Tsinghua-Berkeley Shenzhen Institute
Information Theory and Statistical Learning
Fall 2020

Homework 3

Chenyu Tian October 27, 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.*

3.1. (a)

$$
\begin{aligned}
&\sum_{x\in\mathcal{X}} P_{\mathrm{x}}(x) D\left(P_{\mathrm{y}|\mathrm{x}=x} \| P_{\mathrm{y}|\mathrm{x}=x_0}\right) - D\left(P_{\mathrm{y}} \| P_{\mathrm{y}|\mathrm{x}=x_0}\right)\\
=&\sum_{x\in\mathcal{X}} P_{\mathrm{x}}(x) \sum_{y\in\mathcal{Y}} P_{\mathrm{y}|\mathrm{x}=x}(y) \log \frac{P_{\mathrm{y}|\mathrm{x}=x}(y)}{P_{\mathrm{y}|\mathrm{x}=x_0}(y)} - \sum_{y\in\mathcal{Y}} P_{\mathrm{y}}(y) \log \frac{P_{\mathrm{y}}(y)}{P_{\mathrm{y}|\mathrm{x}=x_0}(y)}\\
=&\sum_{x\in\mathcal{X}} P_{\mathrm{x}}(x) \sum_{y\in\mathcal{Y}} \frac{P_{\mathrm{xy}}(x,y)}{P_{\mathrm{x}}(x)} \log \frac{P_{\mathrm{xy}}(x,y)P_{\mathrm{x}}(x_0)}{P_{\mathrm{x}}(x)P_{\mathrm{xy}}(x_0,y)} - \sum_{y\in\mathcal{Y}} P_{\mathrm{y}}(y) \log \frac{P_{\mathrm{y}}(y)P_{\mathrm{x}}(x_0)}{P_{\mathrm{xy}}(x_0,y)}\\
=&\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{\mathrm{xy}}(x,y)\log P_{\mathrm{xy}}(x,y) - \sum_{x\in\mathcal{X}} P_{\mathrm{x}}(x)\log P_{\mathrm{x}}(x)\\
+&\sum_{x\in\mathcal{X}}\sum_{y\in} \log P_{\mathrm{x,y}}(x,y) \log \frac{P_{\mathrm{x}}(x_0)}{P_{\mathrm{xy}}(x_0,y)} - \sum_{y\in\mathcal{Y}} P_{\mathrm{y}}(y) \log \frac{P_{\mathrm{y}}(y)P_{\mathrm{x}}(x_0)}{P_{\mathrm{xy}}(x_0,y)}\\
=&\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{\mathrm{xy}}(x,y)\log P_{\mathrm{xy}}(x,y) - \sum_{x\in\mathcal{X}} P_{\mathrm{x}}(x)\log P_{\mathrm{x}}(x) - \sum_{y\in} P_{\mathrm{y}}(y)\log P_{\mathrm{y}}(y)\\
=&H(\mathrm{x}) + H(\mathrm{y}) - H(\mathrm{x},\mathrm{y})\\
=&I(\mathrm{x};\mathrm{y})
\end{aligned}
$$

(b) There exist x_0, x_0' that satisfies $\sup_{x,x'\in\mathcal{X}} D\left(P_{\mathrm{y}|\mathrm{x}=x} \| P_{\mathrm{y}|\mathrm{x}=x'}\right)$. Base on (a), for any distribution P_{x},

$$
\begin{aligned}
&\sup_{x,x'\in\mathcal{X}} D\left(P_{\mathrm{y}|\mathrm{x}=x} \| P_{\mathrm{y}|\mathrm{x}=x'}\right)\\
\geq&\sum_{x\in X} P_{\times}(x) D\left(P_{\mathrm{y}|x=x} \| P_{\mathrm{y}|\mathrm{x}=x_0}\right)\\
=&I(x;y) + D\left(P_{\mathrm{y}} \| P_{\mathrm{y}|\mathrm{x}=x_0}\right)\\
\geq&I(x;y)
\end{aligned}
\tag{1}
$$

Thus, $\sup_{P_x} I(\mathrm{x};\mathrm{y}) \leq \sup_{x,x'\in X} D\left(P_{\mathrm{y}|\mathrm{x}=x} \| P_{\mathrm{y}|\mathrm{x}=x'}\right)$.

