Problem Set 6 — *Due Friday, December 10, before class starts*

For the Exercise Sessions on Nov 27 and Dec 3

Last name	First name	SCIPER Nr	Points

Problem 1: 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

$$P_T(T = t) = \sum_{i=1}^{n} P_{T|Y}(T = t|Y = i) P_Y(Y = i) \tag{1}$$

$$= P_{T|Y}(T = t|Y = t) P_Y(Y = t) + \sum_{i=1, i \neq t}^{n} P_{T|Y}(T = t|Y = i) P_Y(Y = i) \tag{2}$$

$$= p\frac{1}{n} + (n-1)\frac{1-p}{n-1}\frac{1}{n} \tag{3}$$

$$= \frac{1}{n} \tag{4}$$

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

$$H(T) = -\sum_{t=1}^{n} P_T(T = t) \log P_T(T = t) = \log n \tag{5}$$

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

$$H(T|X) = H(T,Y|X) = H(Y|X) + H(T|X,Y) = H(T|Y) = -p\log p - (n-1)\frac{1-p}{n-1}\log\frac{1-p}{n-1} \tag{6}$$

Therefore,

$$I(X;T) = H(T) - H(T|X) = \log n + p\log p + (1-p)\log\frac{1-p}{n-1} \tag{7}$$

(a2)

$$|\mathbb{E}[X_T]| \leq \sqrt{2I(X;T)} \tag{8}$$

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

$$\begin{aligned}
\mathbf{Pr}(Y=5) &= \mathbf{Pr}(X_5 > X_1, X_5 > X_2, X_5 > X_3, X_5 > X_4) && (9)\\
&= \mathbb{E}_{X_5}[\mathbf{Pr}(X_5 > X_1, X_5 > X_2, X_5 > X_3, X_5 > X_4)|X_5] && (10)\\
&= \mathbb{E}_{X_5}[\mathbf{Pr}(x_5 \geq X_1|X_5 = x_5)\mathbf{Pr}(x_5 \geq X_2|X_5 = x_5)\mathbf{Pr}(x_5 \geq X_3|X_5 = x_5)\mathbf{Pr}(x_5 \geq X_4|X_5 = x_5)] && (11)\\
&= \mathbb{E}_{X_5}[\mathbf{Pr}(x_5 \geq X_1|X_5 = x_5)^4] \quad [X_1, X_2, X_3, X_4 \text{ are i.i.d}] && (12)\\
&= \int_{-\infty}^{\infty} \mathbf{Pr}(X_1 \leq x)^4 \mathbf{Pr}(X_5 = x)dx && (13)\\
&= \int_{-\infty}^{\infty} (1-Q(x))^4 \frac{1}{\sqrt{8\pi}} e^{-\frac{x^2}{8}} dx && (14)\\
&\simeq 0.31 && (15)
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{Pr}(T=5) &= \mathbf{Pr}(T=5, Y=5) + \mathbf{Pr}(T=5, Y \neq 5) && (16)\\
&= \mathbf{Pr}(T=5|Y=5)\mathbf{Pr}(Y=5) + \sum_{i\neq 5}\mathbf{Pr}(T=5|Y=i)\mathbf{Pr}(Y=i) && (17)\\
&\simeq 0.3 \times 0.31 + 4 \times \frac{1-0.3}{4} \times 0.1725 = 0.2137 && (18)
\end{aligned}$$

Problem 2: Dependence and large error events

In the lecture notes we have seen how to bound the expected generalization error using information measures. With this exercise we will work on large error events and provide bounds on the probabilities of such events. The setting is the same: we observe n iid samples $D = (X_1, \ldots, X_n)$ (according to some unknown distribution P) and based on this observation we will choose a hypothesis $w \in W$. We also consider the usual definition of empirical and population risk, *i.e.* given a loss function ℓ, some hypothesis w, $L_D(w) = \frac{1}{n}\sum_{i=1}^n \ell(w, X_i)$, and $L_P(w) = \mathbb{E}_P[\ell(w,X)]$. We are interested in controlling the following quantity:

$$\mathbb{P}\left(|L_P(W) - L_D(W)| > \epsilon\right). \tag{19}$$

(a) Suppose that the loss is such that $\ell(w,x) \in \{0,1\}$ for every $w \in W$ and $x \in \mathcal{X}$. Suppose also that $|\mathcal{W}| < \infty$, *i.e.*, the number of hypotheses is finite.

