1 Distribution

$\text{Bin}(n,p): P_X(x) = \binom{n}{x} p^x q^{n-x}$ for $x \in [n]_0$.
$\mu = np, \sigma^2 = npq, H(X) = \frac{1}{2}\log(2\pi enpq) + O(\frac{1}{n})$.
$\text{Pois}(\lambda): P_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x \in \mathbb{N}_0$.
$\mu = \sigma^2 = \lambda$.
$\text{Geo}(p): P_X(x) = q^{x-1}p$ for $x \in \mathbb{N}$.
$\mu = \frac{1}{p}, \sigma^2 = \frac{q}{p^2}, H(X) = \frac{-q\log q - p\log p}{p}$.
$\text{Exp}(\lambda): f_X(x) = \lambda e^{-\lambda x}$ for $x \in \mathbb{R}_0^+$.
$\mu = \frac{1}{\lambda}, \sigma^2 = \frac{1}{\lambda}$.
$\mathcal{N}(\mu,\sigma^2): f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.
$h(X) = \frac{1}{2}\log(2\pi e\sigma^2)$.
$\text{Lap}(\mu,b): f_X(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$.
$\sigma^2 = 2b^2, h(X) = \log(2be)$.

2 Markov Chain

$X_1 - X_2 - \cdots - X_n := \forall n, x^n,\ P_{X_{n+1}|X^n}(x_{n+1}|x^n) = P_{X_{n+1}|X_n}(x_{n+1}|x_n)$.
Stationary: $P_{X_1,\ldots,X_n} = P_{X_{1+l},\ldots,X_{n+l}},\ \forall n, l \in \mathbb{N}$.

3 Central Limit Theorem

Khinchin WLLN: $X_1, X_2, \ldots,$ are i.i.d. with $E[|X_i|] < \infty$, then $\forall \epsilon > 0,\ \lim_{n\to\infty} \Pr\{|\bar{X}_n - \mu| \geq \epsilon\} = 0$.
Central limit theorem: $X_1, X_2, \ldots,$ are i.i.d. with $E[|X_i|] < \infty$, then $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\text{d}} Z \sim N(0,1)$.
Berry-Esseen: $X_1, X_2, \ldots,$ are i.i.d. with $E[|X_i - \mu|^3] = \rho_3 < \infty$. Let $Z_n := \frac{S_n - n\mu}{\sqrt{n}\sigma}, Z \sim N(0,1)$. Then $|F_{Z_n}(z) - F_Z(z)| \leq c\frac{\rho_3}{\sigma^3}n^{-1/2},\ \forall z \in \mathbb{R}, n \in \mathbb{N}$ for constant $c \in (0.4, 0.5)$.

4 Representing An i.i.d. Sequence Almost Losslessly

DMS: discrete memoryless source. $\mathcal{B}(n,\epsilon)$ is an ϵ-high-probability set: $\Pr\{S^n \in \mathcal{B}(n,\epsilon)\} \geq 1-\epsilon$
s^n is δ-typical: $|\frac{1}{n}\sum_{i=1}^n \log P_S(s_i) + H(S)| \leq \delta$.
δ-typical set $\mathcal{A}_\delta^{(n)}(S) := \{s^n | s^n \text{ is } \delta\text{-typical}\}$.
Properties of typical sequences and typical sets:

- $\forall s^n \in \mathcal{A}_\delta^{(n)}(S), 2^{-n(H(S)+\delta)} \leq \Pr\{S^n = s^n\} \leq 2^{-n(H(S)-\delta)}$.
- $\Pr\{S^n \in \mathcal{A}_\delta^{(n)}(S)\} \geq 1-\epsilon$ for n large enough.
- $|\mathcal{A}_\delta^{(n)}(S)| \leq 2^{n(H(S)+\delta)}$.
- $|\mathcal{A}_\delta^{(n)}(S)| \geq (1-\epsilon)2^{n(H(S)-\delta)}$ for n large enough.

$s^n \to b^k \to \hat{s}^n$: (n,k) code.
(n,k,ϵ) code: (n,k) code with $P_e^{(n)} := \Pr\{S^n \neq \hat{S}^n\} \leq \epsilon$.
$k^*(n,\epsilon)$: the smallest k s.t. $\exists(n,k,\epsilon)$ code.
$R^*(\epsilon) := \lim_{n\to\infty} \frac{k^*(n,\epsilon)}{n}$.
A lossless source coding theorem for DMS: $R^*(\epsilon) = H(S),\ \forall \epsilon \in (0,1)$.
AEP (Asymptotic Equipartition Property): Entropy determines the asymptotic size of a typical set, and determines the probability of a typical sequence asymptotically.

5 Entropy

$H(X|Y) = \sum_y P_Y(y)H(X|Y=y) = \sum_{x,y} P_{X,Y}(x,y)\log\frac{1}{P_{X|Y}(x,y)}$.
$0 \leq H(X) \leq \log|\mathcal{X}|$, where $H(X) = \log|\mathcal{X}| \iff$

X is uniform distributed over $\mathcal{X}$.

$H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$.

$H(X|Y) \le H(X)$, but $H(X|Y=y)$ may $> H(X)$.

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

$H(X|Y,Z) \le H(X|Y)$.

The above still holds for h.

Exercise 4: $H(X,Y,Z) \le H(X,Y) + H(X,Z) - H(X)$.

Concavity of Entropy: $H(\mathbf{p}) := -\sum_{i=1}^{d} p_i \log p_i$ is concave in $\mathbf{p}$.

That is, $H(\lambda\mathbf{p_1} + (1-\lambda)\mathbf{p_2}) \ge \lambda H(\mathbf{p_1}) + (1-\lambda)H(\mathbf{p_2})$.

Fano's inequality: $H(U|V) \le H_b(P_e) + P_e \log|\mathcal{U}|$, where $P_e := \Pr\{U \ne V\}$.

$\Rightarrow \Pr\{U \ne V\} \ge \dfrac{H(U|V)-1}{\log|\mathcal{U}|}$.

Exercise 5: if U, V both take values in $\mathcal{U}$, then $H(U|V) \le H_b(P_e) + P_e \log(|\mathcal{U}|-1)$.

6 Representing A Sequence with Memory Almost Losslessly

Entropy rate:

- $\mathrm{H}(\{X_i\}) := \lim_{n\to\infty} \frac{1}{n} H(X_1,\dots,X_n)$ if exists.
- $\tilde{\mathrm{H}}(\{X_i\}) := \lim_{n\to\infty} H(X_n|X^{n-1})$ if exists.

H and $\tilde{\mathrm{H}}$ may be different: consider $X_1, X_3, \dots$ are i.i.d. and $X_{2k} = X_{2k-1}$.

If $\{X_i\}$ is stationary, then $H(X_n|X^{n-1})$ is decreasing in n.

If $\{X_i\}$ is stationary, then $\mathrm{H}(\{X_i\}) = \tilde{\mathrm{H}}(\{X_i\})$.

Stationary ergodic processes: $\frac{1}{n}\sum_{l=0}^{n-1} f(X_{k_1+l},\dots,X_{k_m+l}) \overset{\text{a.s.}, L^1}{\to} \mathrm{E}[f(X_{k_1},\dots,X_{k_m})]$ as $n \to \infty$.

Shannon-McMillan-Breiman theorem: if $\{S_i\}$ is stationary ergodic, then $\frac{1}{n}\log\frac{1}{P(S^n)} \overset{\text{a.s.}, L^1}{\to} \mathrm{H}(\{S_i\})$ as $n \to \infty$.

A Lossless Source Coding Theorem for Ergodic DSS: For a discrete stationary ergodic source $\{S_i\}$, $R^*(\epsilon) = H(\{S_i\}) \forall \epsilon \in (0,1)$.

Let $\mathcal{X}$ be the state space of a Markov process.

1. A Markov process is irreducible if $\forall x, y \in \mathcal{X}$, it is possible to reach to start at x and reach y in a finite number of steps.

2. The period of a state is the g.c.d. of the # of times that a state can return to itself. A Markov process is aperiodic if all states have period = 1.

3. A Markov process is homogeneous (or time-invariant) if $\forall n > 1$, $P_{X_n|X_{n-1}} = P_{X_2|X_1}$. Hence, a homogeneous Markov process is completely defined by its initial state distribution P_{X_1} and transition probability $P_{X_2|X_1}$.

4. A steady-state distribution $\pi : \mathcal{X} \to [0,1]$ is one such that the distribution does not change after one transition: $\pi(x) = \sum_{y\in X} \pi(y) P_{X_{n+1}|X_n}(x|y)$, $\forall x \in \mathcal{X}$, $n \in \mathbb{N}$. For a finite-alphabet homogeneous Markov process, steady-state distribution always exists, and it is unique if the process is irreducible.

5. For a finite-alphabet homogeneous Markov process that is both irreducible and aperiodic, $\lim_{n\to\infty} \Pr\{X_{n+1} = y | X_1 = x\} = \pi(y)$, $\forall x, y \in \mathcal{X}$, where $\pi(\cdot)$ is the unique steady-state distribution. If $P_{X_1} = \pi$, the Markov process becomes a stationary process.

For a homogeneous, irreducible, and aperiodic Markov process $\{X_i\}$, $\mathrm{H}(\{X_i\}) = \tilde{\mathrm{H}}(\{X_i\}) = H(X_2|X_1)|_{P_{X_1}=\pi} = \sum_{x\in\mathcal{X}} \pi(x) H(X_2|X_1 = x)$, where π is the unique steady-state distribution.

7 Information for Continuous Distributions

The covariance of n-dimensional X is k, then $h(X) \le h(X^G) = \frac{1}{2}\log((2\pi e)^n \det(k))$.

8 Learning a Bit of Information

$\pi_{1|0}(\phi)$: false alarm, false positive, false rejection, type I error.

$\pi_{0|1}(\phi)$: miss detection, false negative, false acceptance, type II error.

$\mathcal{A}_\theta(\phi)$: acceptance region of H_θ.

Likelihood ratio $LR(x) := \frac{P_1(x)}{P_0(x)}$, Log likelihood ratio $LLR(x) := \log LR(x)$.

Likelihood ratio test (LRT) with parameter $\tau \in \mathbb{R}_0^+$ is $\phi_\tau^{LRT}(x) := \mathbb{I}\{LR(x) > \tau\}$.

(Randomized) LRT $\phi_{\gamma,\tau}(x) = \begin{cases} 1, \text{ if } LR(x) > \tau \\ \gamma, \text{ if } LR(x) = \tau \\ 0, \text{ if } LR(x) < \tau \end{cases}$.

Neyman-Pearson problem: minimize $\pi_{0|1}(\phi)$ subject to $\pi_{1|0}(\phi) \le \epsilon$.

Neyman-Pearson: LRT is optimal.

Generalized to n i.i.d.: $\phi_{\eta_n,\gamma_n}^n(x^n) = \begin{cases} 1/0, \text{ if } \sum_{i=1}^n LLR(x_i) \gtrless \eta_n \\ \gamma_n, \text{ if } \sum_{i=1}^n LLR(x_i) = \eta_n \end{cases}$.

Chernoff-Stein lemma: $\lim_{n\to\infty} -\frac{1}{n}\log \omega_{0|1}^*(n,\epsilon) = D(P_0\|P_1)$.

Typical set:

9 Information Divergence

$D(P\|Q) := \sum_a P(a)\log\frac{P(a)}{Q(a)}$.

$D(P\|Q) \ge 0$, with equality $\iff P(x) = Q(x),\ \forall x$.

$D(P_{Y|X}\|Q_{Y|X}|P_X) := \mathrm{E}_{X\sim P_X}[D(P_{Y|X}(\cdot|X)\|Q_{Y|X}(\cdot|X))]$.

Chain rule for information divergence: $D(P_{X,Y}\|Q_{X,Y}) = D(P_{Y|X}\|Q_{Y|X}|P_X) + D(P_X\|Q_X)$.

$D(P_Y\|Q_Y) \le D(P_{Y|X}\|Q_{Y|X}|P_X)$, with equality iff $D(P_{X|Y}\|Q_{X|Y}|P_Y) = 0$.

