EE376A - Information Theory
Final, Thursday March 22nd

Instructions:

- You have **three hours**, 12:15PM - 3:15PM
- The exam has 5 questions, totaling 100 points.
- Please start answering each question on a new page of the answer booklet.
- You are allowed to carry the textbook, your own notes and other course related material with you. Electronic reading devices [including kindles, laptops, ipads, etc.] are allowed, provided they are used solely for reading pdf files already stored on them and not for any other form of communication or information retrieval.
- Calculators are allowed for numerical computations.
- You are required to provide a sufficiently detailed explanation of how you arrived at your answers.
- You can use previous parts of a problem even if you did not solve them.
- As throughout the course, entropy (H) and Mutual Information (I) are specified in bits.
- log is taken in base 2.
- Good Luck!

1. ~~**Universal Compression**~~ *(20 points)*
In this problem, we describe a lossless compression scheme that asymptotically (for large n) achieves entropy for any iid source. Let x^n be a particular sequence, where each symbol is in alphabet $\mathcal{X} = \{1, 2, 3, \ldots, |\mathcal{X}|\}$. Let P_{x^n} be the empirical distribution of the sequence x^n. Consider the compressor C for the sequence x^n:

- In the first step, the compressor encodes the empirical distribution P_{x^n} of the sequence, using a fixed-length code.
- In the second step, the compressor outputs the index of the sequence in the type class $\mathcal{T}(P_{x^n})$, using $\lceil \log_2 |\mathcal{T}(P_{x^n})| \rceil$ bits.

(a) Describe the operations of the decoder D, when a sequence x^n is compressed using the compressor C.

(b) Let $L(x^n)$ be number of bits required to encode a sequence x^n using the compressor. Show that:

$$L(x^n) \leq |\mathcal{X}| \log_2(n+1) + nH(P_{x^n}) + 2$$

(c) Let the sequence X^n be generated i.i.d according to the distribution $q(x)$. We define the rate of the compressor to be R:

$$R = \frac{\mathbb{E}[L(X^n)]}{n}$$

Show that for any distribution $q(x)$, the rate R converges to $H(q)$ as $n \to \infty$.

(d) Let $f : \mathcal{X} \to \mathbb{R}$ be an arbitrary function, and let $\bar{f}(x^n) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$. Show that it is possible to compute $\bar{f}(x^n)$ from the compressed sequence without decoding it completely. How many bits of the compressed sequence need to be read for computing $\bar{f}(x^n)$?

Solution:

(a) The decoder decodes the empirical distribution P_{x^n} from the first fixed-length code, and then using the index in the second part to find the sequence in $\mathcal{T}(P_{x^n})$.

(b) The number of types is at most $(n+1)^{|\mathcal{X}|}$, thus the fixed-length code is of length at most $\lceil \log_2 (n+1)^{|\mathcal{X}|} \rceil \leq |\mathcal{X}| \log_2(n+1) + 1$. We also know from class that $|\mathcal{T}(P_{x^n})| \leq 2^{nH(P_{x^n})}$, and thus the code in the second step has length at most $\lceil \log_2 |\mathcal{T}(P_{x^n})| \rceil \leq nH(P_{x^n}) + 1$. Summing up gives the desired answer.

(c) Note that $H(P)$ is concave in P, we have

$$\begin{aligned} R = \frac{\mathbb{E}[L(X^n)]}{n} &\leq \frac{|\mathcal{X}| \log_2(n+1) + 2}{n} + \mathbb{E}H(P_{X^n}) \\ &\leq \frac{|\mathcal{X}| \log_2(n+1) + 2}{n} + H(\mathbb{E}P_{X^n}) \\ &= \frac{|\mathcal{X}| \log_2(n+1) + 2}{n} + H(q) \overset{n\to\infty}{\to} H(q). \end{aligned}$$

On the other hand, $R \geq H(q)$ for any lossless code with source distribution $q(x)$, so the rate converges to $H(q)$.

(d) We only need to know the type of P_{x^n} to compute $\bar{f}(x^n)$. Hence, only $|\mathcal{X}|\log_2(n+1)+1$ bits at the beginning of the compressed sequence need to be read.

