

Data processing inequality

(11)

- no clever manipulation of data can improve inferences that can be made from the data.

Def: r.v.s X, Y, Z forms a Markov chain $X \rightarrow Y \rightarrow Z$ if Z is conditionally independent of X given Y , i.e.

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

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

Remark: (i) $X \rightarrow Y \rightarrow Z \Rightarrow Z \rightarrow Y \rightarrow X$

(ii) if $Z = g(Y)$, then $X \rightarrow Y \rightarrow Z$

Data processing inequality (thm 2.8.1)

If $X \rightarrow Y \rightarrow Z$, then $I(X; Y) \geq I(X; Z)$

Proof: $I(X; Y, Z) = I(X; Z) + I(X; Y | Z)$ (chain rule)
 $= I(X; Y) + I(X; Z | Y)$ (chain rule)

Since X, Z are conditionally independent given Y , $I(X; Z | Y) = 0$

Hence $I(X; Z) + I(X; Y | Z) = I(X; Y)$.

But $I(X; Y | Z) \geq 0 \Rightarrow I(X; Z) \leq I(X; Y)$.

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

Cor: $I(X; Y | Z) \leq I(X; Y)$ if $X \rightarrow Y \rightarrow Z$.

Fano inequality

(12)

- gives lower bound on error prob of estimating from a "noisy" version of random variable.

(Theorem 2.11.1):

Let $X \sim p(x)$ and $(X, Y) \sim p(x, y) = p(x) p(y|x)$.

Let $\hat{X} = g(Y)$ for some function g with range \mathcal{X} .

Then $X \rightarrow Y \rightarrow \hat{X}$.

Define prob. of error $P_e \triangleq \Pr \{ \hat{X} \neq X \}$.

Then

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

Remarks: (i) since $H(P_e) = -P_e \log P_e - (1-P_e) \log(1-P_e) \leq 1$ and $\log(|\mathcal{X}| - 1) \leq \log |\mathcal{X}|$, the inequality can be weakened to a more convenient form

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

(ii) $P_e = 0 \Rightarrow H(X|Y) = 0$ (X is ^{essentially} a fn of Y)

Interpretation

X is the r.v. to be estimated

Y is the noisy observation of X .

\hat{X} is the estimate based on observing Y

Note the Fano inequality applies to all estimator of X .

Proof: Define error r.v.

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X \text{ (with prob. } P_e) \\ 0 & \text{if } \hat{X} = X \text{ (with prob. } 1-P_e) \end{cases}$$

Consider $H(E, \hat{X} | Y)$ by applying chain rule in 2 different ways

$$\begin{aligned} H(E, \hat{X} | Y) &= H(\hat{X} | Y) + \underbrace{H(E | \hat{X}, Y)}_{(i)} \\ &= \underbrace{H(E | Y)}_{(ii)} + \underbrace{H(\hat{X} | E, Y)}_{(iii)} \end{aligned}$$

(i) E is a fn of \hat{X} and $\hat{X} = g(Y)$, i.e. a fn of \hat{X} & Y
 $\Rightarrow H(E | \hat{X}, Y) = 0$ (similar to example on page 4)

(ii) $H(E | Y) \leq H(E) = H(P_e)$
 \uparrow conditioning reduces entropy (Thm 2.6.5)

$$\begin{aligned} \text{(iii)} \quad H(\hat{X} | E, Y) &= P_r(E=0) H(\hat{X} | Y, E=0) + P_r(E=1) H(\hat{X} | Y, E=1) \\ &= (1-P_e) \underbrace{H(\hat{X} | Y, E=0)}_{(iv)} + P_e \underbrace{H(\hat{X} | Y, E=1)}_{(v)} \end{aligned}$$

(iv) Given $E=0$, $\hat{X} = g(Y) = X \Rightarrow X$ is fn of Y
 $\Rightarrow H(\hat{X} | Y, E=0) = 0$

$$\text{(v)} \quad H(\hat{X} | Y, E=1) = \sum_y p(y | E=1) H(\hat{X} | Y=y, E=1)$$

$$\begin{aligned} \text{But } H(\hat{X} | Y=y, E=1) &= -\sum_x p(x | Y=y, E=1) \log p(x | Y=y, E=1) \\ &= -\sum_{x \neq \hat{x}} p(x | Y=y, E=1) \log p(x | Y=y, E=1) \end{aligned}$$

since $p(x = \hat{x} | Y=y, E=1) = 0$ where $\hat{x} = g(y)$

Thus $H(\hat{X} | Y=y, E=1) \leq \log(|\mathcal{X}| - 1)$ (Thm 2.6.4)

$$\Rightarrow H(\hat{X} | Y, E=1) \leq \log(|\mathcal{X}| - 1).$$

Asymptotic Equipartition Property

(14)

- + Law of large numbers in information theory.
- useful in many proofs.

