Problem Set 2 — *Due Friday, October 26, before class starts*

For the Exercise Sessions on Oct 12 and 19

Last name	First name	SCIPER Nr	Points

Problem 1: Information Measures for Continuous Random Variables

Recommended Reading: Chapter 8 of the book by T. M. Cover and J. A. Thomas. "Elements of Information Theory," Second Edition, Wiley, 2006. It is available for free download from the EPFL library.

Find the differential entropy $h(X)$ for the case where X is an exponentially distributed random variable of mean $1/\lambda$.

Solution

Let X be an exponentially distributed random variable with mean $\frac{1}{\lambda}$. Then the pdf of X is

$$f(x) = \lambda e^{-\lambda x} \text{ for } x \geq 0 \tag{1}$$

The differential entropy of X is

$$\begin{aligned}
h(X) &= -\int_0^\infty f(x)\log f(x)dx &(2)\\
&= -\int_0^\infty \lambda e^{-\lambda x}\log(\lambda e^{-\lambda x})dx &(3)\\
&= -\left(\int_0^\infty \lambda e^{-\lambda x}\log\lambda dx + \int_0^\infty \lambda e^{-\lambda x}(-\lambda x)dx\right) &(4)\\
&= -\log\lambda\int_0^\infty \lambda e^{-\lambda x}dx + \int_0^\infty x\lambda e^{-\lambda x}dx &(5)\\
&= -\log\lambda + 1 &(6)
\end{aligned}$$

where the last step is because $\int_0^\infty \lambda e^{-\lambda x}dx = \int_0^\infty f(x)dx = 1$ and $\int_0^\infty x\lambda e^{-\lambda x}dx = \mathbb{E}[X] = \frac{1}{\lambda}$.

Problem 2: MMSE Estimation

Consider the scenario where $p(x|d) = de^{-dx}$, for $x \geq 0$ (and zero otherwise), that is, the observed data x is distributed according to an exponential with mean $1/d$. Moreover, the desired variable d itself is also exponentially distributed, with mean $1/\mu$.

(a) Find the MMSE estimator of d given x, and calculate the corresponding mean-squared error incurred by this estimator.

(b) Find the MAP estimator of d given x.

Solution

(a) Since d is exponentially distributed random variable with mean $1/\mu$, we have

$$p(d) = \mu e^{-\mu d} \tag{7}$$

Then the probability $p(x,d) = p(d)p(x|d) = \mu e^{-\mu d} d e^{-dx} = \mu d e^{-(\mu+x)d}$. Thus, we have

$$p_X(x) = \int_d p(x,d) = \frac{\mu}{(\mu+x)^2} \tag{8}$$

Given x, the probability of $p(d|x) = (\mu+x)^2 d e^{-(\mu+x)d}$, which is Gamma distribution $\Gamma(2, \mu+x)$. The MMSE estimator of d given x, $\hat{d}_{MMSE}(x)$, satisfies

$$\hat{d}_{MMSE}(x) = E[d|X=x] = \frac{2}{\mu+x} \tag{9}$$

and the mean-squared error is

$$\begin{aligned}
\mathcal{E} &= E_D[(d-\hat{d}_{MMSE})^2] && (10)\\
&= E_X[E_D[(d-\hat{d}_{MMSE}(x))^2|X=x]] && (11)\\
&= \int E_D[(d-\hat{d}_{MMSE}(x))^2|X=x]p_X(x)dx && (12)\\
&= \int E_D[(d^2 - 2d\hat{d}_{MMSE}(x) + \hat{d}^2_{MMSE}(x)|X=x]p_X(x)dx && (13)\\
&= \int (E_D[d^2|X=x] - \hat{d}^2_{MMSE}(x))p_X(x)dx && (14)\\
&= \int \left(\frac{6}{(\mu+x)^2} - \frac{4}{(\mu+x)^2}\right) p_X(x)dx && (15)\\
&= \int \frac{2\mu}{(\mu+x)^4}dx && (16)\\
&= \frac{2}{3\mu^2} && (17)
\end{aligned}$$

where $E_D[d\hat{d}_{MMSE}(x)|X=x] = \hat{d}^2_{MMSE}(x)$ and $E_D[d^2|X=x] = \mathrm{Var}(D|X) + E[D|X] = \frac{6}{(\mu+x)^2}$ is because $p(d|x)$ is Gamma distribution.

