Semih Yagli

09/29/2016
Thursday

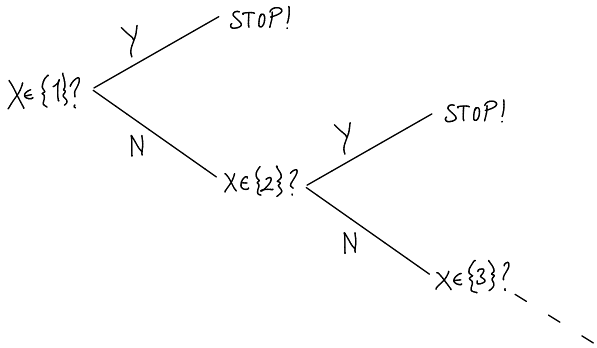
ELE538 - Information Theoretic Security
Homework - 1

2.1

(a) $X \sim Geom(1/2)$ $\quad P_X(k) = \left(\frac{1}{2}\right)^k$, $\quad \mathbb{E}[X] = 2$

$$H(X) = \sum_{k=1} P_X(k) \log_2 2^k = \mathbb{E}[X] = 2 \text{ bits}$$

(b)



Algorithm:

1. Set $i=1$
2. Ask "Is $X \in \{i\}$?". Wait for answer.
3. If the answer is "No" increment i and go to step 2.
   If the answer is "Yes" STOP.

$$\mathbb{E}[\# \text{ of Y/N questions}] = \sum_{k=1}^{\infty} k\left(\tfrac{1}{2}\right)^k = \mathbb{E}[X] = 2 \overset{*}{=} H(X)$$

Note: * suggests that the above scheme of Y/N questions is optimal.

2.4

(a) By chain rule for entropy.
(b) $g(X)$ is a deterministic function of X. Thus, given X, there is nothing random regarding $g(X) \Rightarrow H(g(X)|X)=0$
(c) By chain rule of entropy.
(d) Conditional entropy is non-negative. Equality holds iff g^{-1} is well-defined (i.e, when g is bijection)

2.8

→ Drawing $k \geq 2$ balls with replacement has higher entropy here is why:

Let X_i denote the random variable of drawing a ball from the urn the i-th time w/ replacement, and Y_i denote the random variable of drawing a ball from the urn the i-th time w/out replacement.

Note that $X_1, X_2, \dots, X_k$ are i.i.d random variables, whereas Y_i depends on Y^{i-1} $\forall i \geq 2$. Note further that X_i and Y_i are identically distributed

$$H(X_1, \dots, X_k) = \sum_{i=1}^{k} H(X_i) \quad \text{and} \quad H(Y_1, \dots, Y_k) = \sum_{i=1}^{k} H(Y_i \mid Y^{i-1})$$

Since conditioning reduces entropy, we have $H(Y_i) \geq H(Y_i \mid Y^{i-1})$ $\forall i \geq 2$ and since Y_i and X_i are identically distributed it follows that $H(X_i) = H(Y_i) \geq H(Y_i \mid Y^{i-1})$ ∎

2.9

(a) Let $\rho(X, Y) = H(X \mid Y) + H(Y \mid X)$

(i) $\rho(X, Y) \geq 0$ as conditional entropy is non-negative

(ii) $\rho(X, Y) = \rho(Y, X)$ is clear from the definition

(iii) If equality $X = Y$ means $\exists f$ a bijection such that $f(X) = Y$, then

- when $X = Y$; $H(X \mid f(X)) + H(f(X) \mid X) = 0$ and
- $\rho(X, Y) = 0 \Rightarrow H(X \mid Y) = 0$ and $H(Y \mid X) = 0 \Rightarrow \exists f$, a bijection such that $f(X) = Y$.

(iv) $H(X \mid Y) + H(Y \mid X) + H(Y \mid Z) + H(Z \mid Y) \geq H(X, Y \mid Z) + H(Y, Z \mid X)$ (*)

$\geq H(X \mid Z) + H(Z \mid X)$ (**)

where (*) follows from the followings

$$H(X \mid Y) \geq H(X \mid Y, Z)$$
$$H(Z \mid Y) \geq H(Z \mid Y, X)$$

and (**) follows from the fact that joint entropy is no-less than marginal entropies.
(in other words, (**) follows because $H(Y \mid X, Z) \geq 0$)

(b) Note that $I(X; Y) \overset{(1)}{=} H(X) - H(X \mid Y) \overset{(2)}{=} H(Y) - H(Y \mid X) \overset{(3)}{=} H(X) + H(Y) - H(X, Y)$.

