas Cover & Joy Thomas

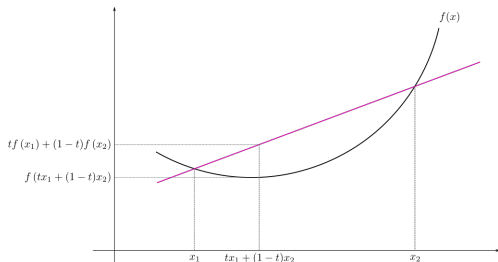
# Convex Functions

## Definition (Convexity of functions)

- A function  $f(x)$  is *convex* over an interval  $(a, b)$  if and only if

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

holds for any  $x_1, x_2 \in (a, b)$  and  $\lambda \in [0, 1]$ .



**Figure:** Any *chord* of a convex function is always above the function itself: The 0th-order condition of convexity.

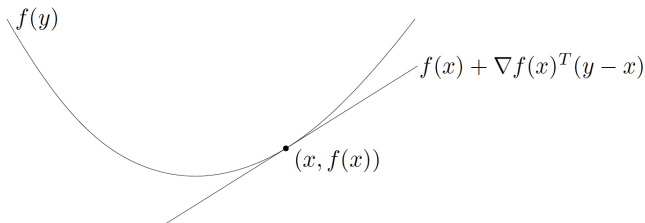
# Convex Functions

## Definition (Convexity of functions)

- A function  $f(\cdot)$  that is *differentiable everywhere* in  $(a, b)$  is *convex* if and only if

$$f(y) \geq f(x) + f'(x)(y - x)$$

for any  $x, y \in (a, b)$ .



**Figure:** Tangent lines are always global under-estimator of the function: The 1st-order condition of convexity.

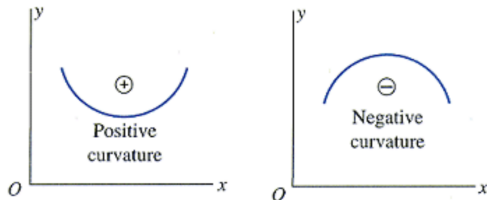
# Convex Functions

## Definition (Convexity of functions)

- A function  $f(x)$  that is *twice differentiable* over  $(a, b)$  is *convex* if and only if

$$f''(x) \geq 0$$

for any  $x \in (a, b)$ .



**Figure:** Convex functions always curve upward (positive curvature).: The 2nd-order condition of convexity.

**Examples.** Convex functions:  $ax + b$ ,  $|x|$ ,  $x^2$ ,  $x^4$ ,  $e^{\pm x}$ ,  $x \log x$ .

If  $f(x)$  is *convex*, then  $-f(x)$  is *concave*. Affine functions are both convex and concave.

# Convexity in High-dimensional Spaces

## Definition (Convex function)

A function  $f(\mathbf{x}) : \mathbb{R}^n \mapsto \mathbb{R}$  is *convex* iff for any  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ , we have one of the following:

- 0th-order condition:  $f(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y})$  for any  $\lambda \in [0, 1]$ .
  - 1st-order condition:  $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$
  - 2nd-order condition:  $\mathbf{H}_f := \nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$ ; i.e., the *Hessian* matrix is positive semi-definite (all eigenvalues are nonnegative), where  $\mathbf{H}_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$ .
- 
- Strictly convex: if strict inequality always holds when  $\mathbf{x} \neq \mathbf{y}$  and  $\lambda \in (0, 1)$ .
  - Strongly convex:  $\mathbf{H}_f \succeq a\mathbf{I}$  for some constant  $a > 0$  (the Hessian is positive definite).
  - Geometrically, strict/strong convexity implies that the function has no flat part and curves upward everywhere  $\implies$  unique minimizer.

# Convex Sets

## Definition (Convex sets)

A set  $S \subseteq \mathbb{R}^n$  is convex if and only if  $\lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2 \in S, \forall \mathbf{x}_1, \mathbf{x}_2 \in S, \lambda \in [0, 1]$ .

