

Homework 4

Hanmo Chen

November 19, 2020

-
- **Acknowledgments:** For Problem 1, I refer to https://en.wikipedia.org/wiki/Incomplete_gamma_function for Incomplete Gamma function None
 - **Collaborators:** I finish this homework by myself.
 - *I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.*

Hanmo Chen

4.1. (a) Because $x_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $y = \sum_{i=1}^n \frac{x_i^2}{\sigma^2} \sim \chi_n^2$.

$\mathbb{P}(Y \geq n\alpha^2/\sigma^2) = \frac{\Gamma(\frac{n}{2}, \frac{n\alpha^2}{2\sigma^2})}{\Gamma(\frac{n}{2})}$ where $\Gamma(s, x)$ denotes the upper Incomplete Gamma function.

So

$$\begin{aligned} -\frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n x_i^2 \geq \alpha^2\right) &= -\frac{1}{n} \log \mathbb{P}\left(Y \geq \frac{n\alpha^2}{\sigma^2}\right) \\ &= -\frac{1}{n} \log \frac{\Gamma(\frac{n}{2}, \frac{n\alpha^2}{2\sigma^2})}{\Gamma(\frac{n}{2})} \end{aligned} \quad (1)$$

To find the asymptotic property, using Sanov's theorem,

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n x_i^2 \geq \alpha^2\right) = \inf_{\mathbb{E}_P[X^2] \geq \alpha^2} D(P \parallel \mathcal{N}(0, \sigma^2)) \quad (2)$$

Suppose the distribution P has pdf $f(x)$, it can be seen as an optimization problem with constraints, that is,

$$\begin{aligned} \min \quad & D(P \parallel \mathcal{N}(0, \sigma^2)) = \int f(x) \log \frac{f(x)}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}} \\ \text{s.t.} \quad & \int f(x) x^2 \geq \alpha^2 \\ & \int f(x) = 1 \end{aligned} \quad (3)$$

Define

$$J(f) = \int f(x) \log \frac{f(x)}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}} + \lambda \left(\int f(x) x^2 - \alpha^2 \right) + \mu \left(\int f(x) - 1 \right) \quad (4)$$

And let $\frac{\partial J}{\partial f} = 0$ we have

$$\frac{\partial J}{\partial f} = \log f(x) + \lambda x^2 + \mu = 0 \quad (5)$$

So $f(x) = \exp^{-\mu - \lambda x^2}$, which is normal distribution and satisfies $\mathbb{E}[X^2] \geq \alpha^2$. So $P^* = \mathcal{N}(0, \alpha^2)$ and

$$\begin{aligned} \lim_{n \rightarrow \infty} -\frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n x_i^2 \geq \alpha^2\right) &= D(P^* \| \mathcal{N}(0, \sigma^2)) \\ &= \int_{\mathbb{R}} f(x) \log \frac{\frac{1}{\sqrt{2\pi\alpha^2}} e^{-\frac{x^2}{2\alpha^2}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}} dx \quad (6) \\ &= \ln \frac{\sigma}{\alpha} + \frac{1}{2} \left(\frac{\alpha^2}{\sigma^2} - 1 \right) \end{aligned}$$

(b) Using the conclusion from (a), $P^* = \mathcal{N}(0, \alpha^2)$.

4.2. (a) To prove the following lemma,

$$\left(\frac{n}{e}\right)^n \leq n! \leq n \left(\frac{n}{e}\right)^n \quad (7)$$

Which is equivalent to,

$$n \ln n - n \leq \ln(n!) \leq (n+1) \ln n - n \quad (8)$$

For the left part, notice that $\ln(1 + \frac{1}{i}) < \frac{1}{i}$ for $i \geq 1$ which leads to,

$$(i+1) \ln(i+1) - i \ln i - 1 < \ln(i+1) \quad (9)$$

Sum for $i = 1, 2, \dots, n-1$, we have

$$n \ln n - (n-1) < \sum_{i=1}^{n-1} \ln(i+1) = \ln(n!) \quad (10)$$

For the right part, it holds only when $n \geq 7$. It is easy to check $n = 7$. So $\ln(7!) \leq 8 \ln 7 - 7$

And for $n \geq 8$, because $\ln(1+x) > \frac{x}{x+1}$ for $x > 0$, $\ln(1 + \frac{1}{i}) > \frac{1}{i+1}$, $\ln i < (i+1) \ln(i+1) - i \ln i - 1$.