AEP: (Thm 3.1.1)

If X_1, X_2, \dots are iid $\sim p(x)$, then

$$-\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \rightarrow H(X) \text{ in prob.}$$

Proof: $-\frac{1}{n} \log p(X_1, X_2, \dots, X_n) = -\frac{1}{n} \sum_{i=1}^n \log p(X_i) \xrightarrow{\text{weak law of large numbers}} -E \log p(X)$

Interpretation: $\Pr\{p(X_1, X_2, \dots, X_n) \approx 2^{-nH(X)}\} \approx 1$ for large n .

Typical set $A_\epsilon^{(n)}$

Def. $A_\epsilon^{(n)}$ w.r.t. $p(x)$ is set of seqs. $(x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ with:

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \dots, x_n) \leq 2^{-n(H(X)-\epsilon)}$$

- seqs that have prob. close to AEP.

Props: (Thm 3.1.2).

(i) If $(x_1, x_2, \dots, x_n) \in A_\epsilon^{(n)}$, then $H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, \dots, x_n) \leq H(X) + \epsilon$

(ii) $\Pr\{A_\epsilon^{(n)}\} \geq 1 - \epsilon$ for sufficiently large n .

(iii) $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$

(iv) $|A_\epsilon^{(n)}| \geq (1-\epsilon) 2^{n(H(X)-\epsilon)}$ for sufficiently large n .

Proof: (iii) $1 = \sum_{x \in \mathcal{X}^n} p(x^n) \geq \sum_{x \in A_\varepsilon^{(n)}} p(x^n)$ (15)

$$\geq \sum_{x \in A_\varepsilon^{(n)}} 2^{-n(H(\mathcal{X})+\varepsilon)} = |A_\varepsilon^{(n)}| 2^{-n(H(\mathcal{X})+\varepsilon)}$$

(iv) For sufficient large n , $1 - \varepsilon \leq \Pr\{A_\varepsilon^{(n)}\} = \sum_{x \in A_\varepsilon^{(n)}} p(x^n)$

$$\leq \sum_{x \in A_\varepsilon^{(n)}} 2^{-n(H(\mathcal{X})-\varepsilon)}$$

$$= |A_\varepsilon^{(n)}| 2^{-n(H(\mathcal{X})-\varepsilon)}$$

Interpretation: (i) A randomly picked a sequence will be in the typical set with a very high probability.

(iii) There are approximately $2^{nH(\mathcal{X})}$ sequences in the typical set.

An application of AEP: Data compression

Want to develop an efficient representation of sequences of r.v.s.

Result: (Thm 3.2.1)

Let \mathcal{X}^n be i.i.d. $\sim p(x)$. Let $\varepsilon > 0$. Then there exists a code which maps sequences x^n of length n into binary strings st. the mapping is 1-to-1 and

$$E\left[\frac{1}{n} l(x^n)\right] \leq H(\mathcal{X}) + \varepsilon \quad \text{for sufficiently large } n,$$

where $l(x^n)$ = length of binary sequence that represents x^n

Interpretation: Can represent \mathcal{X}^n using $nH(\mathcal{X})$ bits on average.

Proof: Divide all seqs. in \mathcal{X}^n into 2 sets

(16)

- (i) those in $A_\varepsilon^{(n)}$
- (ii) those in $A_\varepsilon^{(n)c}$

Order seqs in each set according to, say, lexicographic order and represent each seq in each set by the index of seq in the set. Add an additional bit to determine the set, say 0 for $A_\varepsilon^{(n)}$ & 1 for $A_\varepsilon^{(n)c}$.