Donsker-Varadhan theorem: $D(P\|Q) = \max_{f:\mathcal{X}\to\mathbb{R}} E_{X\sim P}[f(X)] - \log \mathrm{E}_{X\sim Q}[2^{f(X)}]$ s.t. $E_{X\sim Q}[2^{f(X)}] < \infty$.

10 Error Exponents and Chernoff Information

P_0, P_1 are given.

$P_\lambda(a) := \frac{P_0(a)^{1-\lambda}P_1(a)^\lambda}{\sum_b P_0(b)^{1-\lambda}P_1(b)^\lambda}$.

Exercise 6: $D(P_\lambda\|P_0)$ is a continuous and strictly increasing function of λ for $\lambda \in [0,1)$.

$P_e^*(\pi(=(\pi_0,\pi_1)), n) := \min_\phi\{\pi_0\pi_{1|0}^{(n)}(\phi) + \pi_1\pi_{0|1}^{(n)}(\phi)\}$.

$\bar{P}_e^*(n) := \min_\phi\{\max\{\pi_{1|0}^{(n)}, \pi_{0|1}^{(n)}\}\}$.

Chernoff Information: $CI(P_0,P_1) :=$

$\max_{\lambda\in(0,1)} \underbrace{-\log\sum_{a\in\mathcal{X}} P_0(a)^{1-\lambda}P_1(a)^{\lambda}}_{f(\lambda)}$.

Theorem 11: $\lim_{n\to\infty}\{-\frac{1}{n}\log P_e^*(\pi,n)\} = \lim_{n\to\infty}\{-\frac{1}{n}\log \bar{P}_e^*(n)\} = CI(P_0,P_1)$.

11 Deviverling Information Reliably

$BSC(p)$: flip the bit bit i.i.d. with probability $p\in(0,\frac{1}{2})$.

12 Mutual Information

$I(X;Y) = D(P_{X,Y}\|P_X\times P_Y)$.

Exercise 1: $I(X;Y) = \min_{Q_Y:D(P_Y\|Q_Y)<\infty} D(P_{Y|X}\|Q_Y|P_X)$.

$I(X;Y|Z) := H(X|Z) - H(X|Y,Z)$.

Chain rule: $I(X;Y^n) = \sum_{i=1}^{n} I(X;Y_i|Y^{i-1})$.

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

$X - Y - Z$, then $I(X;Y)\geq I(X;Y|Z)$.

13 Noisy Channel Coding Theorem

An (n,k) code with $P_e^{(n)} := \Pr\{W\neq\hat{W}\}\leq\epsilon$ is called an (n,k,ϵ) code.

$k^*(n,k)$ is the largest k s.t. $\exists(n,k,\epsilon)$ code.

$C(\epsilon) := \lim_{n\to\infty}\frac{1}{n}k^*(n,\epsilon)$.

Channel coding theorem for DMC without feedback: $C(\epsilon) = C^I := \max_{P_X} I(X;Y),\ \forall\epsilon\in(0,1)$.

x^n is robust typical sequence: $|\hat{P}_{x^n}(a) - P_X(a)|\leq\epsilon P_X(a)$, where $\hat{P}_{x^n}(a) := \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{x_i = a\}$.

The set of ϵ-robust typical sequence with respect to X: $\mathcal{T}_\epsilon^{(n)}(X)$.

14 Channel Coding with a Cost Constraint

Constraint: $\frac{1}{n}\sum_{i=1}^{n} b(x_i)\leq B$.

$(n,\lceil nR\rceil,B)$ code.

$C(B) := \sup\{R|R:\text{achievable }\}$.

Channel coding for DMC with average input cost constraint: $C(B) = C^I(B) := \max_{P_X:\mathrm{E}_{P_X}[b(X)]\leq B} I(X;Y)$.

The above also holds for CMC.

$C^I(B)$ is non-decreasing, concave, continuous in B.

AWGN (additive with Gaussian noise) channel: noise is Gaussian and independent of others, and constraint: $\frac{1}{n}\sum_{i=1}^{n}|x_i|^2\leq B$.

The capacity of the AWGN channel with input power constraint B and noise variance σ^2 is given by $C(B) = \sup_{X:E[|X^2|]\leq B} I(X;Y) = \frac{1}{2}\log(1+\frac{B}{\sigma^2})$, which is achieved by $X\sim N(0,B)$.

Proposition 2: $X^G\sim N(0,B)$, $Y = X^G + Z$ where $\mathrm{Var}[Z] = \sigma^2$, $Z\perp X^G$, then $I(X^G;Y)\geq\frac{1}{2}\log(1+\frac{B}{\sigma^2})$.

15 Lossy Source Coding

$d(s^n,\hat{s}^n) := \frac{1}{n}\sum_{i=1}^{n} d(s_i,\hat{s}_i)$, where $d(s,\hat{s}) := (s-\hat{s})^2$.

(R,D) achievable: $\exists$ sequence of $(n,\lfloor nR\rfloor)$ codes s.t. $\limsup_{n\to\infty} D^{(n)}\leq D$.

Rate distortion function $R(D) := \inf\{R|(R,D):\text{achievable }\}$.

$D_{\min} := \min_{\hat{s}(s)} \mathrm{E}[d(S, \hat{s}(S))]$.

$D_{\max} := \min_{\hat{s}} \mathrm{E}[d(S, \hat{s})]$.

$R(D) = R^I(D) := \min_{P_{\hat{S}|S}:\mathrm{E}[d(S,\hat{S})]\leq D} I(S; \hat{S})$.

$R^I(D_{\min}) \leq H(S), R^I(D) = 0$ if $D \geq D_{\max}$.

Ber(p) source: $R(D) = \begin{cases} H_b(p) - H_b(D), \text{ if } 0 \leq D \leq \min\{p, 1-p\} \\ 0, \text{ if } D > \min\{p, 1-p\} \end{cases}$.

Gaussian source: $R(D) = \begin{cases} \frac{1}{2}\log(\frac{\sigma^2}{D}), \text{ if } 0 \leq D \leq \sigma^2 \\ 0, \text{ if } D > \sigma^2 \end{cases}$.

$R(D) \leq R^G(D)$.

Information Theory HW1

許博翔

September 21, 2023

Problem 1. I'll prove (b) first, and then use (b) to prove (a) for convenience.

(b) Suppose that $s^n = (s_1, s_2, \ldots, s_n) \in \mathcal{T}_\gamma^{(n)}(S)$.

By the definition of $\mathcal{T}_\gamma^{(n)}(S)$, $\forall a \in \mathbf{S}$, $\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{s_i = a\} - P_S(a)\right| \leq \gamma P_S(a)$.

$\Rightarrow \forall a \in \mathbf{S}$, $\left|\frac{1}{n}\sum_{i=1}^{n} \log(P_S(a))\mathbb{I}\{s_i = a\} - P_S(a)\log(P_S(a))\right| \leq \gamma P_S(a)\log(P_S(a))$.

$\Rightarrow \sum_{a\in\mathbf{S}}\left|\frac{1}{n}\sum_{i=1}^{n} \log(P_S(a))\mathbb{I}\{s_i = a\} - P_S(a)\log(P_S(a))\right| \leq \sum_{a\in\mathbf{S}}\gamma P_S(a)\log(P_S(a))$.

By triangular inequality,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\log(P_S(s_i)) + H(S)\right|$$
$$= \left|\sum_{a\in\mathbf{S}}\frac{1}{n}\sum_{i=1}^{n} \log(P_S(a))\mathbb{I}\{s_i = a\} - \sum_{a\in\mathbf{S}} P_S(a)\log(P_S(a))\right|$$
$$\leq \sum_{a\in\mathbf{S}}\left|\frac{1}{n}\sum_{i=1}^{n} \log(P_S(a))\mathbb{I}\{s_i = a\} - P_S(a)\log(P_S(a))\right|$$
$$\leq \sum_{a\in\mathbf{S}}\gamma P_S(a)\log(P_S(a)) = -\gamma H(S).$$

Taking $\delta = \xi(\gamma) := -\gamma H(S)$, and we get $\left|\frac{1}{n}\sum_{i=1}^{n}\log(P_S(s_i)) + H(S)\right| \leq \delta$, which means $s^n \in \mathcal{A}_\delta^{(n)}(S)$.

$\therefore \mathcal{T}_\gamma^{(n)}(S) \subseteq \mathcal{A}_\delta^{(n)}(S)$.

(a) Recall from (b), we take $\delta = \xi(\gamma) := -\gamma H(S)$.

The 4 properties in the proposition are:

(1) The original property is: $\forall s^n \in \mathcal{A}_\delta^{(n)}(S)$, $2^{-n(H(S)+\delta)} \leq \Pr\{S^n = s^n\} \leq 2^{-n(H(S)-\delta)}$.

$\because$ from (b) we know that $\forall s^n \in \mathcal{T}_\gamma^{(n)}(S),\ s^n \in \mathcal{A}_\delta^{(n)}(S)$.

$\therefore 2^{-n(H(S)+\delta)} \leq \Pr\{S^n = s^n\} \leq 2^{-n(H(S)-\delta)}$.

(2) Let $A_n(a) := \{s^n \in \mathbf{S}^n : |\pi(a|s^n) - P_S(a)| > \gamma P_S(a)\}$.

Since $S \sim P_S$ is a DMS, the random variables $\{X_i\}_{i=1}^{\infty}$ where $X_i := \mathbb{I}\{S_i = a\}$ are i.i.d.

The average of X_i, denote as μ, $= \Pr\{S_i = a\} = P_S(a)$.

$\bar{X_n} = \dfrac{1}{n}\sum_{i=1}^{n} X_i = \dfrac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{S_i = a\} = \pi(a|S^n)$.

Take $\epsilon > \gamma P_S(a)$.

By the weak law of large numbers, $\lim_{n\to\infty} \Pr\{S^n \in A_n(a)\} = \lim_{n\to\infty} \Pr\{|\pi(a|S^n) - P_S(a)| > \gamma P_S(a)\} = \lim_{n\to\infty} \Pr\{|\bar{X_n} - \mu| > \gamma P_S(a)\} \leq \lim_{n\to\infty} \Pr\{|\bar{X_n} - \mu| \geq \epsilon\} = 0$.

$\because \mathcal{T}_\gamma^{(n)}(S) = \mathbf{S}^n \setminus \bigcup_{a\in\mathbf{S}} A_n(a)$.

$\therefore \lim_{n\to\infty} \Pr\{S^n \in \mathcal{T}_\gamma^{(n)}(S)\} = 1 - \lim_{n\to\infty} \Pr\{S^n \in \bigcup_{a\in\mathbf{S}} A_n(a)\} \geq 1 - \lim_{n\to\infty} \sum_{a\in\mathbf{S}} \Pr\{S^n \in A_n(a)\} = 1$.

$\therefore \forall \epsilon > 0$, by the definition of limits, $\Pr\{S^n \in \mathcal{T}_\gamma^{(n)}(S)\} \geq 1 - \epsilon$ for n large enough.

(3) $\because \mathcal{T}_\gamma^{(n)}(S) \subseteq \mathcal{A}_\delta^{(n)}(S)$.

$\therefore |\mathcal{T}_\gamma^{(n)}(S)| \leq |\mathcal{A}_\delta^{(n)}(S)| \leq 2^{n(H(S)+\delta)}$.

(4) By (2), $\forall \epsilon > 0$, for n large enough, there is $1 - \epsilon \leq \Pr\{S^n \in \mathcal{T}_\gamma^{(n)}(S)\} = \sum_{s^n \in \mathcal{T}_\gamma^{(n)}(S)} \Pr\{S^n = s^n\} \overset{(1)}{\leq} \sum_{s^n \in \mathcal{T}_\gamma^{(n)}(S)} 2^{-n(H(S)-\delta)} = |\mathcal{T}_\gamma^{(n)}(S)| 2^{-n(H(S)-\delta)}$.

$\therefore \forall \epsilon > 0$, for n large enough, there is $|\mathcal{T}_\gamma^{(n)}(S)| \geq (1-\epsilon) 2^{n(H(S)-\delta)}$.