~~2. Rate-Distortion function for pairs of random variables~~ ~~*(20 points)*~~

Let X, Y be independent sources, with rate distortion functions $R_X(D)$ and $R_Y(D)$, corresponding to distortion functions $d_X : \mathcal{X}\times\hat{\mathcal{X}} \to \mathbb{R}^+$ and $d_Y : \mathcal{Y}\times\hat{\mathcal{Y}} \to \mathbb{R}^+$ respectively.

We want to perform lossy compression on the product source (X, Y), where the distortion measure $d_{X,Y}$ is given by:

$$d_{X,Y}((x, y),(x', y')) = d_X(x, x') + d_Y(y, y')$$

Let $R(D)$ be the rate distortion function corresponding to the product source (X, Y) and the distortion $d_{X,Y}$.

(a) Show that if X, Y are independent, then for any $\hat{X}, \hat{Y}$:

$$I(X, Y; \hat{X}, \hat{Y}) \geq I(X; \hat{X}) + I(Y; \hat{Y})$$

(b) Show the following lower bound on $R(D)$:

$$R(D) \geq \min_{D_1+D_2\leq D} [R_X(D_1) + R_Y(D_2)]$$

(c) Show that the lower bound on $R(D)$ is achievable, i.e.,

$$R(D) \leq \min_{D_1+D_2\leq D} [R_X(D_1) + R_Y(D_2)]$$

(d) Let X, Y be independent binary random variables, distributed as $X \sim Ber(0.5)$ and $Y \sim Ber(0.3)$. Find the value of $R(D)$ for the product source (X, Y), for $D = 0.4$ where d_X and d_Y are Hamming distortions.
(you can leave the final answers in terms of binary entropy function)

(e) Let X, Y be independent Gaussian random variables distributed as $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(0, 4)$. Find the value of $R(D)$ for the product source (X, Y), for $D = 4$ and mean square distortion:

$$d_{X,Y}((x, y),(x', y')) = (x - x')^2 + (y - y')^2$$

How many bits/symbol are used to describe X?

Solution:

(a) The following chain of inequalities holds:

$$\begin{aligned} I(X, Y; \hat{X}, \hat{Y}) &= H(X, Y) - H(X, Y|\hat{X}, \hat{Y}) \\ &= H(X) + H(Y) - H(X|\hat{X}, \hat{Y}) - H(Y|X, \hat{X}, \hat{Y}) \\ &\geq H(X) + H(Y) - H(X|\hat{X}) - H(Y|\hat{Y}) \\ &= I(X; \hat{X}) + I(Y; \hat{Y}). \end{aligned}$$

(b) Due to the additive structure of $d_{X,Y}$, we have

$$\begin{aligned}
R(D) = R^{(I)}(D) &= \min_{p(\hat{x},\hat{y}|x,y):\mathbb{E}d_{X,Y}((x,y),(\hat{x},\hat{y}))\le D} I(X,Y;\hat{X},\hat{Y}) \\
&\ge \min_{p(\hat{x},\hat{y}|x,y):\mathbb{E}d_{X,Y}((x,y),(\hat{x},\hat{y}))\le D} I(X;\hat{X}) + I(Y;\hat{Y}) \\
&\ge \min_{D_1+D_2\le D}\left(\min_{p(\hat{x},\hat{y}|x,y):\mathbb{E}d_X(x,\hat{x})\le D_1} I(X;\hat{X}) + \min_{p(\hat{x},\hat{y}|x,y):\mathbb{E}d_Y(y,\hat{y})\le D_2} I(Y;\hat{Y})\right) \\
&= \min_{D_1+D_2\le D}\left(\min_{p(\hat{x}|x):\mathbb{E}d_X(x,\hat{x})\le D_1} I(X;\hat{X}) + \min_{p(\hat{y}|y):\mathbb{E}d_Y(y,\hat{y})\le D_2} I(Y;\hat{Y})\right) \\
&= \min_{D_1+D_2\le D} R_X^{(I)}(D_1) + R_Y^{(I)}(D_2) \\
&= \min_{D_1+D_2\le D} R_X(D_1) + R_Y(D_2).
\end{aligned}$$