Geometrically, a convex set contains line segment between any two points in the set.

Convex sets are solid body without holes and curve outward.

**examples** (one convex, two nonconvex sets)

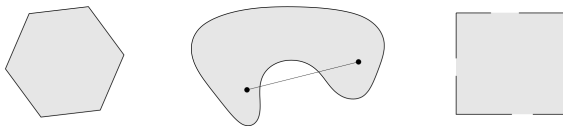


Figure: Convex and nonconvex sets (source: Stephen Boyd, Stanford).

## Epigraph and Sublevel Set

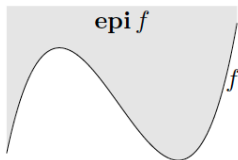
- $\alpha$ -**sublevel set** of  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ :

$$C_\alpha = \{\mathbf{x} \in \text{dom} f : f(\mathbf{x}) \leq \alpha\}$$

sublevel sets of convex functions are convex (converse is false)

- **epigraph** of  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ :

$$\text{epi} f = \{(\mathbf{x}, t) \in \mathbb{R}^{n+1} : \mathbf{x} \in \text{dom} f, f(\mathbf{x}) \leq t\}$$



$f$  is a convex function  $\iff$   $\text{epi} f$  is a convex set

# Convex Optimization Problem

Convex optimization problem in standard form:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{s.to} \quad & f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & \mathbf{a}_i^\top \mathbf{x} = b_i, \quad i = 1, \dots, p \end{aligned}$$

- $f_0, f_1, \dots, f_m$  are convex
- equality constraints are affine (alternatively  $\mathbf{Ax} = \mathbf{b}$ )

**Important properties:**

1. For convex problems, any local solution is also global.
2. If  $f_0()$  is strictly convex, the minimizer is unique.
3. The optimal set  $X_{\text{opt}}$  is convex.



# Jensen's Inequality

## Lemma (Jenson's Inequality)

*If  $X$  is a random variable and  $f(\cdot)$  is a convex function, then*

$$E(f(X)) \geq f(E(X))$$

*Moreover, if  $f(X)$  is strictly convex, equality implies  $X = E(X)$  with probability 1.*

## Jensen's Inequality (cont'd)

### Proof:

1) For a *two-point distribution*  $X \in \{x_1, x_2\}$ , the convexity of  $f(\cdot) \implies$

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2).$$

2) Proof by *induction*: Assume Jensen's inequality holds for a  $(k - 1)$ -mass point distribution. To show the inequality holds for a  $k$ -mass point distribution, define

$p'_i = \frac{p_i}{1-p_k}$  for all  $i = 1, 2, \dots, k - 1$ :

$$\begin{aligned} \sum_{i=1}^k p_i f(x_i) &= p_k f(x_k) + (1 - p_k) \sum_{i=1}^{k-1} p'_i f(x_i) \\ &\geq p_k f(x_k) + (1 - p_k) f\left(\sum_{i=1}^{k-1} p'_i x_i\right) \\ &\geq f\left(p_k x_k + (1 - p_k) \sum_{i=1}^{k-1} p'_i x_i\right) = f\left(\sum_{i=1}^k p_i x_i\right) \end{aligned}$$

where the 1st inequality is from the induction while the 2nd inequality is due to the convexity of  $f(\cdot)$ .

## Gibbs' Inequality (Information Inequality)

### Theorem (Gibbs' Inequality)

Let  $p(x)$  and  $q(x)$  be two probability mass functions. Then,  $D(p||q) \geq 0$  with equality if and only if  $p(x) = q(x)$  for all  $x$ .

**Proof:**  $-D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{q(x)}{p(x)} \leq \log \left( \sum_{x \in \mathcal{X}} p(x) \frac{q(x)}{p(x)} \right) = \log \left( \sum_{x \in \mathcal{X}} q(x) \right) = 0$ .

