1 Distribution

$\mathrm{Bin}(n,p): P_X(x)=\binom{n}{x}p^x q^{n-x}$ for $x\in[n]_0$.

$\mu=np, \sigma^2=npq, H(X)=\frac{1}{2}\log(2\pi enpq)+O(\frac{1}{n})$.

$\mathrm{Pois}(\lambda): P_X(x)=\frac{\lambda^x e^{-\lambda}}{x!}$ for $x\in\mathbb{N}_0$.

$\mu=\sigma^2=\lambda$.

$\mathrm{Geo}(p): P_X(x)=q^{x-1}p$ for $x\in\mathbb{N}$.

$\mu=\frac{1}{p}, \sigma^2=\frac{q}{p^2}, H(X)=\frac{-q\log q-p\log p}{p}$.

$\mathrm{Exp}(\lambda): f_X(x)=\lambda e^{-\lambda x}$ for $x\in\mathbb{R}_0^+$.

$\mu=\frac{1}{\lambda}, \sigma^2=\frac{1}{\lambda}$.

$\mathcal{N}(\mu,\sigma^2): f_X(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

$h(X)=\frac{1}{2}\log(2\pi e\sigma^2)$.

$\mathrm{Lap}(\mu,b): f_X(x)=\frac{1}{2b}e^{-\frac{|x-\mu|}{b}}$.

$\sigma^2=2b^2, h(X)=\log(2be)$.

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: $P_{X_1,\dots,X_n}=P_{X_{1+l},\dots,X_{n+l}},\ \forall n,l\in\mathbb{N}$.

3 Central Limit Theorem

Khinchin WLLN: $X_1,X_2,\dots,$ 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: $X_1,X_2,\dots,$ are i.i.d. with $E[|X_i|]<\infty$, then $\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma}\xrightarrow{d}Z\sim N(0,1)$.

Berry-Esseen: $X_1,X_2,\dots,$ are i.i.d. with $E[|X_i-\mu|^3]=\rho_3<\infty$. Let $Z_n:=\frac{\bar{S}_n-n\mu}{\sqrt{n}\sigma}, Z\sim N(0,1)$. 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 $\mathcal{X}$.

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

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

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

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

The above still holds for h.

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

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

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

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

$\Rightarrow\Pr\{U\neq V\}\geq\frac{H(U|V)-1}{\log|\mathcal{U}|}$.

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

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

H and $\bar{\mathrm{H}}$ 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 $\mathrm{H}(\{X_i\})=\bar{\mathrm{H}}(\{X_i\})$.

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

Shannon-McMillan-Breiman theorem: if $\{S_i\}$ is stationary ergodic, then $\frac{1}{n}\log\frac{1}{P(S^n)}\xrightarrow{a.s.,L^1}\mathrm{H}(\{S_i\})$ 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: $\pi(x)=\sum_{y\in\mathcal{X}}\pi(y)P_{X_{n+1}|X_n}(x|y),\ \forall x\in\mathcal{X},\ n\in\mathbb{N}$. 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 $\{X_i\}$, $\mathrm{H}(\{X_i\})=\bar{\mathrm{H}}(\{X_i\})=H(X_2|X_1)|_{P_{X_1}=\pi}=\sum_{x\in\mathcal{X}}\pi(x)H(X_2|X_1=x)$, where π is the unique steady-state distribution.

7 Information for Continuous Distributions

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

8 Learning a Bit of Information

$\pi_{10}(\phi)$: false alarm, false positive, false rejection, type I error.

$\pi_{01}(\phi)$: 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 $\pi_{01}(\phi)$ subject to $\pi_{10}(\phi)\leq\epsilon$.

Neyman-Pearson: LRT is optimal.

Generalized to n i.i.d.: $\phi^n_{q_0,\gamma_n}(x^n)=\begin{cases}1, \text{if } \sum_{i=1}^n LLR(x_i)\gtrless\eta_n\\ \gamma_n, \text{if } \sum_{i=1}^n LLR(x_i)=\eta_n\end{cases}$.

Chernoff-Stein lemma: $\lim_{n\to\infty}-\frac{1}{n}\log\omega_{01}^n(n,\epsilon)=D(P_0\|P_1)$.

Typical set:

9 Information Divergence

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

$D(P\|Q)\geq 0$, 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: $D(P_{XY}\|Q_{XY})=D(P_{Y|X}\|Q_{Y|X}|P_X)+D(P_X\|Q_X)$.

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

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

10 Error Exponents and Chernoff Information

P_0, P_1 are given.

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

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

$P_{e0}^{*n}(\pi(=(\pi_0,\pi_1)),n):=\min\{\pi_0\pi_{10}^n(\phi)+\pi_1\pi_{01}^n(\phi)\}$.

$\bar{P}_e^{*n}(n):=\min_\phi\{\max\{\pi_{10}^n,\pi_{01}^n\}\}$.

Chernoff Information: $CI(P_0,P_1):=\max_{0\leq\lambda\leq 1}-\log\sum_{a\in\mathcal{X}}P_0(a)^{1-\lambda}P_1(a)^\lambda$.

$f(\lambda)$

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

11 Deviverling Information Reliably

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

12 Mutual Information

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

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

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

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

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

$R^I(D_{\min})=H(S), R^I(D)=0$ 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: $R(D)=\begin{cases}\frac{1}{2}\log(\frac{\sigma^2}{D}), \text{if } 0\leq D\leq\sigma^2\\ 0, \text{if } D>\sigma^2\end{cases}$.

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

Homework 1 Solution

TA: Heng-Chien Liou *

r11942067@ntu.edu.tw

1. (Another kind of typical sequences) [18]

In this problem, let us consider another kind of typical sequences defined as follows.

Definition. For $\gamma\in(0,1)$, a sequence s^n is called γ-typical with respect to a DMS $S\sim\mathsf{P}_S$ if

$$|\pi(a|s^n)-\mathsf{P}_S(a)|\leq\gamma\mathsf{P}_S(a),\ \forall a\in\mathcal{S},$$

where $\pi(a|s^n):=\frac{1}{n}\sum_{i=1}^n\mathbb{1}\{s_i=a\}$. The γ-typical set

$$\mathcal{T}_\gamma^{(n)}(S):=\{s^n\in\mathcal{S}^n|\,s^n\text{ is }\gamma\text{-typical with respect to }S\}.$$

In the following, please assume $|\mathcal{S}|<\infty$.

a) Show that the typical sequence and typical set defined above also satisfy AEP (Proposition 1 of Unit 1) with $\mathcal{A}_\delta^{(n)}(S)$ replace by $\mathcal{T}_\gamma^{(n)}(S)$, and the δ in properties 1, 3, and 4 replaced by something, denoted by $\xi(\gamma)$, depending on γ. Specify this $\xi(\gamma)$. [6]

b) Show that $\mathcal{T}_\gamma^{(n)}\subseteq\mathcal{A}_\delta^{(n)}$ where $\delta=\xi(\gamma)$ found in a). [6]

c) Find an alphabet $\mathcal{S}$, a reference probability mass function P_S, and γ such that $\forall\delta'>0, n\in\mathbb{N}, \mathcal{A}_{\delta'}^{(n)}\nsubseteq\mathcal{T}_\gamma^{(n)}$. [6]

Remark. From b) and c) we see that the typicality defined in this problem is *stronger* than that defined in the lecture. Hence, they are called strong typicality and weak typicality respectively in the literature.

Solution:

a) **Property 1** $\forall s^n\in\mathcal{T}_\gamma^{(n)}$, we have $\forall a\in\mathcal{S}, |\pi(a|s^n)-\mathsf{P}_S(a)|\leq\gamma\mathsf{P}_S(a)$

$$\frac{1}{n}\log\Pr\{S^n=s^n\}=\sum_{a\in\mathcal{S}}\pi(a|s^n)\log\mathsf{P}_S(a)$$

*With contributions by Wen-Shao Ho and Chen-Hao Hsiao.

$$\begin{cases}\leq\sum_{a\in\mathcal{S}}(1-\gamma)\mathsf{P}_S(a)\log\mathsf{P}_S(a)=-(1-\gamma)\mathrm{H}(S)\\ \geq\sum_{a\in\mathcal{S}}(1+\gamma)\mathsf{P}_S(a)\log\mathsf{P}_S(a)=-(1+\gamma)\mathrm{H}(S)\end{cases}.$$

Hence, with $\xi(\gamma)=\gamma\mathrm{H}(S)$

$$\Pr\{S^n=s^n\}=\sum_{a\in\mathcal{S}}\pi(a|s^n)\log\mathsf{P}_S(a)$$

$$\begin{cases}\leq 2^{-n(1-\gamma)\mathrm{H}(S)}=2^{-n(\mathrm{H}(S)-\xi(\gamma))}\\ \geq 2^{-n(1+\gamma)\mathrm{H}(S)}=2^{-n(\mathrm{H}(S)+\xi(\gamma))}\end{cases}.$$

Property 2 Consider any $\epsilon>0$. By basic definition on set and probability,

$$\begin{aligned}\Pr\{s^n\in\mathcal{T}_\gamma^{(n)}\}&=\Pr\{\forall a\in\mathcal{S},|\pi(a|s^n)-\mathsf{P}_S(a)|\leq\gamma\mathsf{P}_S(a)\}\\&=\Pr\bigcap_{a\in\mathcal{S}}\{|\pi(a|s^n)-\mathsf{P}_S(a)|\leq\gamma\mathsf{P}_S(a)\}\\&=1-\Pr\bigcup_{a\in\mathcal{S}}\{|\pi(a|s^n)-\mathsf{P}_S(a)|>\gamma\mathsf{P}_S(a)\}\end{aligned}$$

For given $a\in\mathcal{S}$, by the weak law of large number, $\exists n_a\in\mathcal{N}$ such that $\forall n\geq n_a$

$$\Pr\{|\pi(a|s^n)-\mathsf{P}_S(a)|>\gamma\mathsf{P}_S(a)\}\leq\frac{\epsilon}{|\mathcal{S}|}.$$

Select $n_0=\max\{n_a:a\in\mathcal{S}\}$, then $\forall n\geq n_0$, by union bound

$$\Pr\bigcup_{a\in\mathcal{S}}\{|\pi(a|s^n)-\mathsf{P}_S(a)|>\gamma\mathsf{P}_S(a)\}\leq\sum_{a\in\mathcal{S}}\Pr\{|\pi(a|s^n)-\mathsf{P}_S(a)|>\gamma\mathsf{P}_S(a)\}$$

$$\leq\sum_{a\in\mathcal{S}}\frac{\epsilon}{|\mathcal{S}|}=\epsilon$$

Property 3 by Property 1,

$$1\geq\Pr\{s^n\in\mathcal{T}_\gamma^{(n)}\}=\sum_{s^n\in\mathcal{T}_\gamma^{(n)}}\Pr\{S^n=s^n\}\geq|\mathcal{T}_\gamma^{(n)}|2^{-n(\mathrm{H}(S)+\xi(\gamma))}$$

Property 4 by Property 1 and 2, $\forall n\geq n_0$

$$1-\epsilon\leq\Pr\{s^n\in\mathcal{T}_\gamma^{(n)}\}=\sum_{s^n\in\mathcal{T}_\gamma^{(n)}}\Pr\{S^n=s^n\}\leq|\mathcal{T}_\gamma^{(n)}|2^{-n(\mathrm{H}(S)-\xi(\gamma))}$$

b) $\forall s^n=(s_1,s_2,\dots,s_n)\in\mathcal{T}_\gamma^{(n)}$, by the previous argument and the property of memorylessness:

$$-(1+\gamma)\mathrm{H}(S)\leq\frac{1}{n}\sum_{i=1}^n\log\mathsf{P}_S(s_i)=\frac{1}{n}\log\Pr\{S^n=s^n\}\leq-(1-\gamma)\mathrm{H}(S).$$

$\mathcal{T}_\gamma^{(n)}\subseteq\mathcal{A}_\delta^n$ with $\delta=\xi(\gamma)=\gamma\mathrm{H}(S)$

c) For some $\gamma<1$, consider $\mathcal{S}=\{0,1\}$ and $\mathsf{P}_S(0)=\mathsf{P}_S(1)=\frac{1}{2}$. Consider $s^n=0^n$. $\mathrm{H}(S)=1\Rightarrow 0^n\in\mathcal{A}_\delta^n,\forall\delta>0$. But $0^n\notin\mathcal{T}_\gamma^{(n)}$.

Grading Policy:

a) Property 1 worths [2] points, Property 2 worths [3] points, and Property 3, 4 worth [1] point in total.

b) Correct arguments about set relations [4] and correct arithemtic details [2].

c) A correct example [3], and justification [3].

2. (Finer asymptote for lossless source coding achievability) [20]

Consider a discrete memoryless source $S_i\overset{\text{i.i.d.}}{\sim}\mathsf{P}_S,\ i=1,2,\dots,$ where P_S is the PMF of the source S. Let $\mathrm{R}>\mathrm{H}(S)$, that is,

$$\mathrm{R}=\mathrm{H}(S)+\delta,$$

where $\delta>0$ denotes a constant. Then, it can be shown as a corollary of the lossless source coding theorem in our lecture that there exists a sequence of $(n,\lfloor nR\rfloor)$ codes such that $\forall\epsilon>0$,

$$\mathsf{P}_e^{(n)}\leq\epsilon\quad\text{for } n \text{ sufficiently large.}\qquad(\dagger)$$

Notably, the gap to the fundamental limit, δ, is a constant not depending on ϵ.