Since $|A_\varepsilon^{(n)}| \leq 2^{n(H(\mathcal{X})+\varepsilon)}$, indexing of $A_\varepsilon^{(n)}$ needs no more than $n(H(\mathcal{X})+\varepsilon) + 1$ bits. $\Rightarrow l(x^n) \leq n(H(\mathcal{X})+\varepsilon) + 2$

Indexing of \mathcal{X}^n needs no more than $n \log |\mathcal{X}| + 1$ bits. $\Rightarrow l(x^n) \leq n \log |\mathcal{X}| + 2$ for $x^n \in A_\varepsilon^{(n)c}$

For a sufficiently large n , $\Pr\{A_\varepsilon^{(n)}\} \geq 1 - \varepsilon$.

$$\begin{aligned}
 E[l(\mathcal{X}^n)] &= \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n) l(x^n) + \sum_{x^n \in A_\varepsilon^{(n)c}} p(x^n) l(x^n) \\
 &\leq \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n) (n(H(\mathcal{X})+\varepsilon) + 2) + \sum_{x^n \in A_\varepsilon^{(n)c}} p(x^n) (n \log |\mathcal{X}| + 2) \\
 &= \Pr\{A_\varepsilon^{(n)}\} [n(H(\mathcal{X})+\varepsilon) + 2] + (1 - \Pr\{A_\varepsilon^{(n)}\}) (n \log |\mathcal{X}| + 2) \\
 &\leq n(H(\mathcal{X})+\varepsilon) + 2 + \varepsilon (n \log |\mathcal{X}| + 2) \\
 &= n(H(\mathcal{X}) + \varepsilon')
 \end{aligned}$$

where $\varepsilon' = \varepsilon + \frac{2}{n} + \varepsilon \log |\mathcal{X}| + \frac{2\varepsilon}{n}$

Entropy Rate

(17)

- measures how fast the entropy of a seq. of r.v.'s grows.

Def: $\{Z_n\}$ seq. of r.v.'s (stochastic process)

$$H(\mathcal{X}) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} H(Z_1, Z_2, \dots, Z_n)$$

when limit exists. (per-symbol entropy)

Examples (i) iid r.v.s. Z_1, Z_2, \dots

$$H(Z_1, Z_2, \dots, Z_n) = n H(Z_1) \Rightarrow H(\mathcal{X}) = \lim_{n \rightarrow \infty} \frac{1}{n} H(Z_1, \dots, Z_n) = H(Z_1).$$

(ii) independent r.v.s

$$\frac{1}{n} H(Z_1, Z_2, \dots, Z_n) = \frac{1}{n} \sum_{i=1}^n H(Z_i)$$

limit may/may not exist.

$$\underline{\text{Def:}} \quad H'(\mathcal{X}) \triangleq \lim_{n \rightarrow \infty} H(Z_n | Z_{n-1}, \dots, Z_1)$$

(conditional entropy of last symbol)

Thm 4.2.1: If $\{Z_n\}$ stationary r.p., $H(\mathcal{X}) = H'(\mathcal{X})$.

Proof: (i) $H(Z_{n+1} | Z_n, \dots, Z_1) \leq H(Z_{n+1} | Z_n, \dots, Z_2)$ (conditioning reduces entropy)
 $= H(Z_n | Z_{n-1}, \dots, Z_1)$ (stationarity)

i.e., $H(Z_n | Z_{n-1}, \dots, Z_1)$ decreasing and it is bounded below (by 0).

thus $H'(\mathcal{X})$ exists.

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

↑ chain rule

(18)

Cesàro mean: If $a_n \rightarrow a$ and $b_n = \frac{1}{n} \sum_{i=1}^n a_i$, then $b_n \rightarrow a$.

Let $a_n = H(X_n | X_{n-1}, \dots, X_1)$ and $b_n = \frac{1}{n} \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1)$

$$\text{Then } H(\mathcal{X}) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, \dots, X_n) = H'(\mathcal{X})$$

Usage: For stationary r.p., $H(\mathcal{X})$ acts like $H(X)$ for i.i.d.