1. Show that for every **fixed** $w \in W$ $\mathbb{P}\left(|L_P(w) - L_D(w)| > \epsilon\right) \leq 2\exp(-2n\epsilon^2)$;

2. Show that
$$\mathbb{P}\left(|L_P(W) - L_D(W)| > \epsilon\right) \leq |W| \cdot 2\exp(-2n\epsilon^2); \tag{20}$$
Hint: denote with $E = \{(d, w) : |L_P(w) - L_d(w)| > \epsilon\}$.
You have that $\mathbb{P}\left(|L_P(W) - L_D(W)| > \epsilon\right) = \mathbb{P}(E) = \sum_{(w,d)\in E} P(w, d)$.
(be careful: $\mathbb{P}\left(|L_P(W) - L_D(W)| > \epsilon | W = w\right)$ is not necessarily $\leq 2\exp(-2n\epsilon^2).Why?$)

(b) Now consider the following information measure, given two discrete random variables X, Y:
$$\mathcal{L}(X \to Y) = \log \sum_y \max_{x:P_X(x)>0} P_{Y|X}(y|x). \tag{21}$$
This quantity is known in the literature as Maximal Leakage and quantifies the leakage of information between X and Y.

1. Show that if the alphabet of Y (denoted with $\mathcal{Y}$) is finite then
$$\mathcal{L}(X \to Y) \leq \log |\mathcal{Y}|,$$
which distributions achieve the bound with equality?
2. It is possible to show that
$$\mathcal{L}(X \to Y) \geq 0,$$
which distributions achieve the bound with equality?
3. Let X be a binary random variable and let Y be an observation of X after passing through a Binary Symmetric Channel with parameter $\delta \leq \frac{1}{2}$. More precisely we have $P_{Y|X=x}(x) = 1-\delta$, for $x \in \{0, 1\}$.
What is the maximal leakage $\mathcal{L}(X \to Y)$?
Which values of δ allow you to achieve the bounds in $(1), (2)$ with equality?
4. Suppose further that the space of samples $\mathcal{D}$ is finite. Denote with $E_w = \{d : (d, w) \in E\}$, for $w \in \mathcal{W}$; Show that:
$$\mathbb{P}\left(|L_P(W) - L_D(W)| > \epsilon\right) \leq \exp(\mathcal{L}(D \to W)) \max_{w\in\mathcal{W}} \mathbb{P}(E_w);$$
5. Conclude that
$$\mathbb{P}\left(|L_P(W) - L_D(W)| > \epsilon\right) \leq 2\exp(\mathcal{L}(D \to W) - 2n\epsilon^2);$$
6. Compare the two bounds retrieved in *(a2)* and *(b4)*, what do you notice? Is one of the two better than the other? When are they equal? What conclusions can you draw?

Solution

(a1) For a given $w \in W$, we have, by assumption, that $\ell(w, X_i) - \mathbb{E}[\ell(w, X)]$ is a 0-mean Bernoulli random variable. Hence, $L_D(w) - L_P(w)$ is a $\frac{1}{4}$-sub-Gaussian random variable. By Hoeffding's inequality for σ^2-sub-Gaussian (Lemma 5.5 in the Lecture Notes):
$$\mathbb{P}\left(|L_P(w) - L_D(w)| > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right) = 2\exp(-2n\epsilon^2) \tag{22}$$

(a2) Using the hint, we have that:

$$
\begin{aligned}
\mathbb{P}((|L_P(W) - L_D(W)| > \epsilon) &= \mathbb{P}(E) && (23)\\
&= \sum_{(w,d)\in E} P_{WD}(w,d) && (24)\\
&= \sum_{w\in W} \sum_{d\in E_w} P_{W|D=d}(w) P_D(d) && (25)\\
&\leq \sum_{w\in W} \sum_{d\in E_w} P_D(d) && (26)\\
&= \sum_{w\in W} P_D(E_w) && (27)\\
&\leq \sum_{w\in W} 2\exp(-2n\epsilon^2) && (28)\\
&= 2|W|\exp(-2n\epsilon^2). && (29)
\end{aligned}
$$

Where we denoted with $E_w = \{d : (d,w) \in E)\}$. The important thing to notice here is that splitting the summation in this way, and considering E_w for a given $w \in W$ is, in a way, equivalent to fixing the hypothesis w, just like we assumed in *(a1)*. This, along with upper-bounding $P_{W|D=d}(w)$ by 1, allows us to 'ignore' the dependence between W and D. Better bounds can be obtained, as we will soon see, by actually expoliting the dependence and not trivially upper-bounding the conditional probabilities.