(b) MAP estimator is

$$\begin{aligned}
\hat{d}_{MAP}(x) &= \arg\max_d p(d|x) && (18)\\
&= \arg\max_d (\mu+x)^2 d e^{-(\mu+x)d} && (19)\\
&= \arg\max_d d e^{-(\mu+x)d} && (20)\\
&= \frac{1}{\mu+x} && (21)
\end{aligned}$$

as $\frac{\partial}{\partial d} d e^{-(\mu+x)d} = 0$, when $d = \frac{1}{\mu+x}$.

Problem 3: Tweedie's Formula

For the special case where $X = D + N$, where N is Gaussian noise of mean zero and variance σ^2, *Tweedie's formula* says that the conditional mean (that is, the MMSE estimator) can be expressed as

$$\mathbb{E}\left[D \mid X = x\right] = x + \sigma^2 \ell'(x), \tag{22}$$

where

$$\ell'(x) = \frac{d}{dx} \log f_X(x), \tag{23}$$

where $f_X(x)$ denotes the marginal PDF of X. In this exercise, we derive this formula.

(a) Assume that $f_{X|D}(x|d) = e^{\alpha d x - \psi(d)} f_0(x)$ for some functions $\psi(d)$ and $f_0(x)$ and some constant α (such that $f_{X|D}(x|d)$ is a valid PDF for every value of d). Define

$$\lambda(x) = \log \frac{f_X(x)}{f_0(x)}, \tag{24}$$

where $f_X(x)$ is the marginal PDF of X, i.e., $f_X(x) = \int f_{X|D}(x|\delta) f_D(\delta) d\delta$. With this, establish that

$$\mathbb{E}\left[D \mid X = x\right] = \frac{1}{\alpha} \frac{d}{dx} \lambda(x). \tag{25}$$

(b) Show that the case where $X = D + N$, where N is Gaussian noise of mean zero and variance σ^2, is indeed of the form required in Part *(a)* by finding the corresponding $\psi(d), f_0(x)$, and α. Show that in this case, we have

$$\frac{f_0'(x)}{f_0(x)} = -\frac{x}{\sigma^2}, \tag{26}$$

and use this fact in combination with Part *(a)* to establish Tweedie's formula.

Solution

This formula is due to M. C. K. Tweedie, "Function sof a statistical variate with given means, with special reference to Laplacian distributions," *Proc. Camb. Phil. Soc.,* Vol. 43 (1947), pp.41-49.

(a) Simply plugging in, we find

$$\begin{aligned}
\frac{d}{dx}\lambda(x) &= \frac{d}{dx} \log \left(\frac{\int f_{X|D}(x|\delta) f_D(\delta) d\delta}{f_0(x)} \right) && (27)\\
&= \frac{d}{dx} \log \left(\frac{\int e^{\alpha\delta x - \psi(\delta)} f_0(x) f_D(\delta) d\delta}{f_0(x)} \right) && (28)\\
&= \frac{d}{dx} \log \int e^{\alpha\delta x - \psi(\delta)} f_D(\delta) d\delta && (29)\\
&= \frac{1}{\int e^{\alpha\delta x - \psi(\delta)} f_D(\delta) d\delta} \int \alpha\delta e^{\alpha\delta x - \psi(\delta)} f_D(\delta) d\delta && (30)
\end{aligned}$$

But since we know that

$$\int e^{\alpha\delta x - \psi(\delta)} f_0(x) f_D(\delta) d\delta = f_X(x), \tag{31}$$

we can rewrite

$$\frac{d}{dx}\lambda(x) = \frac{f_0(x)}{f_X(x)}\int \alpha\delta e^{\alpha\delta x-\psi(\delta)} f_D(\delta)d\delta \tag{32}$$

$$= \alpha\int \delta\, \underbrace{\frac{e^{\alpha\delta x-\psi(\delta)} f_0(x) f_D(\delta)}{f_X(x)}}_{f_{D|X}(d|x)} d\delta \tag{33}$$

$$= \alpha\mathbb{E}\left[D|\, X = x\right] \tag{34}$$

as claimed.

