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

Homework 1

Chenyu Tian September 19, 2020

- **Acknowledgments:** This template takes some materials from course CSE 547/Stat 548 of Washington University: `https://courses.cs.washington.edu/courses/cse547/17sp/index.html`.
- *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.*

1.1. (a) i.

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[\mathsf{x}|\mathsf{yz}]|\mathsf{y}] &= \mathbb{E}((\sum_i x_i P(\mathsf{x}=x_i|\mathsf{yz}))|\mathsf{y}) \\
&= \sum_j (\sum_i x_i P(\mathsf{x}=x_i|\mathsf{y},\mathsf{z}=z_j))P(\mathsf{z}=z_j) \\
&= \sum_i x_i (\sum_j P(\mathsf{x}=x_i|\mathsf{y},\mathsf{z}=z_j)P(\mathsf{z}=z_j)) \\
&= \sum_i x_i P(\mathsf{x}=x_i|\mathsf{y}) \\
&= \mathbb{E}(\mathsf{x}|\mathsf{y})
\end{aligned}$$

ii.

$$g(\mathsf{y})\,\mathbb{E}[\mathsf{x}|\mathsf{y}] = \sum_i g(\mathsf{y}) x_i \Pr(\mathsf{x}=x_i|\mathsf{y}) = \mathbb{E}[xg(\mathsf{y})|\mathsf{y}]$$

iii.

$$\begin{aligned}
\mathbb{E}[\mathsf{x}\,\mathbb{E}[\mathsf{x}|\mathsf{y}]] &= \mathbb{E}\left[\mathsf{x}\sum_i x_i \Pr(\mathsf{x}=x_i|\mathsf{y})\right] \\
&= \sum_{j,k}(x_j \sum_i x_i \Pr(\mathsf{x}=x_i|\mathsf{y}=y_k)) \Pr(\mathsf{x}=x_j,\mathsf{y}=y_k) \\
&= \sum_{i,j,k} x_i x_j \Pr(\mathsf{x}=x_i|\mathsf{y}=y_k) \Pr(\mathsf{x}=x_j,\mathsf{y}=y_k)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[(\mathbb{E}[\mathsf{x}|\mathsf{y}])^2] &= \mathbb{E}\left[(\sum_i x_i \Pr(\mathsf{x}=x_i|\mathsf{y}))^2\right] \\
&= \sum_k (\sum_i x_i \Pr(\mathsf{x}=x_i|\mathsf{y}=y_k))^2 \Pr(\mathsf{y}=y_k) \\
&= \sum_{i,j,k} x_i x_j \Pr(\mathsf{x}=x_i|\mathsf{y}=y_k) \Pr(\mathsf{x}=x_j|\mathsf{y}=y_k) \Pr(\mathsf{y}=y_k) \\
&= \sum_{i,j,k} x_i x_j \Pr(\mathsf{x}=x_i|\mathsf{y}=y_k) \Pr(\mathsf{x}=x_j,\mathsf{y}=y_k)
\end{aligned}$$

Thus, $\mathbb{E}[(\mathbb{E}[\mathsf{x}|\mathsf{y}])^2] = \mathbb{E}[\mathsf{x}\,\mathbb{E}[\mathsf{x}|\mathsf{y}]]$.

iv.

$$
\begin{aligned}
&\mathbb{E}[\mathrm{Var}(\mathrm{x} \mid \mathrm{y})] + \mathrm{Var}(\mathbb{E}[\mathrm{x} \mid \mathrm{y}]) \\
=&\mathbb{E}[\mathbb{E}[\mathsf{x}^2|\mathsf{y}] - (\mathbb{E}[\mathsf{x}|\mathsf{y}])^2] + \mathbb{E}[(\mathbb{E}[\mathsf{x}|\mathsf{y}])^2] - (\mathbb{E}[\mathbb{E}[\mathsf{x}|\mathsf{y}]])^2 \\
=&\mathbb{E}[\mathbb{E}[\mathsf{x}^2|\mathsf{y}]] - (\mathbb{E}[\mathbb{E}[\mathsf{x}|\mathsf{y}]])^2 \\
=&\mathbb{E}[\mathsf{x}^2] - (\mathbb{E}[\mathsf{x}])^2 \\
=&\mathrm{Var}(\mathrm{x})
\end{aligned}
$$

Thus, $\mathrm{Var}(\mathrm{x}) = \mathbb{E}[\mathrm{Var}(\mathrm{x} \mid \mathrm{y})] + \mathrm{Var}(\mathbb{E}[\mathrm{x} \mid \mathrm{y}])$.