Suppose we would like to achieve ($\dagger$) for a given $\epsilon\in(0,1/2)$. Is it possible to derive a finer asymptote for $\mathrm{R}-\mathrm{H}(S)$, the gap to the fundamental limit? In this problem, we are going to show that $\mathrm{R}-\mathrm{H}(S)=\Theta(n^{-1/2})$ suffices.

a) (Warm-up) Let $\varsigma(S)>0$ denote the standard deviation of $\log\frac{1}{\mathsf{P}_S(S)}$ when $S\sim\mathsf{P}_S$ and $\Phi(\cdot)$ denote the CDF of a standard normal RV. Use the central limit theorem to prove the following

$$\lim_{n\to\infty}\Pr\left\{\prod_{i=1}^n\mathsf{P}_S(S_i)\geq 2^{-n\left(\mathrm{H}(S)+n^{-1/2}\delta\varsigma(S)\right)}\right\}=\Phi(\delta)\qquad[8]$$

With the above, if we define a set of length-n source sequences $\mathcal{B}_\delta^{(n)}(S)$ as follows:

$$\mathcal{B}_\delta^{(n)}(S):=\left\{s^n\,\middle|\,\prod_{i=1}^n\mathsf{P}_S(s_i)\geq 2^{-n\left(\mathrm{H}(S)+n^{-1/2}\delta\varsigma(S)\right)}\right\},$$

one can control the probability of each sequence in $\mathcal{B}_\delta^{(n)}(S)$ from below and hence can control the cardinality of this set from above. Also, we know that $\Pr\{S^n\in\mathcal{B}_\delta^{(n)}(S)\}\to\Phi(\delta)$ as $n\to\infty$ from Part a). It is hence tempting to use label all the sequences in $\mathcal{B}_\delta^{(n)}$ and give up the rest as a source encoding scheme. However, to upper bound the error probability, knowing its limit as $n\to\infty$ is not enough. Berry-Esseen theorem is a standard refinement of the CLT.

b) Show that ($\dagger$) can be attained using the aforementioned scheme if the rate approaches $\mathrm{H}(S)$ from above as $n\to\infty$ in the following manner:

$$\mathrm{R}_n=\mathrm{H}(S)-n^{-1/2}\varsigma(S)\Phi^{-1}(\epsilon)+\zeta_n$$

where $\zeta_n=O(n^{-1})$ denotes a positive sequence tends to zero not slower than n^{-1}. [12]

Remark. The above is not optimal – the optimal asymptote of the rate (when $\varsigma(S)>0$) is

$$\mathrm{R}_n=\mathrm{H}(S)-n^{-1/2}\varsigma(S)\Phi^{-1}(\epsilon)-\frac{\log n}{2n}+O(1/n).$$

Solution:

a) Central limit theorem states that for i.i.d random variables $X_1,\cdots,X_n$ with finite mean μ, variance σ^2. As $n\to\infty$, the random variable $\sqrt{n}(\bar{X}_n-\mu)$ converges in distribution to $\mathcal{N}(0,\sigma^2)$, where $\bar{X}_n:=\frac{1}{n}\sum_{i=1}^n X_i$.
Namely, for any real number z,

$$\lim_{n\to\infty}\Pr\left\{\sqrt{n}(\bar{X}_n-\mu)\leq z\right\}=\Phi\left(\frac{z}{\sigma}\right).$$

Choose $X_i=\log\frac{1}{\mathsf{P}_S(S_i)}$ and $z=\delta\varsigma(S)$ in the Central limit theorem:

$$\begin{aligned}&\lim_{n\to\infty}\Pr\left\{\sqrt{n}\left(-\frac{1}{n}\log\prod_{i=1}^n P_S(S_i)-\mathsf{E}\left[\log\frac{1}{P_S(S)}\right]\right)\leq\delta\varsigma(S)\right\}=\Phi(\delta)\\&\Leftrightarrow\lim_{n\to\infty}\Pr\left\{\left(-\frac{1}{n}\log\prod_{i=1}^n P_S(S_i)\right)\leq H(S)+n^{-1/2}\delta\varsigma(S)\right\}=\Phi(\delta)\\&\Leftrightarrow\lim_{n\to\infty}\Pr\left\{\prod_{i=1}^n P_S(S_i)\geq 2^{-n\left(H(S)+n^{-1/2}\delta\varsigma(S)\right)}\right\}=\Phi(\delta).\end{aligned}$$

b) Note that we can now label all the sequences in $\mathcal{B}_\delta^{(n)}(S)$ and give up the rest, we want to find $\delta(\epsilon)$ such that when n is sufficiently large,

$$\Pr\left\{S^n\in\mathcal{B}_{\delta(\epsilon)}^{(n)}(S)\right\}\geq 1-\epsilon.$$

Combine Berry-Esseen theorem and subproblem a), we know that

$$\Pr\left\{S^n\in\mathcal{B}_{\delta(\epsilon)}^{(n)}(S)\right\}\geq\Phi(\delta(\epsilon))-c\frac{\rho^3}{\varsigma(S)^3}n^{-1/2}.$$

Therefore, it suffices to find $\delta(\epsilon)$ such that

$$\Phi(\delta(\epsilon))-c\frac{\rho^3}{\varsigma(S)^3}n^{-1/2}\geq 1-\epsilon\Leftrightarrow\delta(\epsilon)\geq\Phi^{-1}\left(1-\epsilon+c\frac{\rho^3}{\varsigma(S)^3}n^{-1/2}\right).$$

The order of $\Phi^{-1}\left(1-\epsilon+c\frac{\rho^3}{\varsigma(S)^3}n^{-1/2}\right)$ can be further characterized using Taylor expansion at $1-\epsilon$ on the function $\Phi^{-1}(x)$:

$$\Phi^{-1}(x)=\Phi^{-1}(1-\epsilon)+\frac{\Phi^{-1}(1-\epsilon)}{1!}(x-(1-\epsilon))+O(x^2).$$

Plug in $x = 1-\epsilon+c\frac{\rho^2}{\zeta(S)^3}n^{-1/2}$ we have

$$\Phi^{-1}\left(1-\epsilon+c\frac{\rho^3}{\zeta(S)^3}n^{-1/2}\right) = \Phi^{-1}(1-\epsilon)+O(n^{-1/2}) = -\Phi^{-1}(\epsilon)+O(n^{-1/2}).$$

The result above indicates that if we choose $\delta(\epsilon) = -\Phi^{-1}(\epsilon)+O(n^{-1/2})$, we can guarantee

$$\mathsf{Pr}\left\{S^n \in \mathcal{B}^{(n)}_{\delta(\epsilon)}(S)\right\} \geq 1-\epsilon.$$

Similar to the proof of the properties of the typical set, for n sufficiently large:

$$1 \geq \mathsf{Pr}\left\{S^n \in \mathcal{B}^{(n)}_{\delta(\epsilon)}(S)\right\} \geq |\mathcal{B}^{(n)}_{\delta(\epsilon)}(S)| 2^{-n\left(H(S)+n^{-1/2}\delta(\epsilon)\zeta(S)\right)}$$
$$\Rightarrow |\mathcal{B}^{(n)}_{\delta(\epsilon)}(S)| \leq 2^{n\left(H(S)+n^{-1/2}\delta(\epsilon)\zeta(S)\right)} = 2^{n\left(H(S)-n^{-1/2}\zeta(S)\Phi^{-1}(\epsilon)+O(n^{-1})\right)}.$$

Therefore, the choice

$$R_n = H(S) - n^{-1/2}\zeta(S)\Phi^{-1}(\epsilon)+O(n^{-1}),$$

makes it possible to encode all the sequences in $\mathcal{B}^{(n)}_{\delta(\epsilon)}(S)$ and also satisfies the requirement that $\mathsf{Pr}\left\{S^n \in \mathcal{B}^{(n)}_{\delta(\epsilon)}(S)\right\} \geq 1-\epsilon$.

Grading Policy:

a) Justification for your application of central limit theorem [4], other details [4].

b) Overall logic in scheme and analysis [4], application of Berry-Esseen theorem [4], other details such as mean value theorem or Taylor's theorem [4].

3. (An alternative lossless source coding theorem) [12]

For a discrete memoryless source $\{S_i \,|\, i \in \mathbb{N}\}$, consider a sequence of $(n, \lfloor nR \rfloor)$ source codes indexed by $n = 1, 2, \ldots$ with compression rate $R > 0$.

Prove the following statements.

a) If $R > H(S)$, there exist a sequence of $(n, \lfloor nR \rfloor)$ codes with

$$\lim_{n\to\infty} \mathsf{P}_e^{(n)} = 0.$$

In other words, the probability of error can be driven to zero as $n \to \infty$. [6]

b) If $R < H(S)$, for any sequence of $(n, \lfloor nR \rfloor)$ codes, the sequence of error probabilities must converge to 1, that is,

$$\lim_{n\to\infty} \mathsf{P}_e^{(n)} = 1. \quad [6]$$

Solution:

a) Let $\delta = R - H(S)$. A simple scheme is to use $\mathcal{A}^{(n)}_{\delta/2}$ and standard asymptotic equipartition property for coding. Here we try to prove it using the lossless source coding theorem in the lecture.

Consider a sequence $\{\epsilon_m\}_{m=1}^{\infty} > 0$ such that $\lim_{m\to\infty}\epsilon_m = 0$. For any $n, \epsilon_m > 0$, consider $\exists k^*$ such that for any $k \geq k^*(n,\epsilon_m)$, there exists a (n,k) code with $\mathsf{P}_e^{(n)} < \epsilon_m$. By the lossless source coding theorem in the lecture and the Archimedean property, $\exists n_m > \max\{n_{m-1}, \{n_i\}\}$ such that $\forall n \geq n_m, k^*(n,\epsilon_m) < n\left(\mathrm{H}(S)+\delta/2\right) < \lfloor nR \rfloor$. $\{n_m\}_m^{\infty}$ then is a strictly increasing sequence of integers.

Therefore, we can construct a sequence of $(n, \lfloor nR \rfloor)$ codes as follows: for n such that $n_m \leq n < n_{m+1}$ select $k = k^*(n,\epsilon_m)$ and one of the corresponding (n,k) code $(\mathrm{enc}_n, \mathrm{dec}_n)$ such that $\mathsf{P}_e^{(n)} < \epsilon_m$. By the above justification, $k < \lfloor nR \rfloor$.

$\{(\mathrm{enc}_n, \mathrm{dec}_n)\}_n$ is then a desirable code with the property.

b) Let $\delta = \mathrm{H}(S) - R$. Again, a simple strategy is to use $\mathcal{A}^{(n)}_{\delta/2}$ and standard asymptotic equipartition property to characterize the error probability for any sequence of codes. Here we try to prove it using the lossless source coding theorem in the lecture.

We attempt to prove by contradiction. Suppose there exist a sequence of $(n, \lfloor nR \rfloor)$ codes with error probabilities not converging to 1, namely, for some $\epsilon < 0$, $\forall m \in \mathbb{N}, \exists n'(m) \geq m$ such that $\mathsf{P}_e^{(n'(m))} < 1-\epsilon$.

$\forall m \in \mathbb{N}$, the $(n'(m), \lfloor n'(m)R \rfloor)$ code from the above sequences, by definition, is then a $(n'(m), \lfloor n'(m)R \rfloor, 1-\epsilon)$ code.

$$\mathsf{P}_e^{(n'(m))} < 1-\epsilon, \ \forall m \in \mathbb{N} \quad (1)$$

On the other hand, for such ϵ, by the lossless source coding theorem, we know that $\exists n_{\delta,\epsilon} \in \mathbb{N}$ such that $\forall n \geq n_{\delta,\epsilon}$,

$$k^*(n, 1-\epsilon) > n\left(\mathrm{H}(S) - \delta/2\right) > nR \quad (2)$$

From (1), for $m \geq n_{\delta,\epsilon}$, $\exists n'(m) \geq m \geq n_{\delta,\epsilon}$ and a corresponding $(n'(m), \lfloor n'(m)R \rfloor)$ code with error smaller than $1-\epsilon$. But from (2), for such $n'(m) \geq n_{\delta,\epsilon}$, the smallest length required is $k^*(n'(m), 1-\epsilon) > n'(m)R$. This leads to contradiction.

Grading Policy:

a) A correct scheme [3] and the justification [3]; alternatively, a correct proof of existence [6].

b) A proof that applies for all codes [4] and other details [2].

Homework 2 Simple Solution

TA: Heng-Chien Liou*

1. (Maximum entropy) [22]

In the lecture we show that for a random variable taking values in a *finite* alphabet $\mathcal{X}$, the maximum entropy is $\log|\mathcal{X}|$, and the maximizing distribution is $\mathrm{Unif}(\mathcal{X})$. In this problem, we consider a random variable X taking values in the countable alphabet $\mathbb{N}$ (the set of positive integers) with mean $\mu < \infty$. Please use the steps below to prove that its entropy is upper bounded as follows:

$$\mathrm{H}(X) \leq \mu_X \mathsf{h}_{\mathsf{b}}(\mu_X^{-1}).$$

Notation: The PMF of X is denoted by $\mathsf{P}_X(i) \equiv p_i$, $i \in \mathbb{N}$, and $\mathsf{E}[X] \equiv \mu_X = \sum_{i=1}^{\infty} i p_i$.

a) Use the concavity of $\log(\cdot)$ and Jensen's inequality to show that for any non-negative sequence $\{q_i\}_{i=1}^{\infty}$ with $\sum_{i=1}^{\infty} q_i = 1$,

$$\mathrm{H}(X) \leq -\sum_{i=1}^{\infty} p_i \log q_i. \quad [8]$$

b) Find a non-negative sequence $\{q_i\}_{i=1}^{\infty}$ and α, β such that $\sum_{i=1}^{\infty} q_i = 1$, $\sum_{i=1}^{\infty} i q_i = \mu_X$, and

$$-\log q_i = \alpha i + \beta \quad \forall\, i \in \mathbb{N}. \quad [8]$$

c) Complete the proof by plugging in the $\{q_i\}_{i=1}^{\infty}$ found in Part b) into the upper bound of $\mathrm{H}(X)$ found in Part a). Show that this upper bound on $\mathrm{H}(X)$ is attainable with an appropriate choice of $\{p_i\}_{i=1}^{\infty}$ and hence it is the maximum entropy of random variables taking values in positive integers with expected value being μ_X. [6]

Solution:

a) $\mathrm{H}(X) + \sum_{i=1}^{\infty} p_i \log q_i = \sum_{i=1}^{\infty} p_i \log \frac{q_i}{p_i}$. By Jensen, $\sum_{i=1}^{\infty} p_i \log \frac{q_i}{p_i} \leq \log \sum_{i=1}^{\infty} p_i \frac{q_i}{p_i} = 0$.

b) $\begin{cases} 1 = \sum_{i=1}^{\infty} q_i = \frac{2^{-(\alpha+\beta)}}{1-2^{-\alpha}} \\ \mu_X = \sum_{i=1}^{\infty} i q_i = \frac{2^{-(\alpha+\beta)}}{(1-2^{-\alpha})^2} \end{cases} \Rightarrow \begin{cases} \alpha = \log\frac{\mu_X}{\mu_X-1} \\ \beta = \log(\mu_X - 1) \end{cases}$

c) $\mathsf{H}_{\mathsf{b}}(\frac{1}{\mu_X}) = \frac{1}{\mu_X}\ln\mu_X + \frac{\mu_X-1}{\mu_X}\ln\frac{\mu_X}{\mu_X-1}$

*with contributions by Wen-Shao Ho

$\mathrm{H}(X) \leq -\sum_{i=1}^{\infty} p_i \log q_i = \sum_{i=1}^{\infty} p_i[i\ln\frac{\mu_X}{\mu_X-1} + \log(\mu_X-1)] = \mu_X \mathsf{H}_{\mathsf{b}}(\frac{1}{\mu_X})$

Equality holds iff $p_i = q_i\ \forall i$.

