1 Distribution

$\text{Bin}(n,p): P_X(x) = \binom{n}{x} p^x q^{n-x}$ for $x \in [n]_0$.
$\mu = np, \sigma^2 = npq, H(X) = \frac{1}{2}\log(2\pi enpq) + O(\frac{1}{n})$.
$\text{Pois}(\lambda): P_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x \in \mathbb{N}_0$.
$\mu = \sigma^2 = \lambda$.

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

2 Markov Chain

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

3 Central Limit Theorem

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

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

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

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

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

5 Entropy

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

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

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

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

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

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

The above still holds for h.

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

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

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

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

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

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

6 Representing A Sequence with Memory Almost Losslessly

Entropy rate:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

7 Information for Continuous Distributions

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

8 Learning a Bit of Information

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

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

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

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

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

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

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

Neyman-Pearson: LRT is optimal.

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

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

Typical set:

9 Information Divergence

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

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

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

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

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

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

10 Error Exponents and Chernoff Information

P_0, P_1 are given.

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

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

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

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

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

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

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

11 Deviverling Information Reliably

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

12 Mutual Information

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

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

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

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

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

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

13 Noisy Channel Coding Theorem

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

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

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

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

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

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

14 Channel Coding with a Cost Constraint

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

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

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

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

The above also holds for CMC.

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

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

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

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

15 Lossy Source Coding

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

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

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

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

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

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

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

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

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

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

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

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

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 $\forall\, \delta' > 0, n \in \mathbb{N}$, $\mathcal{A}_{\delta'}^{(n)} \not\subseteq \mathcal{T}_\gamma^{(n)}$. [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 $\forall a \in \mathcal{S}$, $|\pi(a|s^n) - \mathsf{P}_S(a)| \le \gamma \mathsf{P}_S(a)$

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

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

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

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

$$\begin{aligned} \mathsf{Pr}\{S^n = s^n\} &= \sum_{a\in\mathcal{S}}\pi(a|s^n)\log\mathsf{P}_S(a) \\ &\begin{cases} \le 2^{-n(1-\gamma)\mathrm{H}(S)} = 2^{-n(\mathrm{H}(S)-\xi(\gamma))} \\ \ge 2^{-n(1+\gamma)\mathrm{H}(S)} = 2^{-n(\mathrm{H}(S)+\xi(\gamma))}. \end{cases} \end{aligned}$$

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

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

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

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

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

$$\begin{aligned} \mathsf{Pr} \bigcup_{a\in\mathcal{S}} \{|\pi(a|s^n) - \mathsf{P}_S(a)| > \gamma\mathsf{P}_S(a)\} &\le \sum_{a\in\mathcal{S}} \mathsf{Pr}\left\{|\pi(a|s^n) - \mathsf{P}_S(a)| > \gamma\mathsf{P}_S(a)\right\} \\ &\le \sum_{a\in\mathcal{S}} \frac{\epsilon}{|\mathcal{S}|} = \epsilon \end{aligned}$$

Property 3 by Property 1,

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

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

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

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

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

$\mathcal{T}_\gamma^{(n)} \subseteq \mathcal{A}_\delta^n$ 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 $S_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}_S$, $i = 1, 2, \ldots$, 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 $(n, \lfloor n\mathrm{R} \rfloor)$ codes such that $\forall\, \epsilon > 0$,

$$\mathsf{P}_\mathsf{e}^{(n)} \le \epsilon \quad \text{for } n \text{ sufficiently large.} \tag{†}$$

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

Suppose we would like to achieve (†) 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

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

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

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

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 $\mathsf{Pr}\{S^n \in \mathcal{B}_\delta^{(n)}(S)\} \to \Phi(\delta)$ 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 (†) 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

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

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,

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

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

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

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,

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

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

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

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

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

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

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

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:

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

Therefore, the choice

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

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 $(n, \lfloor n\mathrm{R}\rfloor)$ source codes indexed by $n = 1, 2, \ldots$ with compression rate $\mathrm{R} > 0$.

Prove the following statements.

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

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

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

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

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

Solution:

a) Let $\delta = \mathrm{R} - \mathrm{H}(S)$. A simple scheme is to use $\mathcal{A}_{\delta/2}^{(n)}$ and standard aymptotic 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 $\mathsf{P}_{\mathsf{e}}^{(n)} < \epsilon_m$. By the lossless source coding theorem in the lecture and the Archimedean property, $\exists n_m > \max_{i=1}^{m-1}\{n_i\}$ such that $\forall n \geq n_m, k^*(n, \epsilon_m) < n(\mathrm{H}(S) + \delta/2) < \lfloor n\mathrm{R} \rfloor$. $\{n_m\}_m^{\infty}$ then is a strictly increasing sequence of integers.