(c) Consider $\mathbf{S} = \{0, 1\},\ P_S(0) = P_S(1) = \dfrac{1}{2}, \gamma = 0.1$.

For the sequence $s^n = 0^n$, $|\pi(0|s^n) - P_S(0)| = \dfrac{1}{2} \not\leq 0.05 = \gamma P_S(0)$.

$\Rightarrow 0^n \notin \mathcal{T}_\gamma^{(n)}(S)$.

However, $\forall \delta' > 0$, $\left|\dfrac{1}{n}\sum_{i=1}^{n} \log P_S(s_i) + H(S)\right| = \left|\dfrac{1}{n}\sum_{i=1}^{n} \log\dfrac{1}{2} - \log\dfrac{1}{2}\right| = 0 \leq \delta'$.

$\Rightarrow 0^n \in \mathcal{A}_{\delta'}^{(n)}$.

$\therefore \mathcal{A}_{\delta'}^{(n)} \not\subseteq \mathcal{T}_\gamma^{(n)}(S)$.

Problem 2.

(a) Define $X_i = \log \frac{1}{P_S(S_i)}$. Since S_i are i.i.d, X_i are also i.i.d.

Since $P_S(S_i) \le 1$, we get that $\log \frac{1}{P_S(S_i)} \ge 0$.

$\Rightarrow \mathrm{E}[|X_i|] = \mathrm{E}[X_i] = \mathrm{E}[\log \frac{1}{P_S(S_i)}] = H(S) < \infty$.

$$\prod_{i=1}^{n} P_S(S_i) \ge 2^{-n(H(S)+n^{-1/2}\delta\varsigma(S))}$$

$$\iff \prod_{i=1}^{n} \frac{1}{P_S(S_i)} \le 2^{n(H(S)+n^{-1/2}\delta\varsigma(S))}$$

$$\iff \sum_{i=1}^{n} \log \frac{1}{P_S(S_i)} \le n(H(S) + n^{-1/2}\delta\varsigma(S))$$

$$\iff \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - H(S) \le n^{-1/2}\delta\varsigma(S)$$

$$\iff \frac{\sqrt{n}(\overline{X_n} - \mathrm{E}[X_i])}{\varsigma(S)} \le \delta.$$

By central limit theorem, $\frac{\sqrt{n}(\overline{X_n} - \mathrm{E}[X_i])}{\varsigma(S)} \xrightarrow{d} Z \sim N(0,1)$ as $n \to \infty$.

$$\Rightarrow \Pr\left\{\prod_{i=1}^{n} P_S(S_i) \ge 2^{-n(H(S)+n^{-1/2}\delta\varsigma(S))}\right\} = \Pr\left\{\frac{\sqrt{n}(\overline{X_n} - \mathrm{E}[X_i])}{\varsigma(S)} \le \delta\right\}$$

$\to \Pr\{Z \le \delta\} = \Phi(\delta)$ as $n \to \infty$.

(b) Let $Z \sim N(0,1)$, by Berry-Esseen theorem, $|\Pr\{S^n \in \mathcal{B}_\delta^{(n)}(S)\} - \Pr\{Z \le \delta\}| = \left|\Pr\left\{\frac{\sqrt{n}(\overline{X_n} - E[X_i])}{\varsigma(S)} \le \delta\right\} - \Pr\{Z \le \delta\}\right| \le cn^{-1/2}$ for some constant $c > 0$.

$\Rightarrow \Pr\{S^n \in \mathcal{B}_\delta^{(n)}(S)\} \ge \Pr\{Z \le \delta\} - cn^{-1/2} = \Phi(\delta) - cn^{-1/2}$.

Take $\delta = \Phi^{-1}(1 - \epsilon + cn^{-1/2}) = -\Phi^{-1}(\epsilon - cn^{-1/2})$, we get that $\Pr\{S^n \in \mathcal{B}_\delta^{(n)}(S)\} \ge 1 - \epsilon$.

$\Rightarrow \Pr\{S^n \notin \mathcal{B}_\delta^{(n)}(S)\} \le \epsilon$.

Since $\frac{d\Phi^{-1}(x)}{dx} = \left.\frac{1}{\frac{d\Phi(y)}{dy}}\right|_{y=\Phi^{-1}(x)} = \left.\sqrt{2\pi}e^{y^2/2}\right|_{y=\Phi^{-1}(x)} = \sqrt{2\pi}e^{(\Phi^{-1}(x))^2/2}$, there is $\Phi^{-1}(\epsilon - cn^{-1/2}) \approx \Phi^{-1}(\epsilon) - \sqrt{2\pi}e^{(\Phi^{-1}(\epsilon))^2/2}cn^{-1/2} = \Phi^{-1}(\epsilon) - O(n^{-1/2})$ for n sufficiently large.

$\Rightarrow \delta = -\Phi^{-1}(\epsilon) + \zeta_n'$, where $\zeta_n' = O(n^{-1/2})$.

Lemma 2.1. $\exists \zeta_n = O(n^{-1})$ s.t. $nk \le \lfloor n(k + \zeta_n)\rfloor$.

Proof. Consider $\zeta_n = \frac{1}{n}$, we get that $\lfloor n(k + \zeta_n)\rfloor = \lfloor n(k + \frac{1}{n})\rfloor = \lfloor nk\rfloor + 1 \ge$

nk. ■

Since $\displaystyle\sum_{s^n \in \mathcal{B}_\delta^{(n)}(S)} P_{S^n}(s^n) \le \sum_{s^n} P_{S^n}(s^n) = 1$,

and if $s^n \in B$, then $P_{S^n}(s^n) = \displaystyle\prod_{i=1}^{n} P_S(s_i) \ge 2^{-n(H(S)+n^{-1/2}\delta\varsigma(S))}$.

$\therefore |\mathcal{B}_\delta^{(n)}(S)| 2^{-n(H(S)+n^{-1/2}\delta\varsigma(S))} = \displaystyle\sum_{s^n \in \mathcal{B}_\delta^{(n)}(S)} 2^{-n(H(S)+n^{-1/2}\delta\varsigma(S))} \le \sum_{s^n \in \mathcal{B}_\delta^{(n)}(S)} P_{S^n}(s^n) \le 1$.

$\Rightarrow |\mathcal{B}_\delta^{(n)}(S)| \le 2^{n(H(S)+n^{-1/2}\delta\varsigma(S))}$.

By **Lemma (2.1)**, there exists $\zeta_n'' \in O(n^{-1})$ s.t. $n(H(S)+n^{-1/2}\delta\varsigma(S)) \le \lfloor n(H(S)+n^{-1/2}\delta\varsigma(S)+\zeta_n'')\rfloor$.

Take $R = H(S)+n^{1/2}\varsigma(S)\delta+\zeta_n'' = H(S)-n^{-1/2}\varsigma(S)\Phi^{-1}(\epsilon)+n^{-1/2}\varsigma(S)\zeta_n'+\zeta_n''$.

Since $n^{-1/2}\varsigma(S)\zeta_n' = O(n^{-1})$, we get that $R = H(S)-n^{-1/2}\varsigma(S)\Phi^{-1}(\epsilon)+\zeta_n$ for some $\zeta_n = O(n^{-1})$.

Therefore, $\mathcal{B}_\delta^{(n)}(S)$ is an $(n, \lfloor nR \rfloor)$ code with $P_e^{(n)} \le \epsilon$.

Problem 3.

(a) Let $\delta \in (0, R-H(S))$, and $\mathcal{A}_\delta^{(n)}(S)$ be the δ-typical set defined in Definition 1.

By the third property of Proposition 1, we know that $|\mathcal{A}_\delta^{(n)}(S)| \le 2^{n(H(S)+\delta)} \overset{H(S)+\delta<R \Rightarrow n(H(S)+\delta)<nR-1 \text{ for } n \text{ large enough}}{<} 2^{\lfloor nR \rfloor}$ for n large enough.

$\Rightarrow \mathcal{A}_\delta^{(n)}(S)$ is an $(n, \lfloor nR \rfloor)$ code.

By the second property of Proposition 1, we know that $\forall \epsilon > 0,\ \exists N$ s.t. $\forall n \ge N,\ P_e^{(n)} = \Pr\{S^n \notin \mathcal{A}_\delta^{(n)}(S)\} \le \epsilon$.

Since $P_e^{(n)} \ge 0$, therefore by the definition of limits, $\displaystyle\lim_{n\to\infty} P_e^{(n)} = 0$.

$\therefore$ such sequence exists, and it is $\mathcal{A}_\delta^{(n)}(S)$.

(b) For a given $(n, \lfloor nR \rfloor)$ code, let $\mathcal{B}^{(n)}$ denote the range of the decoding function.

Let $\delta \in (0, H(S)-R)$, and $\mathcal{A}_\delta^{(n)}(S)$ be the δ-typical set defined in Definition 1.

By the first property of Proposition 1, we know that $\forall s^n \in \mathcal{A}_\delta^{(n)}(S),\ \Pr\{S^n = s^n\} \le 2^{-n(H(S)-\delta)}$.

$\Rightarrow \Pr\{S^n \in \mathcal{A}_\delta^{(n)}(S) \cap \mathcal{B}^{(n)}\} = \displaystyle\sum_{s^n \in \mathcal{A}_\delta^{(n)}(S)\cap\mathcal{B}^{(n)}} \Pr\{S^n = s^n\}$

$$\leq \sum_{s^n \in \mathcal{A}_\delta^{(n)}(S) \cap \mathcal{B}^{(n)}} 2^{-n(H(S)-\delta)} \leq \sum_{s^n \in \mathcal{B}^{(n)}} 2^{-n(H(S)-\delta)}$$

$$= |\mathcal{B}^{(n)}| 2^{-n(H(S)-\delta)} \leq 2^{\lfloor nR \rfloor - n(H(S)-\delta)} \leq 2^{-n(H(S)-R-\delta)}.$$

Since $H(S) - R - \delta > 0$ by definition of δ, we get that

$\lim_{n\to\infty} P_e^{(n)} = \lim_{n\to\infty} \Pr\{S^n \in \mathcal{A}_\delta^{(n)}(S) \cap \mathcal{B}^{(n)}\} \geq \lim_{n\to\infty} (1 - 2^{-n(H(S)-R-\delta)}) = 1.$

On the other hand, $P_e^{(n)} \leq 1$, so there is $\lim_{n\to\infty} P_e^{(n)} = 1$.

Information Theory HW2

許博翔

October 5, 2023

Problem 1.

(a) Define $Q_X(x) = q_x$.

$H(X) + \sum_{i=1}^{\infty} p_i \log q_i = \sum_{i=1}^{\infty} \mathrm{E}\left[\log \frac{Q_X}{P_X}\right] \overset{\because \log \text{ is concave}}{\leq} \log \mathrm{E}\left[\frac{Q_X}{P_X}\right] = \log\left(\sum_{i=1}^{\infty} p_i \cdot \frac{q_i}{p_i}\right) = \log\left(\sum_{i=1}^{\infty} q_i\right) = \log 1 = 0.$

$\therefore H(X) \leq -\sum_{i=1}^{\infty} p_i \log q_i.$

(b) $-\log q_i$ is an arithmetic sequence $\Rightarrow q_i$ is an geometric sequence.

Suppose that $q_i = q_0 r^i$, where $1 < r < 1$ and $q_0 > 0$.

$\because 1 = \sum_{i=1}^{\infty} q_i = \frac{q_0 r}{1-r}$

$\Rightarrow q_0 = \frac{1-r}{r}.$

$\because \mu_X = \sum_{i=1}^{\infty} i q_i = \sum_{i=1}^{\infty} q_0 i r^i = \sum_{i=1}^{\infty}\sum_{j=1}^{i} q_0 r^i = \sum_{j=1}^{\infty}\sum_{i=j}^{\infty} q_0 r^i = \sum_{j=1}^{\infty} \frac{q_0 r^j}{1-r} = \frac{q_0 r}{(1-r)^2}$

$\Rightarrow \frac{1}{1-r} = \mu_X$

$\therefore r = 1 - \frac{1}{\mu_X} = \frac{\mu_X - 1}{\mu_X},\ q_0 = \frac{\frac{1}{\mu_X}}{1 - \frac{1}{\mu_X}} = \frac{1}{\mu_X - 1}.$

$-\log q_i = -\log q_0 r^i = -\log q_0 - i \log r.$

Take $\alpha = -\log r = \log(\mu_X) - \log(\mu_X - 1),\ \beta = -\log q_0 = \log(\mu_X - 1)$ satisfies the conditions.

