Lecture 4: Data-processing, Fano

- Data-processing inequality
- Sufficient statistics
- Fano's inequality

Dr. Yao Xie, ECE587, Information Theory, Duke University

Data processing system

Nature Camera CD

Markovity

- Definition: We say X, Y, Z is a Markov chain in this order, denoted

$$X \to Y \to Z,$$

  if we can write

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

- Special case:

$$X \to Y \to g(Y)$$

- Examples
  - X is binary, you change w.p. p becomes Y, and you further corrupt it and it becomes Z.
  - Bent coin: probability of getting a head is θ. Generate a sequence of independent tosses $X_1, X_2, \cdots$ (Bernoulli(θ) process).

$$\bar{X}_n = \sum_{i=1}^{n} X_n$$

is Markov:

$$\theta \to \{X_1, \cdots, X_n\} \to \bar{X}_n$$

Simple consequences

- $X \to Y \to Z$ iff X and Z are conditionally independent given Y.

  Proof:

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

- This characterization is true for general n-dimensional Markov field.
- Useful for checking Markovity

Best definition of Markovity:

Past and future are conditionally independent given the present.

Reminiscence of

- Quote:

"Yesterday is history. Tomorrow is a mystery. Today is a gift. That's why it is called the present."
– Alice Morse Earle, 1851 – 1911

- If $X \to Y \to Z$ is a Markov chain, then so is $Z \to Y \to X$.

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

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

Data-processing inequality

- No clever manipulation of the data can improve inference

  Theorem. *If* $X \to Y \to Z$, *then the*

  $$I(X;Y) \geq I(X;Z), \quad I(Y;Z) \geq I(X;Z).$$

  Equality iff $I(X;Y|Z) = 0$.

- Discouraging: we process information, then we will loose information

- Encouraging: sometimes we throw away something, equality still holds.

Proof:

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + \underbrace{I(X;Z|Y)}_{0} \end{aligned}$$

Since X and Z are cond. indept. given Y. So

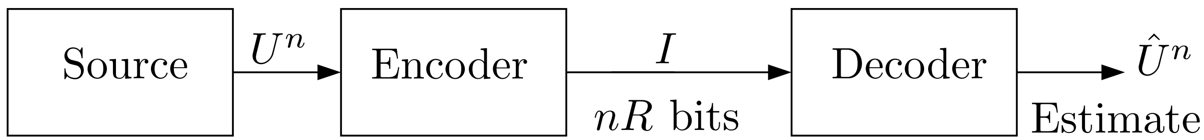
$$I(X;Y) \geq I(X;Z).$$

Equality iff $I(X;Y|Z) = 0$: $X \to Z \to Y$ form a Markov chain. Similarly, can also prove

$$I(Y;Z) \geq I(X;Z).$$

Modeling data-compression systems

Compression system model:



- Encode message W from source using $X^n = (X_1, X_2, \cdots, X_n)$ (sequence of RVs)

- Through a channel, get Y^n,

- Decode to obtain $\hat{W}$.

$$I(W; \hat{W}) \leq I(X; Y).$$

Consequences of data-processing inequality

- Given g, since $X \to Y \to g(Y)$,

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

- If $X \to Y \to Z$,

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

Proof:

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + \underbrace{I(X;Z|Y)}_{0} \end{aligned}$$

  - Dependence of X and Y is decreased (or unchanged) by observing a "downstream" RV Z
  - Counterexample: when X, Y, Z do not form Markov chain, possible $I(X;Y|Z) > I(X;Y)$. X and Y independent coin tosses, $Z = X + Y$. Then $I(X;Y) = 0$, but $I(X;Y|Z) = 1/2$.

Sufficient statistics

- Data-processing inequality clarifies an important idea in statistics - sufficient statistics

- Given a family of distributions $\{f_\theta(x)\}$ indexed by θ

- Let X be sample from f_θ, $T(X)$ be any statistics, then

$$\theta \to X \to T(X)$$

- Data processing inequality

$$I(\theta; T(X)) \leq I(\theta; X)$$

- A statistic is sufficient for θ if it contains all information in X about θ:

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

- Examples: Given $X_1, X_2, \cdots, X_n$, i.i.d. $P(X_i = 1) = \theta$. Sufficient statistic is $T(X_1, \cdots, T_n) = \sum_{i=1}^n X_i$.

$$P\left((X_1, \cdots, X_n) = (x_1, \cdots, x_n) \middle| \sum_{i=1}^n X_i = k\right) = \begin{cases} 1/\binom{n}{k} & \text{if } \sum x_i = k \\ 0 & \text{otherwise.} \end{cases}$$

  Hence $\theta \to \sum X_i \to (X_1, \cdots, X_n)$, $I(\theta; X^n) \leq I(\theta, T)$. Together with data processing inequality: $I(\theta; T) = I(\theta; X^n)$.

- Minimal sufficient statistic is a function of all other sufficient statistic – maximally compresses information about θ in the sample

Fano's inequality

- Fano's inequality (1942) relates P_e to entropy

- Why do we need to relate P_e to entropy $H(X|Y)$? Because when we have a communication system, we send X, receive a corrupted version Y. We want to infer X from Y. Our estimate is $\hat{X}$ and we will make a mistake.

$$P_e = P(\hat{X} \neq X)$$

- Markov: $X \to Y \to \hat{X}$

- Can estimate X from Y with zero probability iff $H(X|Y) = 0$ (Prob. 2.5): only one possible value of y given x (asking native weather).