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

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

b) Let $\delta = \mathrm{H}(S) - \mathrm{R}$. Again, a simple strategy is to use $\mathcal{A}_{\delta/2}^{(n)}$ and standard aymptotic 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 $(n, \lfloor n\mathrm{R} \rfloor)$ codes with error probabilitties not converging to 1, namely, for some $\epsilon < 0$, $\forall m \in \mathbb{N}, \exists n'(m) \geq m$ such that $\mathsf{P}_{\mathsf{e}}^{(n'(m))} < 1 - \epsilon$.

$\forall m \in \mathbb{N}$, 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.

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

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 $\forall n \geq n_{\delta,\epsilon}$,

$$k^*(n, 1-\epsilon) > n(\mathrm{H}(S) - \delta/2) > n\mathrm{R} \tag{2}$$

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 $k^*(n'(m), 1-\epsilon) > n'(m)\mathrm{R}$. 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:

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

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 $\sum_{i=1}^\infty q_i = 1$,

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

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

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

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, $\sum_{i=1}^\infty p_i \log \frac{q_i}{p_i} \le \log \sum_{i=1}^\infty p_i \frac{q_i}{p_i} = 0$.

b)
$$\left\{ \begin{array}{l} 1 = \sum_{i=1}^\infty q_i = \frac{2^{-(\alpha+\beta)}}{1-2^{-\alpha}} \\ \mu_X = \sum_{i=1}^\infty i q_i = \frac{2^{-(\alpha+\beta)}}{(1-2^{-\alpha})^2} \end{array} \right. \implies \left\{ \begin{array}{l} \alpha = \log \frac{\mu_X}{\mu_X - 1} \\ \beta = \log(\mu_X - 1) \end{array} \right.$$

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

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

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 $\alpha \ge 0$. 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 $X_\alpha \in \{2, 3, \dots\}$ 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 $1 < \alpha \le 2$. [7]

Solution:

a) $\int_2^{\infty} \frac{dx}{x(\log x)^{\alpha}} = \begin{cases} \frac{ln2}{(\alpha-1)} & (\alpha > 1) \\ \infty & (0 \le \alpha \le 1) \end{cases}.$

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 $\alpha > 1$.

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

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

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 $\{X_0[i] \mid i \in \mathbb{N}\}$ and $\{X_1[i] \mid i \in \mathbb{N}\}$ taking values in disjoint alphabets $\mathcal{X}_0$ and $\mathcal{X}_1$ respectively. The two processes are independent from each other, that is, $\{X_0[i]\} \perp\!\!\!\perp \{X_1[i]\}$, and they have entropy rates $\mathcal{H}_0$ and $\mathcal{H}_1$ respectively. Let $\{\Theta_i \mid i \in \mathbb{N}\}$ 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, $\forall n.l \in \mathbf{N}$,

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

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

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

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

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

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

Grading Policy:

a) Stationary argument [2], chain rule and entropy calculation [3], asympototic [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 $X_1, X_2, ...$ 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:

$$\begin{aligned} \mathcal{H}_0 : &\quad p = p_0 \\ \mathcal{H}_1 : &\quad p = p_1 \end{aligned}$$

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

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 $\varpi^*_{0|1}(n, \epsilon)$ 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: $\frac{p_1}{p_0}, \frac{1-p_1}{1-p_0}$. Therefore, discuss the range of τ we get

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

[*] 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

$$\begin{aligned} \mathcal{H}_0 : &\quad L \sim \text{Geo}(p_0) \\ \mathcal{H}_1 : &\quad L \sim \text{Geo}(p_1) \end{aligned}$$

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

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

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 $|01|0001|1|01|$. Apperantly, the length of the subsequence between two — is a geometric random variable. Hence, in this subproblem, we are testing $\text{Geo}(p_0)^{\otimes n}$ and $\text{Geo}(p_1)^{\otimes n}$. By Chernoff-Stein lemma,

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

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*

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

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,

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

Solution:

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

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

b) Recall the basic Bayes' theorem.

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

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 $n \geq N_\epsilon$

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

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

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

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

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

Furthermore,

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

Hence, with probability 1,

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

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

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

Grading Policy:

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

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

3. Minimizing information divergence) [22]

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

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

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 $\mathrm{D}(\mathsf{P}\|\mathrm{G}(p))$ 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:

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

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

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

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

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

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. 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 $\{X_i \mid i = 1, 2, \ldots\}$. 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:

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

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

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

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

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

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

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

Solution:

a)

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

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

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

3. (Sum Channel) [16]

Consider l DMC's

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

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

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

The **sum channel** $(\mathcal{X}^{\oplus}, \mathsf{P}^{\oplus}_{Y|X}, \mathcal{Y}^{\oplus})$ 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,