(b1) Starting from the expression:

$$
\begin{aligned}
\mathcal{L}(X \to Y) &= \log \sum_y \max_{x:P_X(x)>0} P_{Y|X}(y|x) && (30)\\
&\leq \log \sum_y 1 = \log |\mathcal{Y}|. && (31)
\end{aligned}
$$

Fixed a distribution P_X a set of distributions $P_{Y|X}$ that achieves the bound with equality is the one induced by a deterministic mapping (*i.e.*, $Y = f(X)$, and f is deterministic).

(b2) Fixed a distribution P_X, if Y is independent from X we have that $P_{Y|X} = P_Y$ and:

$$
\begin{aligned}
\mathcal{L}(X \to Y) &= \log \sum_y \max_{x:P_X(x)>0} P_{Y|X}(y|x) && (32)\\
&= \log \sum_y P_Y(y) = \log 1 = 0. && (33)
\end{aligned}
$$

(b3) The Maximal Leakage in this case is $\mathcal{L}(X \to Y) = \log(2(1-\delta))$.

With $\delta = 0$, we have a determinstic channel, as $P_{Y|X=0}(0) = 1$ and $P_{Y|X=0}(1) = 0$ and we have that $\mathcal{L}(X \to Y) = \log(2) = \log(|\mathcal{Y}|)$, recovering *(b1)* with equality.

With $\delta = 1/2$, we have that Y is independent of X as $P_{Y|X=0}(0) = P_{Y|X=1}(0) = 1/2$ and we have that $\mathcal{L}(X \to Y) = \log(1) = 0$, recovering *(b2)* with equality.

(b4)

$$\begin{aligned}
\mathbb{P}((|L_P(W) - L_D(W)| > \epsilon) &= \mathbb{P}(E) && (34)\\
&= \sum_{(w,d)\in E} P_{WD}(w,d) && (35)\\
&= \sum_{w\in W}\sum_{d\in E_w} P_{W|D=d}(w)P_D(d) && (36)\\
&\leq \sum_{w\in W} \max_{d:P_D(d)>0} P_{W|D=d}(w) \sum_{d\in E_w} P_D(d) && (37)\\
&= \sum_{w\in W} \max_{d:P_D(d)>0} P_{W|D=d}(w)P_D(E_w) && (38)\\
&\leq \max_{w\in W} P_D(E_w) \sum_{w\in W} \max_{d:P_D(d)>0} P_{W|D=d}(w) && (39)\\
&= \max_{w\in W} P_D(E_w) \exp(\mathcal{L}(D\to W)). && (40)
\end{aligned}$$

(b5) As noticed in *(a1)*, for every $w \in W$ we have that $\mathbb{P}\left(|L_P(w) - L_D(w)| > \epsilon\right) \leq 2\exp(-2n\epsilon^2)$. The bound follows from this and *(b2)*.

(b6) Under the assumption that $|W| < \infty$ as seen in *(b1)*, the Maximal Leakage $\mathcal{L}(D \to W) \leq \log|W|$. Thus, the bound in *(b4)* can be tighter that the one in *(a2)*. More precisely: when $W = f(D)$ where f is deterministic, *i.e.* maximum dependence of W on D, we recover the bound in *(a2)*. When W and D are independent, we have that $\mathcal{L}(D \to W) = 0$ and we recover the Hoeffding's bound. Measuring the dependence (or, in other words, the amount of adaptivity) via Maximal Leakage, allows us to go from the classical Hoeffding's bound to the 'worst-case' scenario provided by the union bound in *(a2)*. The first case represents **no** dependence, the second case represents **maximum** dependence, with a whole spectrum of behaviours in the middle. Another interpretation of the bound (specific to the measure) is the following: if the learning algorithm leaks too much information about the training set, then it will overfit. If you limit the amount of leakage, then you generalize (and potentially, with an exponentially decaying bound).