Sum for $i = 7, \dots, n-1$

$$\ln(6!) + \sum_{i=7}^{n-1} \ln i + \ln n < 7 \ln 7 - 7 + n \ln n + -7 \ln 7 + \ln n - (n-7) = (n+1) \ln n - n \quad (11)$$

So

$$\left(\frac{n}{e}\right)^n \leq n! \leq n \left(\frac{n}{e}\right)^n \quad (12)$$

(b) From (a) we have that as $n \rightarrow \infty$,

$$\frac{\ln(n!)}{n} \sim \ln \frac{n}{e} \quad (13)$$

Therefore

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \log \binom{n}{k} &= \lim_{n \rightarrow \infty} \frac{1}{n} \log \frac{n!}{k!(n-k)!} \\ &= \lim_{n \rightarrow \infty} \log \frac{n}{e} - p \log \frac{pn}{e} - (1-p) \log \frac{(1-p)n}{e} \\ &= -p \log p - (1-p) \log(1-p) = H(p) \end{aligned} \quad (14)$$

Another explanation using Sanov's theorem, suppose

$X_1, X_2, \dots, X_n \stackrel{i.i.d}{\sim} \text{Bernoulli}(\frac{1}{2})$ consider

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i = p\right) \quad (15)$$

On the one hand, $\sum_{i=1}^n X_i \sim \text{Binomial}(n, \frac{1}{2})$, so

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i = p\right) = \lim_{n \rightarrow \infty} -\frac{1}{n} \log \left[\binom{n}{k} \frac{1}{2^n} \right] = 1 - \lim_{n \rightarrow \infty} \frac{1}{n} \log \binom{n}{k} \quad (16)$$

On the other hand, using Sanov's theorem,

$$\begin{aligned} \lim_{n \rightarrow \infty} -\frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i = p\right) &= D(\text{Bernoulli}(p) \parallel \text{Bernoulli}(\frac{1}{2})) \\ &= p \log 2p + (1-p) \log(2(1-p)) \\ &= 1 + p \log p + (1-p) \log(1-p) \\ &= 1 - H(p) \end{aligned} \quad (17)$$

Also we can get

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \binom{n}{k} = H(p) \quad (18)$$

Using the same way but using categorical and multinomial distribution instead of Bernoulli and binomial distribution.

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \left(\binom{n}{\lfloor np_1 \rfloor \lfloor np_2 \rfloor \cdots \lfloor np_{m-1} \rfloor} \left(n - \sum_{i=1}^{m-1} \lfloor np_i \rfloor \right) \right) = - \sum_{i=1}^m p_i \log p_i \quad (19)$$

where $\sum_{i=1}^m p_i = 1$.

- 4.3. (a) $\mathbb{E}_p[y] = 0$ means that $p_0 = 1, p_1 = p_2 = 0$. So \mathcal{L}_0 is just a single point $(1, 0, 0)$.

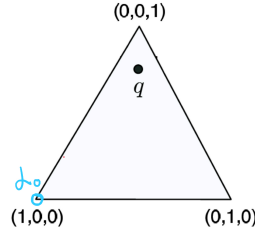


Figure 1: \mathcal{L}_0

- (b) $\mathbb{E}_p[y] = \frac{1}{2}$ means $\mathcal{L}_{\frac{1}{2}} = \{p = (p_0, p_1, p_2) : p_0 + p_1 + p_2 = 1, p_1 + 2p_2 = \frac{1}{2}\}$ which is a line passing $(\frac{1}{2}, \frac{1}{2}, 0)$ and $(\frac{3}{4}, 0, \frac{1}{4})$.