### Corollary (Nonnegativity of mutual information)

For any two random variables  $X$  and  $Y$ , we have  $I(X;Y) \geq 0$  with equality iff  $X$  and  $Y$  are independent.

### Corollary (Conditional mutual Information)

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

with equality iff  $X$  and  $Y$  are conditionally independent given  $Z$ , which is denoted as  $(X \perp\!\!\!\perp Y) \mid Z$ .

## Gibbs' Inequality (cont'd)

### Theorem (Conditioning reduces entropy (information cannot hurt))

$H(X|Y) \leq H(X)$  with equality iff  $X \perp\!\!\!\perp Y$ .

Intuitively, knowing  $Y$  can only reduce the uncertainty in  $X$ . Note that this is true only on the average (expectation) sense. That is,  $H(X|Y = y) > H(X)$  can happen.

### Theorem (Uniform distribution has the maximum entropy)

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

where  $|\mathcal{X}|$  is the cardinality of the set  $\mathcal{X}$  (i.e., the number of elements in the set) with equality iff  $X$  has a uniform distribution over  $\mathcal{X}$ .

**Proof:** Let  $u(x) = \frac{1}{|\mathcal{X}|}$  be the uniform PMF, and  $p(x)$  be the PMF of  $X$  over  $\mathcal{X}$ , respectively. Then,  $D(p||u) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{u(x)} = \log |\mathcal{X}| - H(X) \geq 0$ .

## Gibbs' Inequality (cont'd)

### Theorem (Independence bound on entropy)

Let RVs  $\{X_i\}_{i=1}^n$  be drawn from  $p(x_1, \dots, x_n)$ . Then,  $H(X_1, \dots, X_n) \leq \sum_{i=1}^n H(X_i)$  with equality iff  $\{X_i\}_{i=1}^n$  are independent.

**Proof:** By chain rule:  $H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) \leq \sum_{i=1}^n H(X_i)$ .

### Lemma (Log-sum inequality)

For nonnegative numbers  $a_1, \dots, a_n$  and  $b_1, \dots, b_n$ , we have

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

with equality iff  $\frac{a_i}{b_i} = \text{constant}$ .

**Proof:** using convexity of  $f(t) = t \log t$  ( $f''(t) > 0$ ); see page 31 of Cover's book.

**Example:**  $a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2} \geq (a_1 + a_2) \log \left( \frac{a_1 + a_2}{b_1 + b_2} \right)$ .

# Convexity of Information Measures

## Theorem (Concavity of entropy)

$H(\mathbf{p})$  is a concave function of  $\mathbf{p}$ , where  $\mathbf{p} := [p_1, \dots, p_k]$  denotes an arbitrary  $k$ -point discrete probability mass function.

**Proof 1:** Let  $X_i \sim \mathbf{p}_i (i = 1, 2)$  be two RVs defined on the same set. For any  $\lambda \in [0, 1]$ , define

$$\theta = \begin{cases} 1, & \text{with prob } \lambda \\ 2, & \text{with prob } 1 - \lambda. \end{cases}$$

Hence, the new RV  $X_\theta \sim \lambda \mathbf{p}_1 + (1 - \lambda) \mathbf{p}_2$ .

$$\therefore H(X_\theta) \geq H(X_\theta | \theta) \implies H(\lambda \mathbf{p}_1 + (1 - \lambda) \mathbf{p}_2) \geq \lambda H(\mathbf{p}_1) + (1 - \lambda) H(\mathbf{p}_2).$$

**Proof 2:**  $H(\mathbf{p}) = -\sum_i p_i \log(p_i)$ . Note that each term in the sum is convex and only depends on one  $p_i$ . Hence,  $H(\mathbf{p})$  is concave.

**Proof 3:**  $H(\mathbf{p}) = \log |\mathcal{X}| - D(\mathbf{p} || \mathbf{u})$ , where  $\mathbf{u}$  is the uniform distribution. Since  $D(\mathbf{p} || \mathbf{u})$  is convex in  $\mathbf{p}$ ,  $H(\mathbf{p})$  is concave.

