II.5. Data processing inequality

Def.:

- A "Markov chain" is a (finite or infinite) sequence of random variables $\{X_i\}_{i\in\mathcal{N}}$ for which

$$p(x_n \mid x_{n-1}, \dots, x_1) = p(x_n \mid x_{n-1}) \quad \text{for all } n\in\mathcal{N} \text{ and all } x\text{'s}.$$

- A Markov chain is called "stationary" or "homogeneous" iff

$$p(X_n = a \mid X_{n-1} = b) = p(X_2 = a \mid X_1 = b) \quad \text{for all } n, a, b.$$

remarks:

- Markov chains are often indicated by $X_1 \to X_2 \to X_3 \to \dots$ or, equivalently, $X_1 \leftarrow X_2 \leftarrow X_3 \leftarrow \dots$ (we will write $X_1 - X_2 - X_3 - \dots$)
- The probability distribution characterizing a Markov chain is

$$\begin{aligned} p(x_n, \dots, x_1) &= p(x_n \mid x_{n-1}) \dots p(x_2 \mid x_1)\, p(x_1) && X_1 \to X_2 \to \dots \\ &= p(x_1 \mid x_2) \dots p(x_{n-1} \mid x_n)\, p(x_n) && X_1 \leftarrow X_2 \leftarrow \dots \end{aligned}$$

for $X \to Y \to Z$ this means $p(x,y,z) = \dfrac{p(x,y)\,p(y,z)}{p(y)} \quad \forall x,y,z$

Prop.: X, Y, Z form a Markov chain $X - Y - Z$ iff $I(X:Z\mid Y) = 0$.

proof: $I(X:Z\mid Y) = \sum_{x,y,z} p(x,y,z) \log\left[\dfrac{p(x,y,z)\,p(y)}{p(x,y)\,p(y,z)}\right]$ □

Lemma: If $Z = f(Y)$ for some $f:\mathbb{R}\to\mathbb{R}$, then $X - Y - Z$ is a Markov chain.

proof: $p(x \mid y,z) = p(x\mid y)$ □

Thm.: ("data processing inequality")

If $X - Y - Z$ is a Markov chain, then $H(X|Y) \leq H(X|Z)$ and

$$\boxed{I(X:Y,Z) = I(X:Y) \geq I(X:Z)}$$

proof: using chain rules in two different ways we get:

$$I(X:Y,Z) = \begin{cases} I(X:Z) + \overbrace{I(X:Y|Z)}^{=0} \\ I(X:Y) + \underbrace{I(X:Z|Y)}_{=0} \end{cases}$$

□

interpretation:
- Z contains no more information about X than Y does.
- processing information (about X) cannot increase it.

II.6. Fano's inequality

Quantitative version of: "if Y allows to estimate X well, then $H(X|Y)$ is small."

For random variables X, Y we define $p_e := p(Y \neq X)$

$h(p_e) := H((p_e, 1-p_e))$ "binary entropy"

Thm.: (Fano's inequality) If X, Y are random variables with range$(X) = \mathcal{X}$. Then

$$\boxed{h(p_e) + p_e \log|\mathcal{X}| \geq H(X|Y)}$$

proof: define a random variable $E := \begin{cases} 1, \text{if } Y \neq X \\ 0 \end{cases}$

from the chain rule we obtain:

$$H(E,X|Y) = \underbrace{H(X|Y)}_{\leq h(p_e)} + \underbrace{H(E|X,Y)}_{=0^{(i)}} = \underbrace{H(E|Y)}_{\leq h(p_e)^{(ii)}} + \underbrace{H(X|E,Y)}_{\leq p_e \log|\mathcal{X}|^{(iii)}}$$

(i) E is a function of X and Y $\Rightarrow$ $H(E|X,Y) = 0$

(ii) $H(E|Y) \leq H(E) = h(p_e)$

$\uparrow$ $I(E;Y) \geq 0$

(iii) $H(X|E,Y) = \underbrace{p(E=0)}_{=p_e} \underbrace{H(X|E=0,Y)}_{\leq H(X) \leq \log|\mathcal{X}|} + p(E=1) \underbrace{H(X|E=1,Y)}_{=0}$

□

remark: if $\text{range}(Y) = \text{range}(X)$, we can replace $|\mathcal{X}|$ by $|\mathcal{X}-1|$.
In particular:

Corollary: If $\text{range}(Y) = \text{range}(X) = \{0,1\}$, then

$$\boxed{h(p_e) \geq H(X|Y)}$$

Corollary: Let $X = (X_1, \dots, X_n)$ describe a random n-bit string, Y a random variable, $\{f_i : \mathbb{R} \to \mathbb{R}\}_{i=1}^n$ and $p_e := \frac{1}{n} \sum_{i=1}^n p(X_i \neq f_i(Y))$ the "average bit-error rate". Then

$$\boxed{h(p_e) \geq \frac{1}{n} H(X|Y)}$$

proof: → exercise ...

II.7. Entropy rates

Def.: ∘ A "stochastic process" $\{X_i\}_{i\in\mathbb{N}}$ is a sequence of random variables.
It is called "stationary" iff $\forall n\in\mathbb{N}$

$p(X_{l+1}=x_1,\dots,X_{l+n}=x_n)$ is independent of $l\in\mathbb{N}_0$ for all x's.

∘ The "entropy rates" of a stochastic process are defined as

$$H(\{X_i\}) := \lim_{n\to\infty} \frac{1}{n} H(X_1,\dots,X_n),$$

$$H'(\{X_i\}) := \lim_{n\to\infty} H(X_n \mid X_{n-1},\dots,X_1)$$

if the limits exist.

Thm.: For a stationary stochastic process the entropy rates both exist and

$$\boxed{H(\{X_i\}) = H'(\{X_i\})}$$

proof:

$$H(X_{n+1} \mid X_n,\dots,X_1) \overset{\text{strong sub-additivity}}{\leq} H(X_{n+1} \mid X_n,\dots,X_2)$$

$$\overset{\text{stationarity}}{=} H(X_n \mid X_{n-1},\dots,X_1)$$

→ H' exists since $H(X_n \mid X_{n-1},\dots,X_1)$ is a non-increasing and non-negative sequence.

The chain rule implies:

$$\lim_{n\to\infty} \frac{1}{n} H(X_n,\dots,X_1) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i \mid X_{i-1},\dots,X_1)$$

$$= \lim_{n\to\infty} H(X_n \mid X_{n-1},\dots,X_1)$$

Lemma (Cesaro mean): $a_n \to a \Rightarrow \frac{1}{n}\sum_{i=1}^{n} a_i \to a$

Corollary: For a stationary Markov chain $X_1 - X_2 - \dots$ we have

$$H(\{X_i\}) = H'(\{X_i\}) = H(X_2 | X_1)$$

proof: $H(X_n | X_{n-1}, \dots, X_1) = H(X_n | X_{n-1}) = H(X_2 | X_1)$

↑ Markov ↑ stationary

□