$\therefore$ the answer is $q_i = \frac{(\mu_X - 1)^{i-1}}{(\mu_X)^i},\ \alpha = \log(\mu_X) - \log(\mu_X - 1),\ \beta = \log(\mu_X - 1).$

(c) $-\sum_{i=1}^{\infty} p_i \log q_i = \sum_{i=1}^{\infty} p_i(\alpha i + \beta) = \alpha\mu_X + \beta = \log(\mu_X)\mu_X - \log(\mu_X - 1)\mu_X + \log(\mu_X - 1) = \mu_X(\log(\mu_X) - (1 - \frac{1}{\mu_X})\log(\mu_X - 1)) = \mu_X(\log(\frac{\mu_X}{\mu_X - 1}) - \frac{1}{\mu_X}\log(\frac{\mu_X}{\mu_X - 1}) +$

$\frac{1}{\mu_X}\log(\mu_X)) = \mu_X(-(1-\frac{1}{\mu_X})\log(1-\frac{1}{\mu_X}) - \frac{1}{\mu_X}\log(\frac{1}{\mu_X})) = \mu_X h_b(\mu_X^{-1})$.
$\therefore H(X) \leq \mu_X h_b(\mu_X^{-1})$, and the equation holds when $p_i = q_i$ for all i, that is, $X \sim \text{Geo}(\frac{1}{\mu_X})$ is the geometric distribution.

Problem 2.

(a) $\int_2^\infty \frac{1}{x(\log x)^\alpha}dx = \int_{x=2}^\infty (\log x)^{-\alpha} d(\log x)$

$$= \begin{cases} \frac{1}{1-\alpha}(\log x)^{1-\alpha}|_{x=2}^\infty & \text{, if } \alpha \neq 1 \text{, which converges } \iff 1-\alpha < 0 \iff \alpha > 1, \\ & \text{since } \lim_{y\to\infty} y^a = 0 \text{ for } a < 0 \text{, and } \lim_{y\to\infty} y^a \text{ does not exist for } a > 0\cdot \\ \log\log x|_{x=2}^\infty & \text{, if } \alpha = 1 \text{, which does not converges} \end{cases}$$

$\therefore \sum_{n=2}^\infty \frac{1}{n(\log n)^\alpha}$ converges $\iff \alpha > 1$.

(b) First, we know that the series converges $\iff \alpha > 1$, so we only consider $\alpha > 1$.
$H(X_\alpha) = -\text{E}(\log P_{X_\alpha}) = \sum_{n=2}^\infty \frac{1}{s_\alpha n(\log n)^\alpha}\log(s_\alpha n(\log n)^\alpha) = \sum_{n=2}^\infty \frac{\log s_\alpha}{s_\alpha n(\log n)^\alpha} + \sum_{n=2}^\infty \frac{1}{s_\alpha n(\log n)^{\alpha-1}} + \sum_{n=2}^\infty \frac{\alpha\log\log n}{s_\alpha n(\log n)^\alpha}$.
For $\alpha \leq 2$, since $H(X_\alpha) > \sum_{n=2}^\infty \frac{1}{s_\alpha n(\log n)^{\alpha-1}} \to \infty$ from (a); therefore $H(X_\alpha)$ diverges to ∞.
For $\alpha > 2$, since $H(X_\alpha) < \sum_{n=2}^\infty \frac{\log s_\alpha}{s_\alpha n(\log n)^\alpha} + \sum_{n=2}^\infty \frac{1}{s_\alpha n(\log n)^{\alpha-1}} + \sum_{n=2}^\infty \frac{\alpha\log\log n}{s_\alpha n(\log n)^\alpha}$
$\overset{\log\log n<\log n \text{ for } n\geq 2}{<} \sum_{n=2}^\infty \frac{\log s_\alpha}{s_\alpha n(\log n)^\alpha} + \sum_{n=2}^\infty \frac{1}{s_\alpha n(\log n)^{\alpha-1}} + \sum_{n=2}^\infty \frac{\alpha}{s_\alpha n(\log n)^{\alpha-1}}$
$= \log s_\alpha + \frac{(1+\alpha)s_{\alpha-1}}{s_\alpha} < \infty$,
and $\sum_{n=2}^m \frac{1}{s_\alpha n(\log n)^\alpha}\log(s_\alpha n(\log n)^\alpha)$ is increasing as m increases.
$\Rightarrow H(X_\alpha) = \sum_{n=2}^\infty \frac{1}{s_\alpha n(\log n)^\alpha}\log(s_\alpha n(\log n)^\alpha)$ converges.
$\therefore H(X_\alpha)$ exists if $\alpha > 2$, and diverges to ∞ if $1 < \alpha \leq 2$.

Problem 3. Note that $P_{X_{\Theta_i}[i]}(\theta_i, x_i)$ is defined as $\Pr\{\Theta_i = \theta_i \wedge X_{\theta_i}[i] = x_i\}$, while $P_{X_{\theta_i}[i]}(x_i)$ is defined as $\Pr\{X_{\theta_i}[i] = x_i\}$.
Since $X_{\theta_i}[i]$ and Θ_i are independent, there is $P_{X_{\Theta_i}[i]}(\theta_i, x_i) = P_{\Theta_i}(\theta_i)P_{X_{\theta_i}[i]}(x_i)$.

(a) $\because \forall l, n \in \mathbb{N},\ P_{X_{\Theta_1}[1], X_{\Theta_2}[2], \ldots, X_{\Theta_n}[n]} = P_{X_\Theta[1], X_\Theta[2], \ldots, X_\Theta[n]}$
$\overset{X_\Theta \text{ is stationary no matter } \Theta \text{ is 0 or 1}}{=} P_{X_\Theta[l+1], X_\Theta[l+2], \ldots, X_\Theta[l+n]} = P_{X_{\Theta_{l+1}}[l+1], X_{\Theta_{l+2}}[l+2], \ldots, X_{\Theta_{l+n}}[l+n]}$.

$\therefore \{X_{\Theta_i}[i]\}$ is stationary.

By the definition of entropy rates,

$\lim_{n\to\infty} -\frac{1}{n}\mathrm{E}[\log P_{X_k[1], X_k[2], \ldots, X_k[n]}] = \lim_{n\to\infty} \frac{1}{n} H(X_k[1], X_k[2], \ldots, X_k[n]) = \mathcal{H}_k$.

$\Rightarrow \mathcal{H}(\{X_{\Theta_i}[i]\}) = \lim_{n\to\infty} \frac{1}{n} H(X_{\Theta_1}[1], X_{\Theta_2}[2], \ldots, X_{\Theta_n}[n])$

$= \lim_{n\to\infty} -\frac{1}{n}\mathrm{E}[\log P_{X_\Theta[1], X_\Theta[2], \ldots, X_\Theta[n]}]$

$= \lim_{n\to\infty} -\frac{1}{n}\big(\Pr\{\Theta = 0\}\mathrm{E}[\log \Pr\{\Theta = 0\} P_{X_0[1], X_0[2], \ldots, X_0[n]}]$

$+\Pr\{\Theta = 1\}\mathrm{E}[\log \Pr\{\Theta = 1\} P_{X_1[1], X_1[2], \ldots, X_1[n]}]\big)$

$= \lim_{n\to\infty} -\frac{1}{n}\big((1-q)\mathrm{E}[\log(1-q) + \log P_{X_0[1], X_0[2], \ldots, X_0[n]}] + q\mathrm{E}[\log q + \log P_{X_1[1], X_1[2], \ldots, X_1[n]}]\big)$

$= \lim_{n\to\infty} -\frac{1}{n}((1-q)\log(1-q) + q\log q) + (1-q)\mathcal{H}_0 + q\mathcal{H}_1 = (1-q)\mathcal{H}_0 + q\mathcal{H}_1$.

(b) Suppose $\Theta_1 \sim \mathrm{Ber}(q)$.

Since $\{\Theta_i\}$ is stationary, $\begin{pmatrix} 1-q & q \end{pmatrix} = \begin{pmatrix} 1-q & q \end{pmatrix} \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$.

$\Rightarrow 1 - q = (1-q)(1-\alpha) + q\beta$

$\Rightarrow \alpha(1-q) = q\beta$

$\Rightarrow q = \frac{\alpha}{\alpha+\beta}$.

$\because P_{X_{\Theta_{i+1}}[i+1]|X_{\Theta_i}[i]} \overset{X_j[i] \text{ is independent of } X_{j'}[i'] \text{ for any } (i',j')\neq(i,j)}{=} P_{\Theta_{i+1}|\Theta_i} P_{X_{\theta_{i+1}}[i+1]}$.

$\therefore P_{X_{\Theta_1}[1], X_{\Theta_2}[2], \ldots, X_{\Theta_n}[n]} = \prod_{i=1}^{n} P_{X_{\Theta_i}[i]|X_{\Theta_1}[1], X_{\Theta_2}[2], \ldots, X_{\Theta_{i-1}}[i-1]} = P_{X_{\Theta_1}[1]} \prod_{i=2}^{n} P_{X_{\Theta_i}[i]|X_{\Theta_{i-1}}[i-1]}$

$= P_{\Theta_1} P_{X_{\theta_1}[1]} \prod_{i=2}^{n} P_{\Theta_i|\Theta_{i-1}} P_{X_{\theta_i}[i]} = \left(P_{\Theta_1}\prod_{i=2}^{n} P_{\Theta_i|\Theta_{i-1}}\right)\prod_{i=1}^{n} P_{X_{\theta_i}[i]}$

$\overset{X_j[i] \text{ is independent of } X_{j'}[i'] \text{ for any } (i',j')\neq(i,j)}{=} P_{\Theta_1, \Theta_2, \ldots, \Theta_n} P_{X_{\theta_1}[1], X_{\theta_2}[2], \ldots, X_{\theta_n}[n]}$.

$\Rightarrow \forall l, n \in \mathbb{N},\ P_{X_{\Theta_1}[1], X_{\Theta_2}[2], \ldots, X_{\Theta_n}[n]} = P_{\Theta_1, \Theta_2, \ldots, \Theta_n} P_{X_{\theta_1}[1], X_{\theta_2}[2], \ldots, X_{\theta_n}[n]} \overset{\{X_0[i]\}, \{X_1[i]\}, \{\Theta_i\} \text{ are stationary}}{=}$

$P_{\Theta_{l+1}, \Theta_{l+2}, \ldots, \Theta_{l+n}} P_{X_{\theta_1}[l+1], X_{\theta_2}[l+2], \ldots, X_{\theta_n}[l+n]} = P_{X_{\Theta_{l+1}}[l+1], X_{\Theta_{l+2}}[l+2], \ldots, X_{\Theta_{l+n}}[l+n]}$.

$\therefore \{X_{\Theta_i}[i]\}$ is stationary.