## Convexity of Information Measures (cont'd)

### Theorem (Convexity of relative entropy)

$D(p||q)$  is convex in the pair  $(p, q)$ . That is, if  $(p_1, q_1)$  and  $(p_2, q_2)$  are two pairs of PMFs. Then,  $D(\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2) \leq \lambda D(p_1 || q_1) + (1 - \lambda)D(p_2 || q_2)$  holds for any  $\lambda \in [0, 1]$ .

**Proof:** By log-sum inequality, we have

$$\begin{aligned} & [\lambda p_1(x) + (1 - \lambda)p_2(x)] \log \left( \frac{\lambda p_1(x) + (1 - \lambda)p_2(x)}{\lambda q_1(x) + (1 - \lambda)q_2(x)} \right) \\ & \leq \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1 - \lambda)p_2(x) \log \frac{p_2(x)}{q_2(x)} \end{aligned}$$

Summing this over all  $x \in \mathcal{X}$  completes the proof.

## Convexity of Information Measures (cont'd)

### Theorem (Convexity of mutual information)

*Let  $(X, Y) \sim p(x, y) = p(x)p(y|x)$ .  $I(X; Y)$  is a concave function of  $p(x)$  for fixed  $p(y|x)$  and a convex function of  $p(y|x)$  for fixed  $p(x)$ .*

#### Proof:

Part i) Since  $I(X; Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x)H(Y|X = x)$

- $p(y) = \sum_x p(x, y) = \sum_x p(x)p(y|x) \implies p(y)$  is linear in  $p(x)$  for fixed  $p(y|x)$ .
- $H(Y)$  is concave in  $p(y)$  and hence in  $p(x)$ .
- The second term is linear in  $p(x)$ . Hence, the difference is concave in  $p(x)$ .

Part ii)

- $p(y) = \sum_x p(x)p(y|x) \implies p(y)$  is linear in  $p(y|x)$  for fixed  $p(x)$ .
- Since  $I(X; Y) = D(p(x, y) || p(x)p(y))$  is convex in  $p(x)p(y)$ , and hence convex in  $p(y|x)$  for fixed  $p(x)$ .



# Data Processing Inequality (DPI)

## Definition

Random variables  $X, Y, Z$  are said to form a Markov chain in the  $X \rightarrow Y \rightarrow Z$  if the joint PMF can be written as  $p(x, y, z) = p(x)p(y|x)p(z|y)$ .

**Note:**  $p(x, y, z) = p(x)p(y, z|x) = p(x)p(y|x)p(z|y, x) = p(x)p(y|x)p(z|y)$ .

## Corollary

$X \rightarrow Y \rightarrow Z$  iff  $X$  and  $Z$  are conditionally independent given  $Y$ , denoted as  $(X \perp\!\!\!\perp Z) \mid Y$ . Markovity implies conditional independence since

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x, y)p(z|y)}{p(y)} = p(x|y)p(z|y).$$

## Corollary ( $X \leftrightarrow Y \leftrightarrow Z$ )

$X \rightarrow Y \rightarrow Z \Leftrightarrow Z \rightarrow Y \rightarrow X$ .

## Theorem (Data Processing Inequality (DPI))

If  $X \rightarrow Y \rightarrow Z$  then  $I(X; Y) \geq I(X; Z)$  with equality iff  $I(X; Y|Z) = 0$  (i.e.,  $X \rightarrow Z \rightarrow Y$ ).