Grading Policy:

a) Correct arithmetic [4] and application of Jensen's inequality on the right variable [4].

b) Specify α and β [8].

c) Arithmetic [3] and selection of $\{p_i\}_{i=1}^{\infty}$ [3].

2. (Entropy of a random variable with a infinitely countable support) [14]

In the lecture, we define the entropy for a random variable with a finite alphabet $\mathcal{X}$ (in fact a finite support supp_{P_X} suffices). For a random variable X that has an infinitely *countable* support, sometimes $\mathrm{H}(X)$ is finite and sometimes $\mathrm{H}(X)$ becomes infinite. In this problem we look at an example.

a) Consider an infinite series

$$\sum_{n=2}^{\infty} \frac{1}{n(\log n)^{\alpha}}$$

where $\alpha > 0$. Use the *integral test for convergence* to show that the series converges if and only if $\alpha > 1$. [7]

b) Let s_{α} denote the above series if the series converges. Let us define a random variable $X_{\alpha} \in \{2, 3, \ldots\}$ with PMF

$$\mathsf{P}_{X_{\alpha}}(n) = \frac{1}{s_{\alpha} n(\log n)^{\alpha}}.$$

Show that $\mathrm{H}(X_{\alpha})$ exists if $\alpha > 2$ and it diverges to ∞ if $1 < \alpha \leq 2$. [7]

Solution:

a) $\int_2^{\infty} \frac{dx}{x(\log x)^{\alpha}} = \begin{cases} \frac{(\log 2)^{1-\alpha}}{(\alpha-1)} & (\alpha > 1) \\ \infty & (0 \leq \alpha \leq 1) \end{cases}$.

By the integral test for convergence, the series converges if and only if $\alpha > 1$.

Many students derive wrong integral when $\alpha = 1$.

b) Let $s_{\alpha} := \sum_{n=2}^{\infty} \frac{1}{n(\log n)^{\alpha}}$, for $\alpha > 1$.

$$\mathrm{H}(X_n) = \sum_{n=2}^{\infty} \mathsf{P}_{X_n}(n) \log \frac{1}{\mathsf{P}_{X_n}(n)}$$
$$= \log s_{\alpha} + \frac{1}{s_{\alpha}} \sum_{n=2}^{\infty} \frac{1}{n(\log n)^{\alpha-1}} + \frac{1}{s_{\alpha}} \sum_{n=2}^{\infty} \frac{\alpha \log(\log n)}{n(\log n)^{\alpha}}$$

$\begin{cases} \leq \infty & (\alpha > 2) \\ = \infty & (1 < \alpha \leq 2) \end{cases}$

Grading Policy:

a) Correct statement of the integral test [2] and correct evaluation of the integral [5].

b) Correct evaluation of the relevant integrals [3] and correct argument of the series's convergence and divergence [4].

3. (Mixture of random processes) [14]

In this problem we look at different ways to generate mixtures of random processes, and the entropy rate of the mixture of random processes. Consider two stationary random processes $\{X_0[i] \,|\, i \in \mathbb{N}\}$ and $\{X_1[i] \,|\, i \in \mathbb{N}\}$ taking values in disjoint alphabets $\mathcal{X}_0$ and $\mathcal{X}_1$ respectively. The two processes are independent from each other, that is, $\{X_0[i]\} \perp \{X_1[i]\}$, and they have entropy rates $\mathcal{H}_0$ and $\mathcal{H}_1$ respectively. Let $\{\Theta_i \,|\, i \in \mathbb{N}\}$ be a **stationary** Bernoulli random process, independent of everything else.

a) Let $\Theta_i = \Theta$ for all $i \in \mathbb{N}$, where $\Theta \sim \mathrm{Ber}(q)$. Is the random process $\{X_{\Theta_i}[i]\}$ stationary? What is its entropy rate? [6]

b) Let $\{\Theta_i\}$ be Markov with a probability transition matrix

$$\mathsf{P}_{\Theta_2|\Theta_1} = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}, \text{ for } \alpha, \beta \in (0,1).$$

Suppose that both $\{X_0[i]\}$ and $\{X_1[i]\}$ are i.i.d. processes in this problem. Is the random process $\{X_{\Theta_i}[i]\}$ stationary? What is its entropy rate? [8]

Solution:

a) Since $\{X_0[i] | i \in \mathbf{N}\}$ and $\{X_1[i] | i \in \mathbf{N}\}$ are stationary processes, $\forall n, l \in \mathbf{N}$,

$$\mathsf{P}_{X_{\Theta_1}[1],\ldots,X_{\Theta_n}[n]}$$
$$= (1-q)\mathsf{P}_{X_0[1],\ldots,X_0[n]} + q\mathsf{P}_{X_1[1],\ldots,X_1[n]}$$
$$= (1-q)\mathsf{P}_{X_0[l+1],\ldots,X_0[n]} + q\mathsf{P}_{X_1[1],\ldots,X_1[l+n]}$$
$$= \mathsf{P}_{X_{\Theta_{l+1}}[l+1],\ldots,X_{\Theta_{l+n}}[l+n]}$$

So $\{X_{\Theta_i}[i]\}$ is stationary.
Let $Y_i = X_{\Theta_i}[i]$.

$$\mathcal{H}(X_{\Theta_i}[i]) = \lim_{n\to\infty} \mathrm{H}(Y_n|Y^{n-1})$$
$$= \lim_{n\to\infty} \mathrm{H}(Y_n, \Theta_n | Y^{n-1}, \Theta^{n-1}) \quad (\mathcal{X}_0 \cap \mathcal{X}_1 = \varnothing)$$

$$= \lim_{n\to\infty} \mathrm{H}(Y_n|Y^{n-1}, \Theta^n) + \mathrm{H}(\Theta_n|Y^{n-1}, \Theta^{n-1}) \quad (\textit{chain rule})$$
$$= \lim_{n\to\infty} \mathrm{H}(Y_n|\Theta_n) + \mathrm{H}(\Theta_n|\Theta_{n-1})$$
$$= \lim_{n\to\infty} \mathrm{H}(Y_n|\Theta) + \mathrm{H}(\Theta|\Theta)$$
$$= q\mathcal{H}_1 + (1-q)\mathcal{H}_0 + 0$$

b) $\{X_{\Theta_i}[i]\}$ is stationary. Let $Y_i = X_{\Theta_i}[i]$. $\mathcal{H}(\{X_{\Theta_i}[i]\}) = \lim_{n\to\infty} \mathrm{H}(Y_n|Y^{n-1})$.

$$\mathrm{H}(Y_n|Y^{n-1})$$
$$= \mathrm{H}(Y_n, \Theta_n | Y^{n-1}) \quad (\mathcal{X}_0 \cap \mathcal{X}_1 = \varnothing)$$
$$= \mathrm{H}(Y_n|\Theta_n, Y^{n-1}) + \mathrm{H}(\Theta_n|Y^{n-1}) \quad \textit{chain rule}$$
$$= \mathrm{Pr}\{\Theta_n = 1\}\mathrm{H}(X_1[n]|X_1^{n-1}) + \mathrm{Pr}\{\Theta_n = 0\}\mathrm{H}(X_0[n]|X_0^{n-1}) + \mathrm{H}\left(\Theta_n|Y^{n-1}\right)$$
$$= \mathrm{Pr}\{\Theta_n = 1\}\mathrm{H}(X_1[n]|X_1^{n-1}) + \mathrm{Pr}\{\Theta_n = 0\}\mathrm{H}(X_0[n]|X_0^{n-1}) + \mathrm{H}\left(\Theta_n|\Theta^{n-1}\right) \quad (\mathcal{X}_0 \cap \mathcal{X}_1 = \varnothing)$$
$$= \frac{\alpha}{\alpha+\beta}\mathcal{H}_1 + \frac{\beta}{\alpha+\beta}\mathcal{H}_0 + \mathrm{H}(\Theta_2|\Theta_1)$$
$$= \frac{\alpha}{\alpha+\beta}(\mathcal{H}_1 + \mathsf{H}_{\mathsf{b}}(\beta)) + \frac{\beta}{\alpha+\beta}(\mathcal{H}_0 + \mathsf{H}_{\mathsf{b}}(\alpha))$$

Grading Policy:

a) Stationary argument [2], chain rule and entropy calculation [3], asymptotic [1].

b) Stationary argument [2], chain rule and entropy calculation [6].

Homework 3 Simple Solution

TA: Heng-Chien Liou*

1. (Binary hypothesis testing) [16]

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. Bernoulli p random variables, that is,

$$\mathsf{Pr}\{X_i = 1\} = 1 - \mathsf{Pr}\{X_i = 0\} = p.$$

Based on the observations so far, the goal is of a decision maker to determine which of the following two hypotheses is true:

$$\mathcal{H}_0: \ p = p_0$$
$$\mathcal{H}_1: \ p = p_1$$

where $0 < p_0 < p_1 \leq 1/2$.

a) (Warm-up) Consider the problem of making the decision based on X_1.

Draw the optimal $(\pi_{1|0}, \pi_{0|1})$ trade-off curve. [4]

b) Suppose the decision maker waits until an 1 appears and makes the decision based on the whole observed sequence. Sketch the optimal $(\pi_{1|0}, \pi_{0|1})$ trade-off curve. [4]

c) Now suppose the decision maker waits until in total n 1's appear and makes the decision based on the whole observed sequence. Let $\varpi^*_{\eta}(n,\epsilon)$ denote the minimum type-II error probability subject to the constraint that the type-I error probability is not greater than ϵ, $0 < \epsilon < 1$. Does $\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\varpi^*_{0|1}(n,\epsilon)}$ exist? If so, find it. Otherwise, show that the limit does not exist. [8]

Solution:

a) By Neyman-Pearson theorem, the optimal test is randomized LRT. Note that the likelihood ratio can only take two values: $\frac{p_0}{p_1}, \frac{1-p_0}{1-p_1}$. Therefore, discuss the range of τ we get

$$\begin{cases} \pi_{1|0} = 1, \pi_{0|1} = 0, & 0 \leq \tau < \frac{1-p_0}{1-p_1} \\ \pi_{1|0} = p_0 + \gamma(1-p_0), \pi_{0|1} = (1-\gamma)(1-p_1) = \frac{1-p_0}{1-p_0}(1-\pi_{1|0}), & \tau = \frac{1-p_0}{1-p_0} \\ \pi_{1|0} = p_0, \pi_{0|1} = 1-p_1, & \frac{1-p_0}{1-p_0} < \tau < \frac{p_0}{p_1} \\ \pi_{1|0} = \gamma p_0, \pi_{0|1} = (1-\gamma)p_1 + (1-p_1) = 1 - \frac{p_0}{p_1}\pi_{1|0}, & \tau = \frac{p_0}{p_1} \\ \pi_{1|0} = 0, \pi_{0|1} = 1, & \tau > \frac{p_0}{p_1}. \end{cases}$$

*with contribution by Chen-Hao Hsiao

We can then draw the trade-off curve using the equations derived above.

b) Note that our observation can only be $1, 01, 001, 0001, \cdots$, let L be the length of the observation, we have

$$\mathcal{H}_0: \ L \sim \mathrm{Geo}(p_0)$$
$$\mathcal{H}_1: \ L \sim \mathrm{Geo}(p_1)$$

Similar to a), we can discuss the range of τ and get:

$$\begin{cases} \pi_{1|0} = 0, \pi_{0|1} = 1, \ \tau > \frac{p_1}{p_0} \\ \pi_{1|0} = \sum_{i=1}^{n-1}(1-p_0)^{i-1}p_0 + \gamma(1-p_0)^{n-1}p_0, \\ \pi_{0|1} = \sum_{i=n+1}^{n}(1-p_1)^{i-1}p_1 + (1-\gamma)(1-p_1)^{n-1}p_1, \ \tau = \frac{(1-p_1)^{n-1}p_1}{(1-p_0)^{n-1}p_0} \\ \pi_{1|0} = \sum_{i=1}^{n}(1-p_0)^{i-1}p_0, \pi_{0|1} = \sum_{i=n+1}^{\infty}(1-p_1)^{i-1}p_1, \ \frac{(1-p_1)^{n-1}p_1}{(1-p_0)^{n-1}p_0} > \tau > \frac{(1-p_1)^n p_1}{(1-p_0)^n p_0}. \end{cases}$$

And we can draw the trade-off curve using the equations derived above.

c) The observation can be viewed as n i.i.d. geometric random variables. To see this, for any realization of observation, insert a "$-$" symbol in front of the sequence, also insert a "$-$" right after a "1". For example, if $n = 4$ and the realization is 010001101, we write it as |01|0001|1|01|. Apparently, the length of the subsequence between two — is a geometric random variable. Hence, in this subproblem, we are testing $\mathrm{Geo}(p_0)^{\otimes n}$ and $\mathrm{Geo}(p_1)^{\otimes n}$. By Chernoff-Stein lemma,

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\varpi^*_{0|1}(n,\epsilon)} = \mathrm{D}(\mathrm{Geo}(p_0)\|\mathrm{Geo}(p_1)) = \log\frac{p_0}{p_1} + \left(\frac{1-p_0}{p_0}\right)\log\frac{1-p_0}{1-p_1}.$$

Grading Policy:

a) Specify the optimal trade-off curve [2]; invoke Neyman-Pearson Theorem or directly argue optimality [2].

b) Formulate the problem as detecting geometric random variables (or directly used the mass function) [2] and invoke Neyman-Pearson's to specify the curve [2].

c) Formulate the problem as a hypothesis testing with n instances [3], Chernoff-Stein lemma [2], and calculation [3].