3.2. (a)

$$\begin{aligned}
&H(x,y,z)=H(x)+H(y\mid x)+H(z\mid xy)\\
&H(x,y)=H(x)+H(y\mid x)\\
&H(x,z)=H(x)+H(z\mid x)\geqslant H(x)+H(z\mid xy)\\
&H(y,z)=H(y)+H(z\mid y)\geqslant H(y\mid x)+H(z\mid xy)\\
&H(x;y)+H(y;z)+H(z;x)\geqslant 2H(x;y;z)
\end{aligned} \tag{2}$$

(b) Let (x,y,z) represent the coordinates of uniformly chosen point in S,

$$\begin{aligned}
&\log n=H(x,y,z)=H(x)+H(y\mid x)+H(z\mid xy)\\
&\log n_1\geqslant H(x,y)=H(x)+H(y\mid x)\\
&\log n_2\geqslant H(x,z)=H(x)+H(z\mid x)\geqslant H(x)+H(z\mid xy)\\
&\log n_3\geqslant H(y,z)=H(y)+H(z\mid y)\geqslant H(y\mid x)+H(z\mid xy)\\
&\log n\leqslant \tfrac{1}{2}\log\left(n_1n_2n_3\right)
\end{aligned} \tag{3}$$

3.3. (a) Define $f(p,q)=p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q}-2(p-q)^2\log e$. If we differentiate f with respect to q, we get

$$\begin{aligned}
\frac{df}{dq}&=-\frac{p}{q}+(1-p)\frac{1}{1-q}+4(p-q)\log e\\
&=\frac{-p(1-q)+q(1-p)}{q(1-q)\ln 2}+4(p-q)\log e\\
&=\frac{-p+q}{q(1-q)\ln 2}+4(p-q)\log e\\
&=\frac{p-q}{\ln 2}\left(4-\frac{1}{q(1-q)}\right)
\end{aligned} \tag{4}$$

Since $1\geq q\geq 0$, the maximum value $q(1q)$ can take is $\frac{1}{4}$. The minimum value of $\frac{1}{q(1q)}$ is 4, and $4-\frac{1}{q(1q)}$ will always be negative. Thus, whether this differentiation is increasing or decreasing depends on $p-q$. If $p\geq q$, $\frac{df}{dq}\leq 0$; else, $\frac{df}{dq}\geq 0$. Also, $f(p,q=1)=f(p,q=0)=0$. It means $f(p,q)\geq 0$. Finally, $d(p\|q)\geq 2(p-q)^2\log e$.

(b) For any event E, let $Y=\mathbb{1}\{X\in E\}$ which is Bernoulli with parameter $P(E)$ or $Q(E)$. By data processing inequality, $D(P\|Q)\geq d(P(E)\|Q(E))$. Based on (a), we have $D(P\|Q)\geq d(P(E)\|Q(E))\geq \sup_E(2(P(E)-Q(E))^2\log e)$. Finally, we can prove $TV(P,Q)\leq\sqrt{\frac{D(P\|Q)}{2\log e}}$.

3.4. (a) To minimize the expected error, the simple error cost function can be defined as: $C=\begin{bmatrix}0&1\\1&0\end{bmatrix}$. Following the LRT, we can obtain the likelihood ratio test as:

$$\frac{p_{\mathrm{y}|\mathrm{H}}\left(y\mid H_1\right)}{p_{\mathrm{y}|\mathrm{H}}\left(y\mid H_0\right)}=2y=\frac{1-0}{1-0}\frac{p}{1-p}=\frac{p}{1-p} \tag{5}$$

Thus, the decision rule is

$$\hat{H}=\begin{cases}H_1, & y\geq\frac{p}{2-2p}\\ H_0, & \text{o.w.}\end{cases} \tag{6}$$

(b) Suppose the decision rule now is $\hat{H} = \begin{cases} H_1, & y \geq m \\ H_0, & \text{o.w.} \end{cases}$ Following,

$$\begin{aligned} P_D &\triangleq \mathbb{P}\left(\hat{\mathrm{H}} = H_1 \mid \mathrm{H} = H_1\right) = \int_m^1 2y\mathrm{d}y = 1 - m^2 \\ P_F &\triangleq \mathbb{P}\left(\hat{\mathrm{H}} = H_1 \mid \mathrm{H} = H_0\right) = \int_m^1 1\mathrm{d}y = 1 - m \end{aligned} \tag{7}$$

Therefore, $P_D = P_F\left(2 - P_F\right) = -P_F^2 + 2P_F$.