- So $H(X) + H(Y) - 2I(X; Y) = H(X \mid Y) + H(Y \mid X) = \rho(X, Y)$ Hence, (2.172) follows from (1) and (2)

  Note further that $H(X, Y) \overset{(4)}{=} H(X) + H(Y) - I(X; Y)$.
- Using (2.172), and (4), (2.173) is now verified.
- (2.174) follows from (3) and (2.173)

2.10 (a) Let $f(X) = \mathbb{1}\{X = X_1\}$

Note that $H(X) = H(X, f(X)) = H(f(X)) + H(X \mid f(X))$

$$= H(\alpha) + \alpha H(X_1) + (1-\alpha) H(X_2)$$

where $H(\alpha)$ denotes the binary entropy function.

(b)

$$\frac{d}{d\alpha} H(X) = 0 \Rightarrow \log_2 \frac{1-\alpha}{\alpha} + H(X_1) - H(X_2) = 0 \Rightarrow \alpha = \frac{1}{1 + 2^{-(H(X_1) - H(X_2))}}$$

Note that $H(X)$ is a concave function of α. (because $H(\alpha)$ is concave and $\alpha H(X_1) + (1-\alpha) H(X_2)$ is linear w.r.t. α)
As a result, global maximum is achieved at $\alpha = \frac{1}{1 + 2^{-(H(X_1) - H(X_2))}}$

Let $\xi = H(X_1) - H(X_2)$ and $\alpha^* = \frac{1}{1+2^{-\xi}}$ and note that

$$\begin{aligned} H(X) &\leq H(\alpha^*) + \alpha^* H(X_1) + (1-\alpha^*)H(X_2) \\ &= \log_2(1+2^{-\xi}) + (1-\alpha^*)\xi + \alpha^* H(X_1) + (1-\alpha)^* H(X_2) \\ &= \log_2(1+2^{-H(X_1)+H(X_2)}) + H(X_1) = \log(2^{H(X_1)} + 2^{H(X_2)}) \end{aligned}$$

$$\Rightarrow 2^{H(X)-H(X_1)} \leq 1 + 2^{-H(X_1)+H(X_2)} \Rightarrow 2^{H(X)} \leq 2^{H(X_1)} + 2^{H(X_2)}$$

(Note that since $\frac{d}{d\alpha} H(\alpha) = \log \frac{1-\alpha}{\alpha}$ for any arbitrary logarithm base, the following result comes for free

$$b^{H(X)} \leq b^{H(X_1)} + b^{H(X_2)} \quad \forall b > 1$$)

2.17

(a) X_i are i.i.d $Ber(p)$

(b) $H(X_1, \dots, X_n) \geq H(f(X_1, \dots, X_n))$ where $f(X_1, \dots, X_n) = (Z_1, \dots, Z_K)$ with random K

(c) Chain rule: $H(V, W) = H(W) + H(V|W)$

(d) Z_i are i.i.d $Ber(1/2)$ $H(Z_1, \dots, Z_K | K=k) = k$ bits so $H(P_{Z_1 \dots Z_K|K}) = E_K[H(P_{Z_1 \dots Z_K|K}(\cdots|K))] = E[K]$

(e) $H(K) \geq 0$

A good map f on sequences of length 4:

- To maximize the expected number of pure bits, we combine non-pure bits of same probability. Note that in $(X_1, \dots, X_4)$

0000 → occurs with probability $(1-p)^4$

1111 → occurs with probability p^4

$\left.\begin{matrix} 0011 \\ 0101 \\ 1001 \\ 0110 \\ 1010 \\ 1100 \end{matrix}\right\}$ → occurs with probability $p^2(1-p)^2$

$\left.\begin{matrix} 0001 \\ 0010 \\ 0100 \\ 1000 \end{matrix}\right\}$ → occurs with probability $p(1-p)^3$

$\left.\begin{matrix} 1110 \\ 1101 \\ 1011 \\ 0111 \end{matrix}\right\}$ → occurs with probability $p^3(1-p)$

Therefore, let f be the following map:

0000 → Λ

1111 → Λ

$\left.\begin{matrix} 0001 \\ 1110 \\ 1100 \end{matrix}\right\} \to 00$ $\quad \left.\begin{matrix} 0010 \\ 1101 \\ 0101 \end{matrix}\right\} \to 01$ $\quad \left.\begin{matrix} 0100 \\ 1011 \\ 1010 \end{matrix}\right\} \to 10$ $\quad \left.\begin{matrix} 1000 \\ 0111 \\ 0011 \end{matrix}\right\} \to 11$

1001 → 1

0110 → 0

This should be a good map as we utilize every possible outcome of $(X_1, \dots, X_4)$ except that 0000 and 1111 which are the most unlikely/likely sequence (depending on the value of p).

