Quantum Information Theory
Solutions 3

Exercise 3.1 Smooth min-entropy in the i.i.d. limit

Let (X_i, Y_i) be a sequence of n i.i.d pairs of random variables, meaning that $P_{X_1Y_1...X_nY_n} = P_{XY}^{\times n}$. Also, let $\epsilon_n = \frac{\sigma^2}{n\delta^2}$ for some $\delta > 0$, and σ^2 be the variance of the conditional surprisal $h(X|Y) = -\log_2 P_{X|Y}$. Use the weak law of large numbers to prove the asymptotic equipartition lemma:

$$\lim_{n\to\infty} \frac{1}{n} H_{\min}^{\epsilon_n}(X_1...X_n|Y_1...Y_n)_{P^n} = H(X|Y)_{P_{XY}}.$$
$$\lim_{n\to\infty} \frac{1}{n} H_{\max}^{\epsilon_n}(X_1...X_n|Y_1...Y_n)_{P^n} = H(X|Y)_{P_{XY}}.$$

In exercise sheet 1 we have shown that Chebyshev's inequality for i.i.d. variables given by

$$P\left[\left(\frac{1}{n}\sum_i S_i - \mu\right)^2 > \nu\right] \leq \frac{\sigma^2}{n\nu^2}$$

Setting $S_i = h_P(x_i|y_i) = -\log P_{X|Y}(x_i|y_i)$ we get $\mu = H(X|Y)$ and thus

$$P\left[\left(\frac{1}{n}\sum_i h_P(x_i|y_i) - H(X|Y)\right)^2 < \nu\right] \geq 1 - \frac{\sigma^2}{n\nu^2}$$

for any ν. This knowledge allows us to restrict the set of vector pairs $(\vec{x}, \vec{y})$ to typical outcomes, namely we introduce a subset $\mathcal{G}_\nu$ of $\mathcal{X}^{\times n}$:

$$\mathcal{G}_\nu = \left\{(\vec{x}, \vec{y}) \in \mathcal{X}^{\times n} : \left(\frac{1}{n}\sum_i h_P(x_i|y_i) - H(X|Y)\right)^2 < \nu\right\}.$$

The Chebyshev's inequality can now be restated simply as

$$P_{(\vec{X},\vec{Y})}[\mathcal{G}_\nu] = P_{(\vec{X},\vec{Y})}[(\vec{x}, \vec{y}) \in \mathcal{G}_\nu] \geq 1 - \frac{\sigma^2}{n\nu^2}.$$

Furthermore, let $\mathcal{G}_\nu^c$ denote the complement of $\mathcal{G}_\nu$ in $\mathcal{X}^{\times n}$. As a next step we choose

$$Q_{(\vec{X},\vec{Y})}[(\vec{x}, \vec{y})] = \begin{cases} P_{(\vec{X},\vec{Y})}[(\vec{x}, \vec{y})]/P_{(\vec{X},\vec{Y})}[\mathcal{G}_\nu] & \text{if } (\vec{x}, \vec{y}) \in \mathcal{G}_\nu \\ 0 & \text{if } (\vec{x}, \vec{y}) \in \mathcal{G}_\nu^c \end{cases}.$$

The distribution $Q_{\vec{X}|\vec{Y}}$ is very similar to $P_{\vec{X}|\vec{Y}}$, with exception that it assumes 0 probability for all unlikely events (those in $\mathcal{G}_\nu^c$), and renormalizes all the others. We can show that the distance between the two distributions is small, namely

$$\begin{aligned}
\delta(P_{(\vec{X},\vec{Y})}, Q_{(\vec{X},\vec{Y})}) &= \frac{1}{2}\sum_{(\vec{x},\vec{y})}\left|P_{(\vec{X},\vec{Y})}[(\vec{x}, \vec{y})] - Q_{(\vec{X},\vec{Y})}[(\vec{x}, \vec{y})]\right| \\
&= \frac{1}{2}\sum_{(\vec{x},\vec{y})\in\mathcal{G}_\nu^c} P_{(\vec{X},\vec{Y})}[(\vec{x}, \vec{y})] + \frac{1}{2}\sum_{(\vec{x},\vec{y})\in\mathcal{G}_\nu} P_{(\vec{X},\vec{Y})}[(\vec{x}, \vec{y})]\left(\frac{1}{P_{(\vec{X},\vec{Y})}[\mathcal{G}_\nu]} - 1\right) \\
&= \frac{1}{2}(1 - P_{(\vec{X},\vec{Y})}[\mathcal{G}_\nu]) + P_{(\vec{X},\vec{Y})}[\mathcal{G}_\nu]\left(\frac{1}{P_{(\vec{X},\vec{Y})}[\mathcal{G}_\nu]} - 1\right) \\
&= 1 - P_{(\vec{X},\vec{Y})}[\mathcal{G}_\nu] = \frac{\sigma^2}{n\nu^2}
\end{aligned}$$