(b) In this case, we have

$$f_{X|D}(x|d) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{d^2}{2\sigma^2}} e^{\frac{1}{\sigma^2}xd}. \tag{35}$$

Pattern matching with the desired form

$$f_{X|D}(x|d) = e^{\alpha dx-\psi(d)} f_0(x), \tag{36}$$

it is quickly verified that $\alpha = 1/\sigma^2$, and

$$f_0(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, \tag{37}$$

and thus,

$$f_0'(x) = -\frac{1}{\sqrt{2\pi}\sigma}\frac{2x}{2\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, \tag{38}$$

giving the claimed result.

Putting things together, we have

$$\mathbb{E}\left[D|\, X = x\right] = \frac{1}{\alpha}\frac{d}{dx}\lambda(x) = \sigma^2\left(\frac{d}{dx}\log f_X(x) - \frac{d}{dx}\log f_0(x)\right) \tag{39}$$

$$= \sigma^2\left(\frac{d}{dx}\log f_X(x) - \frac{f_0'(x)}{f_0(x)}\right) \tag{40}$$

$$= \sigma^2\left(\frac{d}{dx}\log f_X(x) + \frac{x}{\sigma^2}\right) \tag{41}$$

$$= x + \sigma^2\frac{d}{dx}\log f_X(x), \tag{42}$$

which is the claimed formula.

Problem 4: FIR Wiener Filter

Consider a (discrete-time) signal that satisfies the difference equation $d[n] = 0.5d[n-1]+v[n]$, where $v[n]$ is a sequence of uncorrelated zero-mean unit-variance random variables. We observe $x[n] = d[n] + w[n]$, where $w[n]$ is a sequence of uncorrelated zero-mean random variables with variance 0.5.

(a) (you may skip this at first and do it later — it is conceptually straightforward) Show that for this signal model, the autocorrelation function of the signal $d[n]$ is

$$\mathbb{E}[d[n]d[n+k]] \quad = \quad \frac{4}{3}\left(\frac{1}{2}\right)^{|k|}, \tag{43}$$

and thus the autocorrelation function of the signal $x[n]$ is

$$\mathbb{E}[x[n]x[n+k]] = \begin{cases} \frac{11}{6}, & \text{for } k=0, \\ \frac{4}{3}\left(\frac{1}{2}\right)^{|k|}, & \text{otherwise.} \end{cases} \tag{44}$$

(b) We would like to find an (approximate) linear predictor $\hat{d}[n+3]$ using only the observations $x[n], x[n-1], x[n-2], \dots, x[n-p]$. Using the Wiener Filter framework, determine the optimal coefficients for the linear predictor. Find the corresponding mean-squared error for your predictor.

(c) We would like to find a linear denoiser $\hat{d}[n]$ using *all* of the samples $\{x[k]\}_{k=-\infty}^{\infty}$. Find the filter coefficients and give a formula for the incurred mean-squared error.

Solution

(a) Let us start by writing out the recursion (where we use the general α, and we can plug in $\alpha = 1/2$ later on):

$$\begin{aligned} d[n] &= \alpha d[n-1] + v[n] && (45)\\ &= \alpha(\alpha d[n-2] + v[n-1]) + v[n] && (46)\\ &= \dots \\ &= \alpha^k d[n-k] + \sum_{i=0}^{k-1} \alpha^i v[n-i] && (47) \end{aligned}$$

In particular, letting k tend to infinity, we thus observe (for $|\alpha| < 1$)

$$d[n] = \sum_{i=0}^{\infty} \alpha^i v[n-i]. \tag{48}$$

Hence, we find, for positive values of k,

$$\begin{aligned} \mathbb{E}[d[n]d[n-k]] &= \mathbb{E}\left[\left(\alpha^k d[n-k] + \sum_{i=0}^{k-1} \alpha^i v[n-i]\right) d[n-k]\right] && (49)\\ &= \alpha^k \mathbb{E}\left[d^2[n-k]\right] + \sum_{i=0}^{k-1} \alpha^i \mathbb{E}\left[v[n-i]d[n-k]\right] && (50)\\ &= \alpha^k \mathbb{E}\left[d^2[n-k]\right], && (51) \end{aligned}$$

because by assumption, the samples $v[i]$ are uncorrelated, hence $\mathbb{E}[v[m]d[n]] = 0$ whenever $m > n$. Moreover,