2. (Asymptotic behavior of posterior probability [12])

Consider a binary hypothesis testing problem

$$\begin{cases} \mathcal{H}_0: \ X_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}_0, \ i = 1, 2, \ldots, n \\ \mathcal{H}_1: \ X_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}_1, \ i = 1, 2, \ldots, n \end{cases}.$$

Under a Bayes setup, the unknown binary parameter Θ is assumed to be *random* and follow a prior distribution defined by the *prior probabilities*

$$\pi_0^{(0)} := \mathsf{Pr}\{\Theta = 0\}, \ \pi_1^{(0)} := \mathsf{Pr}\{\Theta = 1\}.$$

Let the *posterior probabilities* be the conditional distribution of Θ given $X^n = x^n$:

$$\pi_0^{(n)}(x^n) = \mathsf{Pr}\{\Theta = 0 | X^n = x^n\}, \ \pi_1^{(n)}(x^n) = \mathsf{Pr}\{\Theta = 1 | X^n = x^n\}.$$

a) Derive the expressions of $\pi_0^{(n)}(x^n)$ and $\pi_1^{(n)}(x^n)$ in terms of $\pi_0^{(0)}, \pi_1^{(0)}, \mathsf{P}_0, \mathsf{P}_1$. [4]

b) Consider $\pi_0^{(n)}(X^n)$ and $\pi_1^{(n)}(X^n)$ as random variables, because they are functions of the random sequence X^n. Use the Strong Law of Large Numbers to show that if $\mathcal{H}_0$ is true, then with probability 1,

$$\pi_0^{(n)}(X^n) \to 1, \ -\frac{1}{n}\log\pi_1^{(n)}(X^n) \to \mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1) \quad \text{as } n \to \infty. \quad [8]$$

Solution:

a) Denote $x^n = (x_1, x_2, ..., x_i, ..., x_n)$

$$\pi_0^{(n)}(X^n) = \mathsf{Pr}\{\Theta = 0 | X^n = x^n\} = \frac{\mathsf{Pr}\{\Theta = 0, X^n = x^n\}}{\mathsf{Pr}\{X^n = x^n\}}$$
$$= \frac{\pi_0^{(0)}\mathsf{Pr}\{X^n = x^n | \Theta = 0\}}{\pi_0^{(0)}\mathsf{Pr}\{X^n = x^n | \Theta = 0\} + \pi_1^{(0)}\mathsf{Pr}\{X^n = x^n | \Theta = 1\}}$$
$$= \frac{\pi_0^{(0)}\prod_{i=1}^n \mathsf{P}_0(x_i)}{\pi_0^{(0)}\prod_{i=1}^n \mathsf{P}_0(x_i) + \pi_1^{(0)}\prod_{i=1}^n \mathsf{P}_1(x_i)}$$
$$\pi_1^{(n)}(X^n) = \frac{\pi_1^{(0)}\prod_{i=1}^n \mathsf{P}_1(x_i)}{\pi_0^{(0)}\prod_{i=1}^n \mathsf{P}_0(x_i) + \pi_1^{(0)}\prod_{i=1}^n \mathsf{P}_1(x_i)}$$

b) Recall the basic Bayes' theorem.

$$\pi_0^{(n)}(X^n) = \frac{\pi_0^{(0)}\prod_{i=1}^n \mathsf{P}_0(x_i)}{\pi_0^{(0)}\prod_{i=1}^n \mathsf{P}_0(x_i) + \pi_1^{(0)}\prod_{i=1}^n \mathsf{P}_1(x_i)}$$
$$= \frac{\pi_0^{(0)} 2^{-n\left(-\frac{1}{n}\sum_{i=1}^n \log \mathsf{P}_0(x_i)\right)}}{\pi_0^{(0)} 2^{-n\left(-\frac{1}{n}\sum_{i=1}^n \log \mathsf{P}_0(x_i)\right)} + \pi_1^{(0)} 2^{-n\left(-\frac{1}{n}\sum_{i=1}^n \log \mathsf{P}_1(x_i)\right)}}$$
$$= \frac{1}{1 + \left(\frac{\pi_1^{(0)}}{\pi_0^{(0)}}\right) 2^{-n\left(\frac{1}{n}\sum_{i=1}^n \log \frac{\mathsf{P}_0(x_i)}{\mathsf{P}_1(x_i)}\right)}} = \frac{1}{1 + c2^{-n\left(\frac{1}{n}\sum_{i=1}^n \log \frac{\mathsf{P}_0(x_i)}{\mathsf{P}_1(x_i)}\right)}},$$

where we let $c = \pi_1^{(0)}/\pi_0^{(0)}$. For any $\epsilon \in (0, \mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1))$, there is an $N_\epsilon \in \mathbb{N}$ such that, for every $n \geq N_\epsilon$

$$\mathsf{Pr}\left\{\left|\frac{1}{n}\sum_{i=1}^n \log\frac{\mathsf{P}_0(X_i)}{\mathsf{P}_1(X_i)} - \mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)\right| < \epsilon\right\} = 1$$

$$\mathsf{Pr}\left\{\frac{1}{1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\epsilon)}} < \pi_0^{(n)}(X^n) < \frac{1}{1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\epsilon)}}\right\} = 1$$

Hence, with probability 1, $\pi_0^{(n)}(X^n) \to 1$ as $n \to \infty$.

Following this derivation, for every $n \geq N_\epsilon$, with probability 1,

$$-\frac{1}{n}\log\left(\frac{c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\epsilon)}}{1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\epsilon)}}\right) < -\frac{1}{n}\log\pi_1^{(n)}(X^n) < -\frac{1}{n}\log\left(\frac{c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\epsilon)}}{1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\epsilon)}}\right)$$

Furthermore,

$$\begin{aligned}
-\frac{1}{n}\log\left(\frac{c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\epsilon)}}{1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\epsilon)}}\right) &= \mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1) + \epsilon - \frac{\log c}{n} + \frac{1}{n}\log\left(1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\epsilon)}\right) \\
&\leq \mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1) + \epsilon - \frac{\log c}{n} + \frac{1}{n}\log\left(1+c2^{-N_\epsilon(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\epsilon)}\right) \\
&= \mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1) + \epsilon + O\left(\frac{1}{n}\right) \\
-\frac{1}{n}\log\left(\frac{c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\epsilon)}}{1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\epsilon)}}\right) &= \mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1) - \epsilon - \frac{\log c}{n} + \frac{1}{n}\log\left(1+c2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\epsilon)}\right) \\
&\geq \mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1) - \epsilon - \frac{\log c}{n} + \frac{1}{n}\log\left(1+c2^{-N_\epsilon(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\epsilon)}\right) \\
&= \mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1) - \epsilon + O\left(\frac{1}{n}\right)
\end{aligned}$$

Hence, with probability 1,

$$\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1) - \epsilon \leq \lim_{n\to\infty} -\frac{1}{n}\log\pi_1^{(n)}(X^n) \leq \mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1) + \epsilon,$$

where $\epsilon \in (0, \mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1))$ arbitrarily. Hence,

$$\mathsf{Pr}\left\{\lim_{n\to\infty} -\frac{1}{n}\log\pi_1^{(n)}(X^n) = \mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)\right\} = 1$$

Grading Policy:

a) $\pi_0^{(n)}(x^n)$ [2] and $\pi_1^{(n)}(x^n)$ [2].

b) Application of SLLN and argumnet for almost sure convergence [4], convergence for $\pi_0^{(n)}(X^n)$ [2], convergence for $-(1/n)\log\pi_1^{(n)}(X^n)$ [2].

3. Minimizing information divergence) [22]

a) Let $\mathcal{P}(\mathbb{N})$ denote the collection of all probability distributions over $\mathbb{N}$ and $\mathsf{G}(p) \in \mathcal{P}(\mathbb{N})$ be a geometric distribution with parameter $p \in (0,1)$:

$$X \sim \mathsf{G}(p) \iff \mathsf{Pr}\{X = n\} = (1-p)p^{n-1},\ n \in \mathbb{N} = \{1, 2, \ldots\}.$$

Under the constraint that $\mathsf{P} \in \mathcal{P}(\mathbb{N})$ and $\mathsf{E}_{X\sim\mathsf{P}}[X] = \sum_{x=1}^{\infty} x\mathsf{P}(x) = \mu > 1$, find the minimum value of $\mathrm{D}(\mathsf{P}\|\mathsf{G}(p))$ and a minimizing distribution. [12]

b) For m discrete probability distributions $\mathsf{P}_1, \mathsf{P}_2, \ldots, \mathsf{P}_m$ with the same support $\mathcal{X}$, consider the following minimization problem:

$$\min_{\mathsf{Q}\in\mathcal{P}(\mathcal{X})} \sum_{i=1}^{m} \mathrm{D}(\mathsf{P}_i\|\mathsf{Q}),$$

where $\mathcal{P}(\mathcal{X})$ denotes the collection of probability distributions over $\mathcal{X}$. Find a minimizer to the above problem. [10]

Solution:

a) Expanding the Kullback-Leibler divergence over countable alphabet gives:

$$\begin{aligned}
\mathrm{D}(\mathsf{P}\|\mathsf{G}(p)) &= \sum_{x=1}^{\infty} \mathsf{P}(x)\log\frac{\mathsf{P}(x)}{(1-p)p^{x-1}} \\
&= \sum_{x=1}^{\infty} \mathsf{P}(x)\log\mathsf{P}(x) - \log(1-p) - \log p\sum_{x=1}^{\infty}\mathsf{P}(x)(x-1) \\
&= -\mathrm{H}(\mathsf{P}) - \log(1-p) - (\mu-1)\log p \\
&\geq -\mu\mathrm{h_b}\left(\frac{1}{\mu}\right) - \log(1-p) - (\mu-1)\log p,
\end{aligned}$$

where the inequality comes from Problem 1 of Homework 2, and $\mathsf{P}^\star = \mathsf{G}\left(1-\mu^{-1}\right)$ is the minimizing distribution.

b) Let $\overline{\mathsf{P}} = \frac{1}{m}\sum_{i=1}^{m}\mathsf{P}_i$, $\overline{\mathsf{P}} \in \mathcal{P}(\mathcal{X})$.

$\forall \mathsf{Q} \in \mathcal{P}(\mathcal{X})$,

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\mathrm{D}(\mathsf{P}_i\|\mathsf{Q}) - \frac{1}{m}\sum_{i=1}^{m}\mathrm{D}\left(\mathsf{P}_i\middle\|\overline{\mathsf{P}}\right) \\
=&\frac{1}{m}\sum_{i=1}^{m}\sum_{x\in\mathcal{X}}\mathsf{P}_i(x)\log\frac{\overline{\mathsf{P}}(x)}{\mathsf{Q}(x)} \\
=&\sum_{x\in\mathcal{X}}\left(\frac{1}{m}\sum_{i=1}^{m}\mathsf{P}_i(x)\right)\log\frac{\overline{\mathsf{P}}(x)}{\mathsf{Q}(x)} \\
=&\sum_{x\in\mathcal{X}}\overline{\mathsf{P}}(x)\log\frac{\overline{\mathsf{P}}(x)}{\mathsf{Q}(x)} \\
=&\mathrm{D}\left(\overline{\mathsf{P}}\middle\|\mathsf{Q}\right) \geq 0.
\end{aligned}$$

Hence, $\overline{\mathsf{P}}$ is a minimizer.

Grading Policy

a) Fine the tight lower bound [8] and find minimizer to achieve minimum [4].

b) Specify the minimizer [3] and justify it [7].

Homework 4 Simple Solution

TA: Heng-Chien Liou*

1. (Mutual information) [8]

a) How much information does the length of a sequence give about the content of a sequence? In this problem, let us consider a sequence of i.i.d. $\mathrm{Ber}(1/2)$ random variables $X_1, X_2, \ldots$

Let N_0 be a random variable taking value at 6 with probability 1/3 and at 12 with probability 2/3. Furthermore, N_0 is independent of the sequence $\{X_i \,|\, i = 1, 2, \ldots\}$. Also let N_1 denote the length of the sequence when the first "1" appears. Obviously N_1 is also a random variable.

Compute $\mathrm{I}(N_0; X_1, X_2, \ldots, X_{N_0})$ and $\mathrm{I}(N_1; X_1, X_2, \ldots, X_{N_1})$. [4]

b) Consider a sequence of n binary random variables $(X_1, X_2, \ldots, X_n)$. Each sequence with an even number of 1's has probability $2^{-(n-1)}$, and each sequence with an odd number of 1's has probability 0. Compute the following:

$$\mathrm{I}(X_1; X_2),\ \mathrm{I}(X_2; X_3|X_1),\ \mathrm{I}(X_3; X_4|X_1, X_2),\ \ldots,\ \mathrm{I}(X_{n-1}; X_n|X_1, X_2, \ldots, X_{n-2}).$$ [4]

Solution:

a) By definition:

$$\begin{aligned}
\mathrm{I}(N_0; X_1, X_2, \cdots, X_{N_0}) &= \mathrm{H}(N_0) - \mathrm{H}(N_0|X_1, X_2, \cdots, X_{N_0}) = \mathrm{H}(N_0) \\
&= \frac{1}{3}\log 3 + \frac{2}{3}\log\left(\frac{3}{2}\right) = \log 3 - \frac{2}{3}.
\end{aligned}$$

Similarly, $\mathrm{I}(N_1; X_1, X_2, \cdots, X_{N_1}) = \mathrm{H}(N_1)$. We only need to calculate $\mathrm{H}(N_1)$. By problem 4 b) of HW1, $\mathrm{H}(N_1) = 2$.

b) We will first show that $X_1, X_2, \cdots, X_{n-1}$ are mutually independent. Note that for all $x_1, \cdots, x_{n-1} \in \{0,1\}^{n-1}$,

$$\begin{aligned}
\mathsf{Pr}\{X_1 = x_1, \cdots, X_{n-1} = x_{k-1}\} &= \mathsf{Pr}\{X_1 = x_1, \cdots, X_{n-1} = x_{k-1}, X_n = 0\} \\
&\quad + \mathsf{Pr}\{X_1 = x_1, \cdots, X_{n-1} = x_{k-1}, X_n = 1\} \\
&= 2^{-(n-1)}
\end{aligned}$$

since either $x_1 \oplus \cdots \oplus x_{n-1} \oplus 0 = 0$ or $x_1 \oplus \cdots \oplus x_{n-1} \oplus 1 = 0$.