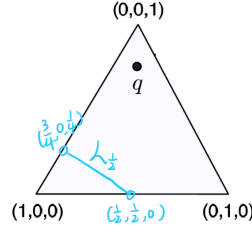


Figure 2: $\mathcal{L}_{\frac{1}{2}}$

- (c) The exponential family is $\mathcal{E} = \{\tilde{q} : \tilde{q} = qe^{sf(y) - \alpha(s)}\}$. Following Pythagoras's Identity, let $f(y) = y$ so \mathcal{E} is orthogonal to $\mathcal{L}_{\frac{1}{2}}$. $\mathcal{E} = \{\tilde{q} : \tilde{q} = qe^{sy - \alpha(s)}\}$. Denote $\lambda = e^s$, so $\tilde{q}_0 = \frac{1}{1+\lambda+4\lambda^2}$, $\tilde{q}_1 = \frac{\lambda}{1+\lambda+4\lambda^2}$, $\tilde{q}_2 = \frac{4\lambda^2}{1+\lambda+4\lambda^2}$. Notice that \mathcal{E} passes $(1, 0, 0)$ and $(0, 0, 1)$ and $\tilde{q}_1 \leq \frac{1}{5}$.
- (d) Using the Lagrange-Multiplier method we can induce that the I-projection p^* of q onto $\mathcal{L}_{\frac{1}{2}}$ belongs to \mathcal{E} . So $p^* \in \mathcal{L}_{\frac{1}{2}} \cap \mathcal{E}$. By $\tilde{q}_1 + 2\tilde{q}_2 = \frac{1}{2}$ we can solve $\lambda = \frac{1}{4}$, $p^* = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$.
- (e) As figure 5

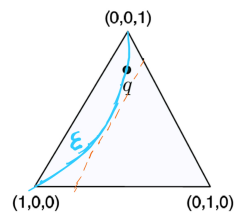


Figure 3: \mathcal{E}

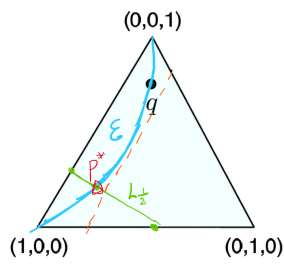


Figure 4: p^*

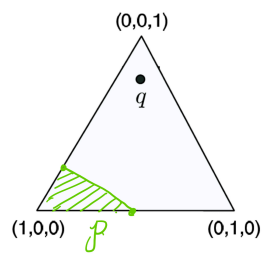


Figure 5: \mathcal{P}

- (f) First, for any $p \in \mathcal{P}$, it belongs to some $\mathcal{L}_\gamma = p : \mathbb{E}_p[y] = \gamma$ and $\gamma \leq \frac{1}{2}$. So $D(p\|q) \geq D(p_\gamma^*\|q)$ where p_γ^* is the I-projection of q onto \mathcal{L}_γ . And $p_\gamma^* \in \mathcal{E}$. Thus,

$$p^* = \arg \min_{p \in \mathcal{P}} D(p\|q) = \arg \min_{p \in \mathcal{P} \cap \mathcal{E}} D(p\|q) \quad (20)$$

For $\tilde{q}_s \in \mathcal{P} \cap \mathcal{E}$, $\gamma = \mathbb{E}_{\tilde{q}}[y] = \tilde{q}_1 + 2\tilde{q}_2 = \frac{\lambda + 8\lambda^2}{1 + \lambda + 4\lambda^2}$, $\lambda = e^s$.

$$\frac{d\gamma}{d\lambda} = \frac{1 + 16\lambda + 4\lambda^2}{(1 + \lambda + 4\lambda^2)^2} \geq 0 \quad (21)$$

So γ strictly increases with λ , then γ strictly increases with s , vice versa. And when $\gamma = \frac{1}{2}$, $\lambda = \frac{1}{4}$, $s = -\ln 4$

And $D(\tilde{q}_s\|q) = s\mathbb{E}_{\tilde{q}_s}[y] - \alpha(s)$.

$$\frac{\partial D(\tilde{q}_s\|q)}{\partial s} = s\text{Var}_{\tilde{q}_s}[y] \leq 0, \quad \text{for } s \leq -\ln 4 < 0 \quad (22)$$

So $\frac{\partial D(\tilde{q}_s\|q)}{\partial \gamma} \leq 0$ for $\gamma \leq \frac{1}{2}$. To minimize $D(\tilde{q}_s\|q)$, $\gamma^* = \frac{1}{2}$, $p^* = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$

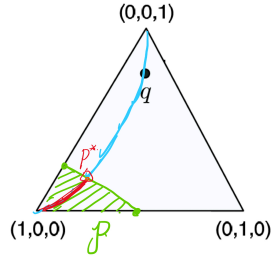


Figure 6: p^*

4.4. (a) $\forall p \in \mathcal{P}$

$$D(q\|p) = \sum_{x=0}^{\infty} q(x) \log \frac{q(x)}{p(x)} \quad (23)$$

For $x \geq M$, $q(x) > 0$, $q(x) = 0$, $q(x) \log \frac{q(x)}{p(x)} = \infty$, so $D(q\|p) = \infty$