$$\begin{aligned} \mathbb{E}\left[d^2[n-k]\right] &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \alpha^i v[n-k-i]\right)^2\right] && (52)\\ &= \sum_{i=0}^{\infty} \alpha^{2i} \mathbb{E}\left[(v[n-k-i])^2\right] + \sum_{i=0}^{\infty} \sum_{j=0, j\neq i}^{\infty} \alpha^{i+j} \mathbb{E}\left[v[n-k-i]v[n-k-j]\right] && (53)\\ &= \sum_{i=0}^{\infty} \alpha^{2i} \mathbb{E}\left[(v[n-k-i])^2\right], && (54) \end{aligned}$$

since, by assumption, $\mathbb{E}[v[n-i]v[n-j]] = 0$ for $i \neq j$. Moreover, also by assumption, $\mathbb{E}\left[(v[n])^2\right] = 1$, for all n, hence, using the formula for the geometric series,

$$\mathbb{E}\left[d^2[n-k]\right] = \frac{1}{1-\alpha^2}. \tag{55}$$

Combining, we thus find, for positive values of k,

$$R_d[k] \quad = \tfrac{1}{1-\alpha^2}\alpha^k. \tag{56}$$

Finally, since $R_d[-k] = R_d[k]$, we conclude

$$R_d[k] \quad = \tfrac{1}{1-\alpha^2}\alpha^{|k|}. \tag{57}$$

Plugging in $\alpha = 1/2$, we thus find

$$R_d[k] \quad = \tfrac{4}{3}\left(\tfrac{1}{2}\right)^{|k|}. \tag{58}$$

Similarly, for $x[n]$ satisfying difference equation $x[n] = d[n] + w[n]$:

$$\begin{aligned} \mathbb{E}[x[n]x[n-k]] &= \mathbb{E}[(d[n]+w[n])(d[n-k]+w[n-k])] && (59)\\ &= \underbrace{\mathbb{E}[d[n]d[n-k]]}_{=R_d[k]} + \underbrace{\mathbb{E}[d[n]w[n-k]]}_{=0} + \underbrace{\mathbb{E}[d[n-k]w[n]]}_{=0} + \underbrace{\mathbb{E}[w[n]w[n-k]]}_{=0.5\delta[k]}, && (60) \end{aligned}$$

where $\mathbb{E}[d[n]w[n-k]] = 0$ and $\mathbb{E}[d[n-k]w[n]] = 0$ since $d[n]$ and $w[n]$ are uncorrelated by assumption, and since $\mathbb{E}[w[n]] = 0$. We have $\mathbb{E}[w[n]w[n-k]] = 0.5\delta[k]$ because the samples of $w[n]$ are uncorrelated and of variance 0.5. Hence,

$$R_x[k] \quad = \begin{cases} \tfrac{4}{3} + \tfrac{1}{2} = \tfrac{11}{6}, & k = 0, \\ \tfrac{4}{3}\left(\tfrac{1}{2}\right)^{|k|}, & \text{otherwise.} \end{cases} \tag{61}$$

(b) From Part *(a),* we already have the covariance matrix of the observations. What we still need is the correlation between the observations and the desired signal, for $\ell \geq 0$,

$$\begin{aligned} \mathbb{E}[d[n+3]x[n-\ell]] &= \mathbb{E}[d[n+3](d[n-\ell]+w[n-\ell])] \\ &= \mathbb{E}[d[n+3]d[n-\ell]] = \mathbb{E}[d[n]d[n-\ell-3]] = \frac{4}{3}\left(\frac{1}{2}\right)^{\ell+3} = \frac{1}{3}\left(\frac{1}{2}\right)^{\ell+1}. && (62) \end{aligned}$$