(b) i.

$$
\begin{aligned}
\mathrm{cov}(\underline{x}) =& \mathbb{E}[(\underline{x} - \mathbb{E}[\underline{x}])(\underline{x} - \mathbb{E}[\underline{x}])^{\mathrm{T}}] \\
=& \mathbb{E}[\underline{x}\underline{x}^{\mathrm{T}}] - \mathbb{E}[\underline{x}]\,\mathbb{E}[\underline{x}]^{\mathrm{T}} \\
=& \mathbb{E}[\mathbb{E}[\underline{x}\underline{x}^{\mathrm{T}}|\mathsf{y}]] - \mathbb{E}[\mathbb{E}[\underline{x}|\mathsf{y}]]\,\mathbb{E}[\mathbb{E}[\underline{x}|\mathsf{y}]]^{\mathrm{T}} \\
=& \mathbb{E}[\mathrm{Cov}[\underline{x}|\mathsf{y}] + \mathbb{E}[\underline{x}|\mathsf{y}]\,\mathbb{E}[\underline{x}|\mathsf{y}]^{\mathrm{T}}] - \mathbb{E}[\mathbb{E}[\underline{x}|\mathsf{y}]]\,\mathbb{E}[\mathbb{E}[\underline{x}|\mathsf{y}]]^{\mathrm{T}} \\
=& \mathbb{E}[\mathrm{Cov}[\underline{x}|\mathsf{y}]] + \mathbb{E}[\mathbb{E}[\underline{x}|\mathsf{y}]\,\mathbb{E}[\underline{x}|\mathsf{y}]^{\mathrm{T}}] - \mathbb{E}[\mathbb{E}[\underline{x}|\mathsf{y}]]\,\mathbb{E}[\mathbb{E}[\underline{x}|\mathsf{y}]]^{\mathrm{T}} \\
=& \mathbb{E}[\mathrm{Cov}[\underline{x}|\mathsf{y}]] + \mathrm{Cov}[\mathbb{E}[\underline{x}|\mathsf{y}]]
\end{aligned}
$$

ii. If $\exists \underline{c} \in \mathbb{R}^k, c \neq \underline{0}$, $\underline{c}^{\mathrm{T}}\underline{x} = d$ and d is a constant. $\underline{c}^{\mathrm{T}}\,\mathbb{E}[\underline{x}] = d$
Then $\underline{c}^{\mathrm{T}}\underline{x}\underline{x}^{\mathrm{T}}\underline{c} = d^2$. Also, $\underline{c}^{\mathrm{T}}\,\mathbb{E}[\underline{x}\underline{x}^{\mathrm{T}}]\underline{c} = d^2$.

$$
\begin{aligned}
\underline{c}^{\mathrm{T}}\,\mathrm{cov}(\underline{x})\underline{c} =& \underline{c}^{\mathrm{T}}\,\mathbb{E}[\underline{x}\underline{x}^{\mathrm{T}}]\underline{c} - \underline{c}^{\mathrm{T}}\,\mathbb{E}[\underline{x}]\,\mathbb{E}[\underline{x}]^{\mathrm{T}}\underline{c} \\
=& \underline{c}^{\mathrm{T}}\,\mathbb{E}[\underline{x}\underline{x}^{\mathrm{T}}]\underline{c} - \underline{c}^{\mathrm{T}}\,\mathbb{E}[\underline{x}]\,\mathbb{E}[\underline{x}]^{\mathrm{T}}\underline{c} \\
=& 0
\end{aligned}
$$