(c) i. The maximal value of PD can be achieved by solving the maximization problem:

$$\begin{aligned} P_D^{\max}(\varepsilon) &= \max_{P_D \geq (1+\varepsilon)P_F} P_D \\ &= \max_{P_F(2-P_F) \geq (1+\varepsilon)P_F} P_F\left(2 - P_F\right) \\ &= \max_{P_F \leq 1-\varepsilon} P_F\left(2 - P_F\right) \\ &= (1-\varepsilon)(1+\varepsilon) \\ &= 1 - \varepsilon^2 \end{aligned} \tag{8}$$

ii.

$$\begin{aligned} P_D^{\max}(\varepsilon) = 1 - \varepsilon^2 > 0 \quad &\text{and} \quad \varepsilon > 0 \\ \Rightarrow 0 < \varepsilon < 1 & \end{aligned} \tag{9}$$

iii. Based on (b), let $m = \frac{p}{2-2p}$,

$$\begin{aligned} &P_D \geq (1+\varepsilon)P_F \\ \Leftrightarrow &P_F \leq 1 - \varepsilon \\ \Leftrightarrow &1 - m \leq 1 - \varepsilon \\ \Leftrightarrow &\frac{p}{2-2p} \geq \varepsilon \\ \Leftrightarrow &p \geq \frac{2\varepsilon}{2\varepsilon + 1} \end{aligned} \tag{10}$$

So, $p \geq \frac{2\varepsilon}{2\varepsilon+1}$.

3.5. (a) i. Given that $y = \underline{y}$, define $\text{cost}\left(H_i\right) = \text{cost}\left(\hat{H} = H_i \mid \underline{\mathrm{y}} = \underline{y}\right) = C_{i1}\pi_1(\underline{y}) + C_{i2}\pi_2(\underline{y}) + C_{i3}\pi_3(\underline{y})$.

$$\begin{aligned} \text{cost}\left(H_1\right) &= \pi_2(\underline{y}) + 2\pi_3(\underline{y}) \\ \text{cost}\left(H_2\right) &= \pi_1(\underline{y}) + 2\pi_3(\underline{y}) \\ \text{cost}\left(H_3\right) &= 2\pi_1(\underline{y}) + 2\pi_2(\underline{y}) \end{aligned} \tag{11}$$

, and the optimal decision rule is $\hat{H} = \underset{H \in \{H_1, H2, H3\}}{\arg\min} \text{cost}(H)$.

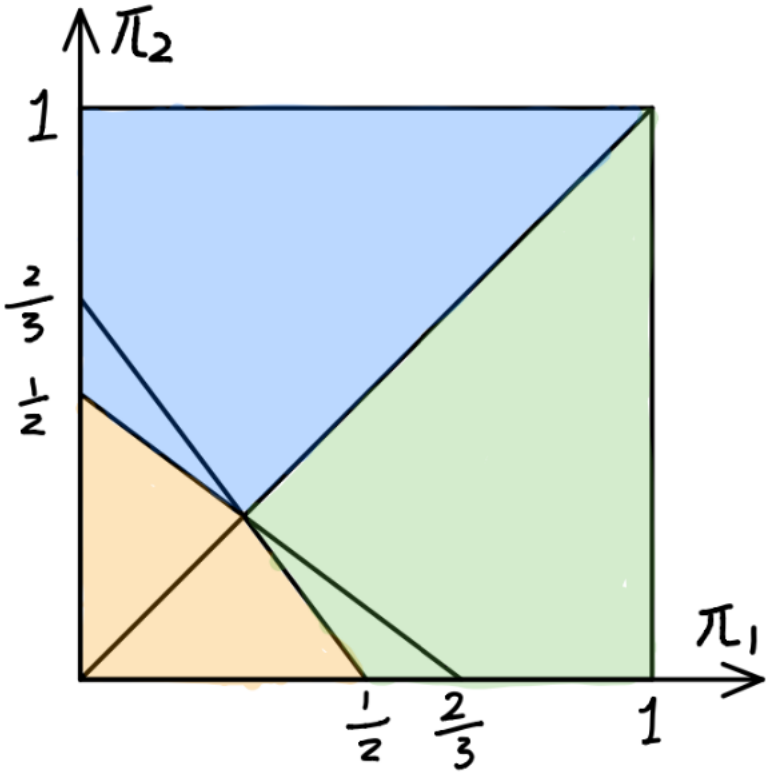
ii. Substitute $\pi_3(\underline{y}) = 1 - \pi_1(\underline{y}) - \pi_2(\underline{y})$. Then,

