Tsinghua-Berkeley Shenzhen Institute
INFORMATION THEORY AND STATISTICAL LEARNING
Fall 2020

Homework 4

Hanmo Chen November 18, 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 $\mathrm{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$, $\mathrm{y} = \sum_{i=1}^{n} \frac{\mathrm{x}_i^2}{\sigma^2} \sim \chi_n^2$.

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

So

$$\begin{aligned} -\frac{1}{n}\log\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}\mathrm{x}_i^2 \geqslant \alpha^2) &= -\frac{1}{n}\log\mathbb{P}(Y \geqslant \frac{n\alpha^2}{\sigma^2}) \\ &= -\frac{1}{n}\log\frac{\Gamma(\frac{n}{2}, \frac{n\alpha^2}{2\sigma^2})}{\Gamma(\frac{n}{2})} \end{aligned} \tag{1}$$

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

$$\lim_{n\to\infty} -\frac{1}{n}\log\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}\mathrm{x}_i^2 \geqslant \alpha^2) = \inf_{\mathbb{E}_P[X^2]\geqslant\alpha^2} D(P\|\mathcal{N}(0,\sigma^2)) \tag{2}$$

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

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

Define

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

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

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

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

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

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

4.2. (a) To prove the following lemma,

$$\left(\frac{n}{e}\right)^n \leqslant n! \leqslant n\left(\frac{n}{e}\right)^n \tag{7}$$

Which is equivalent to,

$$n \ln n - n \leqslant \ln(n!) \leqslant (n+1)\ln n - n \tag{8}$$

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

$$(i+1)\ln(i+1) - i \ln i - 1 < \ln(i+1) \tag{9}$$

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

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

For the right part, it holds only when $n \geqslant 7$. It is easy to check $n = 7$. So $\ln(7!) \leqslant 8 \ln 7 - 7$
And for $n \geqslant 8$, 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 $i = 7, \cdots, n-1$

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

So

$$\left(\frac{n}{e}\right)^n \leqslant n! \leqslant n\left(\frac{n}{e}\right)^n \tag{12}$$

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

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

Therefore

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

Another explanation using Sanov's theorem, suppose $X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} Bernoulli(\frac{1}{2})$ consider

$$\lim_{n\to\infty} -\frac{1}{n}\log\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}X_i = p) \tag{15}$$

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

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

On the other hand, using Sanov's theorem,

$$\begin{aligned}\lim_{n\to\infty} -\frac{1}{n}\log\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}X_i = p) &= D(Bernoulli(p)\|Bernoulli(\frac{1}{2}))\\ &= p\log 2p + (1-p)\log(2(1-p))\\ &= 1+p\log p+(1-p)\log(1-p)\\ &= 1-H(p)\end{aligned} \tag{17}$$

Also we can get

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

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

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

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

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

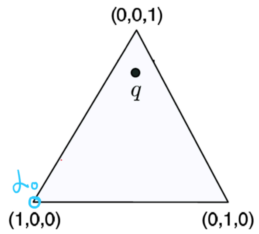

Figure 1: $\mathcal{L}_0$

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

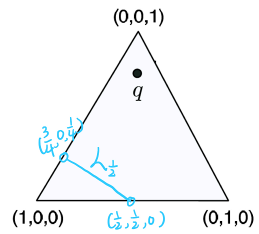

Figure 2: $\mathcal{L}_{\frac{1}{2}}$

(c) The exponential family is $\mathcal{E} = \left\{ \tilde{q} : \tilde{q} = q e^{s f(\mathrm{y}) - \alpha(s)} \right\}$. Following Pythagoras's Identity, let $f(y) = y$ so $\mathcal{E}$ is orthogonal to $\mathcal{L}_{\frac{1}{2}}$.
$\mathcal{E} = \left\{ \tilde{q} : \tilde{q} = q e^{s\mathrm{y} - \alpha(s)} \right\}$. 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 $\tilde{q}_1 \leqslant \frac{1}{5}$.

(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

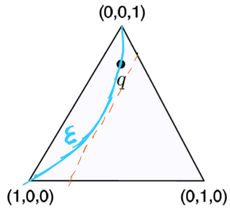

Figure 3: $\mathcal{E}$

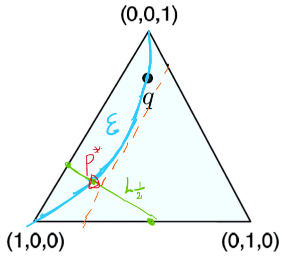

Figure 4: p^*

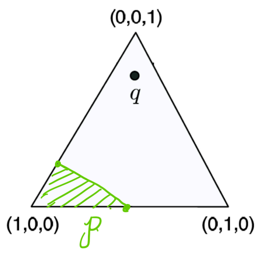

Figure 5: $\mathcal{P}$

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

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

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

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

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

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

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

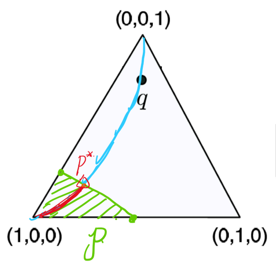

Figure 6: p^*

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

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

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

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

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

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

(c) To find the I-projection,

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

Using Lagrange Multiplier,

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

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

So

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

And

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

(d)

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

Using Lagrange Multiplier,

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

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

So

$$p_i = \begin{cases} \dfrac{(1-\varepsilon)q_i}{Q(M-1)}, & i = 0, 1, 2, \cdot, M-1 \\ \dfrac{\varepsilon q_i}{1 - Q(M-1)}, & i = M, \cdots \end{cases} \tag{33}$$

And

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

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

(e) Define a indication function $f(\mathrm{y}) = \mathbf{1}(\mathrm{y} \geqslant M)$, then $\mathcal{P}_\varepsilon = \{p : \mathbb{E}_p[f(\mathrm{y})] = \epsilon\}$ is a linear family.