And,

$$
\begin{aligned}
\underline{c}^{\mathrm{T}}\,\mathrm{cov}(\underline{x})\underline{c} =& \underline{c}^{\mathrm{T}}\,\mathbb{E}[(\underline{x} - E[\underline{x}])(\underline{x} - E[\underline{x}])^{\mathrm{T}}]\underline{c} \\
=& \mathbb{E}[\underline{c}^{\mathrm{T}}(\underline{x} - E[\underline{x}])(\underline{x} - E[\underline{x}])^{\mathrm{T}}\underline{c}] \\
=& 0 \iff (\underline{x} - E[\underline{x}])(\underline{x} - E[\underline{x}])^{\mathrm{T}}\underline{c} = \underline{0}
\end{aligned}
$$

It has $\mathrm{cov}(\underline{x})c = \underline{0}$. Thus, $\det(\mathrm{cov}(\underline{x})) = 0$.

1.2. (a)

$$
\mathbb{E}[(y - ax - b)^2] == a^2\,\mathbb{E}[x^2] + \mathbb{E}[y^2] + b^2 - 2a\,\mathbb{E}[xy] - 2b\,\mathbb{E}[y] + 2ab\,\mathbb{E}[x]
$$

Take the derivatives with respect to a and b to find the minimum when the derivatives equal 0. It has

$$
\begin{aligned}
a^*\,\mathbb{E}[x^2] + b^*\,\mathbb{E}[x] =& \mathbb{E}[xy] \\
a^*\,\mathbb{E}[x] + b^* =& \mathbb{E}[y]
\end{aligned}
$$

Considering $\mathrm{var}(x) = \mathrm{var}(y) = \sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}$, it has
$a^* = \frac{\mathbb{E}[xy] - \mathbb{E}[x]\,\mathbb{E}[y]}{\mathrm{var}(x)} = \rho(x, y)$.

(b) If $x \perp y$,

$$\begin{aligned}
\mathbb{E}[f(x)g(y)] &= \sum_{i,j} f(x_i)g(y_j)\Pr(x = x_i, y = y_j) \\
&= \sum_{i,j} f(x_i)g(y_j)\Pr(x = x_i)\Pr(y = y_j) \\
&= \left(\sum_i f(x_i)\Pr(x = x_i)\right)\left(\sum_i g(y_j)\Pr(y = x_j)\right) \\
&= \mathbb{E}[f(x)]\,\mathbb{E}[g(y)]
\end{aligned}$$

$$\begin{aligned}
\rho(g(y), g(y)) =& \frac{\mathbb{E}[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])]}{\sqrt{\operatorname{var}(f(x))\operatorname{var}(g(y))}} \\
=& \frac{\mathbb{E}[f(x)g(y)] - \mathbb{E}[f(x)]\mathbb{E}[g(y)]}{\sqrt{\operatorname{var}(f(x))\operatorname{var}(g(y))}} \\
=& 0
\end{aligned}$$

So, $\forall f, g, \rho(f(x), g(y)) = 0$.
On the other hand, if $\forall f, g, \rho(f(x), g(y)) = 0$. Let $f(x) = x, g(y) = y$. It exists

$$\mathbb{E}[xy] = \mathbb{E}[x]\mathbb{E}[y] \Longleftrightarrow \Pr(x, y) = \Pr(x)\Pr(y)$$

So, $x \perp y$.

1.3. Denote $y = exp(x)$. Because exp(x) is monotonically increasing, it has

$$\begin{aligned}
f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y}\Pr(Y \le y) = \frac{\mathrm{d}}{\mathrm{d}y}\Pr(\ln Y \le \ln y) = \frac{\mathrm{d}}{\mathrm{d}y}\Phi(\ln y) \\
&= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\ln y)^2}{2}\right)\frac{\mathrm{d}\ln y}{\mathrm{d}y} \\
&= \frac{1}{y}\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\ln y)^2}{2}\right)
\end{aligned}$$