(b) $\forall p \in \mathcal{P}$

$$D(p\|q) = \sum_{x=0}^{\infty} p(x) \log \frac{p(x)}{q(x)} \quad (24)$$

For $x \geq M$, $q(x) > 0$, $q(x) = 0$, $p(x) \log \frac{p(x)}{q(x)} = 0$. For $x < M$, $p(x) \log \frac{p(x)}{q(x)} < \infty$, so $D(p\|q) < \infty$

(c) To find the I-projection,

$$\begin{aligned} \min \quad & \sum_{i=0}^{M-1} p_i \log \frac{p_i}{q_i} \\ \text{s.t.} \quad & \sum_{i=0}^{M-1} p_i = 1 \end{aligned} \quad (25)$$

Using Lagrange Multiplier,

$$L = \sum_{i=0}^{M-1} p_i \log \frac{p_i}{q_i} + \lambda \left(\sum_{i=0}^{M-1} p_i - 1 \right) \quad (26)$$

$$\frac{\partial L}{\partial p_i} = 1 + \log \frac{p_i}{q_i} + \lambda = 0, \quad i = 0, 1, 2, \dots, M-1 \quad (27)$$

So

$$p_i = \frac{q_i}{\sum_{j=0}^{M-1} q_j} = \frac{q_i}{Q(M-1)} \quad (28)$$

And

$$D(p^* \| q) = \sum_{i=0}^{M-1} p_i \log \frac{p_i}{q_i} = -\log Q(M-1) \quad (29)$$

(d)

$$\begin{aligned} \min \quad & \sum_{i=0}^{\infty} p_i \log \frac{p_i}{q_i} \\ \text{s.t.} \quad & \sum_{i=0}^{M-1} p_i = 1 - \varepsilon, \\ & \sum_{i=M}^{\infty} p_i = \varepsilon \end{aligned} \quad (30)$$

Using Lagrange Multiplier,

$$L = \sum_{i=0}^{\infty} p_i \log \frac{p_i}{q_i} + \lambda \left(\sum_{i=0}^{M-1} p_i - 1 + \varepsilon \right) + \mu \left(\sum_{i=M}^{\infty} p_i - \varepsilon \right) \quad (31)$$

$$\begin{aligned} \frac{\partial L}{\partial p_i} &= 1 + \log \frac{p_i}{q_i} + \lambda = 0, \quad i = 0, 1, 2, \dots, M-1 \\ \frac{\partial L}{\partial p_i} &= 1 + \log \frac{p_i}{q_i} + \mu = 0, \quad i = M, \dots \end{aligned} \quad (32)$$

So

$$p_i = \begin{cases} \frac{(1-\varepsilon)q_i}{Q(M-1)}, & i = 0, 1, 2, \dots, M-1 \\ \frac{\varepsilon q_i}{1 - Q(M-1)}, & i = M, \dots \end{cases} \quad (33)$$

And

$$D(p_\varepsilon^* \| q) = (1 - \varepsilon) \log \frac{1 - \varepsilon}{Q(M - 1)} + \varepsilon \log \frac{\varepsilon}{1 - Q(M - 1)} \quad (34)$$

$$\lim_{\varepsilon \rightarrow 0} D(p_\varepsilon^* \| q) = -\log Q(M - 1) = D(p^* \| q) \quad (35)$$

- (e) Define a indication function $f(y) = \mathbf{1}(y \geq M)$, then $\mathcal{P}_\varepsilon = \{p : \mathbb{E}_p[f(y)] = \varepsilon\}$ is a linear family.
- (f) Because \mathcal{P}_ε is a linear family, the I-projection p_ε^* belongs to a exponential family $\mathcal{E} = \{\tilde{q} = qe^{sf(y) - \alpha(s)}\}$. And because $f(y) = \mathbf{1}(y \geq M)$. So $\tilde{q}_i = e^{-\alpha(s)} q_i, i = 0, 1, \dots, M - 1$ and $\tilde{q}_i = e^{s - \alpha(s)} q_i, i = M, \dots$. Comparing with the result in (4), the corresponding parameter

$$s^* = \log \frac{\varepsilon Q(M - 1)}{(1 - \varepsilon)(1 - Q(M - 1))} \quad (36)$$