By theorem 11, $\mathcal{H}(\{X_{\Theta_i}[i]\}) = H(X_{\Theta_2}[2]|X_{\Theta_1}[1])$

$= -\sum_{\theta_1, \theta_2, x_1, x_2} P_{X_{\Theta_1}[1]}(\theta_1, x_1) P_{\Theta_2|\Theta_1}(\theta_2|\theta_1) P_{X_{\theta_2}[2]}(x_2)(\log(P_{\Theta_2|\Theta_1}(\theta_2|\theta_1)) + \log(P_{X_{\theta_2}[2]}(x_2)))$

$= -\sum_{\theta_1, \theta_2, x_1, x_2} P_{X_{\Theta_1}[1]}(\theta_1, x_1) P_{\Theta_2|\Theta_1}(\theta_2|\theta_1) P_{X_{\theta_2}[2]}(x_2) \log(P_{\Theta_2|\Theta_1}(\theta_2|\theta_1))$

$$- \sum_{\theta_1,\theta_2,x_1,x_2} P_{X_{\Theta_1}[1]}(\theta_1, x_1) P_{\Theta_2|\Theta_1}(\theta_2|\theta_1) P_{X_{\theta_2}[2]}(x_2) \log(P_{X_{\theta_2}[2]}(x_2)))$$
$$= - \sum_{\theta_1,\theta_2,x_1} P_{X_{\Theta_1}[1]}(\theta_1, x_1) P_{\Theta_2|\Theta_1}(\theta_2|\theta_1) \log(P_{\Theta_2|\Theta_1}(\theta_2|\theta_1))$$
$$+ \sum_{\theta_1,\theta_2,x_1} P_{X_{\Theta_1}[1]}(\theta_1, x_1) P_{\Theta_2|\Theta_1}(\theta_2|\theta_1) H(X_{\theta_2}[2])$$
$$= - \sum_{\theta_1,\theta_2} P_{\Theta_1}(\theta_1) P_{\Theta_2|\Theta_1}(\theta_2|\theta_1) \log(P_{\Theta_2|\Theta_1}(\theta_2|\theta_1))$$
$$+ \sum_{\theta_1,\theta_2} P_{\Theta_1}(\theta_1) P_{\Theta_2|\Theta_1}(\theta_2|\theta_1) H(X_{\theta_2}[2])$$
$$= -(1-q)(1-\alpha)\log(1-\alpha) - (1-q)\alpha\log(\alpha) - q\beta\log(\beta) - q(1-\beta)\log(1-\beta) + H(X_0[2])((1-q)(1-\alpha)+q\beta) + H(X_1[2])((1-q)\alpha+q(1-\beta))$$
$$\overset{\{X_k[i]\}\text{ are i.i.d.}\Rightarrow\mathcal{H}_k=H(\{X_k[i]\})=H(X_k[i])}{=} (1-q)H_b(\alpha) + qH_b(\beta) + \mathcal{H}_0((1-q)(1-\alpha)+q\beta) + \mathcal{H}_1((1-q)\alpha+q(1-\beta))$$
$$= \frac{\beta}{\alpha+\beta}H_b(\alpha) + \frac{\alpha}{\alpha+\beta}H_b(\beta) + \mathcal{H}_0(\frac{\beta}{\alpha+\beta}(1-\alpha) + \frac{\alpha}{\alpha+\beta}\beta) + \mathcal{H}_1(\frac{\beta}{\alpha+\beta}\alpha + \frac{\alpha}{\alpha+\beta}(1-\beta))$$
$$= \frac{\beta}{\alpha+\beta}(H_b(\alpha)+\mathcal{H}_0) + \frac{\alpha}{\alpha+\beta}(H_b(\beta)+\mathcal{H}_1).$$

Information Theory HW3

許博翔

October 19, 2023

Note that in this homework, I'll use the following definition:

Problem 1, 2: if $P = G(p)$, then $P(x) = p(1-p)^{1-x}$.

Problem 3: if $P = G(p)$, then $P(x) = (1-p)p^{1-x}$, which is the definition given in the homework.

$\exp_2(x) := 2^x$.

Problem 1.

(a) Consider $\phi_{\tau,\gamma}(x) := \begin{cases} 1, \text{ if } LR(x) > \tau \\ \gamma, \text{ if } LR(x) = \tau \\ 0, \text{ if } LR(x) < \tau \end{cases}$.

$LR(0) = \dfrac{P_1(0)}{P_0(0)} = \dfrac{1-p_1}{1-p_0}$.

$LR(1) = \dfrac{P_1(1)}{P_0(1)} = \dfrac{p_1}{p_0}$.

$\because p_0 < p_1$.

$\therefore LR(1) = \dfrac{p_1}{p_0} > 1 > \dfrac{1-p_1}{1-p_0} = LR(0)$.

By Neyman-Pearson theorem, $\phi_{\tau,\gamma}$ is optimal.

$\pi_{1|0}(\phi_{\tau,\gamma}) = P_0\{LR(X) > \tau\} + \gamma P_0\{LR(X) = \tau\}$.

$\pi_{0|1}(\phi_{\tau,\gamma}) = P_1\{LR(X) < \tau\} + (1-\gamma) P_1\{LR(X) = \tau\}$.

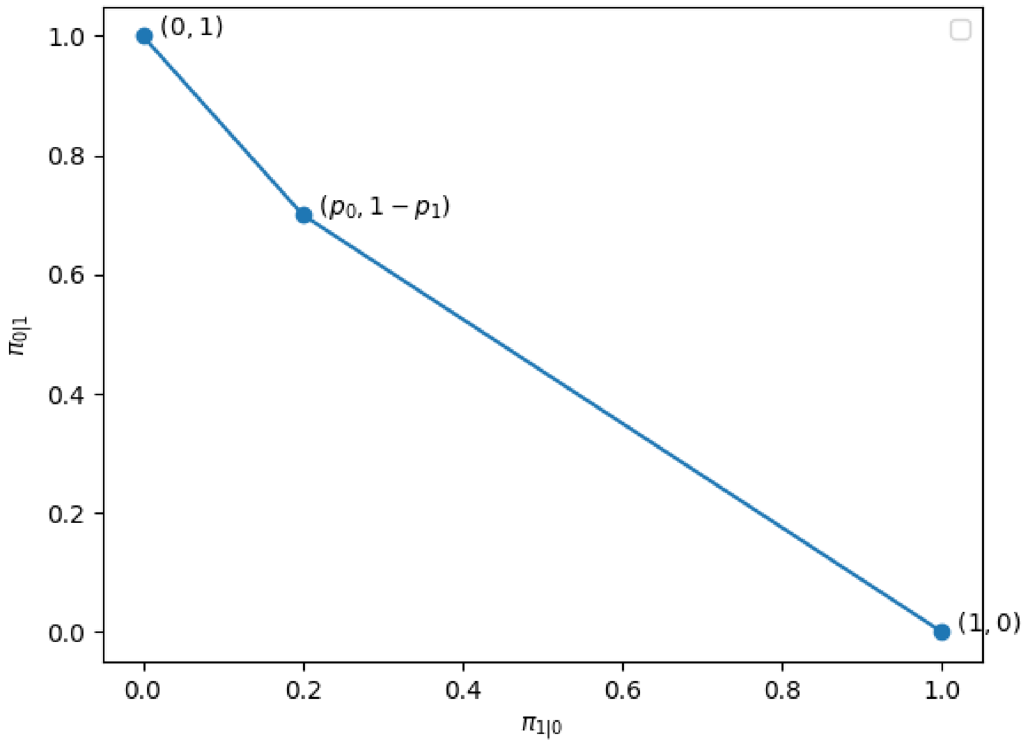
We only need to consider the cases $\tau = LR(x)$ for some x, since other cases can be reduced to these cases by setting γ properly.

For $\tau = LR(0)$, $\pi_{1|0} = P_0(1) + \gamma P_0(0) = p_0 + \gamma(1-p_0)$; $\pi_{0|1} = 0 + (1-\gamma)P_1(0) = (1-\gamma)(1-p_1)$.

For $\tau = LR(1)$, $\pi_{1|0} = 0 + \gamma P_0(1) = \gamma p_0$; $\pi_{0|1} = P_1(0) + (1-\gamma)P_1(1) =$

$1 - p_1 + (1-\gamma)p_1$.

The above forms two segments, and their intersection is $(p_0, 1-p_1)$, which can be calculated by setting γ in the first segment to 0 or in the second segment to 1.



(b) Let Y be the random variable denoting the length of the observed sequence. We can see that $P_Y(y) = p(1-p)^{y-1}$.

$P\{Y > y\} = \sum_{z=y+1}^{\infty} p(1-p)^{z-1} = \frac{p(1-p)^y}{1-(1-p)} = (1-p)^y$.

$P\{Y < y\} = \sum_{z=1}^{y-1} p(1-p)^{z-1} = \frac{p(1-(1-p)^{y-1})}{1-(1-p)} = 1-(1-p)^{y-1}$.

$P_0(y) = p_0(1-p_0)^{y-1}$, $P_1(y) = p_1(1-p_1)^{y-1}$.

Consider $\phi_{\tau,\gamma}(y) := \begin{cases} 1, \text{ if } LR(y) > \tau \\ \gamma, \text{ if } LR(y) = \tau \\ 0, \text{ if } LR(y) < \tau \end{cases}$.

$LR(y) = \frac{P_1(y)}{P_0(y)} = \frac{p_1(1-p_1)^{y-1}}{p_0(1-p_0)^{y-1}}$.

Since $p_0 < p_1$, there is $\frac{1-p_1}{1-p_0} < 1$.

$\Rightarrow LR(y)$ is an decreasing function of y.

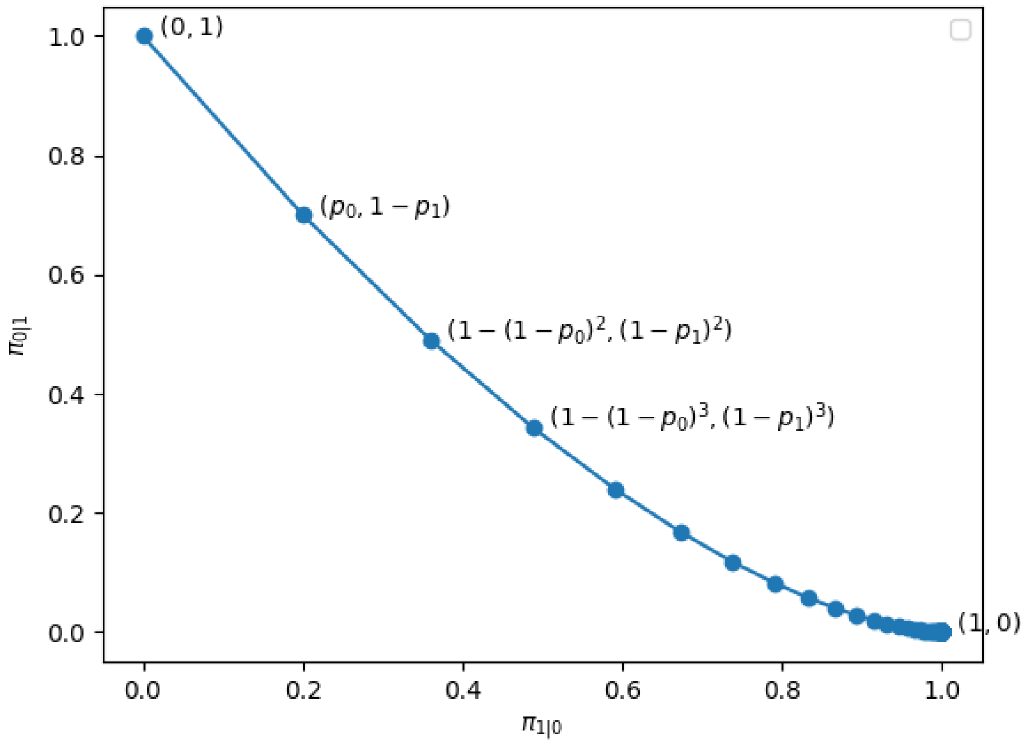
By Neyman-Pearson theorem, $\phi_{\tau,\gamma}$ is optimal.

We only need to consider the cases $\tau = LR(y)$ for some y, since other cases can be reduced to these cases by setting γ properly.

Since $LR(y)$ is decreasing, for $\tau = LR(y)$, $\pi_{1|0}(\phi_{\tau,\gamma}) = P_0\{Y < y\} + \gamma P_0\{Y = y\} = 1 - (1-p_0)^{y-1} + \gamma p_0(1-p_0)^{y-1} = 1 - (1-p_0)^{y-1}(1-\gamma p_0)$.

$\pi_{0|1}(\phi_{\tau,\gamma}) = P_1\{Y > y\} + (1-\gamma)P_1\{Y = y\} = (1-p_1)^y + (1-\gamma)p_1(1-p_1)^{y-1} = (1-\gamma p_1)(1-p_1)^{y-1}$.