(c) For any $D_1, D_2 \ge 0$ with $D_1 + D_2 \le D$, let $p^*(\hat{x}|x), p^*(\hat{y}|y)$ be the minimum achieving distributions of $R_X^{(I)}(D_1), R_Y^{(I)}(D_2)$, respectively. Now consider $p(\hat{x},\hat{y}|x,y) = p^*(\hat{x}|x)p^*(\hat{y}|y)$, then $\mathbb{E}d_{X,Y}((X,Y),(\hat{X},\hat{Y})) = \mathbb{E}d_X(X,\hat{X}) + \mathbb{E}d_Y(Y,\hat{Y}) \le D_1 + D_2 \le D$. Moreover, $(X,\hat{X})$ is independent of $(Y,\hat{Y})$, and thus

$$\begin{aligned}
R(D) = R^{(I)}(D) &\le I(X,Y;\hat{X},\hat{Y}) = I(X;\hat{X}) + I(Y;\hat{Y}) \\
&\le R_X^{(I)}(D_1) + R_Y^{(I)}(D_2) = R_X(D_1) + R_Y(D_2).
\end{aligned}$$

This inequality holds for any $D_1 + D_2 \le D$, and the result follows.

(d) By (b) and (c), we have

$$\begin{aligned}
R(0.4) &= \min_{D_1+D_2\le 0.4} R_X(D_1) + R_Y(D_2) \\
&= \min_{D_1+D_2\le 0.4} H(0.5) - H(\min\{D_1, 0.5\}) + H(0.3) - H(\min\{D_2, 0.3\}) \\
&\ge \min_{D_1+D_2\le 0.4} H(0.5) + H(0.3) - 2H(\frac{\min\{D_1, 0.5\} + \min\{D_2, 0.3\}}{2}) \\
&\ge \min_{D_1+D_2\le 0.4} H(0.5) + H(0.3) - 2H(\frac{D_1 + D_2}{2}) \\
&\ge 1 + H(0.3) - 2H(0.2)
\end{aligned}$$

where we have used the fact that $H(p)$ is increasing on $p \in [0, \frac{1}{2}]$ and concave. The minimum is attained at $D_1 = D_2 = 0.2$.

(e) By (b) and (c), we have

$$R(4) = \min_{D_1+D_2\le 4} R_X(D_1) + R_Y(D_2) = \min_{D_1+D_2\le 4} \frac{1}{2}\log\frac{1}{\min\{D_1, 1\}} + \frac{1}{2}\log\frac{4}{\min\{D_2, 4\}}.$$

If $D_1 \le 1$, by the convexity of $x \mapsto -\log x$ we know that the minimum is achieved at $D_1 = 1, D_2 = 3$. If $D_1 > 1$, we have $D_2 < 3$ and $\log\frac{4}{D_2} > \log\frac{4}{3}$. Hence, $(D_1^*, D_2^*) = (1, 3)$, and $R(4) = \frac{1}{2}\log\frac{4}{3}$. Note that $R_X(D_1^*) = 0$ in this case, no bit is used to describe X_1.

3. Compression with some help *(25 points)*

Consider the lossless source coding problem in Figure 1. The pair (X^n, Y^n) is generated by i.i.d. drawings of the finite alphabet pair (X, Y), that is $p(x^n, y^n) = \prod_{i=1}^n p_{XY}(x_i, y_i)$. We wish to transmit the source sequence X^n near-losslessly when Y^n is available at both the encoder and the decoder. Formally, a $(2^{nR}, n)$ code is defined by an encoder $m(x^n, y^n) \in \{1, 2, \ldots, 2^{nR}\}$ and a decoder $\hat{X}^n(m, y^n)$, and the probability of decoding error is defined as $P_e = P\{\hat{X}^n \neq X^n\}$, where $\hat{X}^n = \hat{X}^n(m(X^n, Y^n), Y^n)$. A rate R is achievable if there exists a sequence of codes with $P_e \to 0$ as $n \to \infty$.