**Proof:** By the chain rule, we expand mutual information in two different ways:

$$I(X; Y, Z) = I(X; Y) + I(X; Z|Y) = I(X; Z) + I(X; Y|Z)$$

$$\because (X \perp\!\!\!\perp Z) \mid Y \implies I(X; Z|Y) = 0 \implies I(X; Y) \geq I(X; Z).$$

## Corollary

1. If  $X \rightarrow Y \rightarrow Z$ , then  $H(X|Z) \geq H(X|Y)$ .
2.  $I(X; Y) \geq I(X; g(Y))$  for any function  $g(\cdot)$ ; i.e., post-processing decreases the info.
3. If  $X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n$ , for any  $i, j, k, l$  such that  $1 \leq i \leq j \leq k \leq l \leq n$ , we have  $I(X_i; X_l) \leq I(X_j; X_k)$ .

**Proof:**

$$1. H(X) - H(X|Z) = \boxed{I(X; Z) \leq I(X; Y)} = H(X) - H(X|Y).$$

## DPI (cont'd)

### Corollary

If  $X \rightarrow Y \rightarrow Z$ , then  $I(X; Y|Z) \leq I(X; Y)$ .

Generally, it is possible to have  $I(X; Y|Z) > I(X; Y)$ ; see the following example when  $X, Y, Z$  do not form a Markov chain.

### Example

Consider  $Z = X + Y$  for two i.i.d.  $X, Y \sim \text{Bern}(0.5)$ , Find  $I(X; Y|Z)$ .

$$\begin{aligned} I(X; Y|Z) &= H(X|Z) - H(X|Y, Z) \\ &= \Pr(Z = 0)H(X|Z = 0) + \Pr(Z = 2)H(X|Z = 2) + \Pr(Z = 1)H(X|Z = 1) \\ &= [\Pr(X = 1, Y = 0) + \Pr(Y = 1, X = 0)] \times H(X|Z = 1) \\ &= 2 \times \frac{1}{4} \times H(1/2, 1/2) = 0.5 > I(X; Y) = 0 \end{aligned}$$

## Sufficient Statistics

- A statistic is any quantity computed from values in a sample which is considered for a statistical purpose; i.e., **statistic is a function of a sample**; e.g., sample mean and sample variance. We estimate things all the time. Estimation help us understand the behavior of a large population with a small sample.
- Example 1: we want to know the average height of students in a school district with a population of 10,000. We take a sample of 100 students, measure them and find that the mean height, which is the sample mean, the statistic (estimator).
- Example 2: we survey the people in a neighborhood about their preference of candidate. Using this limited amount of data, we can estimate the likeliness of a candidate to win an election in a larger area.
- How do we know if every bit of information in data has been explored to estimate the parameter.
- A statistic is **sufficient** w.r.t. a statistical model and its associated unknown parameter if “no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter.”

## Sufficient Statistics (cont'd)

Given a family of distributions  $\{f_\theta(x)\}$ , let  $X$  be a sample drawn from a distribution in this family, and  $T(X)$  be any statistic. Hence,  $\theta \rightarrow X \rightarrow T(X) \Rightarrow I(\theta; X) \geq I(\theta; T(X))$ .

### Definition (Sufficient Statistic)

A function  $T(X)$  is a sufficient statistic relative to the family  $\{f_\theta(x)\}$  if  $X$  is independent of  $\theta$  given  $T(X)$  for any distribution on  $\theta$  ( $\theta \perp\!\!\!\perp X \mid T(X)$ ); i.e.,  $\theta \rightarrow T(X) \rightarrow X$  forms a Markov chain, and

$$I(\theta; X) = I(\theta; T(X)).$$

### Implication

1. Given a set  $X$  of i.i.d. data conditioned on an unknown parameter  $\theta$ , a sufficient statistic is a function  $T(X)$  whose value contains all the information needed to compute any estimate of  $\theta$ .
2. A statistics is sufficient for  $\theta$  if it contains all information in  $X$  about  $\theta$ :  
Once we know  $T(X)$ , the remaining randomness in  $X$  does not depend on  $\theta$ .