Problem 3: 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

$$p(d) = \mu e^{-\mu d} \tag{41}$$

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

$$p_X(x) = \int_d p(x,d) = \frac{\mu}{(\mu+x)^2} \tag{42}$$

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

$$\hat{d}_{MMSE}(x) = E[d|X = x] = \frac{2}{\mu + x} \tag{43}$$

and the mean-squared error is

$$\begin{aligned}
\mathcal{E} &= E_D[(d - \hat{d}_{MMSE})^2] && (44)\\
&= E_X[E_D[(d - \hat{d}_{MMSE}(x))^2|X = x]] && (45)\\
&= \int E_D[(d - \hat{d}_{MMSE}(x))^2|X = x]p_X(x)dx && (46)\\
&= \int E_D[(d^2 - 2d\hat{d}_{MMSE}(x) + \hat{d}^2_{MMSE}(x)|X = x]p_X(x)dx && (47)\\
&= \int (E_D[d^2|X = x] - \hat{d}^2_{MMSE}(x))p_X(x)dx && (48)\\
&= \int \left(\frac{6}{(\mu + x)^2} - \frac{4}{(\mu + x)^2}\right) p_X(x)dx && (49)\\
&= \int \frac{2\mu}{(\mu + x)^4} dx && (50)\\
&= \frac{2}{3\mu^2} && (51)
\end{aligned}$$

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

$$\begin{aligned}
\hat{d}_{MAP}(x) &= \arg\max_d p(d|x) && (52)\\
&= \arg\max_d (\mu + x)^2 d e^{-(\mu+x)d} && (53)\\
&= \arg\max_d d e^{-(\mu+x)d} && (54)\\
&= \frac{1}{\mu + x} && (55)
\end{aligned}$$

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

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

$$\mathbb{E}[d[n]d[n+k]] = \frac{4}{3}\left(\frac{1}{2}\right)^{|k|}, \tag{56}$$

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

$$\mathbb{E}[x[n]x[n+k]] = \begin{cases} \frac{11}{6}, & \text{for } k = 0,\\ \frac{4}{3}\left(\frac{1}{2}\right)^{|k|}, & \text{otherwise.} \end{cases} \tag{57}$$

(b) We would like to find an (approximate) linear predictor $\hat{d}[n+3]$ using only the observations $x[n], x[n-1], x[n-2], \ldots, x[n-p]$. 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):

$$\begin{aligned} d[n] &= \alpha d[n-1] + v[n] && (58)\\ &= \alpha(\alpha d[n-2] + v[n-1]) + v[n] && (59)\\ &= \ldots \\ &= \alpha^k d[n-k] + \sum_{i=0}^{k-1} \alpha^i v[n-i] && (60) \end{aligned}$$

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

$$d[n] = \sum_{i=0}^{\infty} \alpha^i v[n-i]. \tag{61}$$

Hence, we find, for positive values of k,

$$\begin{aligned} \mathbb{E}[d[n]d[n-k]] &= \mathbb{E}\left[\left(\alpha^k d[n-k] + \sum_{i=0}^{k-1} \alpha^i v[n-i]\right) d[n-k]\right] && (62)\\ &= \alpha^k \mathbb{E}\left[d^2[n-k]\right] + \sum_{i=0}^{k-1} \alpha^i \mathbb{E}\left[v[n-i]d[n-k]\right] && (63)\\ &= \alpha^k \mathbb{E}\left[d^2[n-k]\right], && (64) \end{aligned}$$

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

$$\begin{aligned} \mathbb{E}\left[d^2[n-k]\right] &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \alpha^i v[n-k-i]\right)^2\right] && (65)\\ &= \sum_{i=0}^{\infty} \alpha^{2i} \mathbb{E}\left[(v[n-k-i])^2\right] + \sum_{i=0}^{\infty} \sum_{j=0, j\neq i}^{\infty} \alpha^{i+j} \mathbb{E}\left[v[n-k-i]v[n-k-j]\right] && (66)\\ &= \sum_{i=0}^{\infty} \alpha^{2i} \mathbb{E}\left[(v[n-k-i])^2\right], && (67) \end{aligned}$$

since, by assumption, $\mathbb{E}\left[v[n-i]v[n-j]\right] = 0$ 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,

$$\mathbb{E}\left[d^2[n-k]\right] = \tfrac{1}{1-\alpha^2}. \tag{68}$$

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

$$R_d[k] = \tfrac{1}{1-\alpha^2}\alpha^k. \tag{69}$$

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

$$R_d[k] = \tfrac{1}{1-\alpha^2}\alpha^{|k|}. \tag{70}$$

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

$$R_d[k] \quad = \tfrac{4}{3}\left(\tfrac{1}{2}\right)^{|k|}. \tag{71}$$

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

$$\begin{aligned}
\mathbb{E}[x[n]x[n-k]] &= \mathbb{E}[(d[n]+w[n])(d[n-k]+w[n-k])] && (72)\\
&= \underbrace{\mathbb{E}[d[n]d[n-k]]}_{=R_d[k]} + \underbrace{\mathbb{E}[d[n]w[n-k]]}_{=0} + \underbrace{\mathbb{E}[d[n-k]w[n]]}_{=0} + \underbrace{\mathbb{E}[w[n]w[n-k]]}_{=0.5\delta[k]}, && (73)
\end{aligned}$$