As in class, the key ingredients are the covariance matrix of the observations, which is found to be

$$R_x \quad = \begin{bmatrix} \tfrac{11}{6} & \tfrac{2}{3} & \cdots & \tfrac{4}{3}\left(\tfrac{1}{2}\right)^p \\ \tfrac{2}{3} & \tfrac{11}{6} & \cdots & \tfrac{4}{3}\left(\tfrac{1}{2}\right)^{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ \tfrac{4}{3}\left(\tfrac{1}{2}\right)^p & \tfrac{4}{3}\left(\tfrac{1}{2}\right)^{p-1} & \cdots & \tfrac{11}{6} \end{bmatrix} \tag{63}$$

and the correlation between the desired signal and the observations, which is found to be

$$\mathbf{r}_{dx} \quad = \begin{bmatrix} \tfrac{1}{6} & \tfrac{1}{12} & \tfrac{1}{24} & \cdots & \tfrac{1}{3}\left(\tfrac{1}{2}\right)^{p+1} \end{bmatrix}^{\mathsf{T}} \tag{64}$$

Thus, the optimal coefficient should be

$$\mathbf{w} = \begin{bmatrix} w_0 & w_1 & \ldots & w_p \end{bmatrix}^{\mathsf{T}} = R_x^{-1}\mathbf{r}_{dx} \tag{65}$$

As we have seen in class, the estimation error is simply given by

$$\begin{aligned} \mathcal{E} &= \mathbb{E}[d^2[n+3]] - \mathbf{r}_{dx}^H R_x \mathbf{r}_{dx} \\ &= \frac{4}{3} - \mathbf{r}_{dx}^H R_x^{-1}\mathbf{r}_{dx} && (66) \\ & && (67) \end{aligned}$$

(c) The correlation between the observations and the desired signal, for all ℓ,

$$\begin{aligned}
\mathbb{E}[d[n]x[n-\ell]] &= \mathbb{E}[d[n](d[n-\ell]+w[n-\ell])] \\
&= \mathbb{E}[d[n]d[n-\ell]] \qquad (68) \\
&= \frac{4}{3}\left(\frac{1}{2}\right)^{|\ell|}. \qquad (69)
\end{aligned}$$

The covariance matrix of $x[n]$ becomes matrix with infinite dimensions with entry

$$R_x = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \frac{11}{6} & \frac{2}{3} & \frac{1}{3} & \frac{1}{6} & \cdots \\ \cdots & \frac{2}{3} & \frac{11}{6} & \frac{2}{3} & \frac{1}{3} & \cdots \\ \cdots & \frac{1}{3} & \frac{2}{3} & \frac{11}{6} & \frac{2}{3} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{70}$$

and the correlation between the desired signal and the observations, which is found to be

$$\mathbf{r}_{dx} = \begin{bmatrix} \cdots & \frac{1}{3} & \frac{2}{3} & \frac{4}{3} & \frac{2}{3} & \frac{1}{3} & \cdots \end{bmatrix}^{\mathsf{T}} \tag{71}$$

Thus, the optimal coefficient should be

$$\mathbf{w} = R_x^{-1}\mathbf{r}_{dx} \tag{72}$$

and the estimation error is simply given by

$$\begin{aligned}
\mathcal{E} &= \mathbb{E}[d^2[n]] - \mathbf{r}_{dx}^H R_x \mathbf{r}_{dx} \\
&= \frac{4}{3} - \mathbf{r}_{dx}^H R_x^{-1}\mathbf{r}_{dx} \qquad (73) \\
& \qquad (74)
\end{aligned}$$

Problem 5: Bounding The Exploration Bias

(a) Let $X_1, X_2, \ldots, X_n \sim$ i.i.d. $\mathcal{N}(0,1)$. Let $Y = \text{argmax}_i X_i$ and $T \in \{1,2,\ldots,n\}$ is such that $P_{T|Y}(t|y) = \begin{cases} p, & t=y \\ \frac{1-p}{n-1}, & t \neq y \end{cases}$ for some $p \in [0,1]$.

1. Compute $I(X;T)$ where $X = (X_1, X_2, \ldots, X_n)$. (Hint: write $I(X;T) = H(T) - H(T|X)$. What is the marginal distribution of T?)

(b) Let $X_1, \ldots, X_4 \sim$ i.i.d. $\mathcal{N}(0,1)$ and $X_5 \sim \mathcal{N}(0,4)$. Let Y and T be as in part (a) with $p = 0.3$.

1. Show that $\mathbf{Pr}(Y=5) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{8\pi}}(1-Q(x))^4 e^{-x^2/8}dx$ (where $Q(x) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}du$), and find a corresponding numerical approximation (using Mathematica, for example).
2. Using the previous numerical approximation, find the marginal distributions P_Y and P_T.