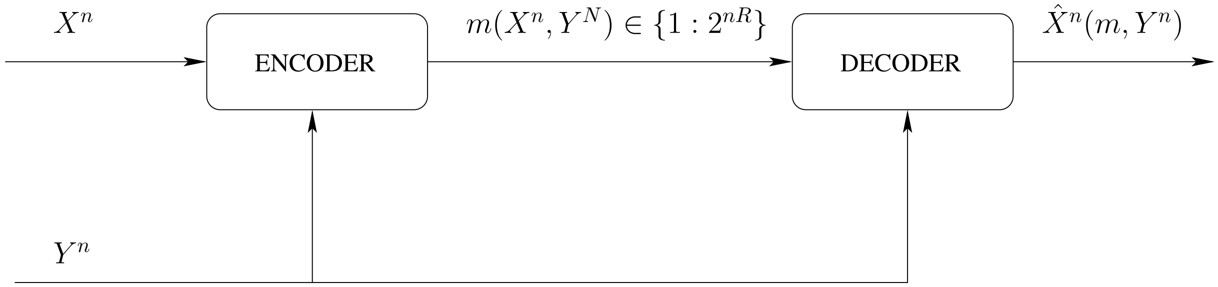

Figure 1: Conditional Lossless Source Coding

(a) Prove that any rate $R > H(X|Y)$ is achievable.
[*Hint*: If $y^n \in T_{\delta'}^{(n)}(Y)$ and $x^n \in T_{\delta}^{(n)}(X|y^n)$ for appropriate $\delta' < \delta$, transmit the index of x^n in $T_{\delta}^{(n)}(X|y^n)$.]

(b) Prove that any rate $R < H(X|Y)$ is not achievable via the following steps:

i. For $M = m(X^n, Y^n)$ argue why

$$I(X^n; M|Y^n) \leq nR.$$

ii. Use the previous step and a relation that you know between conditional entropy and probability of error to deduce that if $R < H(X|Y)$ then one cannot get $P_e \to 0$ as $n \to \infty$.

Now we consider a simple instance of this problem and develop concrete schemes for achieving the optimal rate. Let X be a random variable uniformly distributed on $\{0, 1\}^3$, i.e., X is a sequence of 3 independent unbiased bits. Let $Y = X \oplus Z$, where Z is independent of X and is uniformly distributed on $\{(0,0,0), (0,0,1), (0,1,0), (1,0,0)\}$ (set of binary triplets with at most one 1).

(c) Give a scheme to losslessly compress X into 2 bits when Y is known at both the encoder and the decoder. Specifically, you should describe the encoder $m(x, y) \in \{1, 2, 3, 4\}$ and a decoder $\hat{X}(m, y)$ which satisfy $\hat{X}(m(X, Y), Y) = X$. Is this optimal?

(d) Now, if only the decoder has access to Y, show that random variable X can still be losslessly compressed using 2 bits.
[*Hint*: Partition $\mathcal{X}$ into 4 suitable subsets, and transmit the index of the subset.]

(e) In part (d), can we do better (with less) than 2 bits?

Solution:

(a) Fix any $\delta > \delta' > 0$. By strong AEP, with probability tending to 1, we have $y^n \in T_{\delta'}^{(n)}(Y)$ and $x^n \in T_{\delta}^{(n)}(X|y^n)$. We consider the encoding/decoding scheme as follows:

- Encoding: the compressor sends the index of the sequence x^n in $T_{\delta}^{(n)}(X|y^n)$ if conditional typicality holds; otherwise, just send 1;
- Decoding: find the sequence x^n in $T_{\delta}^{(n)}(X|y^n)$ with the received index.

Note that this scheme has error probability tending to zero. Moreover, $|T_{\delta}^{(n)}(X|y^n)| \le 2^{n(1+\delta)H(X|Y)}$, therefore the rate is at most $R \le (1+\delta)H(X|Y)$. Since $\delta > 0$ is arbitrary, any rate $R > H(X|Y)$ is achievable.

(b) i. Note that $H(M) \le nR$ since $M \in \{1, 2, \cdots, 2^{nR}\}$, we have

$$I(X^n; M|Y^n) = H(M|Y^n) - H(M|X^n, Y^n) = H(M|Y^n) \le H(M) \le nR.$$

ii. Let $p_e = \mathbb{P}(\hat{X}^n \neq X^n)$, Fano's inequality gives

$$\begin{aligned} I(X^n; M|Y^n) &= H(X^n|Y^n) - H(X^n|M, Y^n) \\ &\ge H(X^n|Y^n) - H(X^n|\hat{X}^n) \\ &\ge nH(X|Y) - H(p_e) - np_e \log |\mathcal{X}|. \end{aligned}$$

Combining with the previous question, we see that

$$R \ge H(X|Y) - \frac{H(p_e)}{n} - p_e \log |\mathcal{X}|$$

i.e., any $R < H(X|Y)$ is impossible given $p_e \to 0$.