For each y, it forms a segment, where the intersection of the segments formed by y and $y+1$ is $(1-(1-p_0)^y, (1-p_1)^y)$, which can be calculated by setting γ in the segment formed by y to 1 or in the other segment to 0.



(c) Let Y_i be the random variable denoting the length of the sequence between the $(i-1)$-th 1 and the i-th 1 (including the i-th 1 and excluding the $(i-1)$-th 1). One can see that Y_i are i.i.d. and $Y_i \sim G(p)$.

Clearly, $Z = Y_1 + Y_2 + \cdots + Y_n$ is the random variable of the length of the observed sequence.

Let $Q_0 = G(p_0), Q_1 = G(p_1)$.

From Chernoff-Stein lemma, $\lim_{n\to\infty} -\frac{1}{n}\log \overline{\omega}^*_{0|1}(n,\epsilon) = \mathrm{E}_{Y\sim G(p_0)}[\log \frac{Q_0(Y)}{Q_1(Y)}] = \sum_{i=1}^{\infty} p_0(1-p_0)^{i-1}\log\frac{p_0(1-p_0)^{i-1}}{p_1(1-p_1)^{i-1}} = \sum_{i=1}^{\infty} p_0(1-p_0)^{i-1}\log\frac{p_0}{p_1} + \sum_{i=1}^{\infty}(i-1)p_0(1-p_0)^{i-1}\log\frac{1-p_0}{1-p_1} = p_0\frac{1}{1-(1-p_0)}\log\frac{p_0}{p_1} + p_0\log\frac{1-p_0}{1-p_1}\sum_{i=1}^{\infty}\sum_{j=1}^{i-1}(1-p_0)^{i-1} = \log\frac{p_0}{p_1} + p_0\log\frac{1-p_0}{1-p_1}\sum_{j=1}^{\infty}\sum_{i=j+1}^{\infty}(1-p_0)^{i-1} = \log\frac{p_0}{p_1} + p_0\log\frac{1-p_0}{1-p_1}\sum_{j=1}^{\infty}\frac{(1-p_0)^j}{p_0} = \log\frac{p_0}{p_1} + p_0\log\left(\frac{1-p_0}{1-p_1}\right)\frac{1-p_0}{p_0^2} = \log\frac{p_0}{p_1} + (\frac{1}{p_0}-1)\log\frac{1-p_0}{1-p_1}$.

Problem 2.

(a) $\pi_0^{(n)}(x^n) = \Pr\{\Theta=0|X^n=x^n\} = \Pr\{X_i \overset{\text{i.i.d.}}{\sim} P_0|X^n=x^n\} = \frac{\Pr\{X_i \overset{\text{i.i.d.}}{\sim} P_0 \wedge X^n=x^n\}}{\Pr\{X^n=x^n\}} = \frac{\Pr\{X_i \overset{\text{i.i.d.}}{\sim} P_0 \wedge X^n=x^n\}}{\Pr\{(X_i \overset{\text{i.i.d.}}{\sim} P_0 \wedge X^n=x^n)\vee(X_i \overset{\text{i.i.d.}}{\sim} P_1 \wedge X^n=x^n)\}} = \frac{\pi_0^{(0)}\prod_{i=1}^n P_0(x_i)}{\pi_0^{(0)}\prod_{i=1}^n P_0(x_i) + \pi_1^{(0)}\prod_{i=1}^n P_1(x_i)}$.

Similarly, $\pi_1^{(n)}(x^n) = \frac{\pi_1^{(0)}\prod_{i=1}^n P_1(x_i)}{\pi_0^{(0)}\prod_{i=1}^n P_0(x_i) + \pi_1^{(0)}\prod_{i=1}^n P_1(x_i)}$.

(b) $-\frac{1}{n}\log\pi_0^{(0)}\prod_{i=1}^n P_0(X_i) = -\frac{1}{n}\left(\log\pi_0^{(0)} + \sum_{i=1}^n \log(P_0(X_i))\right) \overset{\text{SLLN}}{\to} -\frac{1}{n}\log\pi_0^{(0)} - \mathrm{E}[\log(P_0(X))] \overset{\log\pi_0^{(0)}\text{ is a constant}}{\to} -\mathrm{E}[\log(P_0(X))] = H(X)$ as $n\to\infty$.

From HW2 we know that $H(X) \le -\sum_{i=1}^{\infty} P_0(i)\log P_1(i)$, with equality $\iff P_1 \sim P_0$.

$-\frac{1}{n}\log\pi_1^{(0)}\prod_{i=1}^n P_1(X_i) = -\frac{1}{n}\left(\log\pi_1^{(0)} + \sum_{i=1}^n \log(P_1(X_i))\right) \overset{\text{SLLN}}{\to} -\frac{1}{n}\log\pi_1^{(0)} - \mathrm{E}[\log(P_1(X))] \overset{\log\pi_1^{(0)}\text{ is a constant}}{\to} -\mathrm{E}[\log(P_1(X))] > H(X)$ as $n\to\infty$.

$\Rightarrow \frac{\pi_1^{(0)}\prod_{i=1}^n P_1(X_i)}{\pi_0^{(0)}\prod_{i=1}^n P_0(X_i)} \to \exp_2(nE[\log(P_1(X))] + nH(X)) = \exp_2(E[\log(P_1(X))] + H(X))^n \overset{\mathrm{E}[\log(P_1(X))]+H(X)<0}{\to} 0$ as $n\to\infty$.

$\therefore \pi_0^{(n)}(X^n) = \frac{\pi_0^{(0)}\prod_{i=1}^n P_0(x_i)}{\pi_0^{(0)}\prod_{i=1}^n P_0(x_i) + \pi_1^{(0)}\prod_{i=1}^n P_1(x_i)} = \frac{1}{1+\frac{\pi_1^{(0)}\prod_{i=1}^n P_1(X_i)}{\pi_0^{(0)}\prod_{i=1}^n P_0(X_i)}} \to \frac{1}{1+0} = 1$ as

$n \to \infty$.

As what we computed above, for any constant $c > 0$, $-\frac{1}{n}\log\frac{\pi_1^{(0)}\prod_{i=1}^{n}P_1(X_i)}{c\pi_0^{(0)}\prod_{i=1}^{n}P_0(X_i)}$

$= H(X) + \mathrm{E}[\log(P_1(X))] + \frac{1}{n}\log c \overset{c\text{ is a constant}}{\to} H(X) + \mathrm{E}[\log(P_1(X))] = D(P_0\|P_1)$.

$\because$ log is an increasing function, and $\frac{\pi_1^{(0)}\prod_{i=1}^{n}P_1(X_i)}{\pi_0^{(0)}\prod_{i=1}^{n}P_0(X_i)} < \frac{\pi_1^{(0)}\prod_{i=1}^{n}P_1(X_i)}{\pi_0^{(0)}\prod_{i=1}^{n}P_0(X_i) + \pi_1^{(0)}\prod_{i=1}^{n}P_1(X_i)}$

$= \pi_1^{(n)}(X^n) < \frac{\pi_1^{(0)}\prod_{i=1}^{n}P_1(X_i)}{2\pi_0^{(0)}\prod_{i=1}^{n}P_0(X_i)}$ when $n \to \infty$.

$\therefore$ by squeeze theorem, $-\frac{1}{n}\log\pi_1^{(n)}(X^n) \to D(P_0\|P_1)$ as $n \to \infty$.

Problem 3.

(a) Let $X \sim P$.

$\mathrm{D}(P\|G(p)) = \sum_{x=1}^{\infty}P(x)\log\frac{P(x)}{Q(x)} = \sum_{x=1}^{\infty}P(x)\log\frac{P(x)}{(1-p)p^{x-1}} = \mathrm{H}(X) - \mathrm{E}[\log((1-p)p^{X-1})] = \mathrm{H}(X) - \log(1-p) - \mathrm{E}[(X-1)\log(p)] = H(X) - \log(1-p) - \log(p)E[X-1] = H(X) - \log(1-p) + \log p - \mu\log p$.

$\frac{d}{dp}\mathrm{D}(P\|G(p)) = \frac{1}{1-p} + \frac{1}{p} - \frac{1}{p}\mu = \frac{1-(1-p)\mu}{p(1-p)}$, which equals to 0 $\iff$ $\frac{1}{1-p} = \mu \iff p = 1 - \frac{1}{\mu}$.

One can also verify that if $p < 1 - \frac{1}{\mu}$, $\frac{d}{dp}\mathrm{D}(P\|G(p)) < 0$ and if $p > 1 - \frac{1}{\mu}$, $\frac{d}{dp}\mathrm{D}(P\|G(p)) > 0$.

$\therefore$ the minimum possible value of $\mathrm{D}(P\|G(p))$ occurs when $p = 1 - \frac{1}{\mu}$, that is, the distribution is $G(1-\frac{1}{\mu})$, and $\mathrm{D}(P\|G(p)) = H(X) - \log\mu + (1-\mu)\log(1-\mu)$.

(b) Let $X_i \sim P_i, Y \sim R$ where $R(y) := \frac{1}{m}\sum_{i=1}^{m}P_i(y)$.

From HW2 we know that $H(R) \le -\sum_{j=1}^{\infty}R(j)\log Q(j)$, with equality $\iff$ $Q \sim R$.

$\Rightarrow \sum_{i=1}^{m}\mathrm{D}(P_i\|Q) = \sum_{i=1}^{m}\left(H(X_i) - \sum_{j=1}^{\infty}P_i(j)\log Q(j)\right)$

$$
\begin{aligned}
&= \sum_{i=1}^{m} H(X_i) - \sum_{j=1}^{\infty} \left(\sum_{i=1}^{m} P_i(j) \right) \log Q(j) \\
&= \sum_{i=1}^{m} H(X_i) - m \sum_{j=1}^{\infty} R(j) \log Q(j) \\
&\geq \sum_{i=1}^{m} H(X_i) - mH(R).
\end{aligned}
$$

$\therefore \min\limits_{Q \in \mathcal{P}(X)} \sum\limits_{i=1}^{m} D(P_i \| Q) = \sum\limits_{i=1}^{m} H(X_i) - mH(R)$, with minimizer $Q = R$, that is, $Q(y) = \dfrac{1}{m} \sum\limits_{i=1}^{m} P_i(y)$.

Information Theory HW3

許博翔

November 2, 2023

Problem 1.

(a) Since N_0 is deterministic from $X_1, X_2, \dots, X_{N_0}$, N_1 is deterministic from $X_1, X_2, \dots, X_{N_1}$, there is $I(N_0; X_1, \dots, X_{N_0}) = H(N_0) = \frac{1}{3}\log 3 + \frac{2}{3}(\log 3 - 1) = \log 3 - \frac{2}{3}$, $I(N_1; X_1, \dots, X_{N_1}) = H(N_1) = \sum_{i=1}^{\infty} \frac{1}{2^i} i = \sum_{i=1}^{\infty} \frac{1}{2^i} \sum_{j=1}^{i} 1 = \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} \frac{1}{2^i} = \sum_{j=1}^{\infty} \frac{1}{2^{j-1}} = 2$.

(b) Let's assume $n \geq 2$ (because for $n = 1$ there is nothing to be computed).

Claim: $X_1, X_2, \dots, X_{n-1}$ are mutually independent Ber$(\frac{1}{2})$.

Proof. $\forall x \in [0,1]^{n-1}$, there is exactly one $x^* \in [0,1]^n$ (which is $(x_1, x_2, \dots, x_{n-1}, x_1 \oplus \dots \oplus x_{n-1})$) s.t. $2 \mid \sum_{i=1}^{n} x_i^*$ and $\forall 1 \leq i \leq n-1,\ x_i^* = x_i$.

$\therefore \Pr((X_1, \dots, X_{n-1}) = x) = \Pr((X_1, \dots, X_n) = x^*) = 2^{-(n-1)}$.

$\Rightarrow (X_1, \dots, X_{n-1})$ is an uniform distribution on $[0,1]^{n-1}$, which means $X_1, X_2, \dots, X_{n-1}$ are mutually independent Ber$(\frac{1}{2})$. ■

Similarly, for any distinct $i_1, i_2, \dots, i_{n-1}$, $X_{i_1}, \dots, X_{i_{n-1}}$ are mutually independent.