Solution

(a1) Since $X_1, \ldots, X_n$ are i.i.d and $Y = \arg\max_i X_i$, we have $P_Y(Y=i) = \frac{1}{n}$, for all $i \in [n]$. The

marginal distribution of T is

$$
\begin{aligned}
P_T(T=t) &= \sum_{i=1}^{n} P_{T|Y}(T=t|Y=i)P_Y(Y=i) && (75)\\
&= P_{T|Y}(T=t|Y=t)P_Y(Y=t) + \sum_{i=1, i\neq t}^{n} P_{T|Y}(T=t|Y=i)P_Y(Y=i) && (76)\\
&= p\frac{1}{n} + (n-1)\frac{1-p}{n-1}\frac{1}{n} && (77)\\
&= \frac{1}{n} && (78)
\end{aligned}
$$

Hence, T is uniformly distributed over $\{1,\dots,n\}$,

$$
H(T) = -\sum_{t=1}^{n} P_T(T=t)\log P_T(T=t) = \log n \tag{79}
$$

Additionally, given $X_1,\dots,X_n$, Y is fixed, which means $H(Y|X)=0$. And given Y, X and T are independent.

$$
H(T|X) = H(T,Y|X) = H(Y|X) + H(T|X,Y) = H(T|Y) = -p\log p - (n-1)\frac{1-p}{n-1}\log\frac{1-p}{n-1} \tag{80}
$$

Therefore,

$$
I(X;T) = H(T) - H(T|X) = \log n + p\log p + (1-p)\log\frac{1-p}{n-1} \tag{81}
$$

(a2)

$$
|\mathbb{E}[X_T]| \leq \sqrt{2I(X;T)} \tag{82}
$$

(b1) Since $Y=5$ means X_5 is the largest one and $X_1,\dots,X_4$ are i.i.d.

$$
\begin{aligned}
\mathbf{Pr}(Y=5) &= \mathbf{Pr}(X_5 > X_1, X_5 > X_2, X_5 > X_3, X_5 > X_4) && (83)\\
&= \mathbb{E}_{X_5}[\mathbf{Pr}(X_5 > X_1, X_5 > X_2, X_5 > X_3, X_5 > X_4)|X_5] && (84)\\
&= \mathbb{E}_{X_5}[\mathbf{Pr}(x_5 \geq X_1|X_5=x_5)\mathbf{Pr}(x_5 \geq X_2|X_5=x_5)\mathbf{Pr}(x_5 \geq X_3|X_5=x_5)\mathbf{Pr}(x_5 \geq X_4|X_5=x_5)] && (85)\\
&= \mathbb{E}_{X_5}[\mathbf{Pr}(x_5 \geq X_1|X_5=x_5)^4] \quad [X_1,X_2,X_3,X_4 \text{ are i.i.d}] && (86)\\
&= \int_{-\infty}^{\infty} \mathbf{Pr}(X_1 \leq x)^4 \mathbf{Pr}(X_5 = x)dx && (87)\\
&= \int_{-\infty}^{\infty} (1-Q(x))^4 \frac{1}{\sqrt{8\pi}} e^{-\frac{x^2}{8}} dx && (88)\\
&\simeq 0.31 && (89)
\end{aligned}
$$

Thus, $\forall i \in \{1,2,3,4\}$, we have $\mathbf{Pr}(Y=i) = \frac{1-\mathbf{Pr}(Y=5)}{4} \simeq 0.1725$.

$$
\begin{aligned}
\mathbf{Pr}(T=5) &= \mathbf{Pr}(T=5, Y=5) + \mathbf{Pr}(T=5, Y\neq 5) && (90)\\
&= \mathbf{Pr}(T=5|Y=5)\mathbf{Pr}(Y=5) + \sum_{i\neq 5}\mathbf{Pr}(T=5|Y=i)\mathbf{Pr}(Y=i) && (91)\\
&\simeq 0.3 \times 0.31 + 4 \times \frac{1-0.3}{4} \times 0.1725 = 0.2137 && (92)
\end{aligned}
$$

Problem 6: Gibbs Algorithm

Let $\mathcal{X}$ be the sample space, $\mathcal{W}$ the hypothesis space, and let $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_+$ be a corresponding loss function. On a dataset $D = (X_1, X_2, \ldots, X_n)$, the empirical risk for a hypothesis w is given by $L_D(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, X_i)$. We saw in class that $I(D; W)$ can be used to bound the generalization error. Hence, we can use it as a *regularizer* in empirical risk minimization.