(c) Since the alphabet of Z has size $|\mathcal{Z}| = 4$, there exists a bijection f between $\mathcal{Z}$ and $\{1, 2, 3, 4\}$. Define encoder $m(x, y) = f(x \oplus y)$ and decoder $\hat{X}(m, y) = f^{-1}(m) \oplus y$. This definition is feasible since $X \oplus Y = Z \in \mathcal{Z}$. Clearly $\hat{X}(m(x, y), y) = f^{-1}(f(x \oplus y)) \oplus y = x$, and the rate is $\log |\mathcal{Z}| = 2$. This is not improvable, for

$$H(X|Y) = H(X) + H(Y|X) - H(Y) = H(X) + H(Z) - H(X \oplus Z) = 2.$$

(d) Split $\{0, 1\}^3$ into four groups: $G_1 = \{(0, 0, 0), (1, 1, 1)\}, G_2 = \{(1, 0, 0), (0, 1, 1)\}, G_3 = \{(0, 1, 0), (1, 0, 1)\}, G_4 = \{(0, 0, 1), (1, 1, 0)\}$. Upon receiving X, the encoder encodes the index of the group which X lies in. The decoder determines $\hat{X}$ to be the closest symbol to the side information Y (in Hamming distance) in the given group. Clearly the rate is 2, and this is lossless because the symbols in each group have minimum distance 3 and can thus correct 1-bit error caused by Z.

(e) No, because 2 bits are optimal even in the setting of (c), where the encoder also has the extra side information Y.

4. Channel Capacity (*15 points*)

Find the capacities of the following channels with the given channel transition matrices $p(y|x)$. Also, give the capacity-achieving input distribution $p(x)$. Justify your answers. (you can leave the final answers in terms of the binary entropy function)

(a) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

$$p(y|x) = \begin{bmatrix} 0 & 1/3 & 2/3 \\ 2/3 & 0 & 1/3 \\ 1/3 & 2/3 & 0 \end{bmatrix}$$

(b) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

$$p(y|x) = \begin{bmatrix} 0 & 1/3 & 2/3 \\ 2/3 & 0 & 1/3 \\ 0 & 2/3 & 1/3 \end{bmatrix}$$

(c) $\mathcal{X} = \{0, 1\}, \mathcal{Y} = \{0, 1, 2\}$

$$p(y|x) = \begin{bmatrix} 0 & 2/3 & 1/3 \\ 1/3 & 2/3 & 0 \end{bmatrix}$$

Solution:

(a) For any input distribution $p(x)$, we have

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\frac{1}{3}) \leq \log 3 - H(\frac{1}{3})$$

with equality iff Y is uniformly distributed on $\mathcal{Y}$. Therefore, the capacity-achieving input distribution is $p(x) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

(b) We can show that $I(X;Y) \leq \log 3 - H(\frac{1}{3})$ as in (a), with equality iff Y is uniformly distributed on $\mathcal{Y}$. This gives the capacity-achieving distribution $p(x) = (0, \frac{1}{2}, \frac{1}{2})$.

(c) For input distribution $(p, 1-p)$, we have $Y \sim (\frac{1-p}{3}, \frac{2}{3}, \frac{p}{3})$, and

$$\begin{aligned} I(X;Y) = H(Y) - H(Y|X) &= -\frac{1-p}{3} \log \frac{1-p}{3} - \frac{p}{3} \log \frac{p}{3} - \frac{2}{3} \log \frac{2}{3} - H(\frac{1}{3}) \\ &\leq 2 \cdot \frac{\log 6}{6} - \frac{2}{3} \log \frac{2}{3} - H(\frac{1}{3}) = \frac{1}{3} \end{aligned}$$

where the inequality follows from the concavity of $x \mapsto -x \log x$. As a result, the capacity-achieving input distribution is $p(x) = (\frac{1}{2}, \frac{1}{2})$.

The capacity can also be computed by observing that the channel is a special case of BEC channel (erasure probability 2/3).