*with contribution by Chen-Hao Hsiao and Wen-Shao Ho

Therefore, for $\mathcal{I} \subseteq \{1, 2, \cdots, n-1\}$, by calculating the marginal distribution, we have

$$\mathsf{Pr}\left\{\bigcap_{i\in\mathcal{I}} X_i = x_i\right\} = \prod_{i\in\mathcal{I}}\mathsf{Pr}\{X_i = x_i\} = 2^{-|\mathcal{I}|}.$$

By the mutual independence derived above, for $k = 1, 2\cdots, n-2$,

$$\begin{aligned}
\mathrm{I}(X_k; X_{k+1}|X_1, \cdots, X_{k-1}) &= \mathrm{H}(X_k|X_1, \cdots, X_{k-1}) - \mathrm{H}(X_k|X_1, \cdots, X_{k-1}, X_{k+1}) \\
&= \mathrm{H}(X_k) - \mathrm{H}(X_k) = 0.
\end{aligned}$$

Also,

$$\begin{aligned}
\mathrm{I}(X_{n-1}; X_n|X_1, \cdots, X_{n-2}) &= \mathrm{H}(X_{n-1}|X_1, \cdots, X_{n-2}) - \mathrm{H}(X_{n-1}|X_1, \cdots, X_{n-2}, X_n) \\
&= \mathrm{H}(X_{n-1}) - \mathrm{H}(X_1 \oplus \cdots \oplus X_{n-2} \oplus X_n|X_1, \cdots, X_{n-2}, X_n) \\
&= 1 - 0 = 1.
\end{aligned}$$

We use $\mathrm{H}(f(X)|X) = 0$ in the second equation.

2. (Data processing) [12]

a) Let $X_1 - X_2 - X_3 - X_4$ form a Markov chain. Prove that

$$\mathrm{I}(X_1; X_3) + \mathrm{I}(X_2; X_4) \leq \mathrm{I}(X_1; X_4) + \mathrm{I}(X_2; X_3).$$ [6]

b) Let $X_1 - X_2 - (X_3, X_4)$ form a Markov chain. Prove that

$$\mathrm{I}(X_1; X_3) + \mathrm{I}(X_1; X_4) \leq \mathrm{I}(X_1; X_2) + \mathrm{I}(X_3; X_4).$$ [6]

Solution:

a)

$$\begin{aligned}
\mathrm{I}(X_1; X_3) + \mathrm{I}(X_2; X_4) &= \mathrm{I}(X_1; X_3) + \mathrm{I}(X_1, X_2; X_4) - \mathrm{I}(X_1; X_4|X_2) && \text{chain rule} \\
&= \mathrm{I}(X_1; X_3) + \mathrm{I}(X_1; X_2; X_4) && \text{markov} \\
&= \mathrm{I}(X_1; X_3) + \mathrm{I}(X_1; X_4) + \mathrm{I}(X_2; X_4|X_1) && \text{chain rule} \\
&\leq \mathrm{I}(X_1; X_3) + \mathrm{I}(X_1; X_4) + \mathrm{I}(X_2; X_3|X_1) && \text{data processing} \\
&= \mathrm{I}(X_1; X_4) + \mathrm{I}(X_1; X_3, X_2) && \text{chain rule} \\
&= \mathrm{I}(X_1; X_4) + \mathrm{I}(X_2; X_3) && \text{markov}
\end{aligned}$$

$$\begin{aligned}
\mathrm{I}(X_1; X_3) + \mathrm{I}(X_1; X_4) &= \mathrm{I}(X_1; X_3, X_4) - \mathrm{I}(X_1; X_4|X_3) + \mathrm{I}(X_1; X_4) && \text{chain rule} \\
&\leq \mathrm{I}(X_1; X_2) - \mathrm{I}(X_1; X_4|X_3) + \mathrm{I}(X_1; X_4) && \text{data processing} \\
&= \mathrm{I}(X_1; X_2) + \mathrm{I}(X_1; X_4)
\end{aligned}$$

$$\begin{aligned}
&- \left(\mathrm{I}(X_1; X_4) + \mathrm{I}(X_3; X_4|X_1) - \mathrm{I}(X_3; X_4)\right) && \text{chain rule} \\
&\leq \mathrm{I}(X_1; X_2) + \mathrm{I}(X_3; X_4) && \text{nonnegative}
\end{aligned}$$

3. (Sum Channel) [16]

Consider l DMC's

$$\left\{\left(\mathcal{X}^{(i)}, \mathsf{P}_{Y|X}^{(i)}, \mathcal{Y}^{(i)}\right) \middle| i = 1, 2, \ldots, l\right\},$$

where DMC $(\mathcal{X}^{(i)}, \mathsf{P}_{Y|X}^{(i)}, \mathcal{Y}^{(i)})$ has channel capacity $\mathrm{C}^{(i)}$, for $1 \leq i \leq l$. The channel input alphabets are disjoint, and so are the channel output alphabets, that is,

$$\mathcal{X}^{(i)} \cap \mathcal{X}^{(j)} = \mathcal{Y}^{(i)} \cap \mathcal{Y}^{(j)} = \emptyset,\ \forall\, i \neq j.$$

The **sum channel** $(\mathcal{X}^{\oplus}, \mathsf{P}_{Y|X}^{\oplus}, \mathcal{Y}^{\oplus})$ associated to these channels is defined as follows:

- Input alphabet is the union $\mathcal{X}^{\oplus} := \cup_{i=1}^{l}\mathcal{X}^{(i)}$ of the individual input alphabets.
- Output alphabet is the union $\mathcal{Y}^{\oplus} := \cup_{i=1}^{l}\mathcal{Y}^{(i)}$ of the respective output alphabets.
- At each time slot the transmitter chooses to use *one and only one* of the l channels to transmit a symbol, that is,

$$\mathsf{P}^{\oplus}(y|x) := \begin{cases} \mathsf{P}_{Y|X}^{(i)}(y|x), & \text{if } x \in \mathcal{X}^{(i)} \text{ and } y \in \mathcal{Y}^{(i)} \\ 0, & \text{otherwise} \end{cases}$$

a) Introduce a random variable I indicating which DMC is used in the sum channel, that is,

$$I = i \quad \text{if } X \in \mathcal{X}^{(i)},\ i = 1, 2, ..., l.$$

Show that for the sum channel $\mathsf{P}_{Y|X}^{\oplus}$, $\mathrm{I}(X;Y) = \mathrm{I}(X;Y|I) + \mathrm{H}(I)$. [4]

b) Find the capacity of the sum channel in terms of $\{\mathrm{C}^{(i)} \,|\, i = 1, 2, \ldots, l\}$. [6]

c) Find the optimal input probability distribution for the sum channel in terms of the optimal input probability distributions for the individual channels. [6]

Solution:

- Let $X^{\oplus}$ be the r.v. of imput sum channel
- Let $Y^{\oplus}$ be the r.v. of output sum channel

$$\begin{aligned}
\mathrm{C}^{\oplus} &= \max_{\mathsf{P}_X^{\oplus}} \mathrm{I}\left(X^{\oplus}; Y^{\oplus}\right) \\
&= \max_{\mathsf{P}_X^{\oplus}} \mathrm{I}\left(X^{\oplus}, Y^{\oplus}, I\right) && \mathcal{X}^{(i)} \text{ disjoint} \\
&= \max_{\mathsf{P}_X^{\oplus}} \left\{\mathrm{I}\left(X^{\oplus}; I\right) + \mathrm{I}\left(X^{\oplus}; Y^{\oplus}|I\right)\right\} && \text{chain rule}
\end{aligned}$$

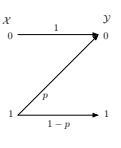
$$\begin{aligned}
&= \max_{\mathsf{P}_X^{\oplus}} \left\{\mathrm{H}(I) + \mathrm{I}\left(X^{\oplus}; Y^{\oplus}|I\right)\right\} && I \text{ is deterministic of } X^{\oplus} \\
&= \max_{\mathsf{P}_X^{\oplus}} \sum_{i=1}^{l}\mathsf{P}_I(i)\left[\log\frac{1}{\mathsf{P}_I(i)} + \mathrm{I}\left(X^{\oplus}; Y^{\oplus}\middle|I=i\right)\right] \\
&= \max_{\mathsf{P}_I} \sum_{i=1}^{l}\mathsf{P}_I(i)\log\frac{2^{\mathrm{C}^{(i)}}}{\mathsf{P}_I(i)} \\
&= \log\sum_{i=1}^{l} 2^{\mathrm{C}^{(i)}} && \text{Jensen's inequality}
\end{aligned}$$

The equality of the Jensen's inequality holds iff $\frac{\mathsf{P}_I(i)}{2^{\mathrm{C}^{(i)}}} = \text{constant}$, $\forall i = 1, \ldots, l$.

Let $\mathsf{P}_X^{(i)}(x)$ be the optimal input probability distribution for the i-th channel. The optimal input probability distribution for the sum channel is

$$\mathsf{P}_{X^{\oplus}}(x) = \sum_{i=1}^{l}\mathsf{Pr}\left(X^{(i)} = x|I = i\right)P_I(i) = \frac{\sum_{i=1}^{l} 2^{\mathrm{C}^{(i)}}\mathsf{P}_X^{(i)}(x)\mathbb{1}\{I(x) = i\}}{\sum_{j=1}^{l} 2^{\mathrm{C}^{(j)}}}, \quad \forall x \in \mathcal{X}^{\oplus}.$$

4. (Z channel) [14]



The Z channel (depicted above) is one of the simplest asymmetric channel with its channel law described as follows.

$$\mathsf{P}_{Y|X} = \begin{bmatrix} 1 & 0 \\ p & 1-p \end{bmatrix}.$$

In the following, let us assume $p = 1/2$.

a) Find the capacity of the Z channel and a capacity achieving input distribution $\mathsf{P}_X^\star$. Also find $\mathsf{P}_Y^\star$, the output distribution induced by the input distribution $\mathsf{P}_X^\star$. [8]

b) Is the capacity achieving input distribution of the Z channel unique? [2]

c) Recall that $\mathrm{C} = \mathrm{D}\left(\mathsf{P}_{Y|X}\middle\|\mathsf{P}_Y^\star\middle|\mathsf{P}_X^\star\right)$ and can be viewed as a weighted average of

$$\left\{\mathrm{D}\left(\mathsf{P}_{Y|X}(\cdot|a)\middle\|\mathsf{P}_Y^\star(\cdot)\right)\middle|\, a \in \mathcal{X}\right\}.$$

For the Z channel, derive $\mathrm{D}\left(\mathsf{P}_{Y|X}(\cdot|0)\middle\|\mathsf{P}_Y^\star(\cdot)\right)$ and $\mathrm{D}\left(\mathsf{P}_{Y|X}(\cdot|1)\middle\|\mathsf{P}_Y^\star(\cdot)\right)$. [4]

Solution:

a) Note that $\mathcal{X}$ is an alphabet of size 2, hence we can model any input distribution as $\mathsf{P}_X = \mathrm{Ber}(q), 0 \leq q \leq 1$. For given $X \sim \mathrm{Ber}(q)$, we then have $Y \sim \mathrm{Ber}\left((1-p)q\right)$.

$$\begin{aligned}
\mathrm{C}(q) &:= \mathrm{I}(X;Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X) \\
&= \mathrm{h_b}\left((1-p)q\right) - \mathsf{P}_X(0)\mathrm{H}(Y|X=0) - \mathsf{P}_X(1)\mathrm{H}(Y|X=1) \\
&= \mathrm{h_b}\left((1-p)q\right) - q\mathrm{h_b}\left(p\right)
\end{aligned}$$

$\mathrm{C}(q)$ can be understood as a real function of single variable. The channel capacity then can be denoted as $\mathrm{C} = \max_{q\in[0,1]}\mathrm{C}(q)$, and can be approached by simple calculus:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}q}\mathrm{C}(q) &= \frac{\mathrm{d}}{\mathrm{d}q}\left[\mathrm{h_b}\left((1-p)q\right) - q\mathrm{h_b}\left(p\right)\right] \\
&= (1-p)\log\frac{1-(1-p)q}{(1-p)q} - \mathrm{h_b}\left(p\right) \\
\mathrm{C}(0) &= 0 \\
\mathrm{C}(1) &= \mathrm{h_b}\left(1-p\right) - \mathrm{h_b}\left(p\right)
\end{aligned}$$

by the extreme value theorem, if we restrict $p = 1/2$, it becomes clear that $q = 2/5$ is the unique maximizer. Hence

$$\mathrm{C} = \max_{\mathsf{P}_X}\mathrm{I}(X;Y) = \max_{q\in[0,1]}\mathrm{C}(q) = \mathrm{C}(2/5) = \mathrm{h_b}\left(1/5\right) - 2/5 = \log 5 - 2.$$

And $\mathsf{P}_X^\star = \mathrm{Ber}(2/5)$ and $\mathsf{P}_Y^\star = \mathrm{Ber}(1/5)$.

b) Yes, it is unique.

c)

$$\begin{aligned}
\mathrm{D}\left(\mathsf{P}_{Y|X}(\cdot|0)\middle\|\mathsf{P}_Y^\star(\cdot)\right) &= 1\log\frac{1}{1-1/5} + 0\log\frac{0}{1/5} = \log 5 - 2 \\
\mathrm{D}\left(\mathsf{P}_{Y|X}(\cdot|1)\middle\|\mathsf{P}_Y^\star(\cdot)\right) &= \frac{1}{2}\log\frac{1/2}{1-1/5} + \frac{1}{2}\log\frac{1/2}{1/5} = \log 5 - 2 \\
\mathrm{D}\left(\mathsf{P}_{Y|X}\middle\|\mathsf{P}_Y^\star\middle|\mathsf{P}_X^\star\right) &= \log 5 - 2
\end{aligned}$$

Homework 5

TA: Heng-Chien Liou*

1. (Information divergence) [10]

Consider two probability density functions $\mathsf{f}(\cdot)$ and $\mathsf{g}(\cdot)$. Let μ_1 and μ_2 denote the mean of f and g respectively. Let σ_1^2 and σ_2^2 denote the variance of f and g respectively.

a) Compute $\mathrm{D}(\mathsf{f}\|\mathsf{g})$ in the following cases: (1) both f and g are Gaussian; (2) both f and g are Laplace. [6]

b) If $\mu_1 = \mu_2$, which of the above cases gives the largest/smallest KL divergence? Your answer may depend on σ_1, σ_2. [2]

c) If $\sigma_1 = \sigma_2$, which of the above cases gives the largest/smallest KL divergence? Your answer may depend on μ_1, μ_2. [2]