## Examples of Sufficient Statistic

### Definition (Sufficient Statistic, SS)

Let  $X$  be a random sample from a distribution with parameter  $\theta$ .  $T(X)$  is a sufficient statistic for  $\theta$  if the conditional probability  $\Pr(X \mid T(X) = k)$  does not depend on  $\theta$ .

Note that from Bayesian's view, we also have  $\Pr(\theta \mid X, T(X) = k) = \Pr(\theta \mid T(X) = k)$

### Example

Given i.i.d.  $X_1, \dots, X_n \sim \text{Bern}(\theta)$ . Let  $X := (X_1, \dots, X_n)$ , a sufficient statistic of  $\theta$  is  $T(X) = \sum_{i=1}^n X_i$ .

**Proof:** We need to show that  $\Pr(X)$  is independent of  $\theta$  given  $T(X)$ .

$$\Pr((X_1, \dots, X_n) = (x_1, \dots, x_n)) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{\sum x_i} (1 - \theta)^{(n - \sum x_i)}.$$

$T(X) \sim B(n, \theta) \implies \Pr(T(X) = k) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}$ . Hence, we get

$$\Pr\left((X_1, \dots, X_n) = (x_1, \dots, x_n) \mid \sum_{i=1}^n X_i = k\right) = \begin{cases} 0, & \text{if } \sum_{i=1}^n x_i \neq k \\ \binom{n}{k}^{-1}, & \text{if } \sum_{i=1}^n x_i = k \end{cases}$$

*Example 2:* For  $X_i \sim \text{Unif}(\theta, \theta + 1)$ , a SS for  $\theta$  is:  $T(X) = (\max\{X_i\}, \min\{X_i\})$ .

# Factorization Theorem

## Theorem (Fisher–Neyman factorization theorem)

A statistic  $T(X)$  is sufficient for  $\theta$  if and only if functions  $g$  and  $h$  can be found such that the probability density or mass function can be factorized as

$$f_{\theta}(\mathbf{x}) = h(\mathbf{x})g_{\theta}(T(\mathbf{x})),$$

where  $h(\cdot)$  does not depend on  $\theta$  while  $g(\cdot)$  which does depend on  $\theta$ , depends on  $\mathbf{x}$  *only through*  $T(\mathbf{x})$ .

- The Bernoulli distribution:

$$f_{\theta}(\mathbf{x}) = \theta^{\sum x_i} (1 - \theta)^{(n - \sum x_i)} = \underbrace{\theta^{T(\mathbf{x})} (1 - \theta)^{(n - T(\mathbf{x}))}}_{g_{\theta}(T(\mathbf{x}))} \times \underbrace{1}_{h(\mathbf{x})}$$

- The uniform distribution  $\text{Unif}(\theta, \theta + 1)$ :

$$f_{\theta}(\mathbf{x}) = \prod_{i=1}^n \mathbb{1}_{\{\theta \leq X_i \leq \theta + 1\}} = \underbrace{\mathbb{1}_{\{\theta \leq \min\{X_i\} \leq \max\{X_i\} \leq \theta + 1\}}}_{g_{\theta}(T(\mathbf{x}))} \times \underbrace{1}_{h(\mathbf{x})}$$

## Minimal Sufficient Statistic

- A statistic  $T(X)$  is a minimal sufficient statistic if it is a function of every other sufficient statistic  $U(X)$ . In terms of DPI,  $\theta \rightarrow T(X) \rightarrow U(X) \rightarrow X$ .
- Minimal sufficient statistic is a statistic that has smallest mutual information with  $X$  while having largest mutual information with  $\theta$ . That is,

$$T(X) \in \arg \min_{U(X)} I(X; U(X)) \text{ s.t. } I(\theta; U(X)) = \max_{T'(X)} I(\theta; T'(X))$$

- For every two sample points  $x$  and  $y$ ,  $\frac{f_{\theta}(x)}{f_{\theta}(y)}$  is independent of  $\theta \iff T(x) = T(y)$ . Then,  $T(X)$  is a minimal sufficient statistic for  $\theta$ .
- Minimal sufficient statistic is not unique. Any one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic.
- Minimal sufficient statistic may not even exist.
- Minimal sufficient statistic most efficiently captures (i.e., maximally compresses) the information about  $\theta$  in the sample; see more discussions here<sup>1</sup>.