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

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

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

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

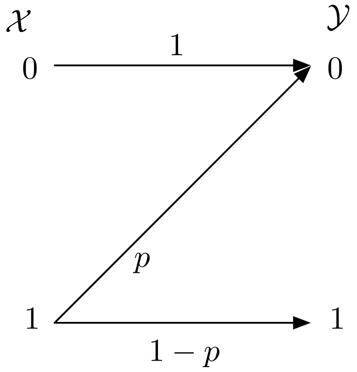
$$
\begin{aligned}
&= \max_{\mathsf{P}_X^{\oplus}} \left\{ \mathrm{H}(I) + \mathrm{I}\left(X^{\oplus}; Y^{\oplus} \middle| I\right) \right\} && I \text{ is deterministic of } X^{\oplus} \\
&= \max_{\mathsf{P}_X^{\oplus}} \sum_{i=1}^{l} \mathsf{P}_I(i) \left[\log \frac{1}{\mathsf{P}_I(i)} + \mathrm{I}\left(X^{\oplus};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

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

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

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

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 P_X^*. Also find P_Y^*, the output distribution induced by the input distribution P_X^*. [8]

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

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

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

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

Solution:

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

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

$\mathrm{C}(q)$ 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:

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

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

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

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

b) Yes, it is unique.

c)

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

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

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

For $\mu_1 \ge \mu_2$

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

Similarly, for $\mu_1 < \mu_2$

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

Hence,

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

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 $\sigma_1 = \sigma_2 = \sigma$,

$$\mathrm{D}(\mathsf{f}\|\mathsf{g}) = \begin{cases} \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2}\log e & \text{Gaussian} \\ -\log e + \log e\left(\frac{\sqrt{2}|\mu_1-\mu_2|}{\sigma} + e^{-\frac{\sqrt{2}|\mu_1-\mu_2|}{\sigma}}\right) & \text{Laplace} \end{cases}$$

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

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

The equation $\frac{y^2}{2} - (-1+\sqrt{2}y + e^{-\sqrt{2}y}) = 0$ has two real solutions, one of them being zero. We denote another one as $y_0 > 0$. For $|\mu_1-\mu_2|/\sigma \le y_0$, 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 $\mathsf{E}[|X|] \le \mathsf{B}$. Find the maximum differential entropy and show that a zero-mean Laplace distributed X attains the maximum value. [6]

Solution:

a)

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

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

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 $\mathsf{E}[|Y|] = b$.

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

Then

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

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

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

3. (Channel Coding with Input-Output Cost Constraint) [10]

In this problem we explore channel coding with input and output cost constraint.

a) Consider a DMC $(\mathcal{X}, \mathsf{P}_{Y|X}, \mathcal{Y})$. Let $b : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ be an input-output cost function. Suppose the channel coding has to satisfy the following average cost constraint: for each codeword x^n,

$$\tfrac{1}{n} \textstyle\sum_{i=1}^n \mathsf{E}_{Y_i}[b(x_i, Y_i)] \le \mathsf{B}.$$

Note that Y_i follows distribution $\mathsf{P}_{Y|X}(\cdot|x_i)$.

Argue that the problem is equivalent to another channel coding problem with a properly defined input-only cost function. Show that the capacity-cost function is

$$\mathsf{C}(\mathsf{B}) = \max_{\mathsf{P}_X : \mathsf{E}_{\mathsf{P}_X \mathsf{P}_{Y|X}}[b(X,Y)] \le \mathsf{B}} \mathrm{I}(X;Y)\,. \qquad [6]$$

Hint: Consider the input-only cost function $\tilde{b}(x) := \mathsf{E}[b(x, Y)]$, and check that the steps in the proof of DMC with input cost in the lecture are still valid.

b) Using discretization techniques, the above DMC result can be extended to continuous memoryless channels. With the extension (no need to prove it here), let us consider an

AWGN channel with *average output power constraint*

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

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

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

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

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

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

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

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

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

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:

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

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 $(S, \boldsymbol{Q}) \sim P$,

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

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

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

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

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

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

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

Solution:

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
\boldsymbol{Q}_{S,\hat{S}} &= \boldsymbol{q}_i = \mathsf{P}_{S|\hat{S}}(\cdot|i)\text{ if }\hat{S} = i \\
\text{equivalently }\boldsymbol{Q}_{S,\hat{S}} &= \sum_{i=1}^{k}\mathsf{P}_{S|\hat{S}}(\cdot|i)\mathbb{1}\left\{\hat{S} = i\right\}
\end{aligned}$$

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, $\mathrm{I}\left(S;\boldsymbol{Q}_{S,\hat{S}}\right)\le\mathrm{I}\left(S;\hat{S}\right)$. 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,

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

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

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. $\mathrm{H}\left(S\middle|\hat{S}\right) = \mathsf{D}$.