Solution:

a) For f and g both being Guassian, we have

$$\begin{aligned}
\mathrm{D}(\mathsf{f}\|\mathsf{g}) &= \int_{\mathbb{R}}\mathsf{f}(x)\log\frac{\mathsf{f}(x)}{\mathsf{g}(x)}\,\mathrm{d}x \\
&= \int_{\mathbb{R}}\mathsf{f}(x)\left(\frac{(x-\mu_2)^2}{2\sigma_2^2}\log e - \frac{(x-\mu_1)^2}{2\sigma_1^2}\log e + \log\frac{\sigma_2}{\sigma_1}\right)\mathrm{d}x \\
&= \log\frac{\sigma_2}{\sigma_1} - \frac{1}{2}\log e + \int_{\mathbb{R}}\mathsf{f}(x)\frac{(x-\mu_2)^2}{2\sigma_2^2}\log e\,\mathrm{d}x \\
&= \log\frac{\sigma_2}{\sigma_1} - \frac{1}{2}\log e + \frac{1}{2\sigma_2^2}\left(\sigma_1^2 + \mu_1^2 - 2\mu_1\mu_2 + \mu_2^2\right)\log e \\
&= \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 - \sigma_2^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2}\log e
\end{aligned}$$

For f and g both being Laplace, we have

$$\begin{aligned}
\mathrm{D}(\mathsf{f}\|\mathsf{g}) &= \int_{\mathbb{R}}\mathsf{f}(x)\log\frac{\mathsf{f}(x)}{\mathsf{g}(x)}\,\mathrm{d}x \\
&= \int_{\mathbb{R}}\mathsf{f}(x)\left(\frac{\sqrt{2}|x-\mu_2|}{\sigma_2}\log e - \frac{\sqrt{2}|x-\mu_1|}{\sigma_1}\log e + \log\frac{\sigma_2}{\sigma_1}\right)\mathrm{d}x
\end{aligned}$$

*with contribution by Chen-Hao Hsiao and Wen-Shao Ho

$$= \log\frac{\sigma_2}{\sigma_1} - \log e + \log e \int_{\mathbb{R}} \mathrm{f}(x)\frac{\sqrt{2}|x-\mu_2|}{\sigma_2}\,\mathrm{d}x$$

For $\mu_1 \geq \mu_2$

$$\begin{aligned}
&\int_{\mathbb{R}} \mathrm{f}(x)\frac{\sqrt{2}|x-\mu_2|}{\sigma_2}\,\mathrm{d}x \\
&= \int_{\mathbb{R}} \frac{|x-\mu_2|}{\sigma_1\sigma_2} e^{-\frac{\sqrt{2}|x-\mu_1|}{\sigma_1}}\,\mathrm{d}x \\
&= \int_{-\infty}^{\mu_2} \frac{-(x-\mu_2)}{\sigma_1\sigma_2} e^{\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}}\,\mathrm{d}x + \int_{-\mu_2}^{\mu_1} \frac{(x-\mu_2)}{\sigma_1\sigma_2} e^{\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}}\,\mathrm{d}x \\
&\quad + \int_{\mu_1}^{\infty} \frac{(x-\mu_2)}{\sigma_1\sigma_2} e^{-\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}}\,\mathrm{d}x \\
&= -\frac{\sqrt{2}(x-\mu_2)-\sigma_1}{2\sigma_2} e^{\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}}\Big|_{-\infty}^{\mu_2} + \frac{\sqrt{2}(x-\mu_2)-\sigma_1}{2\sigma_2} e^{\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}}\Big|_{\mu_2}^{\mu_1} \\
&\quad - \frac{\sqrt{2}(x-\mu_2)+\sigma_1}{2\sigma_2} e^{\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}}\Big|_{\mu_1}^{\infty} \\
&= \frac{\sigma_1}{2\sigma_2} e^{-\frac{\sqrt{2}(\mu_2-\mu_1)}{\sigma_1}} + \frac{\sigma_1}{2\sigma_2} e^{-\frac{\sqrt{2}(\mu_2-\mu_1)}{\sigma_1}} + \frac{\sqrt{2}(\mu_1-\mu_2)-\sigma_1}{2\sigma_2} + \frac{\sqrt{2}(\mu_1-\mu_2)+\sigma_1}{2\sigma_2} \\
&= \frac{\sigma_1}{\sigma_2} e^{-\frac{\sqrt{2}(\mu_2-\mu_1)}{\sigma_1}} + \frac{\sqrt{2}(\mu_1-\mu_2)}{\sigma_2} = \frac{\sigma_1}{\sigma_2} e^{-\frac{\sqrt{2}|\mu_2-\mu_1|}{\sigma_1}} + \frac{\sqrt{2}|\mu_1-\mu_2|}{\sigma_2}
\end{aligned}$$

Similarly, for $\mu_1 < \mu_2$

$$\begin{aligned}
&\int_{\mathbb{R}} \mathrm{f}(x)\frac{\sqrt{2}|x-\mu_2|}{\sigma_2}\,\mathrm{d}x \\
&= \int_{-\infty}^{\mu_1} \frac{-(x-\mu_2)}{\sigma_1\sigma_2} e^{\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}}\,\mathrm{d}x + \int_{-\mu_1}^{\mu_2} \frac{-(x-\mu_2)}{\sigma_1\sigma_2} e^{-\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}}\,\mathrm{d}x \\
&\quad + \int_{\mu_2}^{\infty} \frac{(x-\mu_2)}{\sigma_1\sigma_2} e^{-\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}}\,\mathrm{d}x \\
&= -\frac{\sqrt{2}(x-\mu_2)-\sigma_1}{2\sigma_2} e^{\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}}\Big|_{-\infty}^{\mu_1} + \frac{\sqrt{2}(x-\mu_2)+\sigma_1}{2\sigma_2} e^{\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}}\Big|_{\mu_1}^{\mu_2} \\
&\quad - \frac{\sqrt{2}(x-\mu_2)+\sigma_1}{2\sigma_2} e^{-\frac{\sqrt{2}(x-\mu_1)}{\sigma_1}}\Big|_{\mu_2}^{\infty} \\
&= -\frac{\sqrt{2}(\mu_1-\mu_2)-\sigma_1}{2\sigma_2} - \frac{\sqrt{2}(\mu_1-\mu_2)+\sigma_1}{2\sigma_2} + \frac{\sigma_1}{2\sigma_2} e^{-\frac{\sqrt{2}|\mu_2-\mu_1|}{\sigma_1}} + \frac{\sigma_1}{2\sigma_2} e^{-\frac{\sqrt{2}|\mu_2-\mu_1|}{\sigma_1}} \\
&= -\frac{\sqrt{2}(\mu_1-\mu_2)}{\sigma_2} + \frac{\sigma_1}{\sigma_2} e^{-\frac{\sqrt{2}|\mu_2-\mu_1|}{\sigma_1}} = \frac{\sqrt{2}|\mu_1-\mu_2|}{\sigma_2} + \frac{\sigma_1}{\sigma_2} e^{-\frac{\sqrt{2}|\mu_2-\mu_1|}{\sigma_1}}
\end{aligned}$$

Hence,

$$\mathrm{D}(\mathsf{f}\|\mathsf{g}) = \log\frac{\sigma_2}{\sigma_1} - \log e + \log e\left(\frac{\sqrt{2}|\mu_1-\mu_2|}{\sigma_2} + \frac{\sigma_1}{\sigma_2} e^{-\frac{\sqrt{2}|\mu_2-\mu_1|}{\sigma_1}}\right)$$

b) For $\mu_1 = \mu_2$

$$\mathrm{D}(\mathsf{f}\|\mathsf{g}) = \begin{cases} \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2-\sigma_2^2}{2\sigma_2^2}\log e & \text{Gaussian} \\ \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1}{\sigma_2}\log e - \log e & \text{Laplace} \end{cases}$$

Simple calculations based on σ_2/σ_1 and the property of quadratic forms shows that pair of Gaussian always admit a larger divergence if $\mu_1 = \mu_2$.

c) For $\sigma_1 = \sigma_2 = \sigma$,

$$\mathrm{D}(\mathsf{f}\|\mathsf{g}) = \begin{cases} \frac{(\mu_1-\mu_2)^2}{2\sigma^2}\log e & \text{Gaussian} \\ -\log e + \log e\left(\frac{\sqrt{2}|\mu_1-\mu_2|}{\sigma} + e^{-\frac{\sqrt{2}|\mu_2-\mu_1|}{\sigma}}\right) & \text{Laplace} \end{cases}$$

Let $y = |\mu_1 - \mu_2|/\sigma$, then we can compare

$$\mathrm{D}(\mathsf{f}\|\mathsf{g}) = \begin{cases} \frac{y^2}{2}\log e & \text{Gaussian} \\ -\log e + \log e\left(\sqrt{2}y + e^{-\sqrt{2}y}\right) & \text{Laplace} \end{cases}$$

The equation $\frac{y^2}{2}\log e = (-1+\sqrt{2})y + e^{-\sqrt{2}y})$ has two real solutions, one of them being zero. We denote another one as $y_0 > 0$. For $|\mu_1 - \mu_2|/\sigma \leq y_0$, a pair of Laplace distribution admit a divergence greater than or equal to that of Gaussian distribution. For $|\mu_1 - \mu_2|/\sigma > y_0$, a pair of Gaussian has a larger divergence.

2. (Differential entropy) [10]

a) Consider a Laplace random variable $X \sim \mathsf{Lap}(\mu, b)$, that is, the probability density function of X is $\mathsf{f}_X(x) = \frac{1}{2b}e^{-\frac{|x-\mu|}{b}}$, $x \in \mathbb{R}$. Compute its differential entropy $\mathrm{h}(X)$. [4]

b) Consider a problem of maximizing differential entropy $\mathrm{h}(X)$ subject to the constraint that $\mathsf{E}\|X\| \leq \mathsf{B}$. Find the maximum differential entropy and show that a zero-mean Laplace distributed X attains the maximum value. [6]

Solution:

a)

$$\begin{aligned}
\mathrm{h}(X) &= \mathsf{E}_{\mathsf{X}}\left[\log\frac{1}{\mathsf{f}_X(x)}\right] \\
&= \log 2b + \log e\left(\int_{-\infty}^{\mu} \frac{\mu-x}{2b^2} e^{\frac{-(\mu-x)}{b}}\,\mathrm{d}x + \int_{\mu}^{\infty} \frac{x-\mu}{2b^2} e^{\frac{-(x-\mu)}{b}}\,\mathrm{d}x\right)
\end{aligned}$$

$$\begin{aligned}
&= \log 2b + \log e\left(\int_0^{\infty} \frac{t}{2} e^{-t}\,\mathrm{d}t + \int_0^{\infty} \frac{t}{2} e^{-t}\,\mathrm{d}t\right) \\
&= \log 2b + \log e\left(\int_0^{\infty} te^{-t}\,\mathrm{d}t\right) \\
&= \log 2b + \log e\left(-te^{-t}\big|_0^{\infty} + \int_0^{\infty} e^{-t}\,\mathrm{d}t\right) \\
&= \log 2be
\end{aligned}$$

We can also observe that translation doesn't change differential entropy.

b) Let $X \sim \mathrm{Lap}(0, b)$. Its differential entropy $\mathrm{h}(X) = \log(2be)$ and $\mathbf{E}[\|X\|] = b$.

Let Y be a random variable with $\mathbf{E}[\|Y\|] = b$.

Then we have $\mathrm{h}(X) = \mathsf{E}_{\mathsf{X}}\left[-\log \mathsf{f}_X(X)\right] = \mathsf{E}_{\mathsf{X}}\left[-\log\frac{1}{2b} + \left|\frac{X}{b}\right|\right] = \mathsf{E}_Y\left[-\log \mathsf{f}_Y(Y)\right]$.

Then

$$\begin{aligned}
\mathrm{h}(Y) - \mathrm{h}(X) &= \mathsf{E}_Y\left[\log\frac{f_X(Y)}{f_Y(Y)}\right] \\
&\leq \log \mathsf{E}_Y\left[\frac{\mathsf{f}_X(Y)}{\mathsf{f}_Y(Y)}\right] \quad \text{Jensen's inequality} \\
&= 0
\end{aligned}$$

$\mathrm{Lap}(0, b)$ maximizes $\mathrm{h}(Y)$ for given $\mathbf{E}[\|Y\|] = b$, and the maximum value is $\log(2be)$.

Since $\log(2be)$ is increasing in $0 \leq b \leq \mathsf{B}$, the Laplace distribution $\mathrm{Lap}(0, \mathsf{B})$ also maximizes $\mathrm{h}(S)$ for given $\mathsf{E}\|S\| \leq \mathsf{B}$, and the maximum differential entropy is $\log(2\mathsf{B}e)$.

AWGN channel with *average output power constraint*

$$\tfrac{1}{n}\textstyle\sum_{i=1}^n \mathsf{E}\left[Y_i^2\right] \leq \mathsf{B}.$$

where $Y = X + Z$, $Z \perp\!\!\!\perp X$, and $Z \sim \mathrm{N}(0, \sigma^2)$.