2.37 $D(P_{XYZ} \| P_X P_Y P_Z) = H(X)+H(Y)+H(Z)-H(X,Y,Z)$

The given quantity is equal to 0 iff $P_{XYZ} = P_X P_Y P_Z$, i.e., $D(P_{XYZ} \| P_X P_Y P_Z)=0$ iff X, Y and Z are independent from each other.

$$D(P_{XYZ} \| P_X P_Y P_Z) = E\left[\log \frac{P_{XYZ}(X,Y,Z)}{P_X(X)P_Y(Y)P_Z(Z)}\right] = E\left[\log \frac{P_{XY}(X,Y)}{P_X(X)P_Y(Y)} + \log \frac{P_{Z|XY}(Z|X,Y)\,P_{XY}(X,Y)}{P_Z(Z)\,P_{XY}(X,Y)}\right]$$

$$= I(X;Y) + I(Z;X,Y) \quad (*)$$

[or using similar derivation $(*)$ is equal to: $I(X;Z)+I(Y;X,Z)$ and $I(Y;Z)+I(X;Y,Z)$]

2.41

(a)
$$\begin{aligned}
I(X;Q,A) &= I(X;Q)+I(X;A|Q) && \text{(Chain rule for mutual information)} \\
&= I(X;A|Q) && (X \perp\!\!\!\perp Q) \\
&= H(A|Q) - H(A|Q,X) \\
&= H(A|Q) && (H(A|Q,X)=0 \text{ as } A \text{ is deterministic function of } Q \text{ and } X.)
\end{aligned}$$

Interpretation: The uncertainty removed by (Q,A) pair is the same as the average information of the answer A given the question Q. (Note that information of an outcome x of a discrete r.v. X is given by $i_X(x) = \log \frac{1}{P_X(x)}$.)

(b) $I(X;Q_1,A_1,Q_2,A_2) = I(X;Q_1,A_1) + I(X;Q_2,A_2|Q_1,A_1)$

So it suffices to show that $I(X;Q_2,A_2|Q_1,A_1) \leq H(A_2|Q_2) = I(X;Q_2,A_2)$

In that respect, note that

$$\begin{aligned}
I(X;Q_2,A_2|Q_1,A_1) &= I(X;Q_2|Q_1,A_1) + I(X;A_2|Q_1,A_1,Q_2) && (1) \\
&= I(X;A_2|Q_1,A_1,Q_2) && (2) \\
&= H(A_2|Q_1,A_1,Q_2) - H(A_2|Q_1,A_1,Q_2,X) && (3) \\
&= H(A_2|Q_1,A_1,Q_2) && (4) \\
&\leq H(A_2|Q_2) && (5)
\end{aligned}$$

where (1) follows from chain rule, (2) follows because X does not depend on Q_2, (4) follows because given Q_2 and X, A_2 is deterministic, (5) is due to the fact that further conditioning cannot increase entropy.

3.4 X_i iid $\sim p(x)$ on $\{1,2,\dots,m\}$, $\mu = E[X]$, $H = \sum p(x)\log\frac{1}{p(x)}$

$A_n = \{x^n \in \mathcal{X}^n : |\frac{1}{n}\log\frac{1}{p(x^n)} - H| \leq \varepsilon\}$ and $B^n = \{x^n \in \mathcal{X}^n : |\frac{1}{n}\sum_{i=1}^n x_i - \mu| \leq \varepsilon\}$

(a) $\mathbb{P}[X^n \in A^n] \to 1$ as $n\to\infty$, because of AEP Theorem

(b) $\mathbb{P}[X^n \in A^n \cap B^n] = 1 - \mathbb{P}[X^n \in (A^n)^c \cup (B^n)^c] \geq 1 - (\mathbb{P}[X^n \in (A^n)^c] + \mathbb{P}[X^n \in (B^n)^c])$

$\mathbb{P}[X^n \in (A^n)^c] \to 0$ by AEP Theorem, $\mathbb{P}[X^n \in (B^n)^c] \to 0$ by LLN

It follows that $\mathbb{P}[X^n \in A^n \cap B^n] \to 1$.