(a) First, show that given any joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$ and marginal distribution Q on $\mathcal{Y}$, $D(P_{XY}||P_X P_Y) \leq D(P_{XY}||P_X Q)$.

Since we cannot directly compute $D(P_{DW}||P_D P_W)$, we will use $D(P_{DW}||P_D Q)$ as a proxy, where Q is a distribution on $\mathcal{W}$.

(b) Let

$$P^\star_{W|D} = \underset{P_{W|D}}{\operatorname{argmin}} \left(\mathbb{E}[L_D(W)] + \frac{1}{\beta} D(P_{DW}||P_D Q) \right).$$

1. Show that

$$\min_{P_{W|D}} \left(\mathbb{E}[L_D(W)] + \frac{1}{\beta} D(P_{DW}||P_D Q) \right) = \mathbb{E}_D \left[\min_{P_{W|D=d}} \left(\mathbb{E}[L_d(W)] + \frac{1}{\beta} D(P_{W|D=d}||Q) \right) \right].$$

2. Show that the minimizer on the right-hand side $P^\star_{W|D=d}$ is given by

$$P^\star_{W|D=d} = \frac{e^{-\beta L_d(w)} Q(w)}{\mathbb{E}_Q\left[e^{-\beta L_d(W)}\right]}.$$

This is known in the literature as the Gibbs algorithm. (Hint: Write $\mathbb{E}[\beta L_d(W)] = \mathbb{E}[\log e^{\beta L_d(W)}]$, combine with the KL divergence term and use non-negativity of KL divergence.)

3. Show that $P^\star_{W|D=d}$ is $2\beta/n$-differential private if $\ell \in [0, 1]$.

Solution

(a) For any marginal distribution Q on $\mathcal{Y}$,

$$\begin{aligned}
D(P_{XY}||P_X P_Y) - D(P_{XY}||P_X Q) &= \sum_{x,y} P_{XY}(x,y) \left(\log \frac{P_{XY}(x,y)}{P_X(x) P_Y(y)} - \log \frac{P_{XY}(x,y)}{P_X(x) Q(y)} \right) && (93)\\
&= \sum_{x,y} P_{XY}(x,y) \log \frac{Q(y)}{P_Y(y)} && (94)\\
&= \sum_{y} P_Y(y) \log \frac{Q(y)}{P_Y(y)} && (95)\\
&\overset{(*)}{\leq} \log \sum_{y} P_Y(y) \frac{Q(y)}{P_Y(y)} && (96)\\
&= \log \sum_{y} Q(y) = 0 && (97)
\end{aligned}$$

where $(*)$ is because $\log(x)$ is a concave function of x.

(b1)

$$\min_{P_{W|D}} \left(\mathbb{E}[L_D(W)] + \frac{1}{\beta} D(P_{DW}||P_D Q) \right) \tag{98}$$
$$= \min_{P_{W|D}} \left(\mathbb{E}_D[\mathbb{E}[L_D(W)|D=d]] + \frac{1}{\beta} \sum_{w,d} P_{W|D}(w|d) P_D(d) \log \frac{P_{W|D}(w|d)P_D(d)}{P_D(d)Q} \right) \tag{99}$$
$$= \min_{P_{W|D}} \left(\mathbb{E}_D[\mathbb{E}[L_D(W)|D=d]] + \frac{1}{\beta} \sum_{w,d} P_{W|D}(w|d) P_D(d) \log \frac{P_{W|D}(w|d)}{Q} \right) \tag{100}$$
$$= \min_{P_{W|D}} \left(\mathbb{E}_D[\mathbb{E}[L_D(W)|D=d]] + \mathbb{E}_D[\frac{1}{\beta} D(P_{W|D}||Q)|D=d] \right) \tag{101}$$
$$= \mathbb{E}_D \left[\min_{P_{W|D=d}} \left(\mathbb{E}[L_d(W)] + \frac{1}{\beta} D(P_{W|D=d}||Q) \right) \right] \tag{102}$$