In particular, we can now evaluate the "smooth" min-entropy for any fixed $\epsilon > 0$ and $\nu > 0$:

$$\begin{aligned}
\frac{1}{n} H^{\epsilon_n}_{\min}(\vec{X}|\vec{Y}) &\geq \frac{1}{n} H_{\min}(\vec{X}|\vec{Y})_Q \qquad (1)\\
&= \min_{(\vec{x},\vec{y})\in\mathcal{X}^{\times n}} \frac{1}{n} h_Q(\vec{x}|\vec{y}) \\
&= -\frac{1}{n}\log \max_{(\vec{x},\vec{y})\in\mathcal{X}^{\times n}} Q_{\vec{X}|\vec{Y}}(\vec{x}|\vec{y}) \\
&= -\frac{1}{n}\log \max_{(\vec{x},\vec{y})\in\mathcal{G}_\nu} Q_{\vec{X}|\vec{Y}}(\vec{x}|\vec{y}) \\
&= -\frac{1}{n}\log \max_{(\vec{x},\vec{y})\in\mathcal{G}_\nu} P_{\vec{X}|\vec{Y}}(\vec{x}|\vec{y}) - \frac{1}{n}\log P_{(\vec{X},\vec{Y})}[\mathcal{G}_\nu] \\
&= \min_{(\vec{x},\vec{y})\in\mathcal{G}_\nu} \frac{1}{n}\sum_i h_P(x_i|y_i) \\
&\geq H(X|Y) - \sqrt{\nu}
\end{aligned}$$

The first inequality is a consequence of the fact that our $Q_{\vec{X}|\vec{Y}}$ is not necessarily optimal (as a matter of fact, it could be shown that it actually is). We have ignored the term $\frac{1}{n}\log P_{(\vec{X},\vec{Y})}[\mathcal{G}_\nu]$, because it is very small, since $P_{(\vec{X},\vec{Y})}[\mathcal{G}_\nu] \approx 1$. Now, when we apply the $n \to \infty$ limit, we need to choose ν wisely, so that both $\sqrt{\nu} \to 0$ and $\epsilon = \frac{\sigma^2}{n\nu^2} \to 0$. This can be achieved, for example, by choosing $nu = \frac{\log n}{\sqrt{n}}$.

Now we will briefly outline how to calculate $\lim_{n\to\infty} \frac{1}{n} H^{\epsilon}_{\max}(\vec{X}|\vec{Y})$

Consider $P_{(\vec{X},\vec{Y})}[\mathcal{G}_\nu] = \sum_{(\vec{x},\vec{y})\in\mathcal{G}_\nu} P_{(\vec{X},\vec{Y})}[(\vec{x},\vec{y})]$. One can rearrange the definition of the typical set to show that

$$P_{X_i|Y_i}(x_i|y_i) \geq 2^{-n[H(X|Y)+\sqrt{\nu}]},$$

Let us define a set $\mathcal{X}_y = \{\vec{X} : (\vec{X},\vec{Y}) \in \mathcal{G}_\nu\}$. Then

$$1 = \sum_{\vec{x}} Q_{\vec{X}|\vec{Y}}(\vec{x}) \geq \sum_{\vec{x}} P_{\vec{X}|\vec{Y}}(\vec{x}) \geq |\mathcal{X}_y| P_{X_i|Y_i}(x_i|y_i)$$

Therefore, by solving the above inequality for $|\mathcal{X}_y|$ we find that

$$H_{\max}[\vec{X}|\vec{Y}]_Q = -\log \max_y |\mathcal{X}_y| \leq n(H|Y + \sqrt{\nu}))$$

Finally,

$$H^{\epsilon}_{\max}[\vec{X}|\vec{Y}]_P \leq H_{\max}(X^n)_Q \leq n(H(X|Y) + \sqrt{\nu}).$$

Taking the limits as with $H_{\min}$ gives the desired result

Exercise 3.2 Data Processing Inequality

Random variables X, Y, Z form a Markov chain $X \to Y \to Z$ if the conditional distribution of Z depends only on Y: $p(z|x,y) = p(z|y)$. The goal in this exercise is to prove the data processing inequality, $I(X:Y) \geq I(X:Z)$ for $X \to Y \to Z$.