Evaluate the channel capacity $\mathrm{C}(\mathsf{B})$. [4]

Solution:

a) Follow the hint, let $\tilde{b}(x) := \mathsf{E}[b(x, Y)]$, we have

$$\frac{1}{n}\sum_{i=1}^n \tilde{b}(x_i) = \frac{1}{n}\sum_{i=1}^n \mathsf{E}_{Y_i}\left[b(x_i, Y_i)\right] \leq \mathsf{B}.$$

Therefore, it can be viewed as an input-only cost constraint.
Furthermore, the capacity is

$$\begin{aligned}
\mathrm{C}(\mathsf{B}) &= \max_{\mathsf{P}_X : \mathsf{E}_{\mathsf{P}_X}[\tilde{b}(X)] \leq \mathsf{B}} \mathrm{I}(X;Y) = \max_{\mathsf{P}_X : \mathsf{E}_{\mathsf{P}_X}\left[\mathsf{E}_{Y\sim\mathsf{P}_{Y|X}}[b(X,Y)]\right] \leq \mathsf{B}} \mathrm{I}(X;Y) \\
&= \max_{\mathsf{P}_X : \mathsf{E}_{\mathsf{P}_X\mathsf{P}_{Y|X}}[b(X,Y)] \leq \mathsf{B}} \mathrm{I}(X;Y)\,.
\end{aligned}$$

b) The capacity can be directed calculated via a), note that for random variable Y satisfying $\mathsf{E}_{\mathsf{P}_{X,Y}}[Y^2] \leq \mathsf{B}$:

$$\mathrm{h}(Y) - \mathrm{h}(Y|X) \leq \frac{1}{2}\log\left(2\pi e\mathsf{B}\right) - \mathrm{h}(Z) = \frac{1}{2}\log\left(\frac{\mathsf{B}}{\sigma^2}\right).$$

The equality can be achieved by choosing $X \sim \mathrm{N}(0, \mathsf{B} - \sigma^2)$.
Hence,

$$\mathrm{C}(\mathsf{B}) = \frac{1}{2}\log\left(\frac{\mathsf{B}}{\sigma^2}\right)$$

4. (Compression with guarantee on the cross-entropy loss) [20]

Consider a discrete memoryless source $S \sim \pi$ with a finite alphabet $\mathcal{S} = \{1, 2, ..., k\}$, $|\mathcal{S}| = k < \infty$. The encoder aims to compress the source so that the decoder can give good estimates of the source sequence. In many applications, however, the decoder may not want to give a deterministic estimate. Instead, for each symbol s_i in a length-n sequence s^n, its goal is to produce a *probability vector* $\boldsymbol{q}_i$ in the k-dimensional probability simplex $\mathcal{P}_k$, where the l-th coordinate, $q_i(l)$, stands for the probability of $s_i = l$ that the decoder believes in based on what it receives from the encoder. A standard way to quantify the loss is the empirical cross entropy loss

$$\ell_{\mathrm{CE}}(s^n, \boldsymbol{q}^n) = \sum_{i=1}^n \frac{1}{n}\log\frac{1}{q_i(s_i)}.$$

Note that it can be viewed as the average distortion per symbol when the distortion function is set to be

$$d : \mathcal{S} \times \mathcal{P}_d \to [0, \infty),\ (s, \boldsymbol{q}) \mapsto d(s, \boldsymbol{q}) = \log\frac{1}{q(s)}.$$

Hence, one can study a lossy source coding problem to understand how to represent a memoryless source with the smallest rate so that the decoder can declare an estimation probability vector with the empirical cross entropy loss not greater than a prescribed level D. By the lossy source coding theorem, the rate is given by the following rate distortion function:

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

a) Show that for the lossy source coding problem, $\mathsf{D}_{\min} = 0$ and $\mathsf{D}_{\max} = \mathrm{H}(\pi)$.

b) Show that for any jointly distributed $(S, \boldsymbol{Q}) \sim P$,

$$\mathrm{H}(S|\boldsymbol{Q}) \leq \mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{1}{Q(S)}\right].$$

Then, argue that $\mathrm{R}(\mathsf{D}) \geq \mathrm{H}(\pi) - \mathsf{D}$, for $0 \leq \mathsf{D} \leq \mathrm{H}(\pi)$.

c) Show that for $0 \leq \mathsf{D} \leq \mathrm{H}(\pi)$,

$$\mathrm{R}(\mathsf{D}) \leq \min_{(S,\hat{S}),\ \hat{S}\in\mathcal{S}} \left\{\mathrm{I}\left(S;\hat{S}\right) \,\middle|\, \mathrm{H}\left(S\middle|\hat{S}\right) \leq \mathsf{D} \text{ and } S \sim \pi\right\}.$$

d) Show that for $0 \leq \mathsf{D} \leq \mathrm{H}(\pi)$,

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

Hence, $\mathrm{R}(\mathsf{D}) = \max\{0, \mathrm{H}(\pi) - \mathsf{D}\}$.

Solution:

a) For $\mathsf{D}_{\min}$, one could choose $q(s) = \begin{cases} 1 & l = s \\ 0 & otherwise \end{cases}$

$\mathsf{E}_S[d(S, q(S))] = \mathsf{E}_S[\log 1] = 0$, so $\mathsf{D}_{\min} = 0$.

For $\mathsf{D}_{\max}$, since

$$\begin{aligned}
\mathsf{E}_S\left[\log\frac{1}{\pi(S)}\right] - \mathsf{E}_S\left[\log\frac{1}{q(S)}\right] &= \mathsf{E}_S\left[\log\frac{q(S)}{\pi(S)}\right] \\
&\leq \log \mathsf{E}_S\left[\frac{q(S)}{\pi(S)}\right] = 0 \quad \forall q.
\end{aligned}$$

$\mathsf{D}_{\max} = \min_q \mathsf{E}_S[d(S, q)] = \mathsf{E}_S\left[\log\frac{1}{\pi(S)}\right] = \mathrm{H}(\pi)$.

b) Denote the law of P as $\mathsf{P}_{S,\boldsymbol{Q}}$ and its marginal as $\boldsymbol{Q}$ as $\mathsf{P}_{\boldsymbol{Q}}$

$$\begin{aligned}
\mathrm{H}(S|\boldsymbol{Q}) - \mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{1}{Q(S)}\right] &= \mathrm{h}(S, \boldsymbol{Q}) - \mathrm{h}(\boldsymbol{Q}) - \mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{1}{Q(S)}\right] \\
&= \mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{Q(S)\mathsf{P}_{\boldsymbol{Q}}(\boldsymbol{Q})}{\mathsf{P}_{S,\boldsymbol{Q}}(S,\boldsymbol{Q})}\right] \\
&\leq \log \mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\frac{Q(S)\mathsf{P}_{\boldsymbol{Q}}(\boldsymbol{Q})}{\mathsf{P}_{S,\boldsymbol{Q}}(S,\boldsymbol{Q})}\right] \\
&= \log\left(\sum_{s\in\mathcal{S}}\int_{\boldsymbol{q}\in\mathcal{P}_k} q(s)\mathsf{P}_{\boldsymbol{Q}}(\boldsymbol{q})\right) = 0
\end{aligned}$$

$$\begin{aligned}
\mathrm{R}(\mathsf{D}) &= \inf_{(S,\boldsymbol{Q})} \left\{\mathrm{I}(S;\boldsymbol{Q}) \middle| \mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{1}{Q(S)}\right] \leq \mathsf{D} \text{ and } S \sim \pi\right\} \\
&= \mathrm{H}(\pi) - \sup_{(S,\boldsymbol{Q})} \left\{\mathrm{H}(S|\boldsymbol{Q}) \middle| \mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{1}{Q(S)}\right] \leq \mathsf{D} \text{ and } S \sim \pi\right\} \\
&\geq \mathrm{H}(\pi) - \mathsf{D}, \quad \forall \mathsf{D}_{\min} = 0 \leq \mathsf{D} \leq \mathsf{D}_{\max} = \mathrm{H}(\pi)
\end{aligned}$$

c) The key to this problem is to make the observation that, for all pair of random variable $(S, \hat{S})$ over $\mathcal{S} \times \mathcal{S}$, we can associate a random vector $\boldsymbol{Q}_{S,\hat{S}}$ over $\mathcal{P}_k$ such that

$$\boldsymbol{Q}_{S,\hat{S}} = \boldsymbol{q}_i = \mathsf{P}_{S|\hat{S}}(\cdot|i) \text{ if } \hat{S} = i$$

$$\text{equivalently } \boldsymbol{Q}_{S,\hat{S}} = \sum_{i=1}^k \mathsf{P}_{S|\hat{S}}(\cdot|i)\mathbb{1}\left\{\hat{S} = i\right\}$$

a mixture of k vectors $\boldsymbol{q}_i$ determined by the outcome of $\hat{S}$. Note that this specify a Markov chain $S - \hat{S} - \boldsymbol{Q}_{S,\hat{S}}$. For this Markov chain, by the data processing inequality, $\mathrm{I}\left(S;\boldsymbol{Q}_{S,\hat{S}}\right) \leq \mathrm{I}\left(S;\hat{S}\right)$. And

$$\mathsf{E}_{(S,\boldsymbol{Q}_{S,\hat{S}})}\left[\log\frac{1}{Q_{S,\hat{S}}(S)}\right] = \mathsf{E}_{(S,\hat{S})}\left[\log\frac{1}{\mathsf{P}_{S|\hat{S}}(S|\hat{S})}\right]$$

Hence,

$$\begin{aligned}
\mathrm{R}(\mathsf{D}) &= \inf_{(S,\boldsymbol{Q})} \left\{\mathrm{I}(S;\boldsymbol{Q}) \middle| \mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{1}{Q(S)}\right] \leq \mathsf{D} \text{ and } S \sim \pi\right\} \\
&= \inf_{(S,\hat{S},\boldsymbol{Q})} \left\{\mathrm{I}(S;\boldsymbol{Q}) \middle| \mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{1}{Q(S)}\right] \leq D \text{ and } S \sim \pi\right\} \\
&\leq \min_{(S,\hat{S})} \left\{\mathrm{I}(S;\boldsymbol{Q}) \middle| \mathsf{E}_{(S,\boldsymbol{Q})\sim P}\left[\log\frac{1}{Q(S)}\right] \leq D \text{ and } S \sim \pi \text{ and } \boldsymbol{Q} = \boldsymbol{Q}_{S,\hat{S}}\right\}
\end{aligned}$$

$$\begin{aligned}
&\leq \min_{(S,\hat{S})} \left\{\mathrm{I}\left(S;\hat{S}\right) \middle| \mathsf{E}_{(S,\hat{S})}\left[\log\frac{1}{\mathsf{P}_{S|\hat{S}}(S|\hat{S})}\right] \leq \mathsf{D} \text{ and } S \sim \pi\right\} \\
&= \min_{(S,\hat{S})} \left\{\mathrm{I}\left(S;\hat{S}\right) \middle| \mathrm{H}\left(S\middle|\hat{S}\right) \leq \mathsf{D} \text{ and } S \sim \pi\right\}
\end{aligned}$$

d) $\mathrm{R}(\mathsf{D}) = \mathrm{H}(\pi) - \mathsf{D}$ since

$$\begin{aligned}
\mathrm{I}\left(S;\hat{S}\right) &= \mathrm{H}(S) - \mathrm{H}\left(S\middle|\hat{S}\right) \\
&= \mathrm{H}(\pi) - \mathrm{H}\left(S\middle|\hat{S}\right) \\
&\geq \mathrm{H}(\pi) - \mathsf{D}.
\end{aligned}$$

The equality holds since $0 \leq \mathsf{D} \leq \mathrm{H}(\pi)$, $\exists \mathsf{P}_{S,\hat{S}}$ s.t. $\mathrm{H}\left(S\middle|\hat{S}\right) = \mathsf{D}$.

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,

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

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, $= \Pr\{S_i = a\} = P_S(a)$.

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

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

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

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

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

$\therefore \forall \epsilon > 0$, 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 $\mathbf{S} = \{0, 1\}$, $P_S(0) = P_S(1) = \frac{1}{2}$, $\gamma = 0.1$.

For the sequence $s^n = 0^n$, $|\pi(0|s^n) - P_S(0)| = \frac{1}{2} \nleq 0.05 = \gamma P_S(0)$.

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

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

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

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

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 $P_S(S_i) \leq 1$, we get that $\log \frac{1}{P_S(S_i)} \geqslant 0$.

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

$\prod_{i=1}^{n} P_S(S_i) \geqslant 2^{-n(H(S)+n^{-1/2}\delta_\zeta(S))}$

$\iff \prod_{i=1}^{n} \frac{1}{P_S(S_i)} \leqslant 2^{n(H(S)+n^{-1/2}\delta_\zeta(S))}$

$\iff \sum_{i=1}^{n} \log \frac{1}{P_S(S_i)} \leqslant n(H(S)+n^{-1/2}\delta_\zeta(S))$

$\iff \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - H(S) \leqslant n^{-1/2}\delta_\zeta(S)$

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

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

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

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

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

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

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

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

Since $\frac{d\Phi^{-1}(x)}{dx} = \frac{1}{\frac{d\Phi(y)}{dy}}\Big|_{y=\Phi^{-1}(x)} = \sqrt{2\pi}e^{y^2/2}\Big|_{y=\Phi^{-1}(x)} = \sqrt{2\pi}e^{(\Phi^{-1}(x))^2/2}$, 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. $nk \leqslant \lfloor n(k + \zeta_n) \rfloor$.

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

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

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

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

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

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

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

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

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

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 $|\mathcal{A}_\delta^{(n)}(S)| \leqslant 2^{n(H(S)+\delta)} \overset{H(S)+\delta<R \Rightarrow n(H(S)+\delta)\leqslant nR-1 \text{ for } n \text{ large enough}}{<} 2^{\lfloor nR \rfloor}$ 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. $\forall n \geq N$, $P_e^{(n)} = \Pr\{S^n \notin \mathcal{A}_\delta^{(n)}(S)\} \leqslant \epsilon$.

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

$\therefore$ 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 $\forall s^n \in \mathcal{A}_\delta^{(n)}(S)$, $\Pr\{S^n = s^n\} \leqslant 2^{-n(H(S)-\delta)}$.

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

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

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

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

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

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 $Q_X(x) = q_x$.

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

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

(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 $q_0 > 0$.

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

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

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

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

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

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

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

$\frac{1}{\mu_X}\log(\mu_X)) = \mu_X(-(1 - \frac{1}{\mu_X})\log(1 - \frac{1}{\mu_X}) - \frac{1}{\mu_X}\log(\frac{1}{\mu_X})) = \mu_X h_b(\mu_X^{-1})$.

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

Problem 2.

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

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

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

(b) First, we know that the series converges $\iff \alpha > 1$, so we only consider $\alpha > 1$.

$H(X_n) = -\mathrm{E}[\log P_{X_n}] = \sum_{n=2}^{\infty} \frac{1}{s_n n(\log n)^\alpha} \log(s_n n(\log n)^\alpha) = \sum_{n=2}^{\infty} \frac{\log s_n}{s_n n(\log n)^\alpha} + \sum_{n=2}^{\infty} \frac{1}{s_n n(\log n)^{\alpha-1}} + \sum_{n=2}^{\infty} \frac{\alpha \log\log n}{s_n n(\log n)^\alpha}$.

For $\alpha \leqslant 2$, since $H(X_n) > \sum_{n=2}^{\infty} \frac{1}{s_n n(\log n)^{\alpha-1}} \to \infty$ from (a); therefore $H(X_n)$ diverges to ∞.

For $\alpha > 2$, since $H(X_n) < \sum_{n=2}^{\infty} \frac{\log s_n}{s_n n(\log n)^\alpha} + \sum_{n=2}^{\infty} \frac{1}{s_n n(\log n)^{\alpha-1}} + \sum_{n=2}^{\infty} \frac{\alpha \log\log n}{s_n n(\log n)^\alpha}$

$\overset{\log\log n < \log n \text{ for } n \geqslant 2}{<} \sum_{n=2}^{\infty} \frac{\log s_n}{s_n n(\log n)^\alpha} + \sum_{n=2}^{\infty} \frac{1}{s_n n(\log n)^{\alpha-1}} + \sum_{n=2}^{\infty} \frac{\alpha}{s_n n(\log n)^{\alpha-1}}$

$= \log s_n + \frac{(1+\alpha)s_{n-1}}{s_n} < \infty$,

and $\sum_{n=2}^{m} \frac{1}{s_n n(\log n)^\alpha} \log(s_n n(\log n)^\alpha)$ is increasing as m increases.