4.5. (a)

$$K_x = \text{cov}(\underline{x}) = \mathbb{E} \left[(\underline{x} - \mathbb{E}[\underline{x}]) (\underline{x} - \mathbb{E}[\underline{x}])^T \right] = \mathbb{E} [\underline{x} \underline{x}^T] \quad (37)$$

And $\mathbb{E}[x_1^2] = \mathbb{E}[x_2^2] = \sigma^2, \mathbb{E}[x_1 x_2] = \rho_x \sigma^2$, so

$$K_x = \sigma^2 \begin{bmatrix} 1 & \rho_x \\ \rho_x & 1 \end{bmatrix} \quad (38)$$

Because $\underline{y} = A \underline{x}$

$$K_y = \text{cov}(\underline{y}) = \mathbb{E} \left[(\underline{y} - \mathbb{E}[\underline{y}]) (\underline{y} - \mathbb{E}[\underline{y}])^T \right] = A K_x A^T = \sigma^2 \begin{bmatrix} 1 - \rho_x^2 & 0 \\ 0 & 1 \end{bmatrix} \quad (39)$$

- (b) First we prove that for joint Gaussian distribution (X, Y) , if $\rho(X, Y) = 0$ then x, y are independent. Because

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp \left[-\frac{1}{2(1-\rho^2)} \left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right) \right] \quad (40)$$

Let $\rho = 0$ we have,

$$\begin{aligned} f(x, y) &= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \times \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}} \\ &= f(x)f(y) \end{aligned} \quad (41)$$

So if $\rho(x, y) = 0$, x, y are independent. Using this conclusion, y_1, y_2 are independent. Using the conclusion from Homework 1, $\forall f, g$, $\rho(f(y_1), g(y_2)) = 0$. So $\rho(y_1, g(y_2)) = 0$

(c) $y_1 = x_1 - \rho_{x_1 x_2}, y_2 = x_2$. So $\mathbb{E}[(x_1 - \rho_{x_1 x_2})^2] \leq \mathbb{E}[(x_1 - g(x_2))^2]$ is equivalent to

$$\mathbb{E}[y_1^2] \leq \mathbb{E}[(y_1 - g(y_2))^2] \quad (g'(y_2) = g(y_2) - \rho_{x_1} y_2) \quad (42)$$

Because $\rho(y_1, g(y_2)) = 0$, $\mathbb{E}[y_1 g(y_2)] = \mathbb{E}[y_1] \mathbb{E}[g(y_2)]$

$$\begin{aligned} \mathbb{E}[(y_1 - g(y_2))^2] &= \mathbb{E}[y_1^2] + \mathbb{E}[g(y_2)^2] - 2\mathbb{E}[y_1 g(y_2)] \\ &= \mathbb{E}[y_1^2] + \mathbb{E}[g(y_2)^2] \geq \mathbb{E}[y_1^2] \end{aligned} \quad (43)$$

4.6. (a) i. Because

$$\begin{aligned} \mathbb{E}[(y - \alpha)^2] &= \alpha^2 - 2\alpha\mathbb{E}[y] + \mathbb{E}[y^2] \\ &= (\alpha - \mathbb{E}[y])^2 + \mathbb{E}[y^2] - (\mathbb{E}[y])^2 \end{aligned} \quad (44)$$

Thus,

$$\mathbb{E}[y] = \arg \min_{\alpha \in \mathbb{R}} \mathbb{E}[(y - \alpha)^2] \quad (45)$$

ii. As we can see, when $\alpha = \mathbb{E}[y]$, the corresponding MSE is

$$\mathbb{E}[y^2] - (\mathbb{E}[y])^2 = \text{var}(y) = \min_{\alpha \in \mathbb{R}} \mathbb{E}[(y - \alpha)^2] \quad (46)$$

(b) i. To minimize $\mathbb{E}[(y - f(x))^2]$, we can minimize $\mathbb{E}[(y - f(x))^2 | x = x]$ for every x . Since $f(x)$ is a constant given $x = x$.

$$\begin{aligned} \mathbb{E}[(y - f(x))^2 | x = x] &= (f(x))^2 - 2f(x)\mathbb{E}[y|x] + \mathbb{E}[y^2|x] \\ &= (f(x) - \mathbb{E}[y|x])^2 + \mathbb{E}[y^2|x] - (\mathbb{E}[y|x])^2 \\ &= (f(x) - \mathbb{E}[y|x])^2 + \text{var}(y|x) \end{aligned} \quad (47)$$

Thus,

$$\mathbb{E}[y | x] = \arg \min_{f: \mathcal{X} \rightarrow \mathbb{R}} \mathbb{E}[(y - f(x))^2] \quad (48)$$

ii. For every x , the standard error is $\text{var}(y|x)$, so the MSE is $\mathbb{E}[\text{var}(y|x)]$.