(b2) Given $D = d$, we know that $P_W(w) = \sum_{d'} P_{W|D}(w|d')P_D(d') = P_{W|D}(w|d)$.

$$\arg \min_{P_{W|D=d}} \left(\mathbb{E}[L_d(W)] + \frac{1}{\beta} D(P_{W|D=d}||Q) \right) \tag{103}$$
$$= \arg \min_{P_{W|D=d}} \left(\mathbb{E}[\beta L_d(W)] + D(P_{W|D=d}||Q) \right) \tag{104}$$
$$= \arg \min_{P_{W|D=d}} \left(\mathbb{E}[\log e^{\beta L_d(W)}] + D(P_{W|D=d}||Q) \right) \tag{105}$$
$$= \arg \min_{P_{W|D=d}} \left(\sum_w \log e^{\beta L_d(w)} P_W(w) + \sum_w P_{W|D}(w|d) \log \frac{P_{W|D}(w|d)}{Q(w)} \right) \tag{106}$$
$$= \arg \min_{P_{W|D=d}} \left(\sum_w \log e^{\beta L_d(w)} P_{W|D}(w|d) + \sum_w P_{W|D}(w|d) \log \frac{P_{W|D}(w|d)}{Q(w)} \right) \tag{107}$$
$$= \arg \min_{P_{W|D=d}} \left(\sum_w P_{W|D}(w|d) (\log e^{\beta L_d(w)} + \log \frac{P_{W|D}(w|d)}{Q(w)}) \right) \tag{108}$$
$$= \arg \min_{P_{W|D=d}} \left(\sum_w P_{W|D}(w|d) \log \frac{P_{W|D}(w|d)}{Q(w)e^{-\beta L_d(w)}} \right) \tag{109}$$
$$= \arg \min_{P_{W|D=d}} \left(\sum_w P_{W|D}(w|d) \log \frac{P_{W|D}(w|d)}{Q(w)e^{-\beta L_d(w)}} \frac{\mathbb{E}_Q[e^{-\beta L_d(W)}]}{\mathbb{E}_Q[e^{-\beta L_d(W)}]} \right) \tag{110}$$
$$= \arg \min_{P_{W|D=d}} D\left(P_{W|D} || \frac{Q(w)e^{-\beta L_d(w)}}{\mathbb{E}_Q[e^{-\beta L_d(W)}]} \right) - \log \mathbb{E}_Q[e^{-\beta L_d(W)}] \tag{111}$$
$$= \arg \min_{P_{W|D=d}} D\left(P_{W|D} || \frac{Q(w)e^{-\beta L_d(w)}}{\mathbb{E}_Q[e^{-\beta L_d(W)}]} \right) \tag{112}$$
$$= \frac{Q(w)e^{-\beta L_d(w)}}{\mathbb{E}_Q[e^{-\beta L_d(W)}]} \tag{113}$$

The reason why we added $\mathbb{E}_Q[e^{-\beta L_d(W)}]$ as a normalization term is $P_{W|D}$ has to be a valid pmf, i.e. $\sum_w P_{W|D}(w|d) = 1$. However, the scaled version of Q, $Q(w)e^{-\beta L_d(w)}$, may not be a valid pmf.

(b3) Suppose d and d' differ at j-th entry only. Hence,

$$e^{-\beta L_d(w)} e^{\beta L_{d'}(w)} = e^{-\frac{\beta}{n}(e(w_j,X_j)-e(w_j,X_j'))} \leq e^{\beta/n} \tag{114}$$

Similarly,

$$\frac{\mathbb{E}_Q[e^{-\beta L_d(w)}]}{\mathbb{E}_Q[e^{-\beta L_{d'}(w)}]} \leq \frac{\mathbb{E}_Q[e^{\frac{\beta}{n}} e^{-\beta L_{d'}(w)}]}{\mathbb{E}_Q[e^{-\beta L_{d'}(w)}]} \leq e^{\frac{\beta}{n}} \tag{115}$$

Thus, we have $\frac{P^*_{W|D=b}}{P^*_{W|D=d'}} \leq e^{2\beta/n}$ and $P^*_{W|D=d}$ is $2\beta/n-$ differential private if $l \in [0,1]$.