- Fano's inequality extend this idea: we can estimate X with small P_e if $H(X|Y)$ is small

Theorem. *For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$,*

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

A useful Corollary:

$$P_e \geq \frac{H(X|Y) - 1}{\log|\mathcal{X}|} = \frac{H(X) - I(X;Y) - 1}{\log|\mathcal{X}|}$$

For any two RVs X and Y, if estimator $g(Y)$ takes values in $\mathcal{X}$, we get a slightly stronger inequality

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

Proof of Fano's inequality

- Strategy: we first ignore Y, prove the first inequality; then use data processing inequality:

$$X \to Y \to \hat{X}$$

- Introduce error RV

$$E = \begin{cases} 1, & \text{if } \hat{X} \neq X \\ 0, & \text{if } \hat{X} = X. \end{cases}$$

- Using chain rule to expand $H(E, X|\hat{X})$ in two different ways

$$H(E, X|\hat{X}) = H(X|\hat{X}) + \underbrace{H(E|X, \hat{X})}_{0}$$

$$= \underbrace{H(E|\hat{X})}_{\leq H(E) = H(P_e)} + \underbrace{H(X|E, \hat{X})}_{(*)}$$

$$(*) H(X|E, \hat{X}) = \underbrace{P(E = 0)}_{1 - P_e} \underbrace{H(X|\hat{X}, E = 0)}_{0} + \underbrace{P(E = 1)}_{P_e} H(X|\hat{X}, E = 1)$$

$$= (1 - P_e) \cdot 0 + P_e \underbrace{H(X|\hat{X}, E = 1)}_{\leq H(X)} \leq P_e \log |\mathcal{X}|$$

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

Fano's inequality is sharp

- Suppose there is no knowledge of Y, X must be guessed with only knowledge about its distribution:
  $X \in \{1, \cdots, m\}$, $p_1 \geq \cdots \geq p_m$

- Best guess of X is $\hat{X} = 1$, $P_e = 1 - p_1$

- On the other hand, Fano's inequality

$$H(P_e) + P_e \log(m-1) \geq H(X|X) = H(X),$$

  Left hand side $= -(1-P_e)\log(1-P_e) - P_e \log \frac{P_e}{m-1}$

- Fano's inequality is achieved by $(1 - P_e, \frac{P_e}{m-1}, \cdots, \frac{P_e}{m-1})$

Applications of Fano's inequality

- Prove converse in many theorems (including channel capacity)
- Information theoretic compressed sensing matrix design
- Compressed sensing signal model

$$y = Ax + w$$

  $A \in \mathbb{R}^{M\times d}$: projection matrix for dimension reduction. Signal x is sparse. Want to estimate x from y.

- Find optimal projection matrix $A^\star = \arg\max_A I(x; Ax + w)$.

M. Chen, Bayesian and Information-Theoretic Learning of High Dimensional Data, PhD thesis, Duke University, 2012.

Deviation

Theorem. *If X and X' are i.i.d. with entropy $H(X)$,*

$$P(X = X') \geq 2^{-H(X)}.$$

with equality iff X has uniform distribution.
Proof: Apply Jensen's on $f(x) = 2^x$:

$$2^{-H(X)} = 2^{E \log p(X)} \leq E 2^{\log p(X)} = \sum_x p(x) 2^{\log p(x)} = \sum_x p^2(x) = P(X = X').$$

- $2^{H(X)}$ is the effective alphabet size.
- Corollary: Let X, X' be independent with $X \sim p(x)$, $X' \sim r(x)$, x, $x' \in \mathcal{X}$. Then

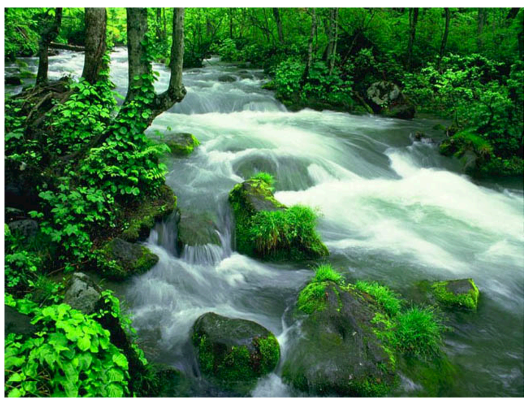
$$P(X = X') \geq 2^{-H(p) - D(p||r)}$$

$$P(X = X') \geq 2^{-H(r) - D(r||p)}$$

- A manifestation of large deviation principle. Can lead to Sanov's Theorem.

Summary

- Data-processing inequality: data processing may (or may not) lose information
- Sufficient statistic preserves information
- When estimate source from observation, error can be bound using Fano's inequality



Nature

Camera

CD

Coin Weighing

Coin weighing strategy for $k = 3$ weighings to find out 1 counterfeit coin in 12 coins?

- n coins, 1 bad (light or heavier), k weighing, possible to tell the bad coin if

$$2n + 1 \leq 3^k \Rightarrow k \geq \log_3(2n + 1)$$

- Information theory interpretation
  - Each weighing result in "lighter", "heavier", "same", $\log_2 3$ bit information
  - Possible state: $2n + 1$, $\log_2(2n + 1)$ bit
  - Need at least the number of weighings

$$k \geq \frac{\log_2(2n + 1)}{\log_2 3} = \log_3(2n + 1)$$

- Express number $-12, \cdots, 12$ in a ternary system with alphabet $-1, 0, 1$
- Negate some columns such that row sums are zero
- Single error correcting Hamming code
- Connection with compressed sensing and group testing

$$y = Ax$$

Weighing strategy

	1	2	3	4	5	6	7	8	9	10	11	12	
3^0	1	-1	0	1	-1	0	1	-1	0	1	-1	0	$\Sigma_1 = 0$
3^1	0	1	1	1	-1	-1	1	0	0	0	-1	-1	$\Sigma_2 = 0$
3^3	0	0	0	0	1	1	-1	1	-1	1	-1	-1	$\Sigma_3 = 0$

$$\log_3(2 \times 12 + 1) = 2.9299$$