(c) If x, y are independent, $\text{var}(y) = \text{var}(y|x)$ so $\text{var}(y) = \mathbb{E}[\text{var}(y|x)]$.

$$x \perp y \implies \text{MSE}(\mathbb{E}[y]) = \text{MSE}(\mathbb{E}[y | x]) \quad (49)$$

And Using the Law of Total Variance,

$$\text{var}(y) = \mathbb{E}[\text{var}(y|x)] + \text{var}(\mathbb{E}[y|x]) \quad (50)$$

$\text{var}(\mathbb{E}[y|x]) = 0$, which means $\mathbb{E}[y|x]$ is a constant, so $\mathbb{E}[y|x] = \mathbb{E}[\mathbb{E}[y|x]] = \mathbb{E}[y]$. So

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x})y] &= \mathbb{E}[\mathbb{E}[f(\mathbf{x})y|\mathbf{x}]] \\
&= \mathbb{E}[f(\mathbf{x})\mathbb{E}[y|\mathbf{x}]] \\
&= \mathbb{E}[f(\mathbf{x})\mathbb{E}[y]] \\
&= \mathbb{E}[f(\mathbf{x})]\mathbb{E}[y]
\end{aligned} \tag{51}$$

So

$$\text{MSE}(\mathbb{E}[y]) = \text{MSE}(\mathbb{E}[y | x]) \implies \forall f, \rho(f(\mathbf{x}), y) = 0 \tag{52}$$

In general $\text{var}(\mathbb{E}[y|\mathbf{x}]) \geq 0$, so

$$\begin{aligned}
\text{MSE}(\mathbb{E}[y]) &= \text{var}(y) = \mathbb{E}[\text{var}(y|\mathbf{x})] + \text{var}(\mathbb{E}[y|\mathbf{x}]) \\
&\geq \mathbb{E}[\text{var}(y|\mathbf{x})] = \text{MSE}(\mathbb{E}[y | x])
\end{aligned} \tag{53}$$

$$4.7. \quad (a) \quad \|\underline{y} - \underline{\alpha}\|_2^2 = (\underline{y} - \underline{\alpha})^T (\underline{y} - \underline{\alpha}) = \underline{y}^T \underline{y} - 2\underline{\alpha}^T \underline{y} + \underline{\alpha}^T \underline{\alpha}$$

$$\begin{aligned}
\mathbb{E}[\|\underline{y} - \underline{\alpha}\|_2^2] &= \mathbb{E}[\underline{y}^T \underline{y}] - 2\underline{\alpha}^T \mathbb{E}[\underline{y}] + \underline{\alpha}^T \underline{\alpha} \\
&= (\underline{\alpha} - \mathbb{E}[\underline{y}])^T (\underline{\alpha} - \mathbb{E}[\underline{y}]) + \mathbb{E}[\underline{y}^T \underline{y}] - \mathbb{E}[\underline{y}]^T \mathbb{E}[\underline{y}]
\end{aligned} \tag{54}$$

So when $\underline{\alpha} = \mathbb{E}[\underline{y}] = \underline{P}_y(\cdot)$, $\min_{\underline{\alpha} \in \mathbb{R}^k} \mathbb{E}[\|\underline{y} - \underline{\alpha}\|_2^2] = \text{cov}(\underline{y})$.

- (b) To minimize $\mathbb{E}[\|\underline{y} - \underline{f}(\mathbf{x})\|_2^2]$, we can minimize $\mathbb{E}[\|\underline{y} - \underline{f}(\mathbf{x})\|_2^2 | \mathbf{x}]$ for every \mathbf{x} . And because $\underline{f}(\mathbf{x})$ is a constant vector given \mathbf{x} , using the conclusion from (a),

$$\mathbb{E}[\underline{y}|\mathbf{x}] = \arg \min_{\underline{f}(\mathbf{x}) \in \mathbb{R}^k} \mathbb{E}[\|\underline{y} - \underline{f}(\mathbf{x})\|_2^2 | \mathbf{x}], \forall \mathbf{x} \tag{55}$$