Entropy Rate of stationary Markov process (Thm 4.24)

Let $\{X_i\}$ be stationary Markov process with stationary distribution μ and transition matrix P ($\mu = \mu P$), then

(stationary Markov process: start an irreducible, aperiodic, ^{time-invariant} Markov process with stationary distribution)

$$H(\mathcal{X}) = - \sum_{i,j} \mu_i P_{ij} \log P_{ij}$$

Proof: $H(\mathcal{X}) \stackrel{\text{stationarity}}{=} H'(\mathcal{X}) = \lim_{n \rightarrow \infty} H(X_n | X_{n-1}, \dots, X_1) \stackrel{\text{Markovity}}{=} \lim_{n \rightarrow \infty} H(X_n | X_{n-1}) = H(X_2 | X_1)$

$$H(X_2 | X_1) = - \sum_i \mu_i \sum_j P_{ij} \log P_{ij}$$

Remark If Markov chain irreducible, aperiodic, stationary distribution exists and will be approached from any initial distribution. Therefore $H(\mathcal{X})$ will still be a good measure

Hidden Markov Model

Let $\{Z_i\}$ be stationary Markov chain

Let $Y_i = \phi(Z_i)$ for some function ϕ .

Since $\{Z_i\}$ stationary, $\{Y_i\}$ stationary.

Thus $H(Y_n | Y_{n-1}, \dots, Y_1) \downarrow \dots H(Y) = H(Y)$

Note that $\{Y_n\}$ may not be a Markov chain.

Can calculate $H(Y)$ by evaluating $H(Y_n | Y_{n-1}, \dots, Y_1)$,

but difficult to tell when to stop.

Want to find a lower bound on $H(Y)$ so can tell when to stop.

Thm 4.4.1 :

(i) $H(Y_n | Y_{n-1}, \dots, Y_1, Z_i) \leq H(Y) \leq H(Y_n | Y_{n-1}, \dots, Y_1)$

(ii) $\lim_{n \rightarrow \infty} H(Y_n | Y_{n-1}, \dots, Y_1, Z_i) = H(Y) = \lim_{n \rightarrow \infty} H(Y_n | Y_{n-1}, \dots, Y_1)$

Proof: (i) Need only to show $H(Y_n | Y_{n-1}, \dots, Y_1, Z_i) \leq H(Y)$

Fix $k \in \{1, 2, 3, \dots\}$,

$$H(Y_n | Y_{n-1}, \dots, Y_1, Z_i) = H(Y_n | Y_{n-1}, \dots, Y_1, Z_i, Z_0, Z_{-1}, \dots, Z_{-k})$$

↑ (Markovity of $\{Z_i\}$)

$$= H(Y_n | Y_{n-1}, \dots, Y_1, Z_i, Z_0, Z_{-1}, \dots, Z_{-k}, Y_0, \dots, Y_{-k})$$

↑ ($Y_i = \phi(Z_i)$)

$$\leq H(Y_n | Y_{n-1}, \dots, Y_1, Y_0, \dots, Y_{-k})$$

(conditioning raises entropy)

$$= H(Y_{n+k} | Y_{n+k-1}, \dots, Y_1)$$

(stationarity of $\{Y_i\}$)

Since true for all k ,

$$H(Y_n | Y_{n-1}, \dots, Y_1, Z_i) \leq \lim_{k \rightarrow \infty} H(Y_{n+k} | Y_{n+k-1}, \dots, Y_1) = H(Y).$$

(ii) Again only need to show $\lim_{n \rightarrow \infty} \{ \underbrace{H(Y_n | Y_{n-1}, \dots, Y_1) - H(Y_n | Y_{n-1}, \dots, Y_1, X_1)}_{= I(X_1; Y_n | Y_{n-1}, \dots, Y_1)} \} = 0$ (20)

Consider $I(X_1; Y_1, \dots, Y_n) \leq H(X_1)$ for all n .

Also $I(X_1; Y_1, \dots, Y_n) = \sum_{i=1}^n I(X_1; Y_i | Y_{i-1}, \dots, Y_1)$
 \uparrow chain rule.

Therefore

$$H(X_1) \geq \lim_{n \rightarrow \infty} I(X_1; Y_1, \dots, Y_n) = \sum_{i=1}^{\infty} I(X_1; Y_i | Y_{i-1}, \dots, Y_1)$$

since $I(X_1; Y_i | Y_{i-1}, \dots, Y_1) \geq 0$.

But this implies that $\lim_{n \rightarrow \infty} I(X_1; Y_i | Y_{i-1}, \dots, Y_1) = 0$.