---

<sup>1</sup><https://bookdown.org/jkang37/stat205b-notes/lecture04.html>



## Fano's Inequality

Consider a Markov chain  $X \rightarrow Y \rightarrow \hat{X}$ . In the context of communication:

- Send  $X$  through a noisy channel that possibly yields the received symbol  $Y \neq X$ .
- Recover  $X$  by post-processing  $Y$  to get an estimate  $\hat{X} = g(Y)$  for some function  $g$ .
- Define the probability of error as  $P_e := \Pr(\hat{X} \neq X)$ .

Fano's inequality: We may estimate  $X$  with small  $P_e$  when  $H(X|Y)$  is small.

For any estimator  $\hat{X}$  such that  $X \rightarrow Y \rightarrow \hat{X}$ , define  $P_e = \Pr(X \neq \hat{X})$ , we have

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y) \implies P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}.$$

If  $\hat{X} \in \mathcal{X}$ , we then have a slightly better upper bound for  $H(X|Y)$ :

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y)^a.$$

---

<sup>a</sup> $H(X|Y)$  was called the 'equivocation', which corresponds to effect of channel noise; i.e., the uncertainty in the channel input from the receiver's perspective. Hence, the inequality relates the equivocation (as an information measure) to the probability of error (as the traditional measure).

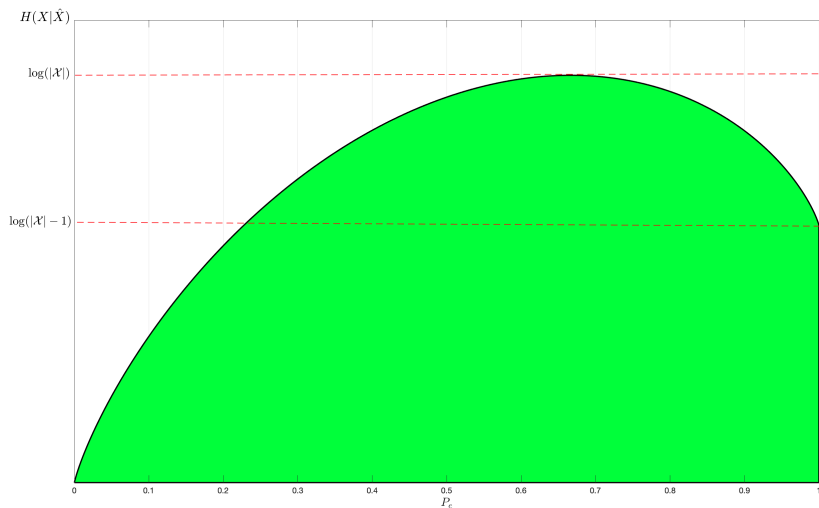
## Fano's Inequality (cont'd)

$$H(X|\hat{X}) \leq H(P_e) + P_e \log(|\mathcal{X}| - 1) \leq H(P_e) + P_e \log |\mathcal{X}|$$

Intuition & implication:

- $H(X|\hat{X})$  is how much uncertainty left in  $X$  after we know its estimate  $\hat{X}$ .  
Intuitively, if  $P_e$  is small, then  $H(X|\hat{X})$  should also be small.
- Coding view:  $H(X|\hat{X})$  quantify how many bits need for measuring  $X$  from  $\hat{X}$ .  
Fano' ineq gives us an upper bound of it. First, use  $H(P_e)$  bits to show if  $\hat{X} = X$ .  
If yes, no more bits needed. If no (with probability  $P_e$ ), we must use  $\log |\mathcal{X}|$  bits to describe  $X$  in the worst case. Furthermore, if  $\hat{X} \in \mathcal{X}$ , then  $X$  can be any one of the other  $|\mathcal{X}| - 1$  symbols in the alphabet.
- Fano's ineq establishes the fundamental limits of data compression and transmission.  
It can be used to characterize when a perfect reconstruction of sent code is not possible, i.e.  $P_e$  is bounded away from zero. This is useful to prove converse results for coding theorems.