Let $1 \leq i \leq n-1$,

$$I(X_i; X_{i+1} | X_1, \dots, X_{i-1}) = H(X_i | X_1, \dots, X_{i-1}) - H(X_i | X_1, \dots, X_{i-1}, X_{i+1})$$
$$\overset{X_1,\dots,X_i \text{ are mutually independent}}{=} H(X_i) - H(X_i | X_1, \dots, X_{i-1}, X_{i+1})$$
$$\overset{X_1,\dots,X_{i+1} \text{ are mutually independent if } i<n-1}{=} \begin{cases} H(X_i) - H(X_i) = 0, \text{ if } i < n-1 \\ H(X_i) - H(X_1 \oplus \dots \oplus X_{i-1} \oplus X_{i+1} | X_1, \dots, X_{i-1}, X_{i+1}) = H(X_i) - 0 = 1, \text{ if } i = n-1 \end{cases}.$$

Problem 2.

(a) $I(X_4; X_1) = I(X_4; X_2) - I(X_4; X_2|X_1)$
$I(X_3; X_1) = I(X_3; X_2) - I(X_3; X_2|X_1) = I(X_3; X_2) - I(X_3; X_2|X_1, X_4) - I(X_4; X_2|X_1)$
$\Rightarrow I(X_1; X_3) + I(X_2; X_4) = I(X_3; X_2) - I(X_3; X_2|X_1, X_4) - I(X_4; X_2|X_1) + I(X_2; X_4) = I(X_2; X_3) + I(X_1; X_4) - I(X_3; X_2|X_1, X_4) \leq I(X_2; X_3) + I(X_1; X_4)$.

(b) It's equivalent to two Markov's chains: $X_1 - X_2 - X_3, X_1 - X_2 - X_4$.
$I(X_3; X_1) = I(X_3; X_2) - I(X_3; X_2|X_1)$
$I(X_4; X_1) = I(X_4; X_2) - I(X_4; X_2|X_1)$
$I(X_1; X_2) + I(X_3; X_4) \geq I(X_1; X_2) + I(X_4; X_1) - I(X_4; X_1|X_3) \geq I(X_2; X_1) + I(X_4; X_1) - I(X_2; X_1|X_3) = I(X_1; X_3) + I(X_1; X_4)$.

Problem 3.

(a) Let $X_i \in \mathcal{X}^{(i)}$.
$I(X;Y) = H(X) - H(X|Y) \overset{I \text{ is deterministic from } X}{=} H(X,I) - H(X|Y) = H(X|I) + H(I) - H(X|Y) \overset{I \text{ is deterministic from } Y}{=} H(X|I) + H(I) - H(X|Y,I) = I(X;Y|I) + H(I)$.

(b) The capacity $= \max_{P_I} I(X;Y) = \max_{P_I} E_{(X,Y)\sim P_{X,Y}}(\log \frac{P_{Y|X}(Y|X)}{P_Y(Y)}) = \max_{P_I} \sum_{i=1}^{l} P_I(i)(I(X_i;Y_i) - \log P_I(i)) = \max_{P_I} \left(\sum_{i=1}^{l} P_I(i) C^{(i)} + H(I) \right)$.

(c) Consider the distribution: $P_J(i) = \frac{2^{C^{(i)}}}{\sum_{j=1}^{l} 2^{C^{(j)}}}$.
$\sum_{i=1}^{l} P_I(i) C^{(i)} + H(I) = \sum_{i=1}^{l} P_I(i) \log \frac{2^{C^{(i)}}}{\sum_{j=1}^{l} 2^{C^{(j)}} P_I(i)} + \sum_{i=1}^{l} P_I(i) \log \sum_{j=1}^{l} 2^{C^{(j)}} = \sum_{i=1}^{l} P_I(i) \log \frac{P_J(i)}{P_I(i)} + \log \sum_{j=1}^{l} 2^{C^{(j)}} = -D(P_I \| P_J) + \log \sum_{j=1}^{l} 2^{C^{(j)}} \geq \log \sum_{j=1}^{l} 2^{C^{(j)}}$,
with equality $\iff D(P_I \| P_J) = 0 \iff P_I = P_J$.
$\therefore$ the capacity $= \log \sum_{j=1}^{l} 2^{C^{(j)}}$, and the distribution P_I is $P_I(i) = \frac{2^{C^{(i)}}}{\sum_{j=1}^{l} 2^{C^{(j)}}}$.

Problem 4.

(a) Suppose that $X \sim \text{Ber}(q)$.
$\Rightarrow P_Y(0) = 1 - q + pq = 1 - \frac{1}{2}q, P_Y(1) = q(1-p) = \frac{1}{2}q$.
$I(X;Y) = H(X) + H(Y) - H(X,Y) = -q \log q - (1-q)\log(1-q) - \frac{1}{2}q \log(\frac{1}{2}q) - (1 - \frac{1}{2}q)\log(1 - \frac{1}{2}q) + (1-q)\log(1-q) + 2 \cdot \frac{1}{2}q \log(\frac{1}{2}q) = -q \log q + \frac{1}{2}q \log(\frac{1}{2}q) - (1 - \frac{1}{2}q)\log(1 - \frac{1}{2}q) = -q - \frac{1}{2}q\log(\frac{1}{2}q) - (1-\frac{1}{2}q)\log(1-\frac{1}{2}q)$.
Let $\frac{dI(X;Y)}{dq} = -1 - \frac{1}{2}\log(\frac{1}{2}q) - \frac{1}{2}\log e + \frac{1}{2}\log(1 - \frac{1}{2}q) + \frac{1}{2}\log e = -1 + \frac{1}{2}\log \frac{1 - \frac{1}{2}q}{\frac{1}{2}q} = 0$.
$\Rightarrow \log \frac{1 - \frac{1}{2}q}{\frac{1}{2}q} = 2$.
$\Rightarrow \frac{1 - \frac{1}{2}q}{\frac{1}{2}q} = 4$.
$\Rightarrow q = \frac{2}{5}$.

$\therefore I(X;Y) \leq -\frac{2}{5} - \frac{1}{5}\log\frac{1}{5} - \frac{4}{5}\log\frac{4}{5} = -\frac{2}{5} - \frac{8}{5} + \log 5 = \log 5 - 2$, with equality iff $P_X^* = \text{Ber}(\frac{2}{5}), P_Y^* = \text{Ber}(\frac{1}{5})$.

(b) Since the equality in (a) is an if and only if condition, so the input distribution is unique.

(c) $D(P_{Y|X}(\cdot|0) \| P_Y^*(\cdot)) = P_{Y|X}(0|0)\log\frac{P_{Y|X}(0|0)}{P_Y^*(0)} = 1\log\frac{1}{1-\frac{1}{2}q} = -\log(1-\frac{1}{2}q) = \log 5 - 2$.

$D(P_{Y|X}(\cdot|1) \| P_Y^*(\cdot)) = P_{Y|X}(0|1)\log\frac{P_{Y|X}(0|1)}{P_Y^*(0)} + P_{Y|X}(1|1)\log\frac{P_{Y|X}(1|1)}{P_Y^*(1)} = \frac{1}{2}\log\frac{1}{2(1-\frac{1}{2}q)} + \frac{1}{2}\log\frac{1}{2(\frac{1}{2}q)} = -\frac{1}{2}(\log(2-q) + \log q) = -\frac{1}{2}(3 - \log 5 + 1 - \log 5) = \log 5 - 2$.

Information Theory HW5

許博翔

November 23, 2023

Problem 1.

(a) (1) From Gaussian integral, we know that $\int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}$.

$\int xe^{-x^2}dx = \int \frac{1}{2}e^{-x^2}d(x^2) = -\frac{1}{2}e^{-x^2} + c.$

$\lim_{x\to\infty} xe^{-x^2} = \lim_{x\to\infty} \frac{x}{e^{x^2}} \overset{\text{L'H}}{=} \lim_{x\to\infty} \frac{1}{2xe^{x^2}} = 0.$

$\lim_{x\to-\infty} xe^{-x^2} = \lim_{x\to-\infty} \frac{x}{e^{x^2}} \overset{\text{L'H}}{=} \lim_{x\to-\infty} \frac{1}{2xe^{x^2}} = 0.$

$\int_{-\infty}^{\infty} x^2e^{-x^2}dx = \int_{-\infty}^{\infty} xe^{-x^2} \cdot xdx = -\frac{1}{2}e^{-x^2}x\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -\frac{1}{2}e^{-x^2} \cdot 1dx = 0 + \frac{1}{2}\sqrt{\pi} = \frac{1}{2}\sqrt{\pi}.$

$f(x) = \frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2}, g(x) = \frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu_2}{\sigma_2})^2}.$

$$D(f\|g) = \int_{-\infty}^{\infty} \frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2} \log\left(\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2} / \frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu_2}{\sigma_2})^2}\right) dx$$
$$= \int_{-\infty}^{\infty} \frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2} \left(\log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2}\log e\left(-(\frac{x-\mu_1}{\sigma_1})^2 + (\frac{x-\mu_2}{\sigma_2})^2\right)\right) dx$$
$$= \int_{-\infty}^{\infty} \frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu_1}{\sigma_1})^2} \left(\log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{1}{2}\log e(\frac{x-\mu_1}{\sigma_1})^2\right.$$
$$\left. + \frac{1}{2}\log e(\frac{x-\mu_1}{\sigma_2})^2 + \log e(\frac{\mu_1-\mu_2}{\sigma_2})(\frac{x-\mu_1}{\sigma_2}) + \frac{1}{2}\log e(\frac{\mu_1-\mu_2}{\sigma_2})^2\right) dx$$
$$= \log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{1}{2}\log e + \frac{1}{2}\log e\frac{\sigma_1^2}{\sigma_2^2} + \frac{1}{2}\log e(\frac{\mu_1-\mu_2}{\sigma_2})^2$$
$$= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\log e}{2\sigma_2^2}(\sigma_1^2 - \sigma_2^2 + (\mu_1-\mu_2)^2).$$

(2) $f(x) = \frac{1}{\sqrt{2}\sigma_1}e^{-\frac{\sqrt{2}|x-\mu_1|}{\sigma_1}}, g(x) = \frac{1}{\sqrt{2}\sigma_2}e^{-\frac{\sqrt{2}|x-\mu_2|}{\sigma_2}}.$

$\int xe^{x}dx = e^{x}x - \int e^{x}dx = (x-1)e^{x} + c.$

$\int xe^{-x}dx = -e^{-x}x - \int -e^{-x}dx = -(x+1)e^{-x} + c.$

$\lim_{x\to\infty} e^{-x}x = 0.$

$\int_{-\infty}^{\infty} |x-a|e^{-|x-b|}dx = \int_{-\infty}^{\infty} |x+b-a|e^{-|x|}dx.$
If $c := a-b \geq 0$, then $\int_{-\infty}^{\infty} |x+b-a|e^{-|x|}dx = \int_{-\infty}^{0} (c-x)e^{x}dx + \int_{0}^{c} (c-x)e^{-x} + \int_{c}^{\infty} (x-c)e^{-x}dx$
$= c+1+(-ce^{-c}+c)+((c+1)e^{-c}-1)+(c+1)e^{-c}-ce^{-c} = 2c+2e^{-c}.$
If $c<0$, then $\int_{-\infty}^{\infty} |x-c|e^{-|x|}dx = \int_{-\infty}^{\infty} |x+c|e^{-|x|}dx = -2c+2e^{c}.$
$\therefore \int_{-\infty}^{\infty} |x-a|e^{-|x-b|}dx = 2|a-b|+2e^{-|a-b|}.$
$D(f\|g) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2}\sigma_1}e^{-\frac{\sqrt{2}|x-\mu_1|}{\sigma_1}} \log\left(\frac{1}{\sqrt{2}\sigma_1}e^{-\frac{\sqrt{2}|x-\mu_1|}{\sigma_1}} / \frac{1}{\sqrt{2}\sigma_2}e^{-\frac{\sqrt{2}|x-\mu_2|}{\sigma_2}}\right)dx$
$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2}\sigma_1}e^{-\frac{\sqrt{2}|x-\mu_1|}{\sigma_1}} \left(\log\left(\frac{\sigma_2}{\sigma_1}\right) + \sqrt{2}\log e(\frac{|x-\mu_2|}{\sigma_2} - \frac{|x-\mu_1|}{\sigma_1})\right)dx$
$= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\log e}{\sigma_1\sigma_2}(\frac{\sigma_1^2}{2})(2\cdot\frac{\sqrt{2}}{\sigma_1}|\mu_1-\mu_2| + 2e^{-\frac{\sqrt{2}}{\sigma_1}|\mu_1-\mu_2|}) - \frac{\log e}{\sigma_1^2}(\frac{\sigma_1^2}{2})2$
$= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1\log e}{\sigma_2}(\frac{\sqrt{2}}{\sigma_1}|\mu_1-\mu_2| + e^{-\frac{\sqrt{2}}{\sigma_1}|\mu_1-\mu_2|}) - \log e.$

(b) The first KL divergence $-$ the second KL divergence $= \frac{\log e}{2\sigma_2^2}(\sigma_1^2-\sigma_2^2) - \frac{\sigma_1\log e}{\sigma_2} - \log e = \frac{\log e}{2}\left((\frac{\sigma_1}{\sigma_2})^2 - 2(\frac{\sigma_1}{\sigma_2})+1\right) = \frac{\log e}{2}(\frac{\sigma_1}{\sigma_2}-1)^2 \geq 0.$
$\therefore$ the first KL divergence $\geq$ the second KL divergence, the equation holds $\iff \sigma_1 = \sigma_2$.