$\Rightarrow H(X_n) = \sum_{n=2}^{\infty} \frac{1}{s_n n(\log n)^\alpha} \log(s_n n(\log n)^\alpha)$ converges.

$\therefore H(X_n)$ exists if $\alpha > 2$, and diverges to ∞ if $1 < \alpha \leqslant 2$.

Problem 3. Note that $P_{X_{\theta_i}|\Theta_i}(\theta_i, x_i)$ is defined as $\Pr\{\Theta_i = \theta_i \wedge X_{\theta_i}[i] = x_i\}$, while $P_{X_{\theta_i}|\Theta_i}(x_i)$ is defined as $\Pr\{X_{\theta_i}[i] = x_i\}$.

Since $X_{\theta_i}[i]$ and Θ_i are independent, there is $P_{X_{\theta_i}|\Theta_i}(\theta_i, x_i) = P_{\Theta_i}(\theta_i)P_{X_{\theta_i}|\Theta_i}(x_i)$.

(a) $\because \forall l, n \in \mathbb{N}$, $P_{X_{\theta_1}[1],X_{\theta_2}[2],\ldots,X_{\theta_n}[n]} = P_{X_\Theta[1],X_\Theta[2],\ldots,X_\Theta[n]}$

$\overset{X_\Theta \text{ is stationary no matter } \Theta \text{ is or not}}{=} P_{X_\Theta[l+1],X_\Theta[l+2],\ldots,X_\Theta[l+n]} = P_{X_{\theta_{l+1}}[l+1],X_{\theta_{l+2}}[l+2],\ldots,X_{\theta_{l+n}}[l+n]}$.

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

By the definition of entropy rates,

$\lim_{n\to\infty} -\frac{1}{n}\mathrm{E}[\log P_{X_1[1],X_1[2],\ldots,X_k[n]}] = \lim_{n\to\infty} \frac{1}{n} H(X_k[1], X_k[2], \ldots, X_k[n]) = \mathcal{H}_k$.

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

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

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

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

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

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

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

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

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

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

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

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

$\therefore P_{X_{\theta_1}[1],X_{\theta_2}[2],\ldots,X_{\theta_n}[n]} = \prod_{i=1}^{n} P_{X_{\theta_i}[i]|X_{\theta_1}[1],X_{\theta_2}[2],\ldots,X_{\theta_{i-1}}[i-1]} = P_{X_{\theta_1}[1]} \prod_{i=2}^{n} P_{X_{\theta_i}[i]|X_{\theta_{i-1}}[i-1]}$

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

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

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

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

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

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

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

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

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

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

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

$= -\sum_{\theta_1,\theta_2} P_{\Theta_1}(\theta_1) P_{\Theta_2|\Theta_1}(\theta_2|\theta_1) \log(P_{\Theta_2|\Theta_1}(\theta_2|\theta_1))$

$+ \sum_{\theta_1,\theta_2} P_{\Theta_1}(\theta_1) P_{\Theta_2|\Theta_1}(\theta_2|\theta_1) H(X_{\theta_2}[2])$

$= -(1-q)(1-\alpha)\log(1-\alpha) - (1-q)\alpha\log(\alpha) - q\beta\log(\beta) - q(1-\beta)\log(1-\beta) + H(X_0[2])((1-q)(1-\alpha+q\beta) + H(X_1[2])((1-q)\alpha + q(1-\beta))$

$\overset{\{X_k[i]\} \text{ are i.i.d.} \Rightarrow \mathcal{H}_k \overset{\text{def}}{=} H(\{X_k[i]\}) = H(X_k[i])}{=} (1-q)H_b(\alpha) + qH_b(\beta) + \mathcal{H}_0((1-q)(1-\alpha) + q\beta) + \mathcal{H}_1((1-q)\alpha + q(1-\beta))$

$= \frac{\beta}{\alpha+\beta} H_b(\alpha) + \frac{\alpha}{\alpha+\beta} H_b(\beta) + \mathcal{H}_0(\frac{\beta}{\alpha+\beta}(1-\alpha) + \frac{\alpha}{\alpha+\beta}\beta) + \mathcal{H}_1(\frac{\beta}{\alpha+\beta}\alpha + \frac{\alpha}{\alpha+\beta}(1-\beta))$

$= \frac{\beta}{\alpha+\beta}(H_b(\alpha) + \mathcal{H}_0) + \frac{\alpha}{\alpha+\beta}(H_b(\beta) + \mathcal{H}_1)$.

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

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

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

$\because p_0 < p_1$.

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

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

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

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

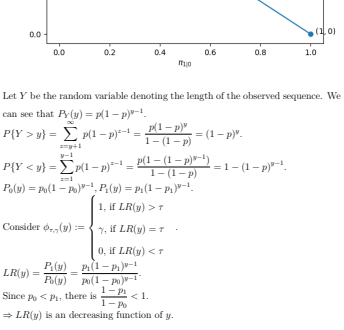
We only need to consider the cases $\tau = LR(x)$ for some x, since other cases can be reduced to these cases by setting γ properly.

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.

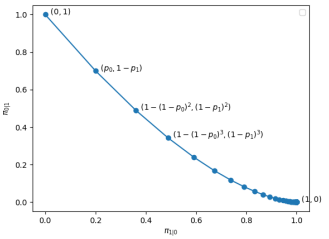
By Neyman-Pearson theorem, $\phi_{\tau,\gamma}$ is optimal.

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 $\tau = LR(y)$, $\pi_{1|0}(\phi_{\tau,\gamma}) = P_1\{Y < y\} + \gamma P_0\{Y = y\} = 1-(1-p_0)^{y-1} + \gamma p_0(1-p_0)^{y-1} = 1-(1-p_0)^{y-1}(1-\gamma p_0)$.

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

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

Problem 2.

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

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 $n\to\infty$.

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

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

$\frac{d}{dp}\mathrm{D}(P\|G(p)) = \frac{1}{1-p} + \frac{1}{p} - \frac{1}{p}\mu = \frac{1-(1-p)\mu}{p(1-p)}$, 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 $\iff Q \sim R$.

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

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

$= \sum_{i=1}^m H(X_i) - m\sum_{j=1}^{\infty} R(j)\log Q(j)$

$\ge \sum_{i=1}^m H(X_i) - mH(R)$.

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

Information Theory HW3

許博翔

November 2, 2023

Problem 1.

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

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

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

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

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

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

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

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

$I(X_i; X_{i+1}|X_1, \ldots, X_{i-1}) = H(X_i|X_1, \ldots, X_{i-1}) - H(X_i|X_1, \ldots, X_{i-1}, X_{i+1}) \overset{X_1,\ldots,X_i \text{ are mutually independent}}{=} H(X_i) - H(X_i|X_1, \ldots, X_{i-1}, X_{i+1}) \overset{X_1,\ldots,X_{i+1} \text{ are mutually independent if } i<n-1}{=}$

$\begin{cases} H(X_i) - H(X_i) = 0, \text{ if } i < n-1 \\ H(X_i) - H(X_1 \oplus \cdots \oplus X_{i-1} \oplus X_{i+1}|X_1, \ldots, X_{i-1}, X_{i+1}) = H(X_i) - 0 = 1, \text{ if } i = n-1 \end{cases}$.

Problem 2.

(a) $I(X_4; X_1) = I(X_4; X_2) - I(X_4; X_2|X_1)$

$I(X_3; X_1) = I(X_3; X_2) - I(X_3; X_2|X_1) = I(X_3; X_2) - I(X_3; X_2|X_1, X_4) - I(X_4; X_2|X_1)$

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

(b) It's equivalent to two Markov's chains: $X_1 - X_2 - X_3, X_1 - X_2 - X_4$.

$I(X_3; X_1) = I(X_3; X_2) - I(X_3; X_2|X_1)$

$I(X_4; X_1) = I(X_4; X_2) - I(X_4; X_2|X_1)$

$I(X_1; X_2) + I(X_3; X_4) \ge I(X_1; X_2) + I(X_4; X_1) - I(X_4; X_1|X_3) \ge I(X_2; X_1) + I(X_4; X_1) - I(X_2; X_1|X_3) = I(X_1; X_3) + I(X_1; X_4)$.

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: $P_J(i) = \frac{2^{C^{(i)}}}{\sum_{j=1}^{l}2^{C^{(j)}}}$.

$\sum_{i=1}^{l} P_I(i)C^{(i)} + H(I) = \sum_{i=1}^{l} P_I(i)\log\frac{2^{C^{(i)}}}{\sum_{j=1}^{l}2^{C^{(j)}}P_I(i)} + \sum_{i=1}^{l} P_I(i)\log\sum_{j=1}^{l}2^{C^{(j)}} = \sum_{i=1}^{l} P_I(i)\log\frac{P_J(i)}{P_I(i)} + \log\sum_{j=1}^{l}2^{C^{(j)}} = -D(P_I\|P_J) + \log\sum_{j=1}^{l}2^{C^{(j)}} \ge \log\sum_{j=1}^{l}2^{C^{(j)}}$, 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 $X \sim \mathrm{Ber}(q)$.

$\Rightarrow P_Y(0) = 1-q+pq = 1-\frac{1}{2}q, P_Y(1) = q(1-p) = \frac{1}{2}q$.

$I(X;Y) = H(X) + H(Y) - H(X,Y) = -q\log q - (1-q)\log(1-q) - \frac{1}{2}q\log(\frac{1}{2}q) - (1-\frac{1}{2}q)\log(1-\frac{1}{2}q) + (1-q)\log(1-q) + 2\cdot\frac{1}{2}q\log(\frac{1}{2}q) = -q\log q + \frac{1}{2}q\log(\frac{1}{2}q) - (1-\frac{1}{2}q)\log(1-\frac{1}{2}q) = -q - \frac{1}{2}q\log(\frac{1}{2}q) - (1-\frac{1}{2}q)\log(1-\frac{1}{2}q)$.

Let $\frac{dI(X;Y)}{dq} = -1 - \frac{1}{2}\log(\frac{1}{2}q) - \frac{1}{2}\log e + \frac{1}{2}\log(1-\frac{1}{2}q) + \frac{1}{2}\log e = -1 + \frac{1}{2}\log\frac{1-\frac{1}{2}q}{\frac{1}{2}q} = 0$.

$\Rightarrow \log\frac{1-\frac{1}{2}q}{\frac{1}{2}q} = 2$.

$\Rightarrow \frac{1-\frac{1}{2}q}{\frac{1}{2}q} = 4$.

$\Rightarrow q = \frac{2}{5}$.

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

(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 $\int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}$.

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

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

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

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

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

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

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

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

$= \log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{1}{2}\log e + \frac{1}{2}\log e\frac{\sigma_1^2}{\sigma_2^2} + \frac{1}{2}\log e(\frac{\mu_1-\mu_2}{\sigma_2})^2$

$= \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\log e}{2\sigma_2^2}(\sigma_1^2 - \sigma_2^2 + (\mu_1-\mu_2)^2)$.

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

$\int xe^xdx = e^xx - \int e^xdx = (x-1)e^x + c$.

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

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

$\int_{-\infty}^{\infty}|x-a|e^{-|x-b|}dx=\int_{-\infty}^{\infty}|x+b-a|e^{-|x|}dx.$

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 $\int_{-\infty}^{\infty}|x-c|e^{-|x|}dx=\int_{-\infty}^{\infty}|x+c|e^{-|x|}dx=-2c+2e^{c}.$

$\therefore\int_{-\infty}^{\infty}|x-a|e^{-|x-b|}dx=2|a-b|+2e^{-|a-b|}.$

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

$\int_{-\infty}^{\infty}\frac{1}{\sqrt{2}\sigma_1}e^{-\frac{\sqrt{2}|x-\mu_1|}{\sigma_1}}\left(\log\left(\frac{\sigma_2}{\sigma_1}\right)+\sqrt{2}\log e|(\frac{|x-\mu_2|}{\sigma_2}-\frac{|x-\mu_1|}{\sigma_1}|\right)dx$

$=\log\left(\frac{\sigma_2}{\sigma_1}\right)+\frac{\log e}{\sigma_1\sigma_2}(\frac{\sigma_1^2}{2})(2\cdot\frac{\sqrt{2}}{\sigma_1}|\mu_1-\mu_2|+2e^{-\frac{\sqrt{2}}{\sigma_1}|\mu_1-\mu_2|})-\frac{\log e}{\sigma_1^2}(\frac{\sigma_1^2}{2})2$

$=\log\left(\frac{\sigma_2}{\sigma_1}\right)+\frac{\sigma_1\log e}{\sigma_2}(\frac{\sqrt{2}}{\sigma_1}|\mu_1-\mu_2|+e^{-\frac{\sqrt{2}}{\sigma_1}|\mu_1-\mu_2|})-\log e.$

(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 $=\frac{\log e}{2}(\mu_1-\mu_2)^2-\log e(|\frac{\sqrt{2}}{\sigma_1}|\mu_1-\mu_2|)+\log e$

$=\frac{\log e}{2}x^2-\log e(|\frac{\sqrt{2}}{\sigma_1}x+e^{-\frac{\sqrt{2}}{\sigma_1}x})+\log e$

$=\log e(|\frac{1}{2}x^2-\frac{\sqrt{2}}{\sigma_1}x-e^{-\frac{\sqrt{2}}{\sigma_1}x}+1).$

$\therefore$ 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 $\text{Lap}(0,\frac{B}{2})$ since the equation in (1) holds $\iff\mu=0$, and the equation in (2) holds.

Problem 3.

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

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.

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

$\therefore$ 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, $C(B)=\max_{P_X:\text{E}_{P_X}\tilde{b}(X)\leq B}I(X;Y)$

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

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

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

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

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

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

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

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 $p_i,q_i\geq 0$, then $\sum_i p_i\log\frac{1}{p_i}\leq\sum_i p_i\log\frac{1}{q_i}.\quad(1)$

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

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

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

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

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

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

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

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

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

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

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

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

Consider the distribution $\mathbf{Q}=\mathbf{q}_{\hat{S}}$:

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 $\mathbf{q}_{\hat{s}}(s)=\mathbb{I}(s=\hat{s})\overset{\mathbf{q}=\mathbf{q}_{\hat{S}}}{=}P_{S|\mathbf{Q}}(s|\mathbf{q}_{\hat{s}})$.

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

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

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

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

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