## Permissible Region of Fano's Inequality



**Figure:** The boundary (blue curve) of the permissible region (green area) is given by the function:  
 $H(X|\hat{X}) = H(P_e) + P_e \log(|\mathcal{X}| - 1)$ .

## Fano's Inequality Proof

**Proof:** Define  $E = \mathbb{1}_{\{\hat{X} \neq X\}}$ . By chain rule, we have

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + \underbrace{H(E|X, \hat{X})}_{=0} \\ &= \underbrace{H(E|\hat{X})}_{\leq H(P_e)} + \underbrace{H(X|E, \hat{X})}_{\leq P_e \log |\mathcal{X}|} \\ &\leq H(P_e) + \Pr(E=0) \underbrace{H(X|\hat{X}, E=0)}_{=0} + \Pr(E=1)H(X|\hat{X}, E=1) \\ &\leq H(P_e) + P_e \log |\mathcal{X}| \end{aligned}$$

$$\implies H(X|Y) \leq H(X|\hat{X}) \leq H(P_e) + P_e \log |\mathcal{X}|.$$

$$\text{If } X, \hat{X} \in \mathcal{X} \implies H(X|\hat{X}, E=1) \leq P_e \log(|\mathcal{X}| - 1).$$

### Corollary

For any two RVs  $X$  and  $Y$ , let  $p = \Pr(X \neq Y)$ . We have  $H(p) + p \log |\mathcal{X}| \geq H(X|Y)$ .

**Proof:** Let  $\hat{X} = Y$  in Fano's inequality.

# Fano's Inequality is Sharp

## Example

Let  $X \in \{1, 2, \dots, m\}$  and  $p_1 \geq p_2 \geq \dots \geq p_m$ . Then the best guess of  $X$  is  $\hat{X} = 1$  and the resulting probability of error is  $P_e = 1 - p_1$ . Fano's inequality becomes

$$H(P_e) + P_e \log(m-1) \geq H(X).$$

The PMF  $(p_1, p_2, \dots, p_m) = \left(1 - P_e, \frac{P_e}{m-1}, \dots, \frac{P_e}{m-1}\right)$  achieves the lower bound with equality. To see this,

$$\begin{aligned} H(X) &= -(1 - P_e) \log(1 - P_e) - (m-1) \times \frac{P_e}{m-1} \log \frac{P_e}{m-1} \\ &= -(1 - P_e) \log(1 - P_e) - P_e \log P_e + P_e \log(m-1) \\ &= H(P_e) + P_e \log(m-1). \end{aligned}$$

# Entropy and Probability of Errors

## Lemma

Consider two independent RVs defined over the same set:  $X \sim p(x)$ ,  $X' \sim r(x')$ . Then,

$$\Pr(X = X') \geq \max \left\{ 2^{-H(p,r)}, 2^{-H(r,p)} \right\}.$$

**Proof:**  $2^{-H(p,r)} = 2^{\mathbb{E}_p(\log_2 r(X))} \leq \mathbb{E}_p \left( 2^{\log_2 r(X)} \right) = \sum_{x \in \mathcal{X}} p(x)r(x) = \Pr(X = X')$ .

## Corollary

Let  $X, X'$  are i.i.d. with entropy  $H(X)$ . Then,

$$\Pr(X = X') \geq 2^{-H(X)},$$

with equality iif  $X$  has a uniform distribution.

*Thank You!*

Email: <zhangy@ucsc.edu>

Homepage: <https://people.ucsc.edu/~yzhan419/>