~~**5. Information Theory and Statistics** (*20 points*)~~

This problem illustrates an application of information-theoretic tools in statistics. Suppose we observe a sample $X \sim \mathcal{N}(\theta, I_d)$, where $\theta \in \mathbb{R}^d$ is an unknown mean vector, and I_d denotes the $d \times d$ identity matrix. An *estimator* $\hat{\theta} = \hat{\theta}(X)$ is a function of X, and we want to find an estimator $\hat{\theta}$ which is close to the true θ. We consider the mean squared error $l(\theta) = \mathbb{E}_\theta \|\hat{\theta}(X) - \theta\|_2^2$, where the expectation is taken with respect to $X \sim \mathcal{N}(\theta, I_d)$.

(a) A natural estimator is $\hat{\theta}(X) = X$. What is $l(\theta)$ in this case? What is the worst-case $l(\theta)$ when θ can be any value in $\mathbb{R}^d$?

In the following, we show that this natural estimator is in fact a *minimax* estimator for estimating θ under mean squared error. By minimax we mean that it achieves the minimum worst-case error possible for any estimator. For this we'll use ideas from channel capacity and rate-distortion. First, we state some results for multivariate Gaussian distributions. These can be derived using similar techniques as those used for univariate Gaussian.

- *Capacity of multivariate AWGN channel*: Consider a channel from θ to X defined as $X = \theta + Z$ where $Z \sim \mathcal{N}(0, I_d)$ with power constraint $\mathbb{E}\|\theta\|_2^2 \leq d\sigma^2$. For this channel,

$$C = \frac{d}{2}\log(1+\sigma^2) \tag{1}$$

- *Rate-distortion function for multivariate Gaussian source*: Consider a source $\theta \sim \mathcal{N}(0, \sigma^2 I_d)$ and distortion metric $d(\theta, \hat{\theta}) = \mathbb{E}\|\theta - \hat{\theta}\|_2^2$. For this setting,

$$R(D) = \frac{d}{2}\log\frac{d\sigma^2}{D} \tag{2}$$

(b) Assume that there exists an estimator $\hat{\theta}$ with $l(\theta) \leq D$ for any $\theta \in \mathbb{R}^d$. Argue why that implies that we must have $R(D) \leq C$, where C and $R(D)$ are as defined in equations (1) and (2), respectively.
[*Hint:* Frame this as a joint source-channel coding problem with appropriate source and channel.]

(c) Conclude from (b) that $D \geq \frac{d\sigma^2}{1+\sigma^2}$. Since that argument holds for any value of σ^2, further conclude that $D \geq d$.

(d) Argue how your results in (b) and (c) imply that the estimator in (a) is a minimax estimator. Specifically, argue why no other estimator can achieve worst-case risk lower than that achieved by $\hat{\theta}(X) = X$.

Solution:

(a) We have $X_i \sim \mathcal{N}(\theta, 1)$ for each $i = 1, 2, \cdots, d$. Hence, $l(\theta) = \sum_{i=1}^d \mathbb{E}_\theta(X_i - \theta)^2 = d$. Since $l(\theta) = d$ for any θ, so is the worst-case risk.

(b) Consider the joint source-channel coding problem with source $\theta \sim \mathcal{N}(0, \sigma^2 I_d)$ and channel $x|\theta \sim \mathcal{N}(\theta, I_d)$. The overall rate is 1, so $R(D) \leq C$ follows from the joint source-channel coding theorem. Alternatively, we can also write

$$R(D) = \min_{p(\hat{\theta}|\theta):\mathbb{E}\|\hat{\theta}-\theta\|_2^2 \leq D} I(\theta; \hat{\theta}) \leq I(\theta; \hat{\theta}) \leq I(\theta; X) \leq \max_{p(\theta):\mathbb{E}\|\theta\|_2^2 \leq d\sigma^2} I(\theta; X) = C$$

for $\theta - X - \hat{\theta}$ forms a Markov chain.

(c) By (b) we have $\frac{d}{2}\log\frac{d\sigma^2}{D} \leq \frac{d}{2}\log(1+\sigma^2)$, which gives $D \geq \frac{d\sigma^2}{1+\sigma^2}$. This inequality holds for any σ^2, we choose $\sigma^2 \to \infty$ to conclude that $D \geq d$.

(d) Part (c) shows that the worst-case risk for any estimator must be no smaller than D. Since the natural estimator $\hat{\theta}(X) = X$ achieves the worst-case risk D, we conclude that this estimator is minimax.