$$\begin{aligned} \text{cost}\left(H_1\right) &= 2 - 2\pi_1(\underline{y}) - \pi_2(\underline{y}) \\ \text{cost}\left(H_2\right) &= 2 - \pi_1(\underline{y}) - 2\pi_2(\underline{y}) \\ \text{cost}\left(H_3\right) &= 2\pi_1(\underline{y}) + 2\pi_2(\underline{y}) \end{aligned} \tag{12}$$

Find the boundary among different hypotheses by setting equal to those equations:

$$\begin{aligned}
&\text{cost}\,(H_1) - \text{cost}\,(H_2) = 0 \Rightarrow \pi_2(\underline{y}) - \pi_1(\underline{y}) = 0 \\
&\text{cost}\,(H_1) - \text{cost}\,(H_3) = 0 \Rightarrow 2 - 4\pi_1(\underline{y}) - 3\pi_2(\underline{y}) = 0 \\
&\text{cost}\,(H_2) - \text{cost}\,(H_3) = 0 \Rightarrow 2 - 3\pi_1(\underline{y}) - 4\pi_2(\underline{y}) = 0
\end{aligned} \tag{13}$$

By setting these equations to inequality, we can obtain the decision regions. The decision regions in the (π_1, π_2) plane is shown below, and green,blue and yellow are the regions for $H1$, $H2$ and $H3$ respectively.



(b) Since the three hypotheses are equally likely a priori and that the

Bayes costs matrix are given, we have:

$$
\begin{aligned}
\operatorname{cost}(H_i) &= \operatorname{cost}\left(\hat{H} = H_i \mid \underline{\mathrm{y}} = \underline{y}\right) \\
&= \sum_{j \neq i} C_{ij} \mathbb{P}\left(\mathrm{H} = H_j \mid \underline{\mathrm{y}} = \underline{y}\right) \\
&= \sum_{j \neq i} C_{ij} \frac{p_{\mathrm{y}|\mathrm{H}}\left(\underline{y} \mid H_j\right) P_{\mathrm{H}}(H_j)}{p_{\mathrm{y}}(\underline{y})} \\
&\propto \sum_{j \neq i} p_{\mathrm{y}|\mathrm{H}}\left(\underline{y} \mid H_j\right) \\
&\propto \sum_{j \neq i} L_j(\underline{y})
\end{aligned}
\tag{14}
$$

$$
\begin{aligned}
&\operatorname{cost}(H_1) \propto L_2(\underline{y}) + L_3(\underline{y}) \\
&\operatorname{cost}(H_2) \propto L_1(\underline{y}) + L_3(\underline{y}) = 1 + L_3(\underline{y}) \\
&\operatorname{cost}(H_3) \propto L_1(\underline{y}) + L_2(\underline{y}) = 1 + L_2(\underline{y})
\end{aligned}
\tag{15}
$$

$$
\begin{aligned}
\log L_i(\underline{y}) &= \log \frac{\frac{1}{\left(2\pi^{\frac{3}{2}}\right)|\sigma|^3} \exp\left(-\frac{1}{\sigma^2}\left(\underline{y} - \underline{m}_i\right)^{\mathrm{T}}\left(\underline{y} - \underline{m}_i\right)\right)}{\frac{1}{\left(2\pi^{\frac{3}{2}}\right)|\sigma|^3} \exp\left(-\frac{1}{\sigma^2}\left(\underline{y} - \underline{m}_1\right)^{\mathrm{T}}\left(\underline{y} - \underline{m}_1\right)\right)} \\
&= \frac{\left(\underline{y} - \underline{m}_1\right)^{\mathrm{T}}\left(\underline{y} - \underline{m}_1\right) - \left(\underline{y} - \underline{m}_i\right)^{\mathrm{T}}\left(\underline{y} - \underline{m}_i\right)}{\sigma^2} \\
&= \frac{2\left(\underline{m}_i - \underline{m}_1\right)^{\mathrm{T}} \underline{y} + \underline{m}_1^2 - \underline{m}_i^2}{\sigma^2} \\
&= \frac{2\ell_i(\underline{y}) + \underline{m}_1^2 - \underline{m}_i^2}{\sigma^2}
\end{aligned}
\tag{16}
$$