(c) $1 \geq \sum_{x^n \in A^n \cap B^n} p(x^n) \geq |A^n \cap B^n| . 2^{-n(H+\varepsilon)} \Rightarrow |A^n \cap B^n| \leq 2^{n(H+\varepsilon)} \quad \forall n$.

↳ By definition of A^n.

(d) Since $\mathbb{P}[X^n \in A^n \cap B^n] \to 1$, for sufficiently large n, we have

$\frac{1}{2} \leq \mathbb{P}[X^n \in A^n \cap B^n] = \sum_{x^n \in A^n \cap B^n} p(x^n) \leq |A^n \cap B^n|\, 2^{-n(H-\varepsilon)} \Rightarrow |A^n \cap B^n| \geq \frac{1}{2} 2^{n(H-\varepsilon)}$ for sufficiently large n.

↳ By definition of A^n

3.10

$$V_n^{1/n} = \left(\prod_{i=1}^n X_i\right)^{1/n} \Rightarrow \frac{1}{n}\log V_n = \frac{1}{n}\sum_{i=1}^n \log X_i \xrightarrow[LLN]{n\to\infty} E[\log X_1] = \int_0^1 \log(x)\,dx = -1$$

$$\lim_{n\to\infty} V_n^{1/n} = \lim_{n\to\infty} e^{\log V_n^{1/n}} = e^{-1} \qquad \text{(because } e^x \text{ is a continuous function)}$$

$$E[V_n] = \int_0^1 \cdots \int_0^1 \int_0^1 (x_1 x_2 \cdots x_n)\,dx_1 dx_2 \cdots dx_n \qquad E[V_n] = (1/2)^n \qquad \lim_{n\to\infty} (E[V_n])^{1/n} = 1/2 > 1/e$$

3.11

(a) $\mathbb{P}[A \cap B] = 1 - \mathbb{P}[A^c \cup B^c] \geq 1 - \mathbb{P}[A^c] - \mathbb{P}[B^c] \geq 1 - \varepsilon_1 - \varepsilon_2$

↳ union bound ↳ $\mathbb{P}[A^c] \leq \varepsilon_1$ and $\mathbb{P}[B^c] \leq \varepsilon_2$

(b)
$$\begin{aligned}
1 - \varepsilon - \delta &\leq \mathbb{P}[A_\varepsilon^{(n)} \cap B_\varepsilon^{(n)}] && \text{(follows from part (a))} \\
&= \sum_{A_\varepsilon^{(n)} \cap B_\varepsilon^{(n)}} p(x^n) && \text{(definition of } \mathbb{P}[A_\varepsilon^{(n)} \cap B_\delta^{(n)}]) \\
&\leq \sum_{A_\varepsilon^{(n)} \cap B_\varepsilon^{(n)}} 2^{-n(H-\varepsilon)} && \text{(follows from the definition of the typical set } A_\varepsilon^{(n)}) \\
&= |A_\varepsilon^{(n)} \cap B_\delta^{(n)}|\, 2^{-n(H-\varepsilon)} && \text{(Term inside summation no longer depends on summation index } x^n) \\
&\leq |B_\delta^{(n)}|\, 2^{-n(H-\varepsilon)} && (A_\varepsilon^{(n)} \cap B_\delta^{(n)}) \subseteq B_\delta^{(n)}
\end{aligned}$$

(c) By part (b), we have $|B_\delta^{(n)}| \geq 2^{n(H-\varepsilon)}(1-\varepsilon-\delta)$, That is for sufficiently large n, $|B_\delta^n| > 2^{n(H-\varepsilon)}$ which is the promised result of Theorem 3.3.1.

3.13

(a) $H(X) = 0.970951$ bits.

(b) $x^{25} \in A_{0.1}^{(25)}$ iff $\quad 0.870951 \leqslant \frac{1}{25} \log \frac{1}{p(x^{25})} \leqslant 1.070951 \quad (*)$

By looking at the given table we see that sequences which have more than 11 one's and less than 19 one's satisfy $(*)$. (This part of the table was correct)

There are $\binom{25}{11}+\binom{25}{12}+\binom{25}{13}+\binom{25}{14}+\binom{25}{15}+\binom{25}{16}+\binom{25}{17}+\binom{25}{18}+\binom{25}{19} = 26366510$ elements on $A_{0.1}^{(25)}$.

Probability of the typical set is $= 0.936246$

(c) Smallest set that has probability 0.9 contains elements with $k \geqslant 13$ one's and 3680638 elements with exactly 12 one's.

So there are 20457854 elements in that set.

(d) The intersection contains 20389448 elements.

The probability of the intersection is roughly ≈ 0.870638