(c) Let $x := |\mu_1-\mu_2|$.
The first KL divergence $-$ the second KL divergence $= \frac{\log e}{2}(\mu_1-\mu_2)^2 - \log e(\frac{\sqrt{2}}{\sigma_1}|\mu_1-\mu_2| + e^{-\frac{\sqrt{2}}{\sigma_1}|\mu_1-\mu_2|}) + \log e$
$= \frac{\log e}{2}x^2 - \log e(\frac{\sqrt{2}}{\sigma_1}x + e^{-\frac{\sqrt{2}}{\sigma_1}x}) + \log e$
$= \log e(\frac{1}{2}x^2 - \frac{\sqrt{2}}{\sigma_1}x - e^{-\frac{\sqrt{2}}{\sigma_1}x} + 1).$
$\therefore$ the first KL divergence is the larger $\iff \frac{1}{2}x^2 - \frac{\sqrt{2}}{\sigma_1}x - e^{-\frac{\sqrt{2}}{\sigma_1}x} + 1 \geq 0.$

Problem 2.

(a) $h(X) = E_{X\sim f_X}(\log\frac{1}{f_X(X)}) = \int_{-\infty}^{\infty} \frac{1}{2b}e^{-\frac{|x-\mu|}{b}}(\log(2b) + \log e\frac{|x-\mu|}{b})dx = \log(2b) + \log e\int_{\mu}^{\infty} \frac{1}{b}e^{-\frac{(x-\mu)}{b}}\frac{x-\mu}{b}dx = \log(2b) + \log e = \log(2be).$

(b) From Problem 1 (a)(2), we know that $\int_{-\infty}^{\infty} |x-a|e^{-|x-b|}dx = 2|a-b| + 2e^{-|a-b|}$.

$E(|X|) = \int_{-\infty}^{\infty} |x|\frac{1}{2b}e^{-\frac{|x-\mu|}{b}}dx = \frac{1}{2b}b^2(2|\mu| + 2e^{-|\mu|}) = b(|\mu| + e^{-|\mu|})$.

Let $g(y) := y + e^{-y}$.

$\Rightarrow g'(y) = 1 - e^{-y} > 0$ when $y > 0$.

$\Rightarrow g(y)$ is strictly increasing on $(0, \infty)$.

$\Rightarrow b(|\mu| + e^{-|\mu|}) = bg(|\mu|) \overset{(1)}{\geq} bg(0) = 2b$.

$\Rightarrow 2b \leq E(|X|) \leq B$.

$\Rightarrow b \overset{(2)}{\leq} \frac{B}{2}$.

$\Rightarrow h(X) = \log(2be) \leq \log Be$, and when the equation holds, the distribution of X is Lap$(0, \frac{B}{2})$ since the equation in (1) holds $\iff$ $\mu = 0$, and the equation in (2) holds.

Problem 3.

(a) Consider $\tilde{b}(x) := \mathrm{E}[b(x,Y)] = \mathrm{E}_{P_{Y|X}}[b(x,Y)]$.

Since $\tilde{b}(x) = \sum_y P_{Y|X}(y|x)b(x,y)$ is a deterministic function of x, $\tilde{b}(x)$ is an input-only cost function.

$\because \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{Y_i}[b(x_i, Y_i)] = \frac{1}{n}\sum_{i=1}^{n} \tilde{b}(x_i)$.

$\therefore$ the cost constraint becomes: $\frac{1}{n}\sum_{i=1}^{n} \tilde{b}(x_i) \leq B$.

Therefore, this problem is equivalent to the channel coding problem with input-cost only function $\tilde{b}$.

From Theorem 1 in Lecture 5, $C(B) = \max_{P_X : \mathrm{E}_{P_X}[\tilde{b}(X)] \leq B} I(X;Y)$

$= \max_{P_X : \mathrm{E}_{P_X}[\mathrm{E}_{P_{Y|X}}[b(X,Y)]] \leq B} I(X;Y) = \max_{P_X : \mathrm{E}_{P_X P_{Y|X}}[b(X,Y)] \leq B} I(X;Y)$.

(b) First, $P_{Y|X}(y|x) = P_Z(y-x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{y-x}{\sigma})^2}$.

Let $b(x,y) := y^2$.

The cost constraint is $\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{Y_i}[b(x_i, Y_i)] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{Y_i}[Y_i^2] \leq B$.

From the formula in Problem 1 (a)(1):

$\tilde{b}(x) := \mathrm{E}[b(x,Y)] = \int_{-\infty}^{\infty} P_{Y|X}(y|x)b(x,y)dy = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{y-x}{\sigma})^2}y^2dy$

$= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{y-x}{\sigma})^2}((y-x)^2 + 2(y-x)x + x^2)dy$

$= \sigma^2 + 0 + x^2 = \sigma^2 + x^2$.

$\Rightarrow$ the cost constraint becomes $\frac{1}{n}\sum_{i=1}^{n}(\sigma^2 + x_i^2) = \frac{1}{n}\sum_{i=1}^{n}\tilde{b}(x_i) \le B$, which is $\frac{1}{n}\sum_{i=1}^{n}|x_i|^2 \le B - \sigma^2$.

From the example of Guasian channel capacity in Lecture 5, we get that $C(B) = \frac{1}{2}\log(1 + \frac{B-\sigma^2}{\sigma^2}) = \frac{1}{2}\log(\frac{B}{\sigma^2})$.

Problem 4. In HW2, we know that if $\sum_i p_i = \sum_i q_i = 1$ where $p_i, q_i \ge 0$, then $\sum_i p_i \log\frac{1}{p_i} \le \sum_i p_i \log\frac{1}{q_i}$. – (1)

(a) $D_{\min} = \min_{\mathbf{q}(s)} \mathrm{E}[d(S, \mathbf{q}(S))] = \min_{\mathbf{q}(s)} \mathrm{E}[\log\frac{1}{\mathbf{q}(S)}] = 0$ if $\mathbf{q}(s) = \mathbb{I}\{S = s\}$.

$D_{\max} = \max_{\mathbf{q}} \mathrm{E}[d(S, \mathbf{q})] = \min_{\mathbf{q}} \mathrm{E}[\log\frac{1}{\mathbf{q}(S)}]$.

$\because \mathrm{E}[\log\frac{1}{\mathbf{q}(S)}] = \sum_s P_S(s)\log\frac{1}{\mathbf{q}(S)} \overset{(1)}{\ge} \sum_s P_S(s)\log\frac{1}{P_S(s)} = H(S) = H(\pi)$, and the equation holds when $\mathbf{q}(s) = P_S(s)$.

$\therefore D_{\max} = H(\pi)$.

(b) $H(S|\mathbf{Q}) = \mathrm{E}_{(S,\mathbf{Q})\sim P}[\log\frac{1}{P_{S|\mathbf{Q}}}] = \sum_{\mathbf{q}} P_{\mathbf{Q}}(\mathbf{q}) \sum_s P_{S|\mathbf{Q}}(s|\mathbf{q}) \log\frac{1}{P_{S|\mathbf{Q}}(s|\mathbf{q})}$

$\overset{(1)}{\le} \sum_{\mathbf{q}} P_{\mathbf{Q}}(\mathbf{q}) \sum_s P_{S|\mathbf{Q}}(s|\mathbf{q}) \log\frac{1}{\mathbf{q}(s)} = \mathrm{E}_{(S,\mathbf{Q})\sim P}\left[\log\frac{1}{\mathbf{Q}(S)}\right]$.

(c) $R(D) = \inf_{(S,\mathbf{Q})}\left\{I(S;\mathbf{Q}) \middle| \mathrm{E}[\log\frac{1}{\mathbf{Q}(S)}] \le D \text{ and } S \sim \pi\right\}$

$= \inf_{(S,\mathbf{Q})}\left\{I(S;\mathbf{Q}) \middle| H(S|\mathbf{Q}) \le \mathrm{E}[\log\frac{1}{\mathbf{Q}(S)}] \le D \text{ and } S \sim \pi\right\}$

$\overset{(2)}{\le} \inf_{(S,\mathbf{Q})}\{I(S;\mathbf{Q})\,|H(S|\mathbf{Q}) \le D \text{ and } S \sim \pi\}$

$\overset{(3)}{\le} \inf_{(S,\mathbf{Q})}\{I(S;\mathbf{Q})\,|H(S|\mathbf{Q}) \le D \text{ and } S \sim \pi \text{ and } \mathbf{Q}(\hat{s}) = 1 \text{ for some } \hat{s} \in \mathcal{S}\}$

$= \min_{(S,\hat{S})}\left\{I(S;\hat{S}) \middle| H(S|\hat{S}) \le D \text{ and } S \sim \pi\right\}$.

(d) Let $\mathbf{q}_{\hat{s}}(s) := \mathbb{I}(s = \hat{s})$.

Consider the distribution $\mathbf{Q} = \mathbf{q}_S$:

The equation in (2) holds $\iff$ the equation in (1) holds $\iff \forall s, \mathbf{q},\ P_{S|\mathbf{Q}}(s|\mathbf{q}) = \mathbf{q}(s)$, which is true because $\forall\mathbf{q}$ with nonzero probability, $\mathbf{q} = \mathbf{q}_{\hat{s}}$ for some $\hat{s}$, and $\mathbf{q}_{\hat{s}}(s) = \mathbb{I}(s = \hat{s}) \overset{\mathbf{Q}=\mathbf{q}_S}{=} P_{S|\mathbf{Q}}(s|\mathbf{q}_{\hat{s}})$.

The equation in (3) holds since $\mathbf{q}_{\hat{s}} = 1$ for $\hat{s} \in S$.

$\therefore$ with this distribution, $R(D) = \min_{(S,\hat{S})} \left\{ I(S;\hat{S}) \,\middle|\, H(S|\hat{S}) \leq D \text{ and } S \sim \pi \right\}$

$= \min_{(S,\hat{S})} \left\{ H(S) - H(S|\hat{S}) \,\middle|\, H(S|\hat{S}) \leq D \text{ and } S \sim \pi \right\}$

$= \min_{(S,\hat{S})} \left\{ H(\pi) - H(S|\hat{S}) \,\middle|\, H(S|\hat{S}) \leq D \text{ and } S \sim \pi \right\}$

$= H(\pi) - D \overset{0 \leq D \leq H(\pi) \text{ is given}}{=} \max(0, H(\pi) - D).$