3.6. (a) Denote two hypothesis as $\begin{matrix} H_0: & x = -1 \\ H_1: & x = 1 \end{matrix}$. Consider simple error cost function the same as 3.4(a), We compute the cost of predicting $\hat{H} = H_1$ as

$$
\begin{aligned}
&\operatorname{cost}(\hat{H} = H_1) \\
=&\mathbb{P}_{\mathrm{H}_0|\underline{\hat{\mathrm{x}}}=\underline{\hat{x}}}(H_0) \\
\propto&\mathbb{P}_{\mathrm{H}_0,\underline{\hat{\mathrm{x}}}=\underline{\hat{x}}}(H_0) \\
=&P_{\mathrm{x}}(-1) \prod_{i=1}^{n} \mathbf{1}\left(\hat{\mathrm{x}}_i = 1\right) p_{\hat{\mathrm{x}}_i|\mathrm{x}}\left(\hat{x}_i \mid -1\right)
\end{aligned} .
\tag{17}
$$

Similarly, $\operatorname{cost}\left(\hat{H} = H_0\right) = P_{\mathrm{x}}(1) \prod_{i=1}^{n} \mathbf{1}\left(\hat{\mathrm{x}}_i = -1\right) p_{\hat{x}_i|\mathrm{x}}\left(\hat{x}_i \mid 1\right)$. Because for all i, y_i are conditionally independent given x, thus the local decision x^i are conditionally independent given x since x^i only depends on y_i. This gives rise to the last equation.

$$
p_{\hat{x}_i|x}\left(\hat{x}_i \mid -1\right) = \begin{cases} \int_0^1 \frac{1}{4} \mathrm{d}y_i = \frac{1}{4}, & \hat{\mathrm{x}}_i \neq -1 \\ \int_{-3}^0 \frac{1}{4} \mathrm{d}y_i = \frac{3}{4}, & \hat{\mathrm{x}}_i = -1 \end{cases}
\tag{18}
$$

Similarly,

$$p_{\hat{x}_i|x}(\hat{x}_i \mid x) = \begin{cases} \frac{1}{4}, & \hat{x}_i \neq x \\ \frac{3}{4}, & \hat{x}_i = x \end{cases} \tag{19}$$

Define m as the number of x^i that equals 1: $m = \sum_i^n \mathbf{1}(\hat{\mathrm{x}}_i = 1)$.
Rewrite costs in the form of m:

$$\begin{aligned} \text{cost}\left(\hat{H} = H_1\right) &= \frac{1}{2}\frac{1}{4}^m \\ \text{cost}\left(\hat{H} = H_0\right) &= \frac{1}{2}\frac{1}{4}^{n-m} \end{aligned} \tag{20}$$

$$\log \frac{\text{cost}\left(\hat{H} = H_1\right)}{\text{cost}\left(\hat{H} = H_0\right)} = (n - 2m)\log 4 \tag{21}$$

So the minimum probability of error decision is
$\hat{\mathrm{x}} = \begin{cases} 1, & m \geq \frac{n}{2} \\ -1, & m \leq \frac{n}{2} \end{cases}$ with $m = \sum_i^n \mathbf{1}(\hat{\mathrm{x}}_i = 1)$.

(b) Calculate the cost of predicting $x^i = 1, 1$ respectively:

$$\begin{aligned} \text{cost}(\hat{x}_1 = 1) &= \sum_{x, \hat{x}_2 \in \{-1,1\}} C(1, \hat{x}_2, x) P_{\mathrm{x}}(x) P_{\hat{x}_2|\mathrm{x}}(\hat{x}_2 \mid x) P_{y_1|x}(y_1 \mid x) \\ \text{cost}(\hat{x}_1 = -1) &= \sum_{x, \hat{x}_2 \in \{-1,1\}} C(-1, \hat{x}_2, x) P_{\mathrm{x}}(x) P_{\hat{x}_2|x}(\hat{x}_2 \mid x) P_{y_1|x}(y_1 \mid x) \end{aligned} \tag{22}$$