Thus:

$$\underline{P}_{y|\mathbf{x}}(\cdot|\mathbf{x}) = \mathbb{E}[\underline{y}|\mathbf{x}] = \arg \min_{\underline{f}:\mathcal{X} \rightarrow \mathbb{R}^k} \mathbb{E}[\|\underline{y} - \underline{f}(\mathbf{x})\|_2^2] \tag{56}$$

4.8. Suppose r, σ^2 is known, denote $\underline{\mathbf{x}} = (\mathbf{x}[0], \mathbf{x}[1], \dots, \mathbf{x}[N-1])^T$,

$$p(\underline{\mathbf{x}}; a) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \prod_{n=0}^{N-1} e^{-\frac{(\mathbf{x}[n] - ar^n)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\sum_{n=0}^{N-1} \frac{(\mathbf{x}[n] - ar^n)^2}{2\sigma^2}\right) \tag{57}$$

And

$$\ln p(\underline{\mathbf{x}}; a) = -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (\mathbf{x}[n] - ar^n)^2 \tag{58}$$

$$\frac{\partial \ln p(\underline{\mathbf{x}}; a)}{\partial a} = -\frac{1}{\sigma^2} \sum_{n=0}^{N-1} r^n (ar^n - \mathbf{x}[n]) \tag{59}$$

And the Fisher Information is

$$I(\theta) = \mathbb{E} \left[\left(\frac{\partial \ln p(x; \theta)}{\partial \theta} \right)^2 \right] \quad (60)$$

And it satisfies the regularity condition $\mathbb{E} \left[\frac{\partial \ln p(x; \theta)}{\partial \theta} \right] = 0$, so the fisher information can be also expressed as,

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2 \ln p(x; \theta)}{\partial \theta^2} \right] \quad (61)$$

And

$$\frac{\partial^2 \ln p(\underline{x}; a)}{\partial a^2} = -\frac{1}{\sigma^2} \sum_{n=0}^{N-1} r^{2n} = -\frac{1 - r^{2N}}{\sigma^2(1 - r^2)} \quad (62)$$

So

$$I(a) = \frac{1 - r^{2N}}{\sigma^2(1 - r^2)} \quad (63)$$

If an efficient estimator exists, then the Maximum Likelihood estimator is the efficient estimator. Let $\frac{\partial \ln p(\underline{x}; a)}{\partial a} = 0$,

$$\hat{a} = \frac{(1 - r^2)}{(1 - r^{2N})} \sum_{n=0}^{N-1} r^n x[n] \quad (64)$$

First,

$$\mathbb{E}[\hat{a}] = \frac{(1 - r^2)}{(1 - r^{2N})} \sum_{n=0}^{N-1} r^n a r^n = a \quad (65)$$

So \hat{a} is an unbiased estimator. And

$$\begin{aligned} \text{var}(\hat{a}) &= \mathbb{E}[(\hat{a} - a)^2] = \mathbb{E}[\hat{a}^2] - a^2 \\ &= \left(\frac{1 - r^2}{1 - r^{2N}} \right)^2 \mathbb{E} \left[\left(\sum_{n=0}^{N-1} r^n x[n] \right)^2 \right] - a^2 \\ &= \left(\frac{1 - r^2}{1 - r^{2N}} \right)^2 \sum_{n=0}^{N-1} r^{2n} \sigma^2 + \left(\frac{1 - r^2}{1 - r^{2N}} \right)^2 \left(\mathbb{E} \left[\sum_{n=0}^{N-1} r^n x[n] \right] \right)^2 - a^2 \\ &\quad (\text{Because } \text{cov}(x_i, x_j) = 0, i \neq j, \text{ and } \text{cov}(x_n, x_n) = \text{var}(x_n) = \sigma^2) \\ &= \left(\frac{1 - r^2}{1 - r^{2N}} \right) \sigma^2 + \left(\frac{1 - r^2}{1 - r^{2N}} \right)^2 \left(\frac{1 - r^{2N}}{1 - r^2} \right) a^2 - a^2 \\ &= \left(\frac{1 - r^2}{1 - r^{2N}} \right) \sigma^2 \\ &= \frac{1}{I(a)} \end{aligned} \quad (66)$$

So $\hat{a} = \frac{(1 - r^2)}{(1 - r^{2N})} \sum_{n=0}^{N-1} r^n x[n]$ is an efficient estimator of a .