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,

$$R_x[k] \quad = \begin{cases} \frac{4}{3} + \frac{1}{2} = \frac{11}{6}, & k = 0, \\ \frac{4}{3}\left(\frac{1}{2}\right)^{|k|}, & \text{otherwise.} \end{cases} \tag{74}$$

(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 $\ell \geq 0$,

$$\begin{aligned}
\mathbb{E}[d[n+3]x[n-\ell]] &= \mathbb{E}[d[n+3](d[n-\ell]+w[n-\ell])] \\
&= \mathbb{E}[d[n+3]d[n-\ell]] = \mathbb{E}[d[n]d[n-\ell-3]] = \frac{4}{3}\left(\frac{1}{2}\right)^{\ell+3} = \frac{1}{3}\left(\frac{1}{2}\right)^{\ell+1}. && (75)
\end{aligned}$$

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

$$R_x \quad = \begin{bmatrix} \frac{11}{6} & \frac{2}{3} & \cdots & \frac{4}{3}\left(\frac{1}{2}\right)^{p} \\ \frac{2}{3} & \frac{11}{6} & \cdots & \frac{4}{3}\left(\frac{1}{2}\right)^{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{4}{3}\left(\frac{1}{2}\right)^{p} & \frac{4}{3}\left(\frac{1}{2}\right)^{p-1} & \cdots & \frac{11}{6} \end{bmatrix} \tag{76}$$

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

$$\mathbf{r}_{dx} \quad = \begin{bmatrix} \frac{1}{6} & \frac{1}{12} & \frac{1}{24} & \cdots & \frac{1}{3}\left(\frac{1}{2}\right)^{p+1} \end{bmatrix}^{\mathsf{T}} \tag{77}$$

Thus, the optimal coefficient should be

$$\mathbf{w} = \begin{bmatrix} w_0 & w_1 & \ldots & w_p \end{bmatrix}^{\mathsf{T}} = R_x^{-1}\mathbf{r}_{dx} \tag{78}$$

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

$$\begin{aligned}
\mathcal{E} &= \mathbb{E}[d^2[n+3]] - \mathbf{r}_{dx}^H R_x \mathbf{r}_{dx} \\
&= \frac{4}{3} - \mathbf{r}_{dx}^H R_x^{-1} \mathbf{r}_{dx} && (79)\\
& && (80)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[d[n]x[n-\ell]] &= \mathbb{E}[d[n](d[n-\ell]+w[n-\ell])] \\
&= \mathbb{E}[d[n]d[n-\ell]] && (81)\\
&= \frac{4}{3}\left(\frac{1}{2}\right)^{|\ell|}. && (82)
\end{aligned}$$

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

$$R_x = \begin{bmatrix} \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \frac{11}{6} & \frac{2}{3} & \frac{1}{3} & \frac{1}{6} & \cdots \\ \cdots & \frac{2}{3} & \frac{11}{6} & \frac{2}{3} & \frac{1}{3} & \cdots \\ \cdots & \frac{1}{3} & \frac{2}{3} & \frac{11}{6} & \frac{2}{3} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{83}$$

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

$$\mathbf{r}_{dx} = \begin{bmatrix} \cdots & \frac{1}{3} & \frac{2}{3} & \frac{4}{3} & \frac{2}{3} & \frac{1}{3} & \cdots \end{bmatrix}^{\mathsf{T}} \tag{84}$$

Thus, the optimal coefficient should be

$$\mathbf{w} = R_x^{-1}\mathbf{r}_{dx} \tag{85}$$

and the estimation error is simply given by

$$\begin{aligned} \mathcal{E} &= \mathbb{E}[d^2[n]] - \mathbf{r}_{dx}^H R_x \mathbf{r}_{dx} \\ &= \frac{4}{3} - \mathbf{r}_{dx}^H R_x^{-1} \mathbf{r}_{dx} \end{aligned} \tag{86}$$

$$\tag{87}$$