Let $\text{cost}(\hat{x}_1 = 1) \leq \text{cost}(\hat{x}_1 = -1)$, and since C strictly increases with the number of errors made by the two sensors, it has

$$\frac{P_{\mathrm{y}_1|\mathrm{x}}(y_1 \mid 1)}{P_{\mathrm{y}_1|\mathrm{x}}(y_1 \mid -1)} \geq \frac{P_{\mathrm{x}}(-1)}{P_{\mathrm{x}}(1)} \frac{\sum_{\hat{x}_2 \in \{-1,1\}} \left(C(1, \hat{x}_2, -1) - C(-1, \hat{x}_2, -1)\right) P_{\hat{x}_2|\mathrm{x}}(\hat{x}_2 \mid -1)}{\sum_{\hat{x}_2 \in \{-1,1\}} \left(C(-1, \hat{x}_2, 1) - C(1, \hat{x}_2, 1)\right) P_{\hat{\mathrm{x}}_2|\mathrm{x}}(\hat{x}_2 \mid 1)} \tag{23}$$

$$\gamma_1 = \frac{P_{\mathbf{x}}(-1)}{P_{\times}(1)} \frac{\sum_{\hat{x}_2 \in \{-1,1\}} \left(C(1, \hat{x}_2, -1) - C(-1, \hat{x}_2, -1)\right) P_{\hat{x}_2|x}(\hat{x}_2 \mid -1)}{\sum_{\hat{x}_2 \in \{-1.1\}} \left(C(-1, \hat{x}_2, 1) - C(1, \hat{x}_2, 1)\right) P_{\hat{x}_2|x}(\hat{x}_2 \mid 1)} \tag{24}$$

with $P_{\hat{\mathrm{x}}_2|\mathrm{x}}(\hat{x}_2 \mid x) = \int 1(\hat{\mathrm{x}}_2 = \hat{x}_2) P_{\mathrm{y}_2|\mathrm{x}}(y_2 \mid x)\,\mathrm{d}y_2$.

(c) The optimal choice for $\hat{x_2^*}$ as $\hat{\mathrm{x}} = \begin{cases} 1, & \frac{P_{\mathrm{y}_2|\mathrm{x}}(y_2|\mathrm{x})}{P_{\mathrm{y}_2|1}(y_2|-1)} \geq \gamma_2 \\ -1, & \mathrm{o} \cdot w \end{cases}$, where
$\gamma_2 = \frac{P_{\mathbf{x}}(-1)}{P_{\mathbf{x}}(1)} \frac{\sum_{\hat{x}_1 \in \{-1,1\}} (C(1,\hat{x}_1,-1) - C(-1,\hat{x}_1,-1)) P_{\hat{x}_1|\mathbf{x}}(\hat{x}_1|-1)}{\sum_{\hat{x}_1 \in \{-1,1\}} (C(-1,\hat{x}_1,1) - C(1,\hat{x}_1,1)) P_{\hat{x}_1|x}(\hat{x}_1|1)}$.

(d) Use the given C to simplify γ_1:

$$\begin{aligned} \gamma_1 &= \frac{P_{\times}(-1)}{P_{\times}(1)} \frac{P_{\hat{x}_2|x}(-1 \mid -1) + (L-1) P_{\hat{x}_2|x}(1 \mid -1)}{P_{\hat{x}_2|x}(1 \mid 1) + (L-1) P_{\hat{x}_2|x}(-1 \mid 1)} \\ &= \frac{P_x(-1)}{P_x(1)} \frac{1 + (L-2) P_{\hat{x}_2|x}(1 \mid -1)}{1 + (L-2) P_{\hat{x}_2|x}(-1 \mid 1)} \end{aligned} \tag{25}$$

The final equation comes from the fact that
$P_{\hat{\mathrm{x}}_2|\mathrm{x}}(1 \mid x) + P_{\hat{\mathrm{x}}_2|\mathrm{x}}(-1 \mid x) = 1$. Therefore, to decouple two sensors,

we only needs to choose $L = 2$, this gives $\gamma_1 = \frac{P_\times(-1)}{P_\times(1)}$. Similarly, $\gamma_2 = \frac{P_\times(-1)}{P_\times(1)}$.