1. *First show the chain rule for mutual information: $I(X:YZ) = I(X:Z) + I(X:Y|Z)$, which holds for arbitrary X, Y, Z. The conditional mutual information is defined as*

$$I(X:Y|Z) = \sum_z p(z) I(X:Y|Z=z) = \sum_z p(z) \sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}.$$

First observe that $\frac{p(x,y|z)}{p(y|z)} = \frac{p(x,y,z)}{p(y,z)} = p(x|y,z)$, which means $I(X{:}Y|Z) = H(X|Z) - H(X|YZ)$. Then

$$I(X{:}YZ) = H(X) - H(X|YZ) = H(X) + I(X{:}Y|Z) - H(X|Z) = I(X{:}Z) + I(X{:}Y|Z).$$

2. *Next show that in a Markov chain $X \to Y \to Z$, X and Z are conditionally independent given Y; that is, $p(x, z|y) = p(x|y)p(z|y)$.*

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x, y)p(z|x, y)}{p(y)} = \frac{p(x|y)p(y)p(z|y)}{p(y)} = p(x|y)p(z|y).$$

3. *By expanding the mutual information $I(X : YZ)$ in two different ways, prove the data processing in equality.*

   There are only two ways to expand this expression:

$$I(X{:}YZ) = I(X{:}Z) + I(X{:}Y|Z) = I(X{:}Y) + I(X{:}Z|Y).$$

   Since X and Z are conditionally independent given Y, $I(X{:}Z|Y) = 0$. Meanwhile, $I(X{:}Y|Z) \geq 0$, since it is a mixture (over Z) of positive quantities $I(X{:}Y|Z = z)$. Therefore $I(X{:}Y) \geq I(X{:}Z)$.

Exercise 3.3 Fano's Inequality

Given random variables X and Y, how well can we predict X given Y? Fano's inequality bounds the probability of error in terms of the conditional entropy $H(X|Y)$. The goal of this exercise is to prove the inequality

$$P_{\text{error}} \geq \frac{H(X|Y) - 1}{\log |X|}.$$

1. *Representing the guess of X by the random variable $\widehat{X}$, which is some function, possibly random, of Y, show that $H(X|\widehat{X}) \geq H(X|Y)$.*

   The random variables X, Y, and $\widehat{X}$ form a Markov chain, so we can use the data processing inequality. It leads directly to $H(X|\widehat{X}) \geq H(X|Y)$.

2. *Consider the indicator random variable E which is 1 if $\widehat{X} \neq X$ and zero otherwise. Using the chain rule we can express the conditional entropy $H(E, X|\widehat{X})$ in two ways:*

$$H(E, X|\widehat{X}) = H(E|X, \widehat{X}) + H(X|\widehat{X}) = H(X|E, \widehat{X}) + H(E|\widehat{X}) \tag{2}$$

   Calculate each of these four expressions and complete the proof of the Fano inequality. Hint: For $H(E|\widehat{X})$ use the fact that conditioning reduces entropy: $H(E|\widehat{X}) \leq H(E)$. For $H(X|E, \widehat{X})$ consider the cases $E = 0, 1$ individually.

   $H(E|X, \widehat{X}) = 0$ since E is determined from X and $\widehat{X}$. $H(E|\widehat{X}) \leq H(E) = h_2(P_{\text{error}})$ since conditioning reduces entropy.

$$\begin{aligned} H(X|E, \widehat{X}) &= H(X|E = 0, \widehat{X})p(E = 0) + H(X|E = 1, \widehat{X})p(E = 1) \\ &= 0(1 - P_{\text{error}}) + H(X|E = 1, \widehat{X})P_{\text{error}} \leq P_{\text{error}} \log |X| \end{aligned}$$

Putting this together we have

$$H(X|Y) \leq H(X|\widehat{X}) \leq h_2(P_{\text{error}}) + P_{\text{error}} \log |X| \leq 1 + P_{\text{error}} \log |X|,$$

where the last inequality follows since $h_2(x) \leq 1$. Rearranging terms gives the Fano inequality.