(f) Because $\mathcal{P}_\varepsilon$ 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 $f(y) = \mathbf{1}(\mathrm{y} \geqslant M)$. So $\tilde{q}_i = e^{-\alpha(s)}q_i, i = 0, 1, \cdots, M-1$ and $\tilde{q}_i = e^{s-\alpha(s)}q_i, i = M, \cdots$. Comparing with the result in (4), the corresponding parameter

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

4.5. (a)

$$K_{\mathrm{x}} = \mathrm{cov}(\underline{\mathrm{x}}) = \mathbb{E}\left[(\underline{\mathrm{x}} - \mathbb{E}[\underline{\mathrm{x}}])\,(\underline{\mathrm{x}} - \mathbb{E}[\underline{\mathrm{x}}])^T\right] = \mathbb{E}\left[\underline{\mathrm{x}}\underline{\mathrm{x}}^T\right] \tag{37}$$

And $\mathbb{E}[\mathrm{x}_1^2] = \mathbb{E}[\mathrm{x}_2^2] = \sigma^2, \mathbb{E}[\mathrm{x}_1\mathrm{x}_2] = \rho_{\mathrm{x}}\sigma^2$, so

$$K_{\mathrm{x}} = \sigma^2\begin{bmatrix} 1 & \rho_{\mathrm{x}} \\ \rho_{\mathrm{x}} & 1 \end{bmatrix} \tag{38}$$

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

$$K_{\mathrm{y}} = \mathrm{cov}(\underline{\mathrm{y}}) = \mathbb{E}\left[\left(\underline{\mathrm{y}} - \mathbb{E}[\underline{\mathrm{y}}]\right)\left(\underline{\mathrm{y}} - \mathbb{E}[\underline{\mathrm{y}}]\right)^T\right] = AK_{\mathrm{x}}A^T = \sigma^2\begin{bmatrix} 1-\rho_{\mathrm{x}}^2 & 0 \\ 0 & 1 \end{bmatrix} \tag{39}$$

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

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

Let $\rho = 0$ we have,

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

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, $\forall f, g$, $\rho(f(\mathrm{y}_1), g(\mathrm{y}_2)) = 0$. So $\rho(\mathrm{y}_1, g(\mathrm{y}_2)) = 0$

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

$$\mathbb{E}\left[y_1^2\right] \leqslant \mathbb{E}\left[(y_1 - g(y_2))^2\right] \quad (g'(y_2) = g(y_2) - \rho_x y_2) \tag{42}$$

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

$$\begin{aligned}\mathbb{E}\left[(y_1 - g(y_2))^2\right] &= \mathbb{E}\left[y_1^2\right] + \mathbb{E}\left[g(y_2)^2\right] - 2\mathbb{E}[y_1 g(y_2)] \\ &= \mathbb{E}\left[y_1^2\right] + \mathbb{E}\left[g(y_2)^2\right] \geqslant \mathbb{E}\left[y_1^2\right]\end{aligned} \tag{43}$$

4.6. (a) i. Because

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

Thus,

$$\mathbb{E}[y] = \underset{\alpha\in\mathbb{R}}{\arg\min}\mathbb{E}\left[(y - \alpha)^2\right] \tag{45}$$

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

$$\mathbb{E}[y^2] - (\mathbb{E}[y])^2 = \mathrm{var}(y) = \min_{\alpha\in\mathbb{R}} \mathbb{E}\left[(y - \alpha)^2\right] \tag{46}$$

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

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

Thus,

$$\mathbb{E}[y \mid x] = \underset{f:\mathcal{X}\to\mathbb{R}}{\arg\min}\mathbb{E}\left[(y - f(x))^2\right] \tag{48}$$

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

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

$$x \perp y \Longrightarrow \mathrm{MSE}(\mathbb{E}[y]) = \mathrm{MSE}(\mathbb{E}[y \mid x]) \tag{49}$$

And Using the Law of Total Variance,

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

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

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

So

$$\mathrm{MSE}(\mathbb{E}[y]) = \mathrm{MSE}(\mathbb{E}[y \mid x]) \Longrightarrow \forall f, \rho(f(\mathrm{x}), \mathrm{y}) = 0 \tag{52}$$

In general $\mathrm{var}(\mathbb{E}\left[\mathrm{y}|\mathrm{x}\right]) \geqslant 0$, so

$$\begin{aligned}\mathrm{MSE}(\mathbb{E}[y]) = \mathrm{var}(\mathrm{y}) &= \mathbb{E}\left[\mathrm{var}(\mathrm{y}|\mathrm{x})\right] + \mathrm{var}(\mathbb{E}\left[\mathrm{y}|\mathrm{x}\right]) \\ &\geqslant \mathbb{E}\left[\mathrm{var}(\mathrm{y}|\mathrm{x})\right] = \mathrm{MSE}(\mathbb{E}[y \mid x])\end{aligned} \tag{53}$$